Faculty of Natural and Mathematical Sciences
Department of Informatics

King's College London
Strand Campus
London
United Kingdom

7CUSMUIP
Individual Project Submission 2021/09

Name: Patrick Stewart
Student Number: 20115524
Degree Programme: Data Science
Project Title: Using Twitter data to understand user perceptions and experiences of the Oxford Street District
Supervisor: Rita Borgo
Word count: 13,940

RELEASE OF PROJECT
Following the submission of your project, the Department would like to make it publicly available via the library electronic resources. You will retain copyright of the project.

☒ I **agree** to the release of my project

☐ I **do not** agree to the release of my project

Signature: PStewart

Date: 03/09/2021

Department of Informatics
King's College London
United Kingdom

7CUSMUIP MSc Project

USING TWITTER DATA TO UNDERSTAND USER PERCEPTIONS AND EXPERIENCES OF THE OXFORD STREET BRAND

Student Name: Patrick Stewart
Student Number: 20115524
Degree Programme: Data Science

Supervisor's Name: Rita Borgo

This dissertation is submitted for the degree of MSc in Data Science

ABSTRACT

Background to the project

In urban research, new social media sources have the potential to provide new insights about citizens, their activities, thoughts and emotions. One of the biggest social network applications is Twitter, which has millions of users expressing their thoughts and sentiments daily. In addition, Twitter provides a data API that is freely accessible to researchers with details provided such as user id, the tweet text, time the tweet was sent, like count and potentially the location the tweet was sent from. Given the data opportunity from Twitter, we decided to investigate how this information can be used in the context of urban planning, with a particular focus on methods to understand user perceptions and experiences of a given area.

Objectives of the project

This research project is sponsored by Westminster City Council, who are looking for new methods to understand how people respond to the Oxford Street District's built environment and the activities within it. We decided that Twitter data would provide an interesting approach to this problem given its spatial (tweet data provides geolocation information if user approval granted) and temporal (exact time of tweet is published in the data) properties.

Therefore, after acquiring a suitable Twitter dataset, we apply sentiment, topic and spatial analysis on the dataset.

1. Sentiment analysis - a lexicon approach is applied to define each tweet as either positive, negative or neutral.
2. Topic analysis - an unsupervised learning technique called latent dirichlet allocation is applied to uncover topics of conversation.
3. Spatial analysis - a clustering technique called k-means++ is applied to highlight key areas where there is a significant concentration of tweets.

As well as the whole dataset, we apply these techniques on various subsections of the data including just positive and negative tweets (topic and spatial only), tweets related to smell and sound (sentiment and spatial only) and tweets before and after COVID. Finally, we present these results using a number of temporal and spatial visualisations.

Contributions

The projects contributions extend both to Westminster City Council and to the wider urban analytics research field. For Westminster City Council, this project provides a clear insight into the uses of twitter data in understanding public experiences and perceptions of an area such as the Oxford Street District. In addition, it gives the opportunity for Westminster City Council to gain insight in real-time that is not possible with traditional sources such as surveys.

In the context of wider research, the application of natural language processing techniques (machine learning techniques applied to text) to short-form microblog sources (e.g. social media) is still in its infancy with the project giving a case study in its usefulness. We aim to

demonstrate that sentiment, topic and spatial analysis techniques undertaken are valid and do provide genuine insight.

Results

To summarise the key results, we observed that change in tweet volumes corresponded to changes in macroeconomic factors such as lockdown conditions leading to reduced tweet volumes. Similarly, this was also reflected in terms of positive and negative sentiment with these significantly reducing and increasing respectively in relation to Government policy action such as lockdown enforcement that has yet to return to pre-pandemic levels. While this could be a cause for concern from an urban planning perspective, analysis of other areas such as Belgravia, London also indicate that positive sentiment as a proportion of the total still remains well below pre-pandemic levels. Turning to topic analysis, key conversations of interest for the dataset included food, celebrations and positive emotions, while when digging deeper into negative sentiment tweets, we uncovered issues related to street complaints. From a spatial perspective, hotspots of tweet activity not surprisingly focussed on areas with a strong presence of shops and food or notable landmarks with examples of this including Carnaby Street and Marble Arch.

Primary conclusions

The results from the analysis undertaken on the dataset used does provide sufficient evidence that both Twitter data and the techniques employed are able to provide novel insights that urban planners can use to improve their decision making. In addition, we also believe our project creates avenues for other research opportunities. Examples of this include investigating the use of supervised topic classification techniques that can more effectively classify important topics (e.g. public health) that have been identified from the topic analysis.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr Rita Borgo, of the Department of Informatics, King's College London, for her support during the project. I would also like to thank my family and friends for their patience during this period of my life. Finally, my thanks also extend to all teachers and staff who work at the Department of Informatics, King's College London.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

1 INTRODUCTION

1.1 Overview

In this chapter, we provide basic background information about the project, the problem that is trying to be solved, the objectives of the research and analysis, motivation for the project and the key results of interest.

1.2 The problem of understanding cities

Cities are complex systems (Theodore (2006)) with a range of dynamic elements that are changing overtime such as residents, tourists and traffic as well as static elements such as buildings and other places of interest (e.g. parks). With respect to continuing to develop all these elements and to create thriving cities, good urban planning is essential with urban analysis key to this. However, with respect to urban analysis, cities are particularly challenging to assess given that they are characterised by high volumes of people, quickly changing needs and other demands that are particular to that area such as having enough green space or maintain a high level of tourism. This makes urban analysis at a high spatial (drill down into very specific regions) or temporal (understanding specific timepoints) scale particularly valuable.

Urban analysis has traditionally been conducted using in-person engagement methods such as surveys, public feedback sessions and counting. However, these are not capable of understanding perception at a high temporal and spatial scale, are often costly, measure a low volume of participants and unlikely to be robust. Although, these traditional methods have more recently been coupled with technical applications (software programs) or other online services that register opinions (e.g. CitizenLab), these solutions still require proactive participation from citizens and other groups, which is difficult to achieve without a concerted marketing effort or rewards scheme to incentivise users (e.g. vouchers for participating in a survey). As a result of this, it is imperative that new methods are utilised to help urban planners make better decisions.

1.3 The opportunity from the web

Easy to access Internet based resources such as social media platforms present new opportunities to understand urban environments. In particular, optional (switched on by the user) geolocation features allow for the generation of highly detailed datasets of users digital footprints and thoughts at very low cost. For example, Twitter allows users to present opinions or status updates, check in to a location or post a photo (Abbasi et al (2015)), while Foursquare allows users to provide recommendations on a place of interest and check in details (Aubrecht et al (2011)) that are timestamped for when the user visited that attraction or place.

With respect to the problem of urban planning, we believe that Twitter is the most suitable source for urban analysis because tweets are geolocated to an exact spot (when approved by the users) as opposed to the likes of Foursquare where visits are tagged to a place, while tweets do not have to be just regarding the specific location but can also refer to other topics such as

the users general wellbeing. In addition, Twitter is one of the most used social media platforms. Using Twitter data, we are able to analyse specific areas in highly granular and temporal detail.

1.4 Objectives as commissioned by Westminster City Council and research aims

This project has been commissioned by Westminster City Council who are keen to better understand user experiences and perceptions of the Oxford Street District using alternative datasets such as those found online. Westminster City Council is currently at the start of a ten-year transformation project for the Oxford Street District to reinvent its high street appeal. Examples of work taking place include Oxford Circus being transformed into two pedestrian friendly piazzas and new attractions being built such as the Marble Arch Mound.

Given Westminster City Council's motivations for the project, we will build a dataset of geolocated tweets data that are within the Oxford Street District since the beginning of 2019. With this dataset, we will look to achieve a number of research aims:

1. How can we identify distinctive sentiment characteristics of tweets to gage users emotions (either positive, negative or neutral) when in the Oxford Street District?
    - We utilised a knowledge-based lexicon method for this approach.
    - The key reasons for using a lexicon method is that it is easy to understand by Westminster City Council and can be adapted in the future by members of that group. In addition, we did not have the resources available to acquire a suitable training data set that could be used for a supervised learning model.
2. How can we identify topics of interest from the tweets dataset?
    - For this we utilise an unsupervised learning topic modelling technique known as latent dirichlet allocation.
    - We used latent dirichlet allocation as we have no preconceptions of what conversations of interest would be and the difficulty in acquiring a suitable training set for a supervised learning method.
3. How can we identify areas within the Oxford Street District that boast a high volume of tweets?
    - For this we use a clustering technique known as k-means++ to identify tweet hot spots within the Oxford Street District.
    - We used k-means++ due to its simplicity compared to other methods and ease of interpretability.
4. What sub-sections of the dataset alongside the full data could be analysed to generate suitable insights?
    - For this, we devise a number of different datasets:
        - Just positive tweets.
        - Just negative tweets.
        - Tweets that refer to a sound.
        - Tweets that refer to a smell.
        - Just tweets before the start of COVID.
        - Just tweets after the start of COVID.
5. How can we communicate and visualise any clear findings from our results that have an impact on the Oxford Street District?

- We use a range of visualisation packages ranging from matplotlib (for graphs) to folium (mapping tool based on OpenStreetMap) and to also highlight any interesting results from our analysis in the report.

1.5 Motivation and significance for undertaking the project

For Westminster City Council, the key motivation for the project was to understand user perceptions and experiences while in the Oxford Street District through alternative data sources. The potential for new and real-time insights are particularly useful for Westminster City Council given urban planning requirements with respect to the ongoing transformation project for the Oxford Street District.

Outside of the uses for Westminster City Council, the project gives a good opportunity to apply natural language processing solutions to assess textual information. Research applying natural language techniques to social media data is still in its infancy, especially when compared to more traditional text sources like news reports and electronic healthcare records. Secondly, there are still some doubts as to the extent of the usability of social media datasets in the field of urban informatics and this project presents a clear opportunity to understand its appropriateness and also identify potential areas of future research and development.

1.6 Main results and contributions from the project

We observe that both positive and negative sentiment as a proportion of the total on a month-on-month basis have decreased and increased respectively since the start of the COVID pandemic and has yet to retrace back to pre-pandemic levels. On the latter point, observing other London regions such as Belgravia show similar trends and suggests that sentiment has been driven by macroeconomic conditions (e.g. lockdown policies) as opposed to any planning policies that have been done or be could in place.

Turning to topic analysis, when looking at the whole dataset it is no surprise to see key topics of interesting including positive events (e.g. celebrations), food and shopping given the Oxford Street District's reputation. More interestingly from an urban planning perspective, when analysing just negative tweets there is evidence of health and nightlife concerns.

Turning to spatial analysis, tweet hotspots tend to centralise on areas with a high concentration of retail and hospitality amenities (e.g. Carnaby Street) or visual landmarks (e.g. Marble Arch). When looking at tweets before and after the start of the COVID pandemic, concentration of tweets has been less focussed on core areas such as Oxford Street.

In terms of contributions, for Westminster City Council we have provided a clear insight into user perceptions and experiences of the Oxford Street District using Twitter data. In addition, this source of data can be updated in real-time, which makes it much more dynamic than traditional urban data sources. From a wider perspective, the application of natural language processing techniques (machine learning techniques applied to text) to short-form microblog sources (e.g. social media) is still in its infancy with the project giving a case study in its usefulness.

1.7 Organisation of the thesis

This dissertation is organized as follows:

- Chapter 1 (Introduction) provides an overview of the project, the objectives and research aims, the wider motivation and significance of the thesis and finally the structure for the report.
- Chapter 2 (Background) provides wider information for the reader to understand the context and techniques used in the project. This includes any definitions of terms and abbreviations, analytics techniques and software libraries.
- Chapter 3 (Related work) provides a critical review of the relevant literature about sentiment analysis and topic analysis, particularly focussed on applications to urban analysis. In addition, visualisation methods that are relevant to sentiment and topic analysis will be briefly discussed.
- Chapter 4 (Approach) describes the approach to generating the dataset, the implementation of the techniques used for the analysis and how these have been analysed and visualised.
- Chapter 5 (Results) evaluates and visualises the results from the analysis. All experiments and analysis will be presented in detail.
- Chapter 6 (Legal, Social, Ethical and Professional issues) gives a discussion of any legal, social, ethical and professional issues from the project.
- Chapter 7 (Conclusion) outlines the conclusions from the thesis. We will also provide a summary of the key results, evaluations of the techniques used and suggest areas of future research.
- Chapter 8 (References) provides a list of source material that has been mentioned in the thesis.
- Chapter 9 (Appendices) provides any other information of relevance.

2 BACKGROUND

2.1 Overview

In this chapter, we introduce wider information for the reader to be able to fully understand the context and techniques used in this project as well as techniques noted in the related work section that need a full explanation.

2.2 Sentiment analysis

Sentiment analysis is a natural language processing technique (i.e. machine learning methods applied to text) used to typically classify textual information by different polarities (e.g. positive, negative and neutral) based on its contents. Classification end results include expressing information by feelings and emotions (angry, sad, happy) and intentions (interested or not interested), although for our study we just focus on polarity. The use of the technique has grown rapidly since the mid-2000s and is now used across a range of industries such as hedge funds using the technique to analyse financial earnings transcripts to better gage the relative sentiment of a company's management team (Askerov et al (2020)). Figure 1 details the rapid uptake in sentiment analysis searches since 20014 (benchmarked to 0) based on Google trends data.

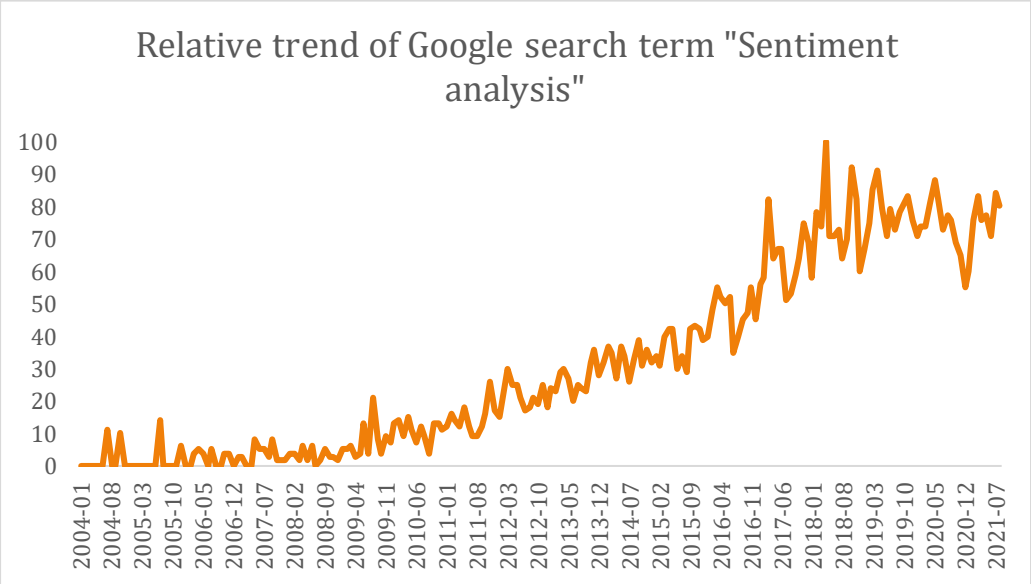


Figure 1 Google trends for the relative popularity of the term "sentiment analysis"

There are three main approaches for sentiment analysis; knowledge-based techniques, statistical methods and hybrid approaches. However, as our thesis only deals with knowledge-based techniques, we only provide detailed information on this.

Knowledge-based techniques make use of pre-defined lexicons (i.e. list of words and expressions) constructed from external knowledge sources with each word or expression defined by a polarity score (e.g. -5 to +5). Using this lexicon, each document or tweet can then be classified by the sum of the positive and negative words contained within the text.

Depending on whether this number is positive, negative or zero determines how we define the document or tweets overall polarity (e.g. greater than 0 would equal a positive polarity).

Knowledge based approaches benefit from users being able to update the lexicon for new words or scores based on their beliefs without any significant data science knowledge. In addition, lexicon-based approaches tend to achieve better results when applying a lexicon to a new domain as opposed to a trained data model using a supervised learning method.

2.3 Topic analysis

Topic analysis refers to techniques that look to extract meaning or build a greater understanding of points of discussion from large swathes of textual information through the identification of recurrent themes or topics. The two main machine learning approaches are topic modelling and topic classification. Topic modelling refers to unsupervised machine learning techniques that cluster groups of texts that are frequently used together. Topic classification refers to supervised learning techniques where a model is trained based on data to classify texts based on characteristics. Our analysis focuses on the use of a specific topic modelling technique known as latent dirichlet allocation.

2.3.1 Latent dirichlet allocation

Latent dirichlet allocation (LDA) is a Bayesian probabilistic model of text documents (Hoffman et al (2010)). The model we use relies on two key assumptions; that similar topics make use of similar words and that tweets can deal with more than one topic. The LDA algorithm is applied on a collection of documents (tweets) with the number of topics pre-specified by the user. From this, each topic is calculated as a collection of words and the words probability of relating to that topic. We can then analyse the topics and key words to better understand what the topic refers to (e.g. events, eating) and subsequently define it. In addition, each document (tweet) can be represented by its most likely topic by the words in the document (tweet). As a result, we can measure the distribution of topics across documents (tweets).

We can summarise the LDA algorithm with the steps taken below.

Equation 1 Latent dirichlet allocation process and equations

1. We assume an appropriate number of topics to start (how this can be accurately defined will be explained in the approach section).
2. Process through each document and randomly assign each word in that document to one of the topics.
3. For each of the documents we process through each word and calculate both the probability of a word given its topic assignment $p(w_j|t_k)$ and probability of a topic for a given document $p(t_k|d_i)$.
    - $p(t_k|d_i) = \frac{\eta_{ik}+\alpha}{N_i-1+K\alpha}$ where η_{ik} is the total number of words in the ith document and in the kth topic, α is a hyperparameter, N_i is the number of words in the ith document and K is the number of topics that was previously defined.
    - $p(w_j|t_k) = \frac{m_{j,k}+\beta}{\sum_{j\in V} m_{j,k}+V\beta}$ where $m_{j,k}$ is the overall assignment (across all documents) of word j to kth topic, V is the vocabulary of all possible words and β is a hyperparameter.

- α and β can be tuned by a number of methods but we use the defaults of 1/number of topics for both parameters.

4. Calculate the new probability of a word given a topic and document $p(w_j|t_k, d_i)$
    - This is achieved using the formula: $p(w_j|t_k, d_i) = p(t_k|d_i) \times p(w_j|t_k)$
5. Now after all this information has been calculated, all words in each document are reassigned to the topic k for which $p(w_j|t_k, d_i)$ is the highest.
6. We then repeat steps 2-5 for a defined number of iterations.
7. Complete

2.3.2 Topic coherence

Topic coherence refers to a method used to score an LDA algorithm by measuring the relative semantic similarity between high scoring words in the topic. We use the following formula to calculate topic coherence (Mimno et al (2011)):

Equation 2 Topic coherence equation

$$C(t;\ V^t) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$

Where t is the number of topics, $V^t = (v_1^{(t)}, \ldots, v_M^{(t)})$ is a list of the most probable words in topic t, $D(v)$ is the document frequency of word v and $D(v, v')$ us the co-document frequency of word v and v'.

2.3.3 Intertopic distance

Intertopic distance relates to the difference in term/word frequency between topics in a two-dimensional space. Principal component analysis is used to rationalise the number of distances into two axis, which can then be visualised and interpreted visually. If two topics have a low intertopic distance, then this suggests that

2.4 Clustering

Clustering refers to the task of grouping a set of data points so that points in the same group are more similar than to points that are in a different group. There are numerous methods of clustering but we use the technique k-means++ for this project. We explain this further below by first explaining k-means, followed by how this is adapted to be k-means++ and finally the method we take in this project (the elbow method) to select the appropriate number of clusters that will be visualised.

2.4.1 K-means clustering

K-means clustering (MacQueen (1967)) is an unsupervised learning algorithm to cluster certain points together based on a certain attribute such as geolocation. The results partition n observations into k clusters with each observation belonging to the nearest centroid cluster. The process is typically formed as follows:

1) The initial centroids (cluster centres) are randomly assigned.
2) Each observation is assigned to its nearest centroid based on a distance measure such as Euclidean distance.
3) The centroids are then re-adjusted to be the mean of all observations assigned to it.
4) Repeat from stage 2 until there is no recognisable change in observation assignment.

2.4.2 The K-means++ adaption

The full technique used in this project is k-means++, which is an adaptation of the algorithm where the initial placement of the centroids (i.e. step 1) is not selected randomly. In this case, the initial centroids are initialised sequentially, with each new centroid having its position at the data point that has the maximum distance from all current centroids that have been previously initialised.

2.4.3 The elbow method

The elbow method is a test used to determine the suitable number of clusters when applying k-means. This runs k-means for a range of pre-defined number of clusters (for example from 1 to 20), and for each of these values the average sum of squared distances between clusters is calculated. As we increase the number of clusters, on average they will be closer together. However, the change in reduction of the sum of squared distances should reduce increasingly quickly and develop a curve like shown below in figure 2. We select a rough point on this curve known as the elbow, where there is a limited increase in sum of squared distances by adding another cluster. This method exists on the idea that one should choose a number of clusters so that adding another cluster does not improve the modelling significantly.

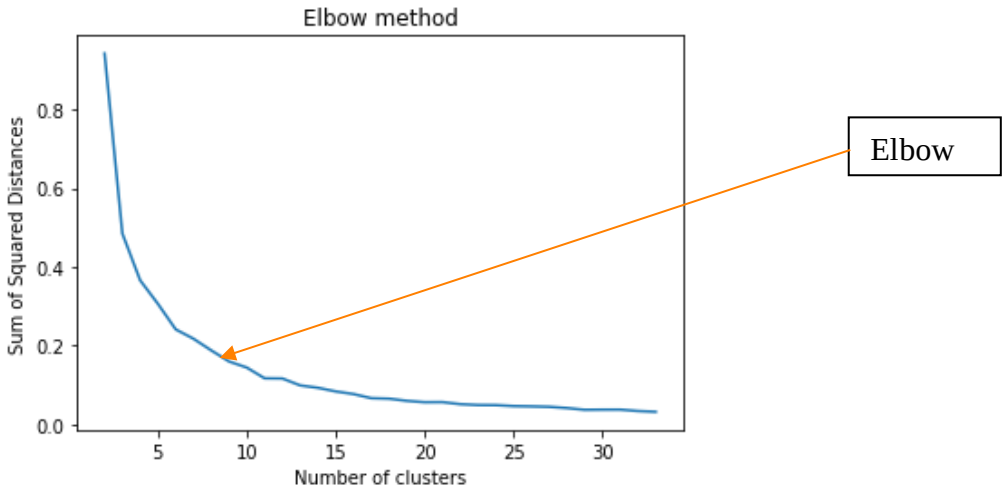


Figure 2 Example of elbow graph

2.5 Terms used in pre-processing

In this section we briefly define a couple of terms that are used when we are processing the twitter data so that it is in an appropriate form to be assessed by natural language processing techniques.

Stemming – Stemming is a text normalisation technique, where morphological variants of a root/base word are made identical. For example, when analysing textual data it is most suitable for terms like "boats" or "boating" to all be simplified to boat. Stemming used in this thesis is achieved using the NLTK package (discussed below).

Lemmatization – Lemmatization is a more complex text normalisation technique than stemming, which looks to consider a language's full vocabulary (typically using a very large term list) to apply a morphological analysis of words. In our thesis, the Gensim package (outlined below) is used for this analysis.

2.6 Python tools used

Requests Python Package - The Requests python package allows users to send HTTP/1.1. requests easily. This is used for Twitter API requests after gaining Academic Research Access. Academic Research access allows a user to download 10m tweets per month from queries of up to 1,024 characters and 50 requests / 15 minutes, per app with metadata such as username, date created at, hashtags, like count and text of tweet. The download can be easily converted into CSV format.

NLTK Python Package – NLTK package stands for natural language toolkit and is an open source Python module focussed on natural language processing. We use the package mainly in the pre-processing of the dataset and removing relevant stopwords that are not useful in understanding sentiment or topics.

Gensim Python Package – The Genism package is an open source Python module focussed on unsupervised learning techniques for text based analytics problems. Gensim is used for LDA analysis.

Preprocessor – Preprocessor is a Python package designed for the preprocessing of twitter texts. It supports the removal of URLs, hashtags, mensions (e.g. @), reserved words and emoticons.

Folium – Folium is a Python geovisualisation package. It allows a user to place geo-visualisations on OpenStreetMap.

pyLDAvis – pyLDAvis is a Python package used to interpret topic models. The outputs and semantic results from the model are presented in the form of an interactive web visualisation.

Scikit-learn – Scikit-learn is a machine learning focussed Python module encompassing areas such as natural language processing, support vector machines, regression, classification and clustering.

Twitter Streaming Application Programming Interface – Platform that allows a user to access twitter data through a user id after filling in relevant information.

3 RELATED WORK

3.1 Overview

In this chapter, we explore various ideas that use alternative datasets such as social media to the application of urban planning. However, we first use prior research to justify why social media data is appropriate for our study and particularly for sentiment and topic analysis (we assume that its use in spatial analysis needs less justification). Turning to the core part of the literature review, we explore ideas that leverage the advantage of fine textual, temporal and spatial information common in social media data sets in the field of urban planning. Finally, we explore prior work on visualisations relevant to the use of sentiment, spatial and topic analysis in urban planning.

3.2 Why can social media data be used for this project?

Despite still being an emerging field, studies have found a strong correlation between the results of sentiment analysis from social media and self-reported life satisfaction or wellbeing (Kramer (2010)). For example, Durahim et al (2015) validated the effect of twitter sentiment by comparing its correlation to stock market (a good proxy for population positivity), while Quercia et al (2012) identified a strong relationship between overall sentiment detected from Twitter data and the wealth status (a good proxy for wellbeing) within the community at large. Given the relationship between Twitter data and wellbeing, we believe that it is appropriate to use the sentiments of tweets in the Oxford Street District as a potential gage for user happiness in an area.

Now we justify why social media can be used for topic analysis. Again, the use of topic analysis on social media is still fairly nascent, although there has been a range of promising work across a range of domains giving evidence of its appropriateness. Examples of this include the successful measurement of public discourse during the COVID pandemic (Xue et al (2020)) and the measurement of road traffic, with this closely correlating to traditional traffic datasets (Hidayatullah et al (2017)).

3.3 Selected research using sentiment analysis on social media datasets for urban planning

Now turning to the use of sentiment analysis in urban planning, Bertrand et al (2013) develop a specifically designed sentiment classifier based on emoticons and certain words to understand the sentiment of a tweet and use the geocoding to map areas of New York based on sentiment. Using this, they were able to analyse how sentiment changes by place over a day, finding that public mood is generally highest in public parks and lowest at transport centres such as bus stops.

Keeping to understanding land use, Cao et al (2018) built a large dataset of geocoded tweets across Massachusetts, USA from between November 2012 and June 2013. A supervised learning approach based on a training dataset was used to quantify and analyse sentiment across different spaces and times. The results showed clear spatiotemporal patterns in user sentiment.

Higher sentiment scores were typically observed in commercial and public places during weekends or evenings.

One of the largest urban studies focussed on using sentiment analysis is from Mitchell et al (2013), whose results shed how social media could potentially be used to estimate real-time changes in a range of core population attributes such as health. This was achieved by combining a large geo-tagged twitter dataset from 2011 and annually surveyed characteristics of all US states and near 400 urban populations. Among a number of results, word choice and message length of tweets were correlated with urban characteristics such as education levels and health. The results showed that social media could be used to estimate changes in important economic measures such as obesity rates.

In terms of prior examples of lexicon-based approaches, Sim et al (2020) investigated the sentiment response and discussion points for two elevated parks in New York and Chicago respectively. A dataset of tweets mentioning the two parks was collected from 2015 to 2019. With this information, user's activities and their sentiment were monitored. The results found that the two parks were enjoyed but there was some concern over issues such as rising house prices.

Keeping to the analysis of specific places of interest, Schwartz et al (2019) used twitter data and a lexicon based approach to investigate how sentiment changed before, during and after visits to one of a number of parks in San Francisco. The results showed that sentiment was much more positive during visits to parks and remained elevated for several hours after. In addition, they analysed vegetative cover across park types with regional parks that typically have greater vegetative cover displaying larger increases in positive sentiment than other types of park such as Civic Plazas and Squares.

3.4 Selected research using topic analysis on social media datasets for urban planning

Now focussing on selected research on the use of topic analysis on alternative datasets, again there has been a number of interesting studies in this area with a diverse range of analytics techniques used. One of the simplest methods used was from Severo (2015) who built a twitter dataset from geocoded tweets in four major European cities. Then using just the twitter hashtags, the most frequent hashtags were identified to highlight the most important points of conversation and which were also presented in word cloud form (discussed further below). Although, a fairly simplistic technique, often just looking at word frequency can give a good understanding of conversations.

Often more complex methods of analysis need to be used to build a finer understanding of content discussion. Lansley et al (2016) builds a large twitter dataset of geotagged tweets from inner London. This dataset is classified into twenty topics using latent dirichlet allocation (explained in background section) with topic examples including describing activities, informal conversations between users and the use of check-in applets. These twenty topics are measured across regions and by time of day with large differences measured. These variations can to some extent be explained by demographic and socio-economic characteristics of users, but the location also has a significant impact.

Content analysis can also be used for the more specific analysis on certain urban structures or functions. For example, Osorio-Arjona et al (2021) gather Twitter posts and user replies from

the Madrid Metro account over a two-month period to better understand user opinions on public transport. They use topic analysis techniques to better understand complaints which is mapped visually to build an understanding of temporal and spatial patterns. In addition, a regression model was applied based on geographic weightings of the population by region to understand which areas are users more likely to complain after controlling for a number of population differences.

3.5 The use of clustering on social media datasets for urban planning

Clustering represents one of the main ways in which geolocated data such as tweets have been analysed from a spatial perspective. For example, Andrienko et al (2013) use a twitter dataset of geolocated tweets in Seattle between August and October 2011. Tweets were matched based on keywords to identify similar semantic contents and refined into a range of topics such as music, fitness and health. These tweets were subsequently clustered by topics to understand areas where there was a high occurrence of similar terms, so that districts could be effectively characterised.

Clustering can also be used in the field of urban informatics to identify land use. For example, Frias-Martinez et al (2014) clustered geographical regions with similar tweeting activity to identify similar urban land uses. Manhattan, London and Madrid were used as test beds with results finding that geolocated tweets can be a powerful source for identifying different areas by their main use (e.g. work, recreation).

In terms of other research where clustering is used for urban analysis, Noulas et al (2011) used spectral clustering (clustering based on graph theory) on geolocated information provided by Foursquare to model crowd activity in London and Manhattan. This was linked to Foursquare location check in details (e.g. hotel) to characterise patterns in movement flows. In similar work, Cranshaw et al (2012) uses clustering techniques to understand land use and social dynamics based on Foursquare data. The results are compared to personal interviews with the users tracked and interestingly show remarkable effectiveness in the approach.

3.6 Related work on visualisations with respect to urban analysis

Finally, while the techniques used to analyse the data present in social media, often the techniques to visualise this, particularly for non-technical audiences, are equally important. When using visualisations in urban analysis, the core idea is that data generated from social media in an urban context is useful to multiple audiences with images that can cater to a number of these groups being the most valuable.

The most typical visualisation of geo-located tweets to date is by using the street structure of a map (e.g. OpenStreetMap) as an overlay with each street highlighted in a certain colour (figure 3 (a)) to emphasise a variable such as sentiment or pedestrian movement (Karduni et al (2017)). Often, visualisations are presented as a data dashboard so that a number of images can be seen in unison.

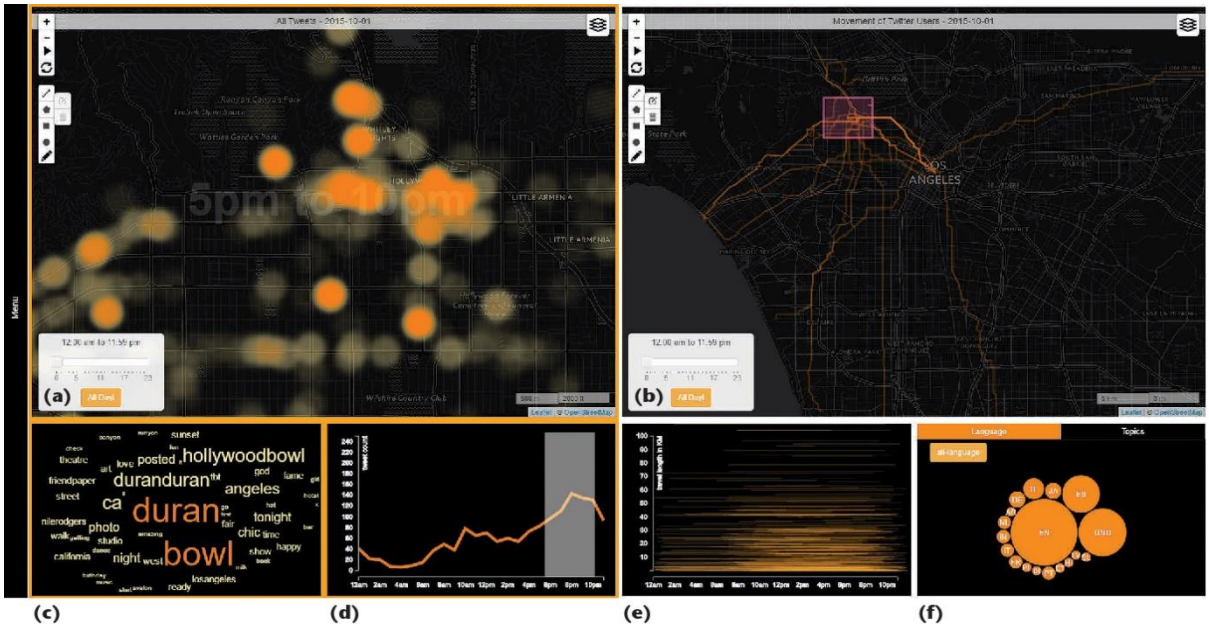


Figure 3 Urban analytics visualisation data dashboard from Karduni et al (2018)

Sentiment visualisations mainly centre on tweet density heatmaps (Karduni et al (2017)) overlayed on top of a traditional map such as from OpenStreetMap (figure 3 (b)). Either separate maps are used for different sentiments or different colours for sentiments on the same map. The user can see clearly the geographic concentration of geolocated tweets for a certain area. Ideally the user should also be able to zoom in on particular areas or zoom out to view a larger proportion of the map, as well as a temporal slider to see how tweet concentration by sentiment changes over time.

The other core visualisation method used are word clouds (Severo (2015)). Word clouds show the most frequent words used within a designated geospatial area. The view is built based on the designated dataset but adaptations include overlaying it on top of a more traditional map and updating the word cloud based on the area of the map currently visualised. An example of this from Severo et al (2015) is provided in figure 4 below.

Figure 4 Example of a word cloud from Severo et al (2015)

Turning to more complex visualisation, Weiler et al (2016) used a clock-face word cloud to visualise frequent terms and the sentiment of those words, with darker red suggesting a more negative sentiment and lighter colour suggesting a more positive sentiment. The clock face can

be drilled down or rolled up for more specific or general detail. An example of this is provided in figure 5 below.

Figure 5 Example a clock-face word cloud graph from Weiler et al (2016)

3.7 What are the opportunities in our study compared to related work?

Firstly, the project represents an opportunity to build on the nascent literature around the use of social media for urban analysis. Firstly, the use of a lexicon-based approach for twitter sentiment analysis offers an interesting contrast to the more typical supervised learning approaches employed. Secondly, there is particularly limited research in combining detailed topic and sentiment analysis to the problem of urban analysis with the project offering a unique approach to this.

Moving away from the actual techniques employed, the project has the opportunity to look at the landscape since the outbreak of COVID and how this differs to prior pre-COVID research. Furthermore, the project is focussed on one specific area, the Oxford Street District, whereas typically this type of research is conducted using a larger geographic area such as a full city. In addition, Oxford Street offers a unique dataset given its cultural significance and high volume of visits from tourists, residents and workers.

4 APPROACH

4.1 Overview

In this section, we outline our approach to analyse the research questions of this thesis. We outline the methods for data acquisition, pre-processing, sentiment analysis, topic analysis, spatial analysis and note some of the visualisations that will be used in the results section. In addition, we defend our reasons for the methods chosen.

4.2 Reminder of problem and solution

As a reminder, we are investigating user experiences and perceptions of the Oxford Street District. We want to measure user views not only on specific places and landmarks (e.g. Foursquare) but general mood and opinions on the area as a whole as well. As a result we analysed that Twitter was the most practical and low cost option to do this. Westminster City Council are also keen to understand how user perceptions and experiences vary over time and by area. As a result, we apply sentiment analysis that can easily measure public mood over time, while spatial analysis understands variation by area within the Oxford Street District. Finally, topic analysis presents the opportunity to understand what the underlying conversations are, which can be filtered by time period or area if appropriate.

4.3 A note on supplementary files for this section

All relevant data and code can be accessed in the supplementary files, although we have included snippets of code in the appendices where appropriate. We have also clarified each section to its relevant python code below.

- "Twitter Data Pull for Oxford Street" – Data acquisition (4.4)
- "Data pre-processing, sound and smell class and sentiment analysis" – Data processing steps (4.6), Smell and sound classification (4.7) and Sentiment analysis classification (4.8).
- "Sentiment visualisations" – Presenting the sentiment results (4.8.3).
- "Topic analysis" – Topic analysis (4.9).
- "Spatial analysis" – Spatial analysis (4.10).

4.4 Data acquisition

The study area for the analysis is the Oxford Street District (figure 6), an area of roughly 2.5miles that attracts 200m visitors a year. In addition, it contains 38,000 residents, while 155,000 people are employed within the district. The twitter data was obtained using the Twitter Streaming Application Programming Interface and the Requests Python Package (see background) between January 2019 and July 2021. The full code and dataset is provided in the supplementary file and is based on Edward (2021).

The relevant area was obtained by searching the bounded box area with longitude and latitude coordinates (-0.163, 51.51), (-0.163, 51.521), (-0.130, 51.510) and (-0.130, 51.521) and only

for geolocated tweets. As you can see from figure 6, Oxford Street District has a slightly odd shape and can not be matched with an exact rectangle so there are some differences between the exact boundaries of the OSD and what has been defined in our study.

Figure 6 Map of the Oxford Street District (diagram from https://osd.london/)

Although there are a range of data columns that have been acquired, we explain the key columns that are used during our analysis:

- "author id" – Each user has a unique author id and is used to make sure that we do not have high concentrations of tweets from the same user.
- "created_at" – Exact date and time by which a tweet was sent.
- "lang" – Language defined by Twitter in which the tweet was written in.
- "coordinates" – Longitude and latitude point at which the tweet was sent from.
- "tweet" – The actual text of the tweet.

In terms of other things to note about the dataset, firstly the Twitter API does not supply any retweets, only organic posts. Secondly geolocated tweets do unfortunately represent a small sample of the overall number of tweets posted in a given period, circa 1-2% based on prior studies (Zhang et al (2016)). Thirdly, it was difficult to distinguish between personal and corporate/government-based Twitter accounts, which may have an effect on the topic and sentiment analysis results.

4.5 Note on coordinates

While most coordinates within the area have geocoordinates, there is a small proportion of tweets that only provide the place_id, although this is still within our bounded box region assigned. As a result, these tweets are still kept in the dataset due to their usefulness for

sentiment and topic analysis, however they will have to be excluded from any spatial analysis given that we cannot directly pinpoint the exact place the tweet was sent from.

4.6 Data processing steps

Figure 7 below outlines all the necessary steps as part of our analysis in converting the dataset into the appropriate form.

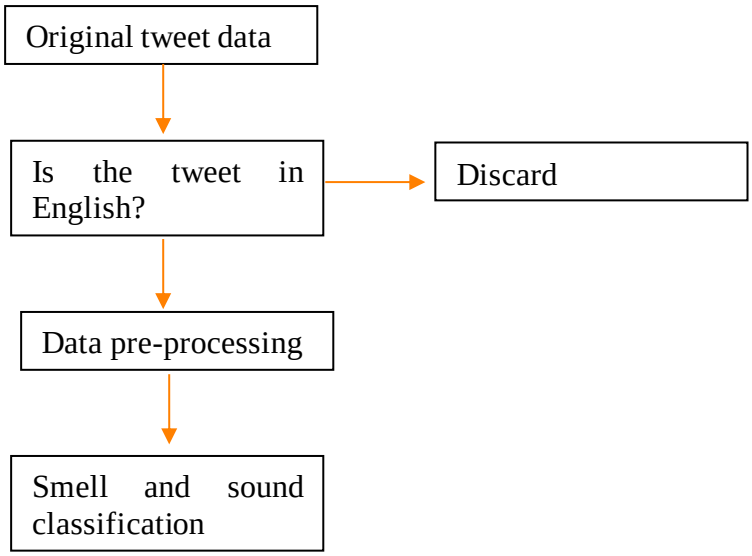


Figure 7 Data pre-processing flowchart

For our study, as applying sentiment and topic analysis techniques across multiple languages has proved challenging, we first only use English language tweets. Luckily Twitter already defines the language of the tweet based on its content in one of the data columns so we can easily select all tweets recognised as English.

Next, we apply appropriate data pre-processing with the steps for this shown as follows:

1. We drop any duplicate texts to remove the risk that any retweets or users double tweeting the same message have been included in the dataset.
2. We convert all text to lower case.
3. We apply the preprocessor python package (outlined in the background section) to clean the tweets of all non-text based items (e.g. @).
4. We remove any other punctuation items.
5. We apply the lemmatization and stemming on the text (explained in background section).
6. We remove any stopwords that would not have been used in the analysis such as “and”, “is” and “at”.

The code for just the data pre-processing is provided in appendix A.

Once the pre-processing has been completed, for the period we are left with 56,550 tweets for the whole period. The tweets contain an average of 10.2 words with the highest number of

words being 39. As we can see from figure 8 below, there is a broad range of tweets with different numbers of words but the heaviest concentration is tweets with 3-12 words.

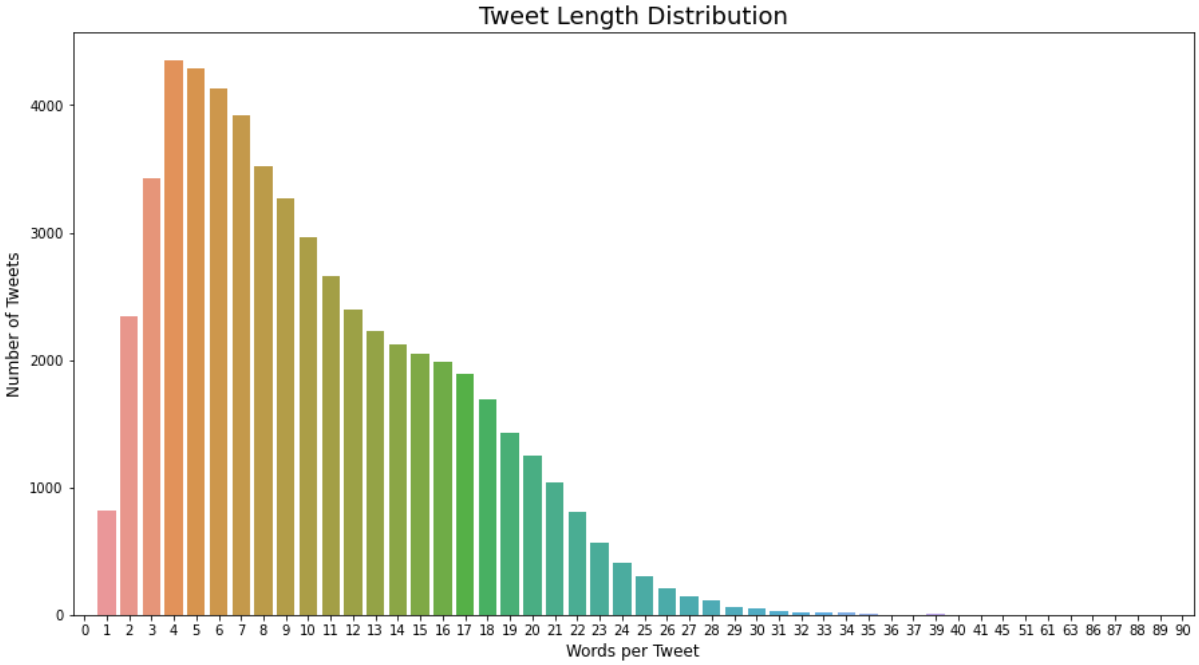


Figure 8 Bar graph showing the concentration of tweets for each number of words

4.7 Smell and sound classification

Once the data has been appropriately pre-processed, we also want to see how many of the tweets have a reference to sound or smell and use this classification in an extra column of the dataset. This will subsequently used later as a different partition of the data to be analysed using the sentiment and spatial analysis techniques.

The reason for devising this partition is as typically councils only receive negative feedback on smells and sounds from the public in the form of complaints. However, it has been pointed out that urban environments contain a range of smells and sounds and our analysis would benefit from understanding how both sentiment and topic analysis varies depending on if the point of discussion contains a smell or sound reference.

To do this, we first acquired two lexicons from the GoodCityLife project focussed on sound (Quercia et al (2016)) and smell (Quercia et al (2015)). The smell lexicon was generated from participants who were exposed to a range of different smellscapes and were asked to record their experiences as well as from social media sources such as Flickr. The sound dictionary was designed by the authors and is based on the most comprehensive research project in the sound field – World Soundscape Project – and from a large crowdsourced web repository of sounds - Freesound.

To identify if a tweet refers to smell or sound, we use a simple word matching solution (displayed in appendix B) where the tokenised text is evaluated to both lexicons with matches noted as referring to sound, smell or both. Once the word matching has finished, we are left with 6,964 tweets that refer to smell and 14,580 tweets that refer to sound out of the total.

4.8 Sentiment analysis classification

4.8.1 Why a lexicon approach was taken?

There were a number of reasons why a lexicon approach was the most suitable measure for this project. The key reason was scalability with both statistical and hybrid methods requiring a suitable training set to be built to be able to train an appropriate sentiment classification model and that the resources required to do this were difficult to acquire given our limited capacity. In addition, this solution is sponsored by Westminster City Council and a lexicon method can be investigated or altered by members of the council without any detailed technical knowledge. Finally, one of the main features of lexicon-based approaches are their generalisation quality, which is particularly relevant to this project as other datasets from the likes of Yelp, Foursquare and Facebook could be added in the future.

4.8.2 Process

As noted in the background section, the sentiment analysis through a lexicon based approach is achieved by scanning each Tweet for keywords defined in the lexicon, with each word rated by an integer depending on its positivity or negativity. The lexicon utilised in this study is by Finn Arup Nielsen (Nielsen (2011)) and contains 3,382 words rated on a scale of -5 (highly negative) to +5 (highly positive). This particular lexicon has already been used in broader studies such as modelling the sentiment characteristics of advertising content (Abrahams et al (2013)) and has also proven to be effective in urban analytics studies (Hollander et al (2015) as well. The additional attractive feature of Nielsen's lexicon over more tradition lexicons such as ANEW (Affective Norms for English Words) is that it was designed for microblogs as opposed to more traditional text sources.

In terms of implementation, this is achieved by scanning the tokenised text of each tweet and identifying the words in a tweet that match the lexicon. If this matches, then the value of the word is added to the overall score of the tweet. Once all the words in a tweet have been reviewed, we take the score and define the sentiment for the tweet to be positive (greater than 0), negative (less than 0) or neutral (equal to 0). The pseudocode for each tweet is as follows:

Input: $Tweet, Lexicon, Lexicon\ value$
Output: $Positive, Negative, Neutral$

1. Begin
2. Counter = 0
3. For each $w \in Tweet$:
4. If $w == Lexicon$:
5. Counter += Lexicon value (-5 to +5)
6. If Counter > 0:
7. return "Positive"
8. If Counter < 0:
9. return "Negative"
10. If Counter == 0:
11. return "Neutral"
12. End

After this is completed, we are left with 29,578 positive tweets, 4,695 negative tweets and 22,277 neutral tweets. The distribution of how these are spread is displayed in figure 9 below.

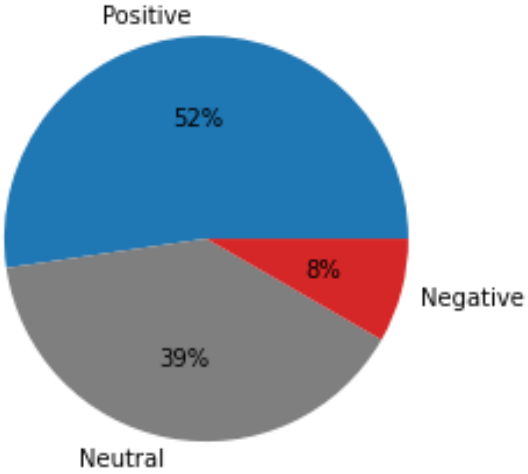


Figure 9 Pie chart showing tweets by sentiment as a percentage of the overall total

4.8.3 Presenting the sentiment results

The sentiment results will be mainly displayed visually based on a time series data set generated from the sentiment analysis. This is shown as $D = \{d_1, d_2, \ldots, d_n | d_x^+, d_x^0, d_x^-\}$ where d_x^+ is the number of positive tweets on that day divided by the total number of tweets, likewise d_x^- is the number of negative tweets on that day divided by the total number of tweets and d_x^0 being the total number of neutral tweets on that day divided by the total. The key line graphs used will illustrate either positive or negative tweets on a monthly basis as a proportion of the total. Figure 10 below provides an example of this.

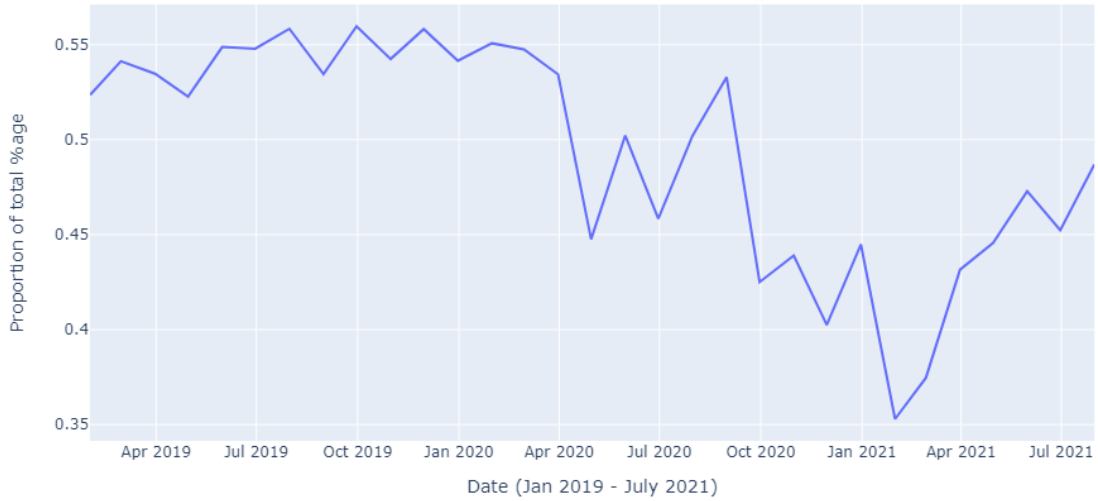


Figure 10 Example of positive sentiment line graph

4.9 Topic analysis

4.9.1 Why we use latent dirichlet allocation?

Topic analysis with respect to twitter remains a very nascent area of research with micro-blog behaviour proving to be quite different to other forms of text (Eisenstein (2013)), which makes gaging the appropriate techniques to understand conversations of interest to be fairly subjective. We found that keyword-based approaches are not suitable as typically there is one to two subjects that dominate, especially in high profile areas such as the Oxford Street District where most discussions focus on hospitality, retail and landmarks. However, evidence from prior academic research does demonstrate that unsupervised machine learning approaches are capable of identifying a much larger subset of topics and distinguishing them accordingly. While, supervised topic classification algorithms (noted in background section) have also shown strong promise, these techniques require a suitable training dataset which we do not have the suitable resources for.

Given the opportunity from unsupervised learning, we apply the algorithm latent dirichlet allocation (explained in the background section) as most research to date has suggested this to be the most appropriate algorithms for this analysis. We use the Gensim package (see background) to perform latent dirichlet allocation in Python.

4.9.2 Identifying the number of topics

With LDA, the number of topics has to be specified by the user. Finding this number is more of an art than a science and we take a trial-and-error approach based on the topic coherence (explained in detail in the background section) and intertopic distance. To roughly gage the appropriate number of topics, we build LDA models for the tweets with topics between 1 and 36. We allow 10 passes for each model and the genism package automatically tunes the values for alpha and beta to be $1/(number\ of\ topics)$. In addition, we do not include any tweets with 3 words or under in it due to data sparsity concerns and remove any additional stop words (e.g. London) that do not add any value to the topic analysis, this is visualised through a word cloud.

4.9.3 Topic coherence score

We subsequently plot the coherence score to ascertain a good tradeoff between the number of topics and coherence values. Figure 11 below shows an example of this, we can see a small plateau between topics 5 and 7 so we would infer this range to be an appropriate number of topics. As a note, this topic number can be adjusted after further analysis using intertopic distance and the underlying words for each topic if suitable.

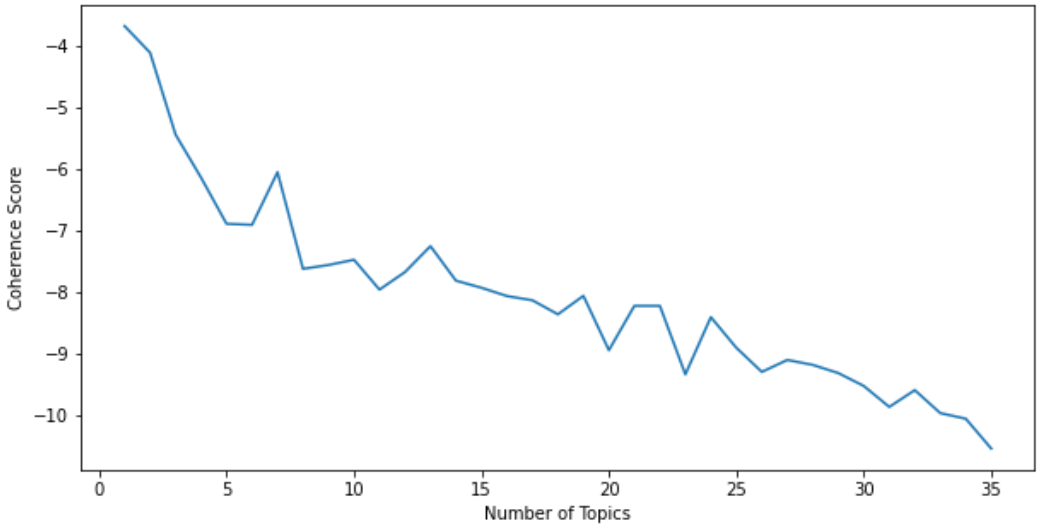


Figure 11 Example of line graph demonstrating number of topics against coherence score

4.9.4 Intertopic distance

To further understand if the appropriate number of topics has been chosen, we look at intertopic distance (see background for more detail), which is a function that is part of the pyLDAvis package for understanding the similarity between topics. Figure 12 below shows an example of this where four is deemed to have the most suitable topic differentiation given that there are appropriate distances between them. We have decided that the required intertopic distance can be flexible if despite low distance between a number of topics, we can still clearly interpret the differences and decide that it is necessary to keep the topics separate.

Figure 12 Example illustration of intertopic distance from the pyLDAvis package

4.9.5 Labelling topics

After topic modelling is completed, we label each topic to build an understanding of the topics and for better development of results. This was done by assessing the most frequently occurring words for each group and identifying key words within this. We also validated the labels by manually checking a number of the tweets for each of the topics to ascertain if the labelling is appropriate compared to the actual tweet.

4.9.6 Presenting topic results

After this has been applied, we focus our analysis on understanding the topics of interest and focus on any key words from topics that we believe to be interesting. In addition, we apply each tweet to its most relevant topic to see how tweet distribution varies by topic. We give an example of this graph in figure 13 below.

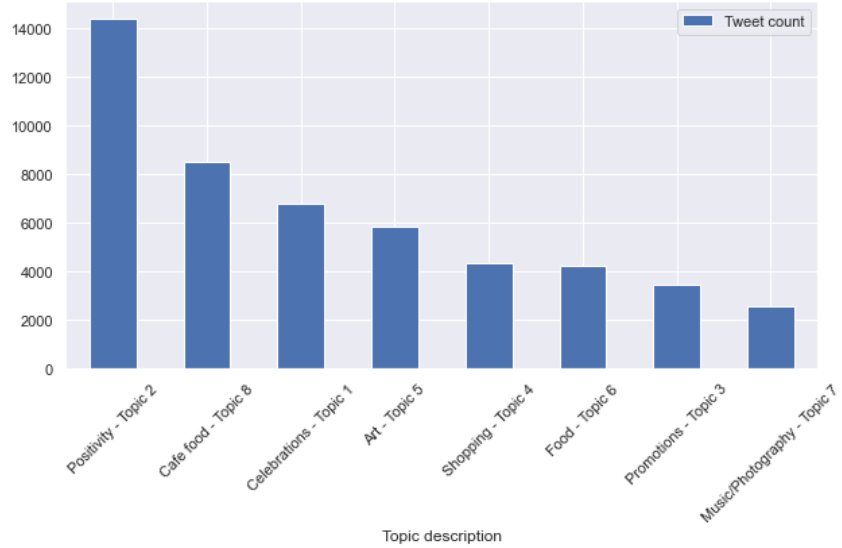


Figure 13 Example bar chart showing tweet distribution over topics

4.10 Spatial analysis

4.10.1 Why we use k-means++ clustering

Our spatial analysis centres on roughly locating areas with high concentration of tweet activity. We have no prior understanding or conceptions of the data set and therefore an unsupervised learning algorithm is most suitable. In this case, the most appropriate solution is a k-means++ clustering algorithm (see background), which clusters the tweets into relative groups as explained in the background section. We also use the k-means++ algorithm for our analysis as the results and process are easily interpretable and has a low computational cost to run.

4.10.2 Deciding on the number of clusters and visualisation

First, we need to decide on the number of clusters. This is achieved using the elbow method (see background), which measures the sum of squared distances between clusters and gives a good reference between number of clusters and distance between them. Once the appropriate number of clusters has been selected using the elbow method (example in figure 14 below), the cluster centres will be visualised using a heat map from the Folium python package overlayed onto OpenStreetMap.

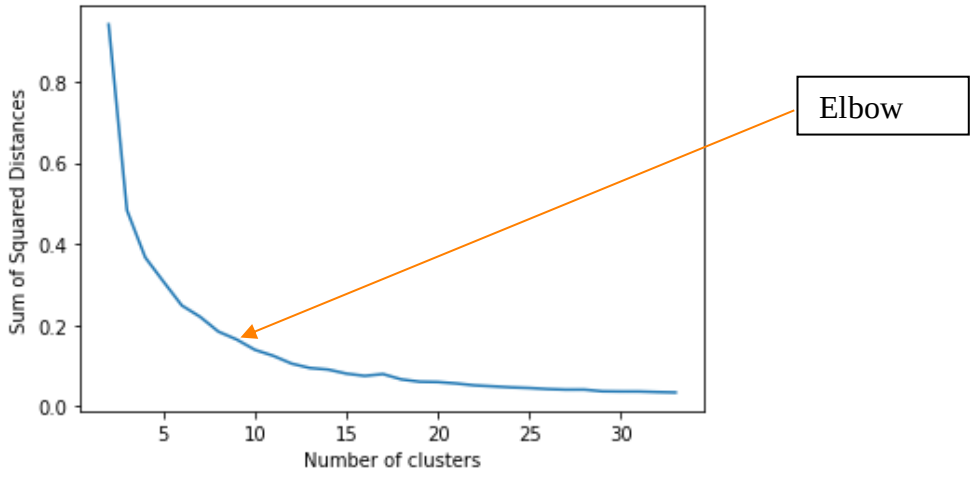


Figure 14 Example of elbow method diagram

4.11 Sub-sections of the data used

We apply our spatial, topic and sentiment analysis on the whole dataset and various sub-sections of the data, with these listed below alongside what is applied.

- Just positive tweets: we apply topic and spatial analysis to understand topics of conversation related to positive views and also where are the hotspots where these positive tweets derive from.
- Just negative tweets: we apply topic and spatial analysis to understand topics of conversation related to negative views and also where are the hotspots where these positive tweets derive from.
- Sound tweets: we apply sentiment and spatial analysis to understand how many positive and negative tweets relate to sound and where these are most concentrated. We do not apply topic analysis due to the obvious similarity between topics.
- Smell tweets: we apply sentiment and spatial analysis to understand how many positive and negative tweets relate to smell and where these are most concentrated. We do not apply topic analysis due to the obvious similarity between topics.
- Post-COVID: we apply topic and spatial analysis to understand conversations of interest and hot spots since the outbreak of COVID. We do not apply sentiment analysis as already covered when analysing the whole dataset.
- Pre-COVID: we apply topic and spatial analysis to understand conversations of interest and hot spots before the outbreak of COVID. We do not apply sentiment analysis as already covered when analysing the whole dataset.

5 RESULTS

5.1 Overview

In this section, we outline the results from our analysis and discuss the key inferences from this. In addition, we note any potential implications and discussion points from a broader urban planning perspective. In terms of the structure of the results section, we first analyse the overall dataset and then deep dive into various data subsections as outlined in the approach section.

5.2 Analysis of the whole dataset

First, we apply sentiment, topic and spatial analysis on the whole dataset to generate appropriate insights.

5.2.1 Sentiment analysis for the whole dataset

After applying the lexicon sentiment procedure on the whole dataset, figure 15 below shows the number of positive, negative and neutral tweets by month. Here, we can see that there has been a clear drop off in overall tweets since the start of the pandemic and was particularly low in the lockdown period at the start of 2021. This is not surprising with figure 16 showing google activity (tracking location data of Android users in establishments such as retail) for the broader area of Westminster, where there has been a clear drop off in activity in every area apart from time spent in residential locations. This also indicates, as expected, that tweet activity is more strongly correlated with sectors such as retail, transit and parks compared to residential activity.

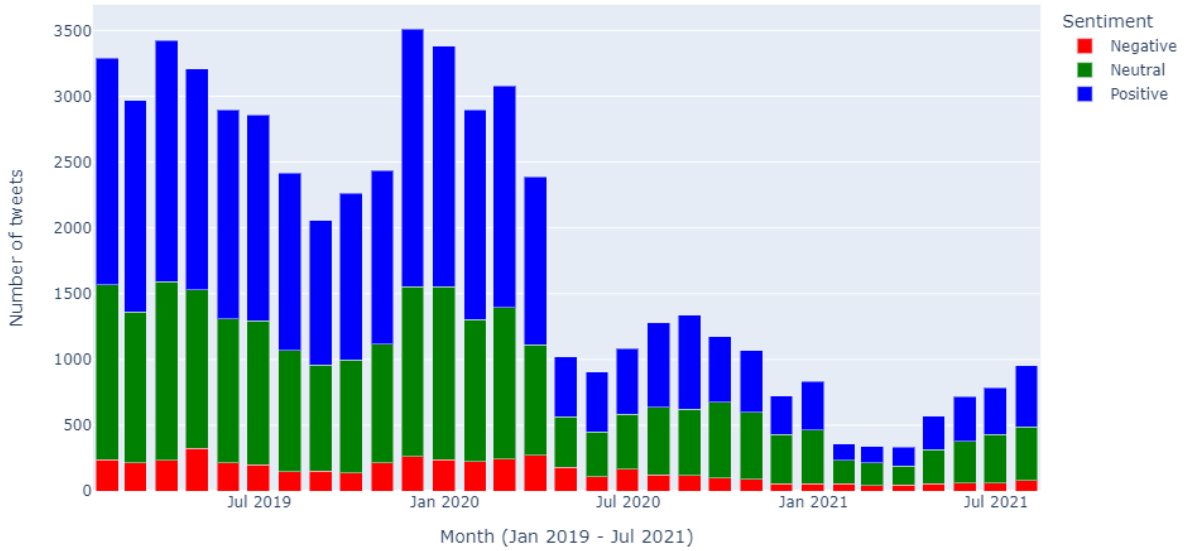


Figure 15 Bar plot showing the tweets by sentiment for each month in the Oxford Street District between January 2019 and July 2021

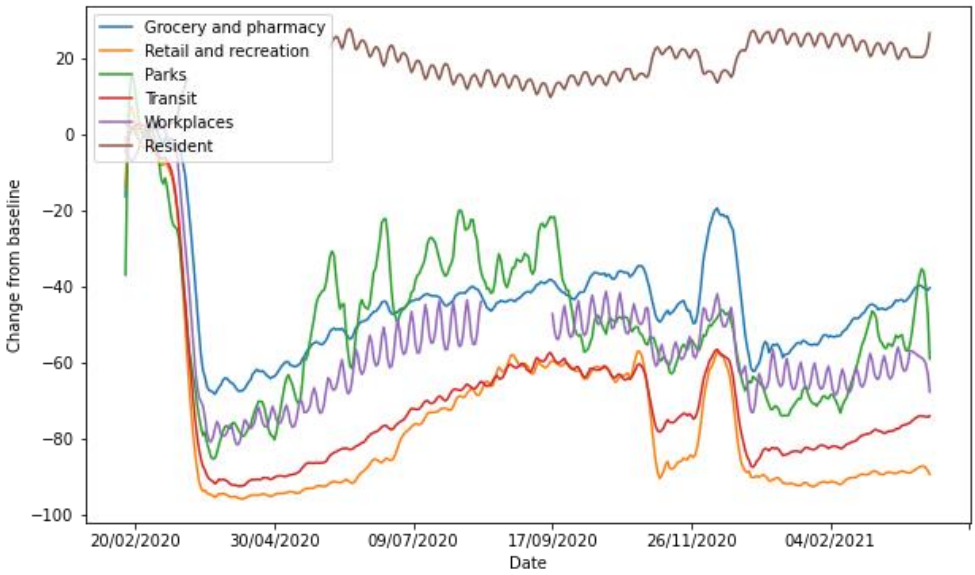


Figure 16 Google activity trends for Westminster from January 2020 to May 2021

In addition, we compare these changes in overall tweets to another similar bounded area in London, Belgravia, in particular to see if this trend of reduced tweet activity since the start of the pandemic is mirrored in more residential focussed areas. From figure 17 below, we can see that there are considerably less tweets per month and that the drop off since the pandemic has been less severe. This is likely driven by the Oxford Street District's high concentration of landmarks, retail and hospitality venues which naturally drives a much higher level of tweets during normal periods, while will fall more significantly in lockdown periods.

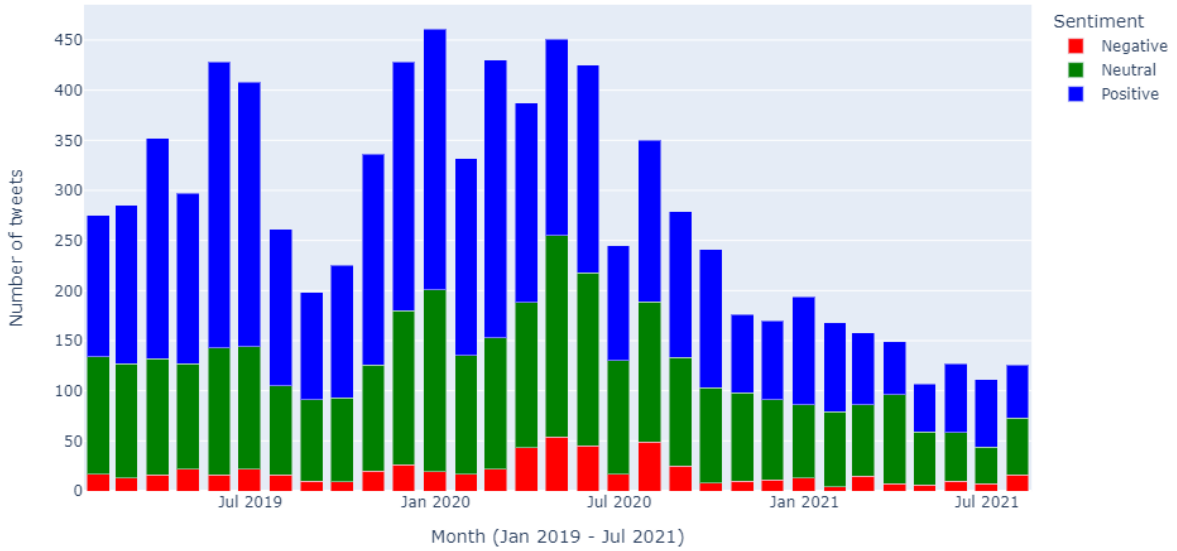


Figure 17 Bar plot showing the tweets per month by sentiment in Belgravia

Next, we drill down into sentiment proportions of the total for the Oxford Street District to see how this changes on a month on month basis. From figure 18 below, we can see that positive sentiment drops at the time of the first COVID outbreak, briefly recovers and falls even lower during the winter lockdown. Although there is somewhat of a recovery in July, the proportion of positive tweets is still significantly below pre-COVID levels of circa 55%.

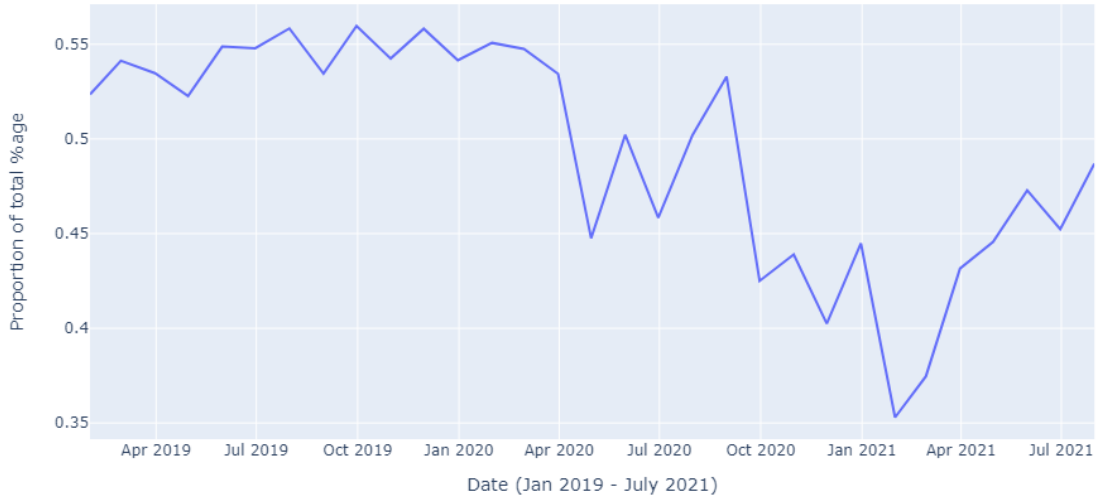


Figure 18 Line graph showing change in positive sentiment for all tweets

Next, we drill into negative sentiment to see how this changes as a proportion of the total on a month on month basis. From figure 19 below, we can see how the COVID pandemic in March 2020 and the lockdown period during the winter of 2021 led to sharp increases in the proportion of negative sentiment. However, unlike positive sentiment, the relative proportion of negative tweets quickly has stabilised to a steady trend of between 6-10% of the total by the Summer of 2021.

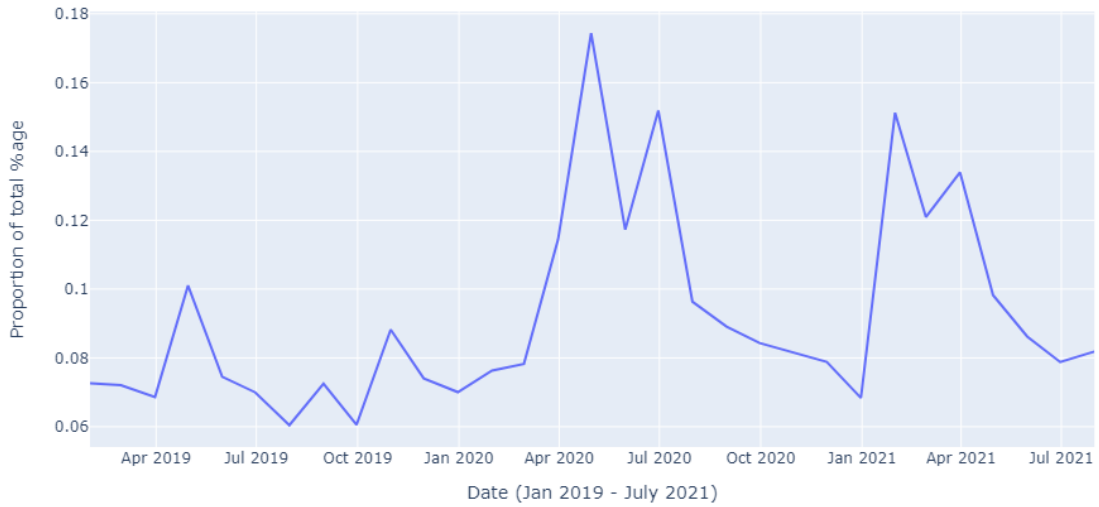


Figure 19 Line graph showing change in negative sentiment for all tweets

Next, we compare the Oxford Street District to Belgravia to see if the proportion changes are mirrored. We can infer from figure 20 below that again the proportion of positive tweets has fallen since the start of the COVID pandemic, mirroring that of Oxford Street with little sign on reforming back to pre-pandemic levels of 55-60% of the total. Turning to negative Sentiment, in figure 21 below, we can see that the proportion of negative tweets for Belgravia rose after the breakout of the pandemic before returning to normal levels. Although in July negative sentiment rose sharply, after viewing the raw data this is likely related to England's loss in the final of the Euro 2020 football tournament and the very low data count of just over 100 tweets likely leading to increased volatility.

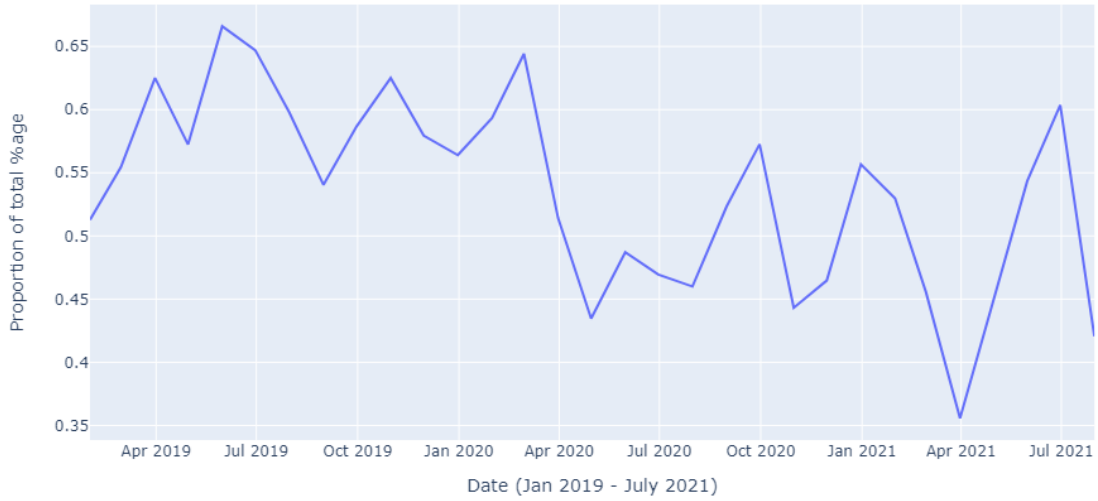


Figure 20 Line graph showing change in positive sentiment for all Belgravia tweets

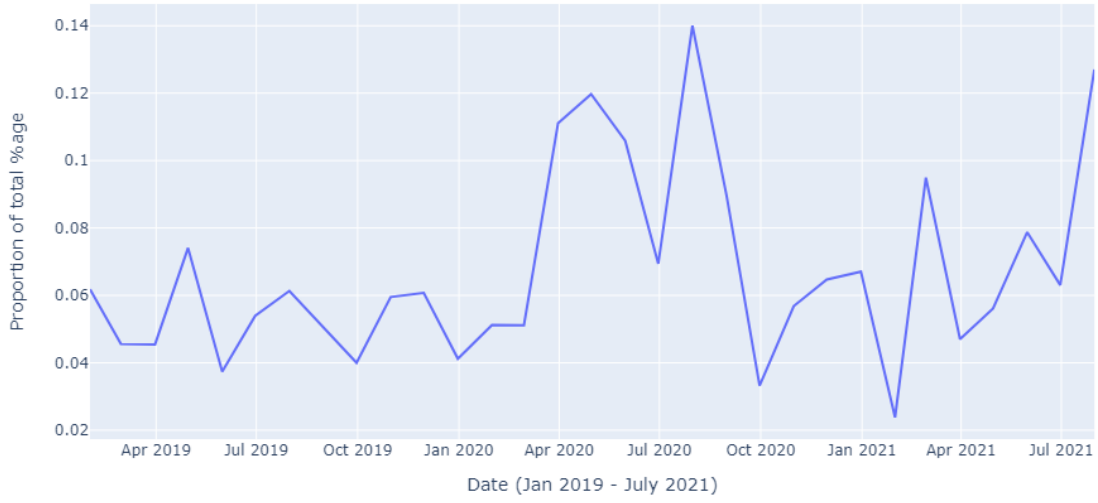


Figure 21 Line graph showing change in negative sentiment for all Belgravia tweets

From an urban planning perspective, we observe that there is clear evidence that satisfaction by users has been lower in the Oxford Street District since the start of the pandemic. However, we observe that this is driven more by macro-economic factors than anything specific to the region with other areas such as Belgravia demonstrating similar trends. With most of the new

Oxford Street Initiatives yet to be employed or just opening (e.g. Marble Arch Mound), we are currently not able to analyse how these effect sentiment in the region.

5.2.2 Topic analysis when looking at all tweets

Next, we look at all tweets to see what are the key conversational topics and what are the actual quantities for each of the probable topics.

Firstly, assessing a word cloud (figure 22) of all the tweets, we can see a clear dominance of positive terms related to the physical environment (e.g. street, London) and positive emotions (e.g. great).

Figure 22 Word cloud showing most frequent words for all tweets

We can dig further into this using latent dirichlet allocation. After the appropriate topic coherence and intertopic distance analysis, we are able to split out six core topics of interest with only topic 1 and 4 demonstrating (figure 23) a strong similarity, although we decided there was significant difference in the words to command separate topics.

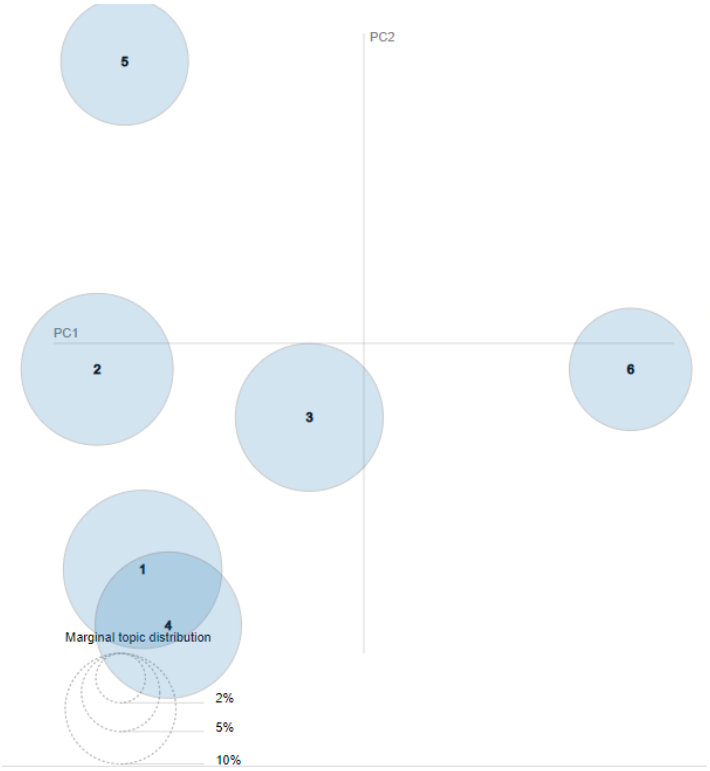


Figure 23 Visualisation of intertopic distance between the eight topics specified

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Day	Posted	Regent	Like	New	Food
Happy	Photo	Oxford	Good	Shop	Tea
Time	London	Soho	People	Fashion	Lunch
Life	Night	Bar	Today	Store	Menu
Birthday	Thank	Central	World	Book	Vegan

Table 1 Words of note for each topic

Looking at the key words of note that we have selected for each of the topics (table 1), we can define the six topics fairly clearly:

- Topic 1 focuses on good times and celebrations
- Topic 2 focuses on photography
- Topic 3 focuses on street observations
- Topic 4 focuses on positive emotions
- Topic 5 focuses on shop promotions and shopping
- Topic 6 focuses on food

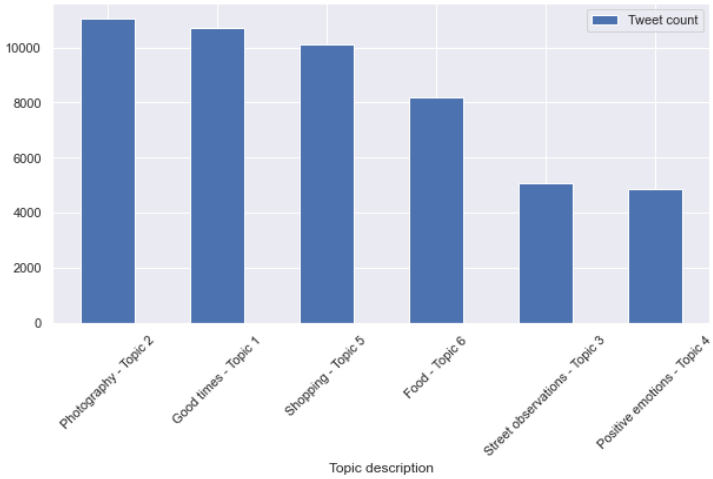


Figure 24 Bar plot showing distribution of tweets per topic

From the bar plot distributions shown in figure 24 above, it is not surprising to see topics such as photography, celebratory events, food and shopping being such a key component of the topic analysis given Oxford Street District's location. From a planning perspective, it is clear that the main ways of maintaining appeal for Twitter users is through a visually pleasing environment (e.g. high photography potential) with lots of retail and eating opportunities.

5.2.3 Spatial analysis when looking at all tweets

Next, we look to analyse how the tweets are clustered by various regions within the Oxford Street District. Based on our analysis shown in figure 25 below, we determine the elbow to be eight clusters.

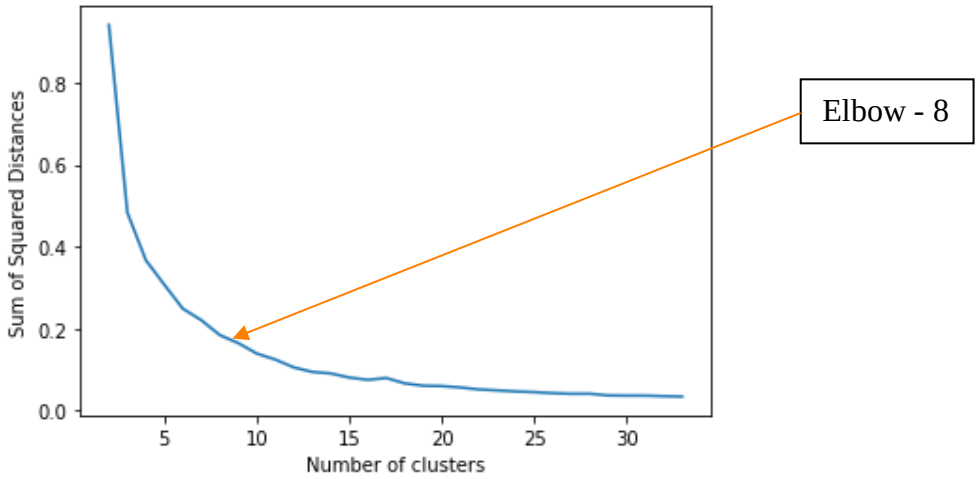


Figure 25 Graph visualising the elbow method for all tweets

Visualising these clusters in heat map form in figure 26 below (with the number of tweets linked to that cluster signalling its intensity), we can see that there are a particularly large number of tweets in the Carnaby Street region. This is no surprise given the high number of eateries in the regions. The second most clustered region was Regents Street, while other obvious clusters include Marble Arch (visual landmark) and the area near Oxford Street/Bond Street

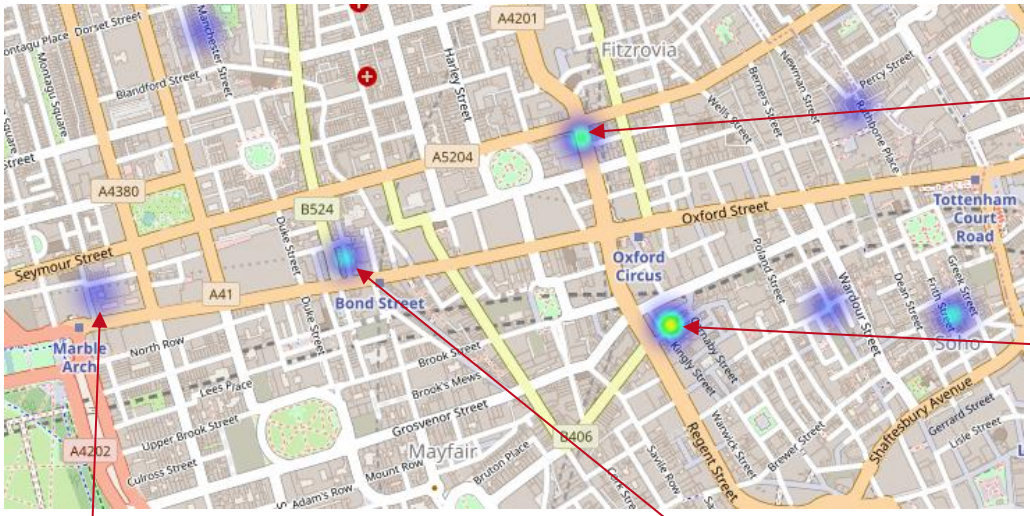


Figure 26 Heatmap of clusters

From an urban analysis perspective, it is no surprise that tweet interest is focussed on key hospitality and urban landmark areas. In particular, we can see that the likes of Carnaby Street, which have a high number of restaurants, shops and visually pleasing photo opportunities are a particularly interesting area for Twitter users in the Oxford Street District.

5.3 Analysis of just tweets with positive sentiment

We now analyse just tweets with a positive sentiment. As we already analysed positive sentiment from a temporal perspective in section 5.2, we focus our analysis on topic and spatial analysis.

5.3.1 Topic analysis when looking at just positive tweets

Again, we will apply latent dirichlet allocation but this time only on tweets with a positive sentiment. First to illustrate frequent terms, the word cloud in figure 27 below heavily features positive terms such as love, amazing as well as other activities associated with positive wellbeing such as music.

Figure 27 Word cloud for just tweets with a positive sentiment

Based on the latent dirichlet allocation analysis undertaken, we identify four core topics of interest with a strong intertopic distance between each of them as visualised in figure 28 below.

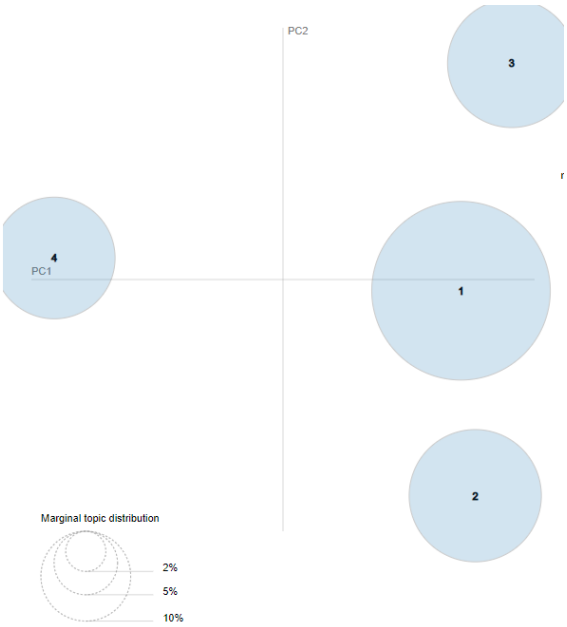


Figure 28 Intertopic distance visualisation for the dour topics identified

Topic 1	Topic 2	Topic 3	Topic 4
Day	Food	Street	Shop
Great	Good	Love	Art
Time	Lunch	Christmas	Beauty
Amazing	Favourite	Beautiful	Hair
Happy	Coffee	Club	Available

Table 2 Selection of key words for each of the four topics

Highlighting words of note for each topics relevant word list (key words shown in table 2 above), we categorise each topic as follows with the subsequent distribution shown in figure 29 below:

- Topic 1 focuses on positive events.
- Topic 2 focuses on food.
- Topic 3 focuses on street positivity.
- Topic 4 focuses on art/shopping/promotions

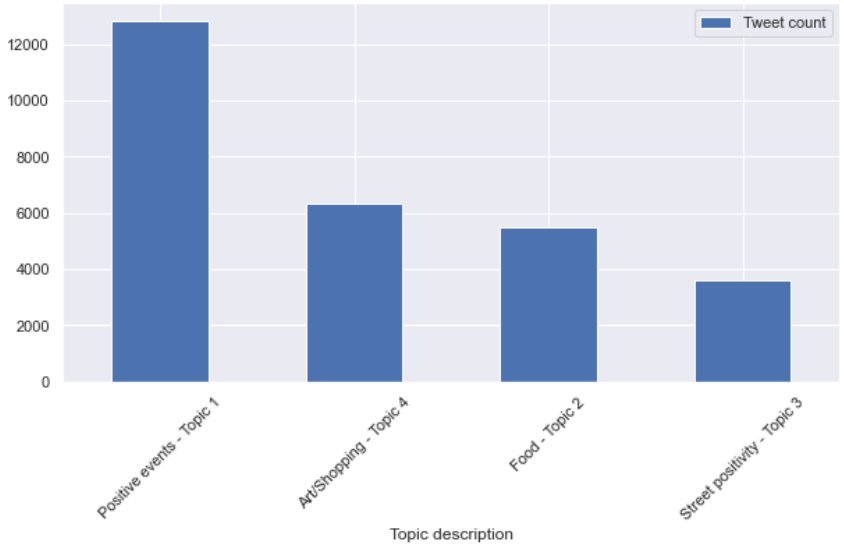


Figure 29 Bar plot showing distribution of positive tweets per topic

From an urban planning perspective, it is no surprise to see overall that positive events and emotions are the major topic of interest given the dataset, although more usefully we can still see plenty of appreciation for art, shopping food and the street at large which are not surprisingly linked to positive wellbeing. Therefore we would highlight these as being some of the key focuses to keep wellbeing high.

5.3.2 Spatial analysis when just looking at positive tweets

Based on the elbow method when just applied to positive tweets (figure 30) and given that eight clusters were used for the whole dataset, we decide on eight spatial clusters again to provide a

suitable comparison. As we can see, the sum of squared distances will be lower as a reflection of the lower volume of data that has been measured.

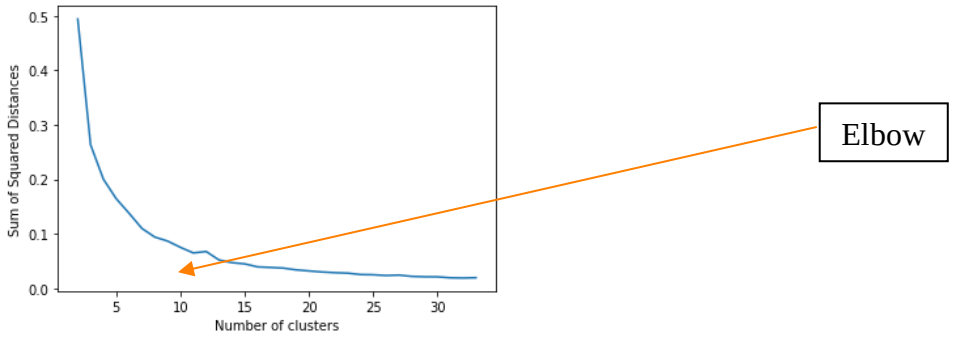
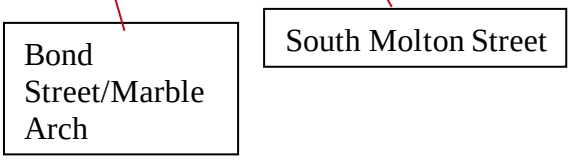


Figure 30 Line graph visualising the elbow method for all positive tweets

Visualising these clusters on a map (figure 31), we again see a strong tweet presence near Carnaby Street, while other highlights include a slight higher emphasis on shopping venues such as the introduction on a centroid at South Molton Street and a particularly high centroid concentration at Soho. Meanwhile, there is now a centroid focussed on Harley Street, an area known for its medical treatment.

From an urban planning perspective, it again highlights that particular areas of Twitter satisfaction when in the Oxford Street District are based around retail and hospitality but less evidence of satisfaction at visually pleasing locations, although this may reflect a limitation of the data with photos on Twitter that only have a brief message with it unlikely to be captured as positive sentiment.

Figure 31 Heat map of positive tweet clusters



5.4 Analysis of tweets with a negative sentiment

Next, we analyse tweets with just a negative sentiment, again focusing just on topic and spatial analysis.

5.4.1 Topic analysis when only looking at tweets with a negative sentiment

Despite a small dataset of negative tweets, we have identified four clear topics of conversation (figure 32).

Figure 32 Intertopic distance visualisation for negative tweets

Topic 1	Topic 2	Topic 3	Topic 4
Sad	Ill	Bar	Hard
Street	Mask	Club	Missed
People	Missed	Don't	Photo
Miss	Radio	Stop	Disappointed
Protest	BBC	Bloody	Stop

Table 3 Selection of key words from each of the four topics

Based on relevant words for each topic (table 3 shows a number of key words), we derive the following four topics.

- Topic 1 focuses on street issues.
- Topic 2 focuses on personal health/political concerns from radio tweets.
- Topic 3 focuses on nightlife complaints.
- Topic 4 centres on personal complaints.

From the distribution of topics in figure 33 below and from an urban planning perspective, we can see that street issues are the most common, which contain tweets related to the high profile Extinction Rebellion and Anti-Lockdown protests. There is also clear evidence of health concerns from the population at large with the word "mask" relevant for topic 2, indicating that perhaps there are health concerns when people are in the Oxford Street District. For example, people not wearing masks or not enough ventilation.

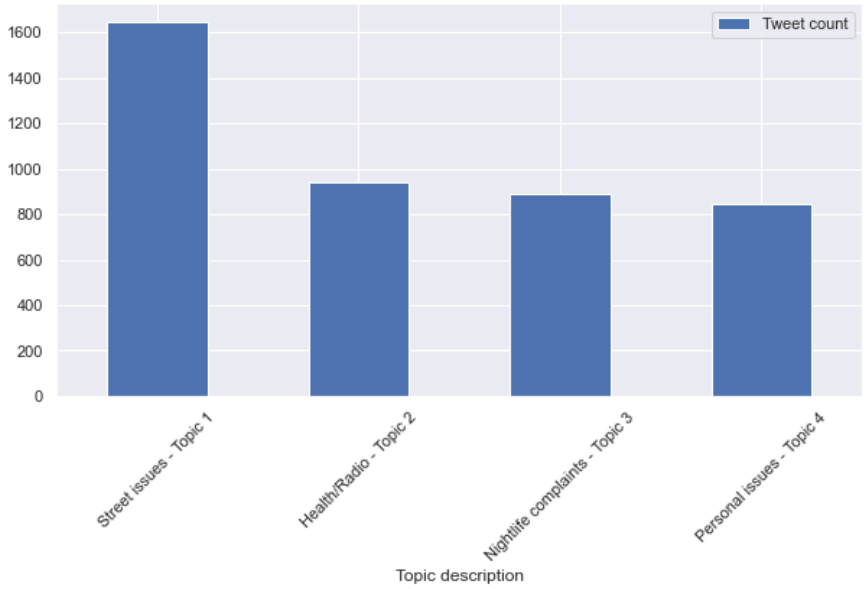


Figure 33 Bar plot showing distribution of negative tweets per topic

5.4.2 Spatial analysis of negative tweets

Turning to spatial analysis of negative tweets, from figure 34 we again see a much lower sum of squared distances testament to the much lower dataset. Based on the elbow method, we decide on seven clusters with very little change through the introduction of further clusters.

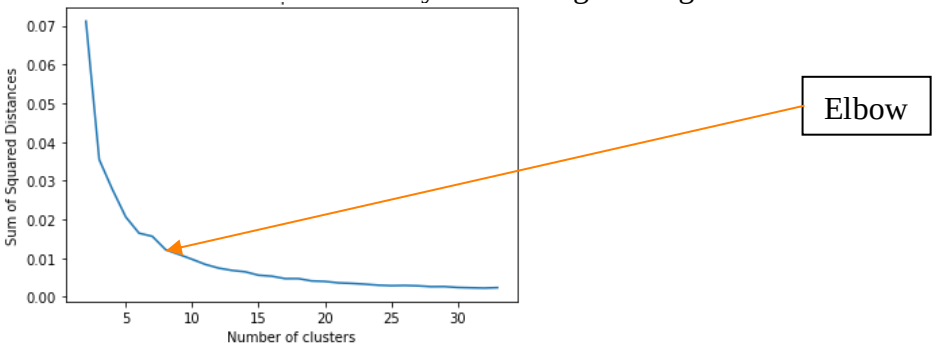


Figure 34 Line graph visualising the elbow method for all negative tweets (k = number of clusters)

Plotting this on the map below (figure 35), there is limited significant difference between negative and positive tweets from a spatial perspective. However, one of the clusters is now situated near Manchester Square, potentially suggesting there are complaints around this area, while the highest concentrated location of tweets is near Oxford Circus, whereas for the whole dataset and just positive tweets we observed the most frequent cluster was near Carnaby Street.

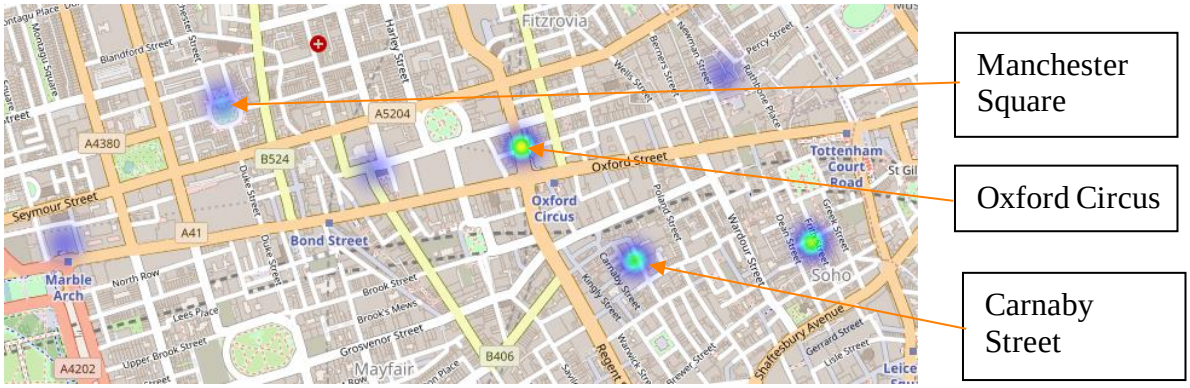


Figure 35 Spatial map of the seven clusters based on negative tweets

5.5 Analysis of smell tweets

Next, we present the results from just tweets that have been identified as having a smell characteristic. As these tweets are dominated by food, topic analysis proved to have limited use so instead we focus our analysis on temporal and spatial factors.

5.5.1 Temporal analysis of smell tweets

After appropriate lexicon based filtering outlined in the approach section, we are left with 6,964 tweets that relate to smell. Like when analysing all tweets, we can see clear drop off in tweet frequency (figure 36) after the onset of the COVID pandemic, once again, this recovers slightly in the summer period before regressing again in the winter period of 2021.

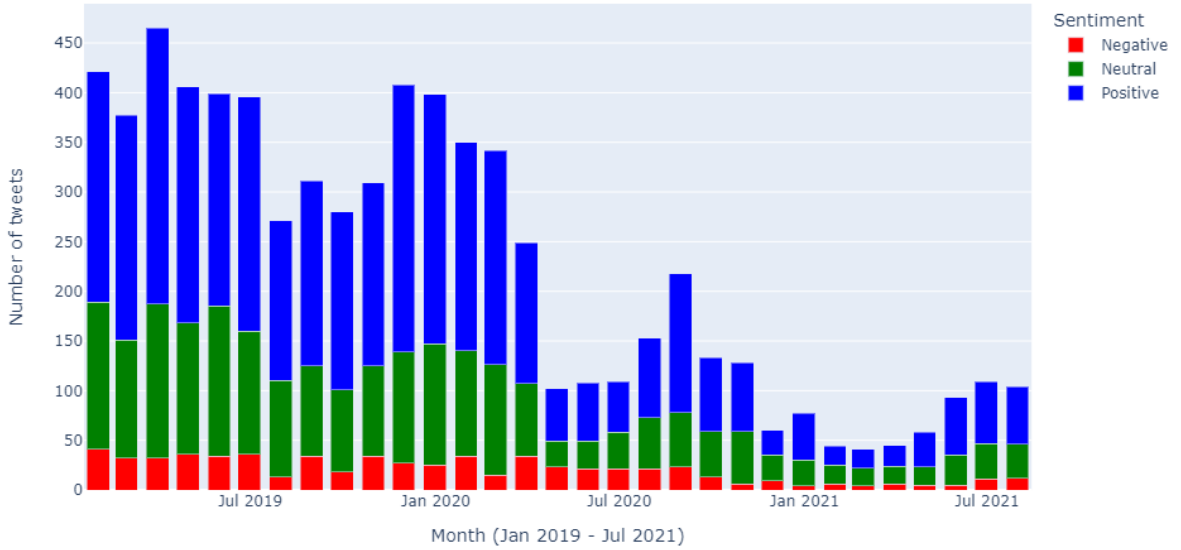


Figure 36 Bar plot showing monthly smell tweets by sentiment

Turning to proportional changes, we can see (figure 37 below) that the proportion of positive smell tweets is much higher compared to the whole dataset (circa 60% vs circa 55%). However, again we see clear negative spikes towards the start of the COVID pandemic and the winter lockdown.

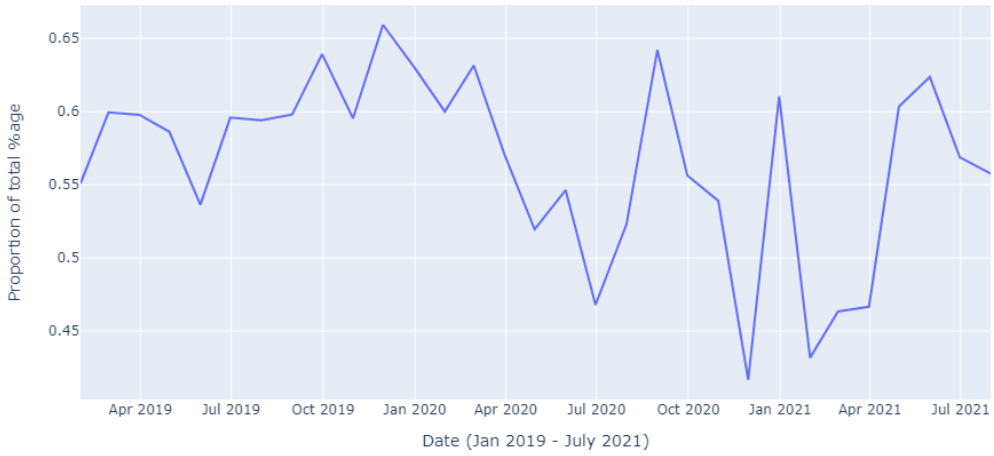


Figure 37 Line graph showing change in positive sentiment for smell tweets

As for the negative tweet proportion (figure 38 below), we can see a real spike negative tweets at the onset of the COVID pandemic and into the summer months, however this stabilises again at a proportion of circa 10% of the total. This is roughly in line with the total tweets, where excluding any significant spikes experience roughly 5-10% negative tweets of the overall total.

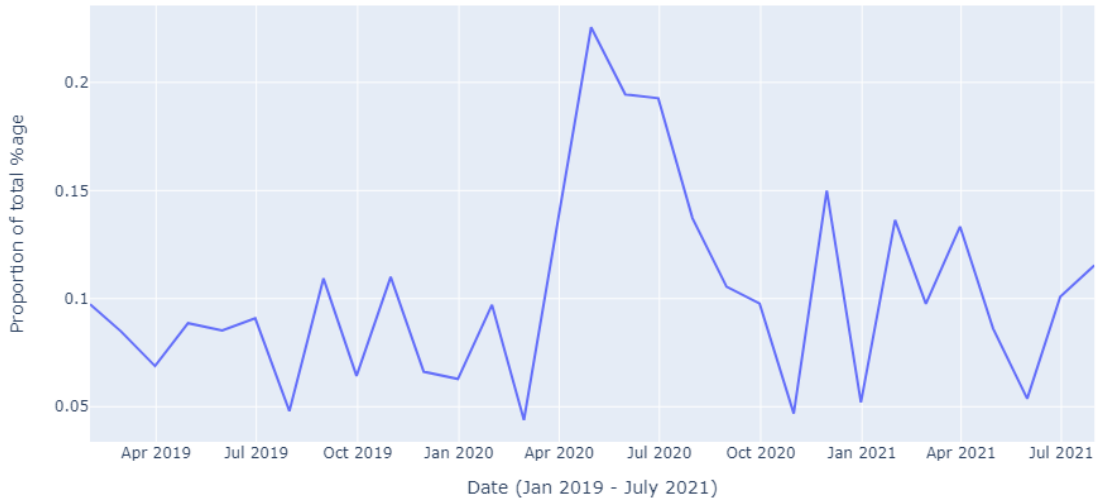


Figure 38 Line graph showing change in negative sentiment for smell tweets

From an urban planning perspective, these changes after the start of COVID pandemic are harder to interpret. Figure 39 below shows a word cloud for smell tweets, which displays a clear emphasis on food and not on areas where we suspect there to be more negative sentiment like politics, public safety and healthcare fears. Instead, we suspect the reason for these changes relates to the fact that tweets can discuss more than one topic and the smaller number of tweets per month post-pandemic means larger fluctuations are more likely. In addition, sentiment may have been affected by closure of food venues.

Figure 39 Word cloud of smell tweets

5.5.2 Spatial analysis of smell tweets

From figure 40, we decided on the number of clusters to be seven based on the elbow method. Looking at the tweet clusters in figure 41, we can observe the core concentrations of tweets were on Oxford Street, Carnaby Street and Soho.

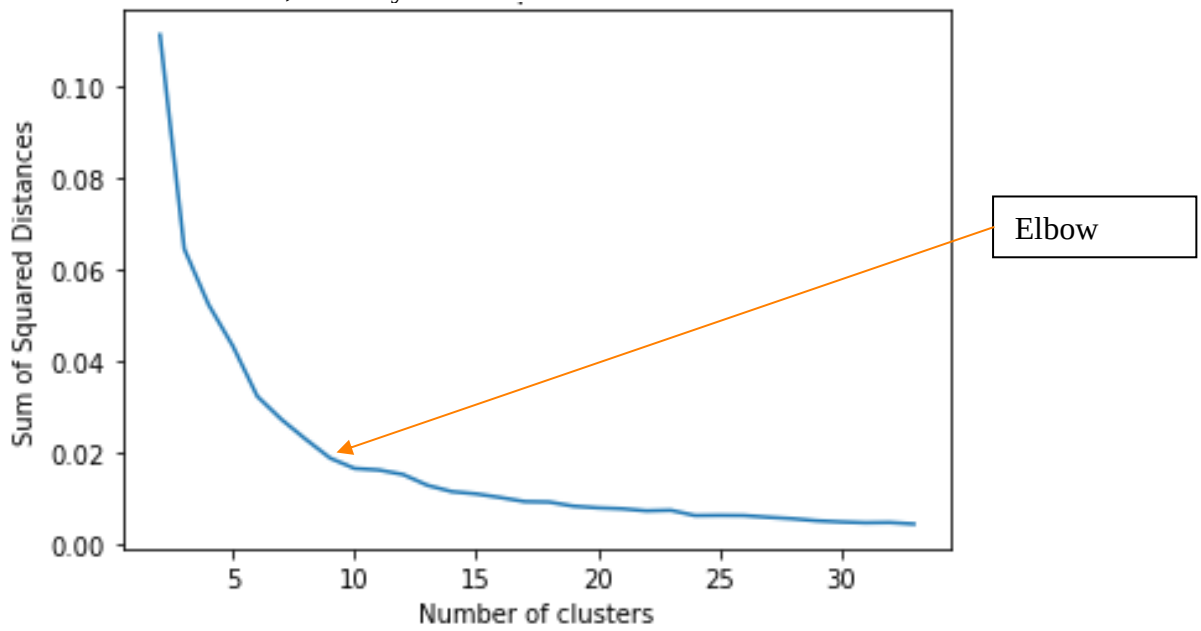


Figure 40 Graph of elbow method for smell tweets

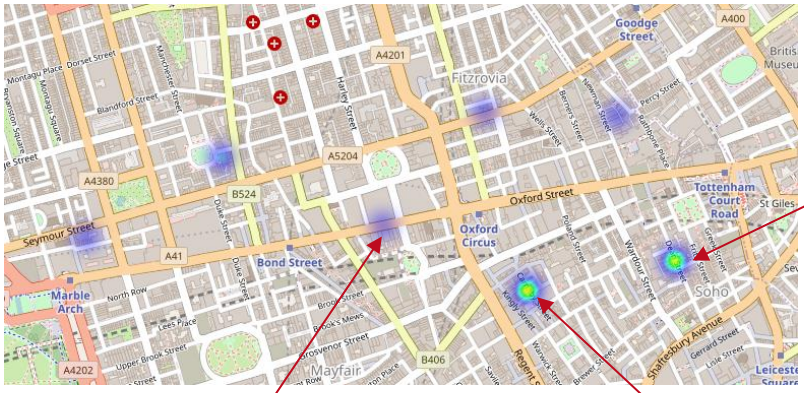


Figure 41 Spatial map of the smell clusters

Most of the centroids identified in figure 41 are to be expected such as Carnaby Street, given the regions and the fact that this sub-section of the data focuses heavily on food. Meanwhile other sectors such as landmarks have less of a focus.

5.6 Analysis of sound tweets

Next, we present the results from tweets that have been identified as having a sound characteristic. We analyse this data from a sentiment and spatial perspective.

5.6.1 Temporal analysis of sound tweets

After appropriate lexicon-based filtering, we are left with 14,580 relevant tweets. As can be seen by figure 42 below, there is again a clear drop in total tweets from the onset of COVID, with this decreasing further during lockdown of 2021 and only recently showing signs of recovery.

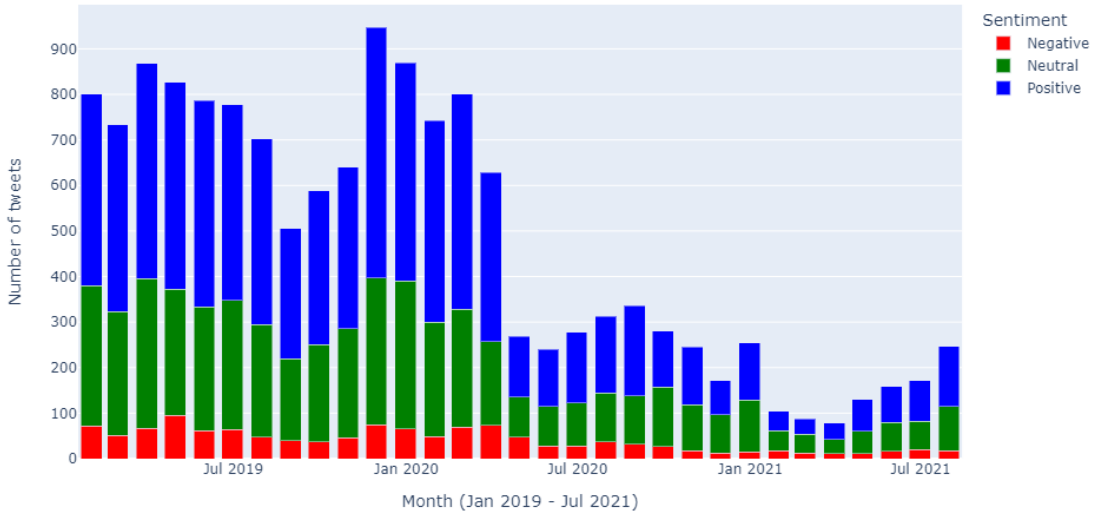


Figure 42 Barplot of sound tweets by sentiment on a monthly basis

Turning to changes in positive sentiment (figure 43), we can see a general trend of positive tweets hovering around 55%, slightly lower than when assessing all tweets. Again, this falls during the initial COVID period before falling further during winter lockdown. Some stability is now being shown.

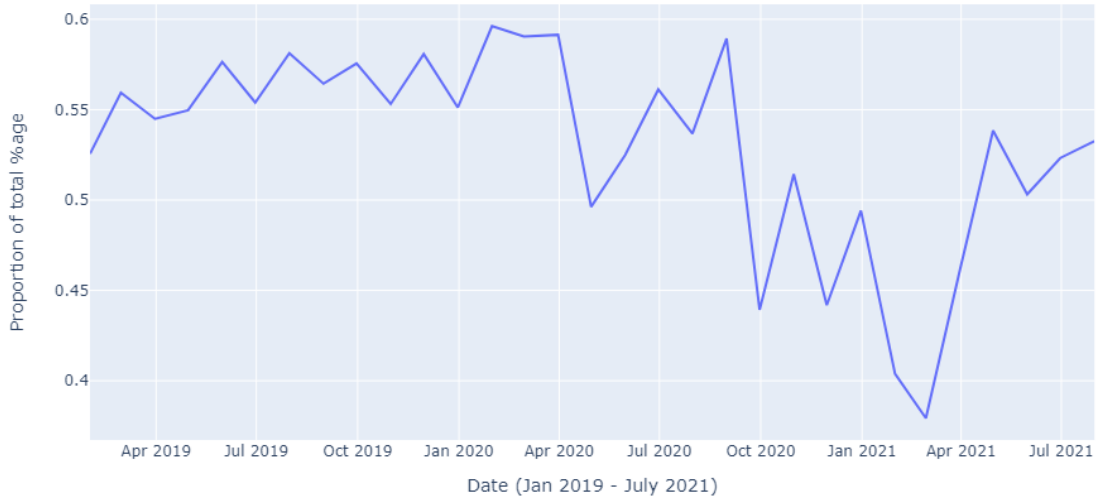


Figure 43 Line graph showing change in positive sentiment for sound tweets

Likewise, for negative tweets (figure 44) hover around 6-10% of the total before spiking at the start of COVID, before recovering and again and then finally spiking one more time during the winter lockdown.

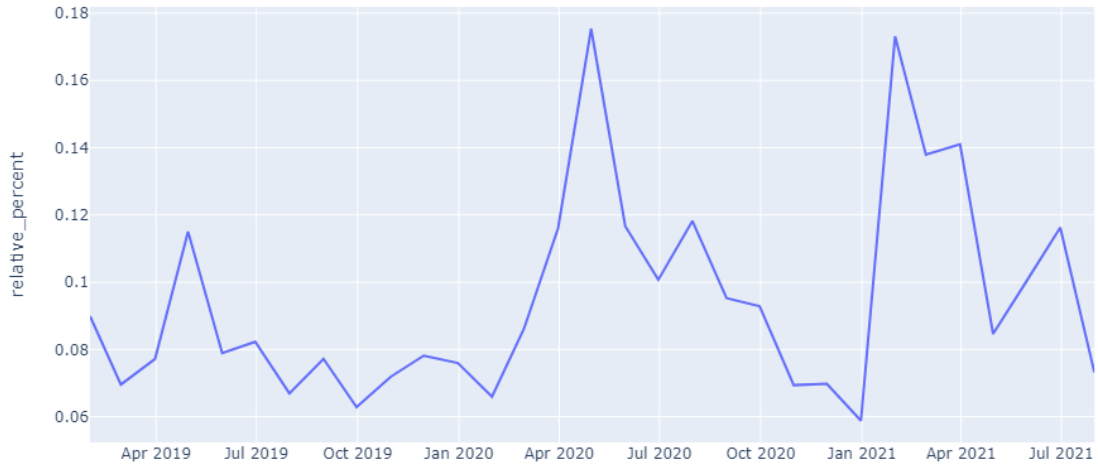


Figure 44 Line graph showing change in negative sentiment for sound tweets

Like in the case of smell, these changes after the start of COVID pandemic are harder to interpret. Figure 45 below shows a word cloud for sound tweets, which displays a clear emphasis on music. Again, we suspect the reason for these changes relates to the fact that tweets can discuss more than one topic and the smaller number of tweets per month post-pandemic means larger fluctuations are more likely. However, we suspect the first spike at the start of the pandemic will have also been influenced by the closure of music and nightlife venues.

Figure 45 Word cloud for sound tweets

5.6.2 Spatial analysis of sound tweets

From figure 46 and using the elbow method, we decided on the number of clusters to be six.

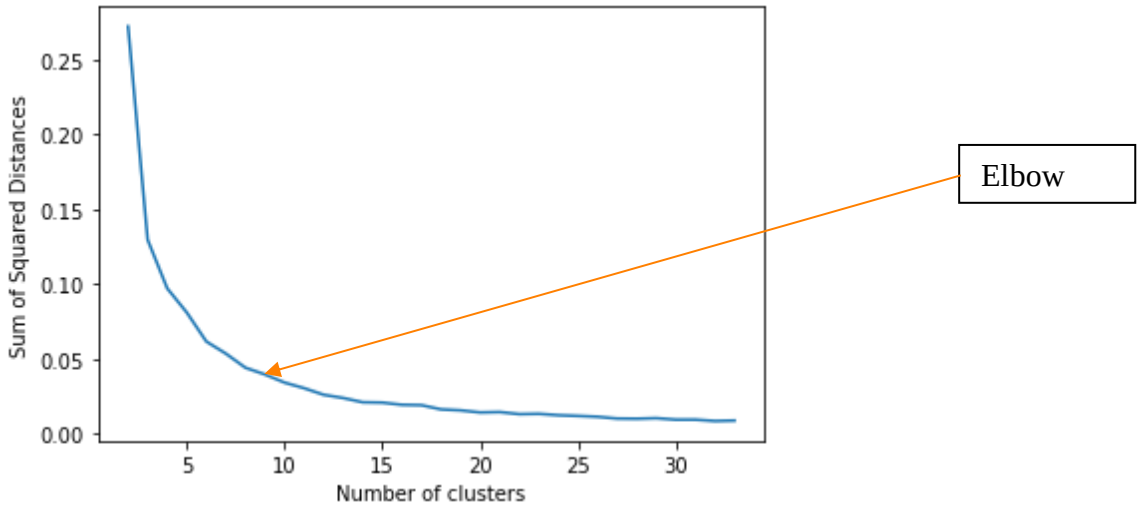


Figure 46 Graph of elbow method for sound tweets

Looking at the spatial clusters in figure 47, we can see a much lower emphasis on Carnaby street compared to looking at smell tweets. In addition, there is now a high concentration in Soho Square, just off Oxford Street, where there is a high concentration of nightlife venues. Other venues of note include The Wallace Collection by Manchester Square, Marble Arch and central Oxford Street, while like when analysing smell tweets, there is no centroid close to Oxford Circus.

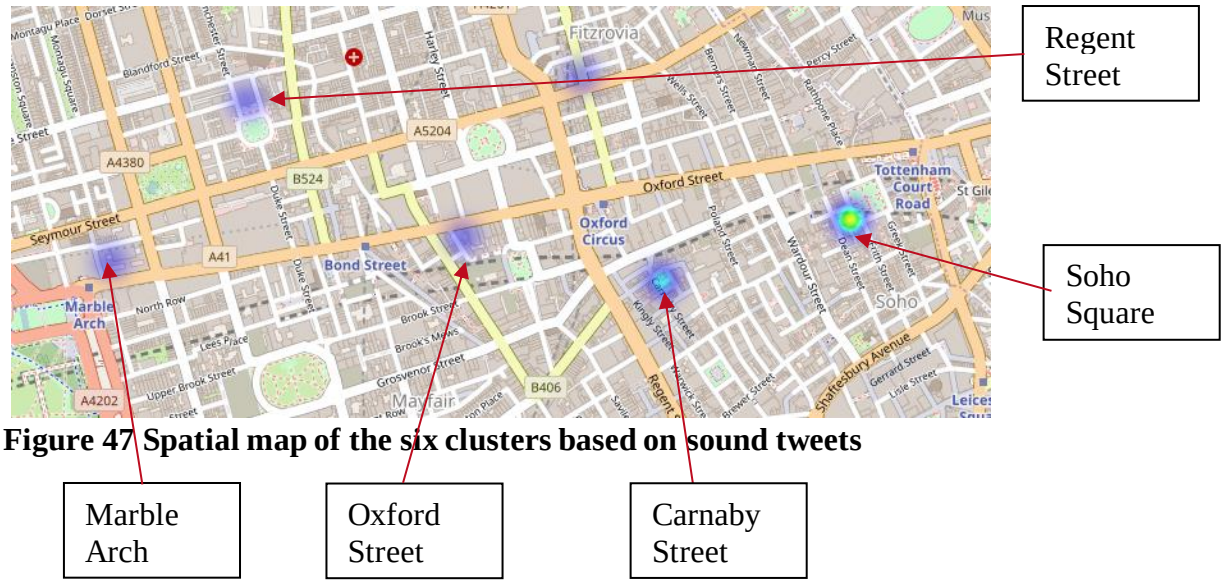


Figure 47 Spatial map of the six clusters based on sound tweets

From an urban planning perspective, this spatial analysis gives a clear insight into the hotspots for music and nightlife activities.

5.7 Pre-COVID versus post-COVID topics

Finally, we compare topic analysis for the pre-COVID versus post-COVID environment. This will give us a good understanding how health concerns and lockdown conditions have changed conversational points of interest.

5.7.1 Pre-COVID topic analysis

From just looking at tweets before the onset of COVID, we have identified four core topics of interest, with a reasonable intertopic distance between them (figure 48).

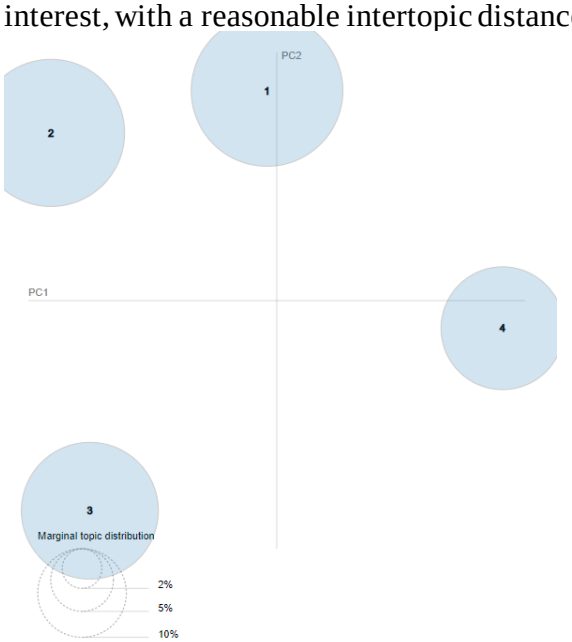


Figure 48 Intertopic distance visualisation between the four topics

Topic 1	Topic 2	Topic 3	Topic 4
Food	Thank	New	Street
Time	Great	BBC	Blood
Good	Amazing	Join	Club
Like	Thanks	March	Store
Best	Beautiful	Radio	Party

Table 4 Selected key words for each of the four topics

Based on table 4, we summarise the four key topics as follows:

- Topic 1 focuses on celebrations and food.
- Topic 2 focuses on positive emotions.
- Topic 3 focuses on politics and radio tweets.
- Topic 4 focuses on street activities

From the distribution of topics in figure 49 we can see that positive things like events, street activities and emotions were the main driver of tweets before COVID, alongside discussions around politics and radio.

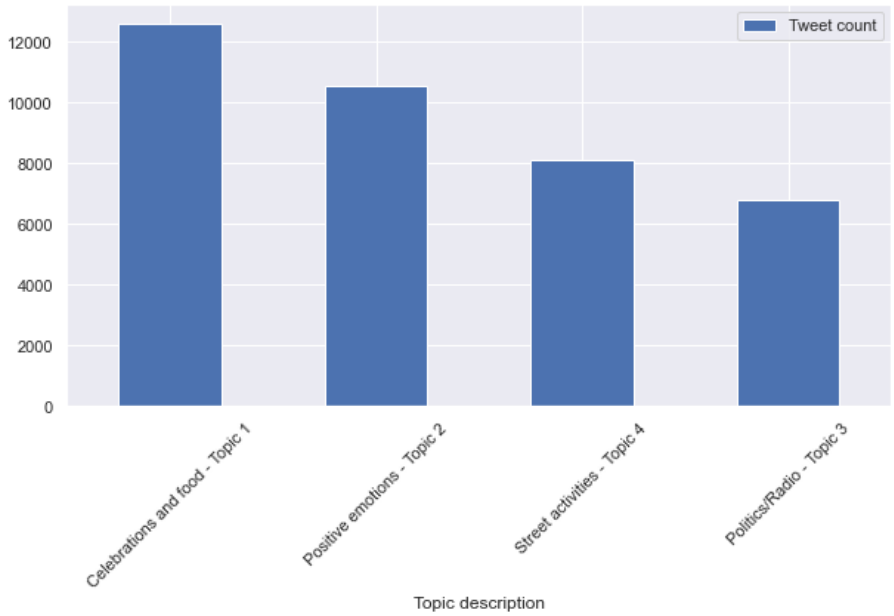


Figure 49 Tweet distribution per topic for pre-COVID

5.7.2 Post-COVID topic analysis

With the low data count from just looking at tweets since the outbreak of COVID, it is not surprising that only three core topics can be clearly identified from latent dirichlet allocation (figure 50).

Figure 50 Intertopic distance between the three topics

Topic 1	Topic 2	Topic 3
New	Like	Posted
Today	Good	Street
Thank	Shop	Photo
Radio	Love	New
BBC	People	Bar

Table 5 Selected words for the three topics

Based on table 5, we summarise these three topics as follow and plot their distribution (figure 51):

- Topic 1 focuses on positive emotions and radio tweets
- Topic 2 focuses on shopping and shop promotions
- Topic 3 focuses on street activities and photography

From an urban planning perspective the noticeable change is the reduced focus on food and celebration events. This is not surprising given the lockdown period. Meanwhile, since the start of the pandemic, shopping has become a more noticeable topic, driven both from actual shoppers and stores maintaining social media activity levels to entice shoppers back to Oxford Street District when possible.

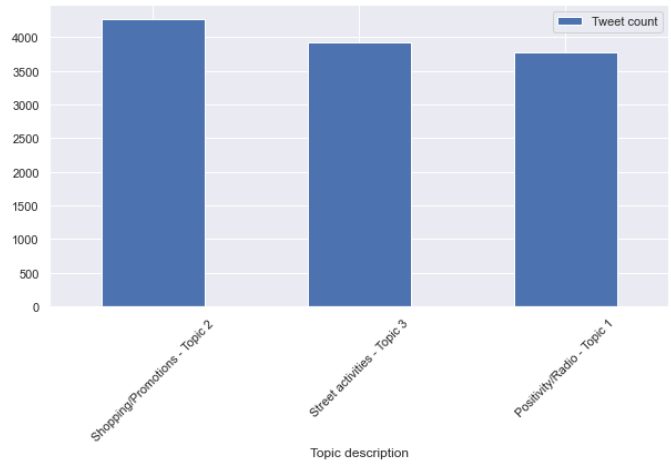


Figure 51 Bar plot showing distribution of post-COVID tweets per topic

5.7.3 Pre-COVID versus post-COVID spatial analysis

Finally, we compare if there are differences between cluster hot spots before and after the COVID pandemic outbreak. First looking at pre-pandemic, from figure 52, we devise a suitable cluster number of nine for pre-pandemic, while for post-pandemic (figure 53) we also assign a cluster number of nine to give an equal comparison between the two.

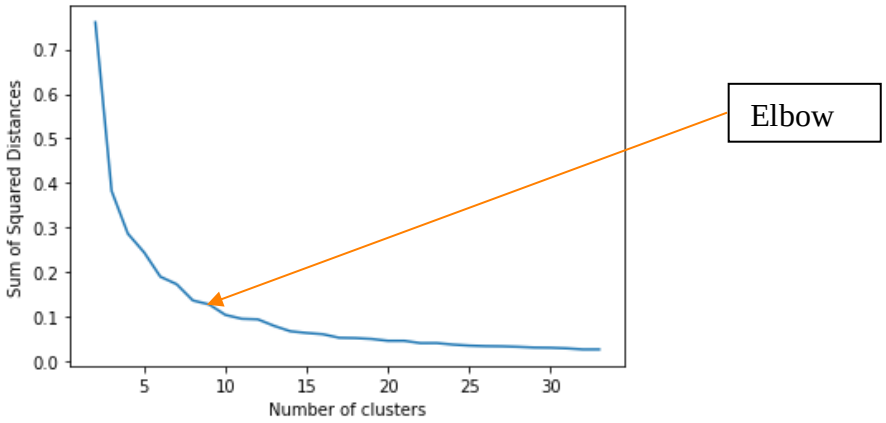


Figure 52 Graph for elbow method for pre-pandemic tweets

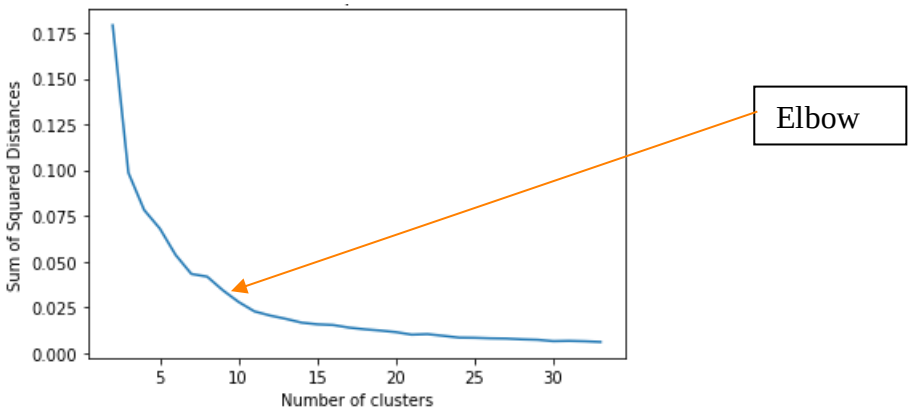


Figure 53 Graph for elbow method for post-pandemic tweets

Next looking at the actual spatial clustering, like when looking at the whole dataset, we can observe that pre-pandemic (figure 54), tweets were heavily concentrated in high profile shopping and food areas (e.g. Carnaby Street) and places of interest (e.g. Marble Arch).

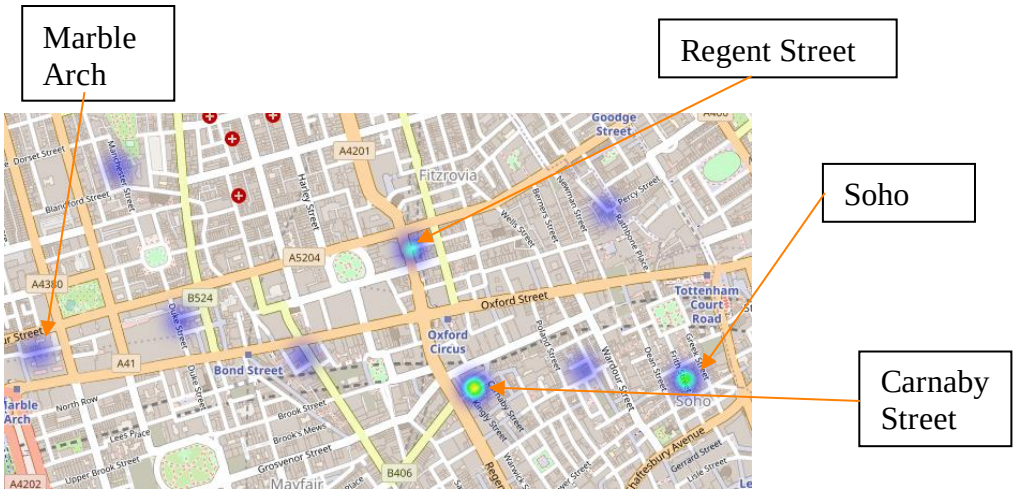


Figure 54 Map showing spatial clusters based on tweets pre-pandemic

However, when looking at spatial clusters post-pandemic (figure 55), centroids are less focussed on the core areas of Bond Street and Oxford Street and no concentration now near Marble Arch. This shows that tweets have become more spread out since the COVID pandemic. This is likely driven by a much lower proportion of tweets coming from visitors and other tourists.

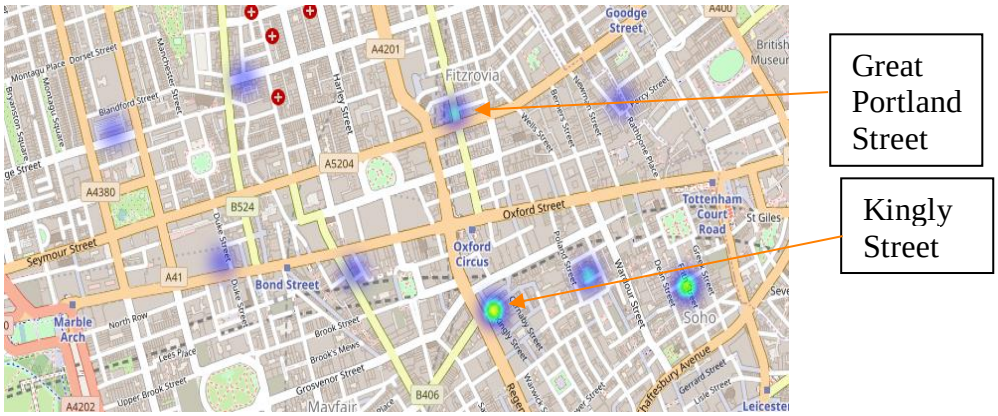


Figure 55 Map showing spatial clusters based on tweets post-pandemic

5.8 Discussion

5.8.1 Results summary

The major objective of our project was to demonstrate the usefulness of twitter data to the application of understanding user perceptions and experiences of the Oxford Street District. From the analysis undertaken, we were able to garner a number of useful insights, which we can briefly summarise in table 6 below:

Analysis undertaken	**Key results**
Sentiment analysis	• There was a clear drop/increase in positive/negative sentiment after the start of the pandemic, which has only now shown signs of retracting back to pre-pandemic level. However, this trend mirrors other London regions such as Belgravia and suggests that there are not underlying planning decisions that have affected sentiment as well post or pre-pandemic. • When analysing tweets that are just related to sound and smell, despite again seeing clear changes in sentiment proportions just after the start of the pandemic, these have more or less returned to pre-pandemic levels.
Topic analysis	• When understanding the reasons for why most people on Twitter come to the Oxford Street District, the most common topics from latent dirichlet allocation included celebrations, photography and food. • When analysing just tweets with negative sentiment, there was indications of concern regarding public health, nightlife and street issues.
Spatial analysis	• When looking at the whole twitter dataset, we observe that tweet hot spots are centralised in typical areas well known for their visually pleasing look, food or retail amenities. Obvious examples of this include Carnaby Street. • When assessing the twitter landscape before and after the outbreak of the COVID pandemic, we observe that tweets have become less centrally focussed on the core tourist and visitor areas such as Bond Street and Oxford Street as well as landmarks such as Marble Arch.

Table 6 Key results from analysis

5.8.2 A note on limitations

The key limitation from the project is from the Twitter data itself. Sloan and Morgan (2015) found that Twitter users in the UK compared to the broader population are more likely to be younger with jobs in areas such as administration and professional occupations. Meanwhile, Twitter users have a lower representation in areas such as lower supervisory and manual labour and in older age groups. Therefore, the sentiment, views and spatial trends of the Twitter population do not necessarily correlated to the general public. In addition, there were found to be significant demographic differences (gender, age, class and language) between those who do and do not enable location services and those who do not geotag their tweets. This likely means that the views ascertained from geolocated data are not representative of the whole Twitter population.

5.8.3 Broader implications and discussion points from an urban analysis perspective?

Finally, we note any broader points that either relate to both urban planning or the techniques employed.

1. The effectiveness of the lexicon method

While some research has questioned the suitability of using lexicons, the project has proved some effectiveness in this method, with changes in positive and negative sentiment broadly mapping to macroeconomic factors such as the enforcement of lockdown policies. This is particularly useful as often for councils and urban planning groups, the use of a training data set to perform statistical machine learning based methods are often not easily available. As a result, the use of lexical methods offers a practical and more scalable method for urban planners to use.

2. Economic indication

At an area level, the use of twitter can be used to gain an insight into the economics of the region. Tweet volumes clearly dropped in line with other comparable datasets such as google trends and was in line with other macro factors such as the outbreak of COVID and the enforcement of lockdown policies. This indicator was fairly consistent both when using overall tweets, positive tweets as a proportion of the overall total or negative tweets as a proportion of the overall total.

This result is particularly profound as urban planners often struggle to gain an understanding as to how an area is performing currently in terms of user participation. Given the data from Twitter is updated in real time, this presents a potential solution to this problem.

3. A note on urban hotspots

Spatial analysis through the use of k-means++ analysis was able to highlight key user hotspots under certain criteria. This gives an indication as to where users are reacting more with their environment and identifies particular places which garner a lot of public interest. Going

forward, there are number of planning initiatives within the Oxford Street District such as the Marble Arch Mound. Therefore this type of spatial analysis can be used to understand if when these developments are completed, do we see a significant impact in terms of key urban hot spots from a Twitter perspective.

4. Topics of interest

While, in certain instances topics were overlapping, overall latent dirichlet allocation proved effective in drilling down into topic conversations. It is important to note that as with all machine learning solutions, outcomes are not always perfect with the technique better suited to giving simple insights into topics. Urban planners can subsequently complement this with supervised learning models to drill down more accurately into topics of particular interest after building a suitable training dataset.

6 LEGAL, SOCIAL, ETHICAL AND PROFESSIONAL ISSUES

During my project we have abided by all the rules appropriate the BCS Code of Conduct.

6.1 Public interest

We have ensured due regard for privacy of others and the environment as well as confined within all data protection laws. All data has been anonymised so no individual can be identified with the tweet or tweet details apart from if a relevant username is used. All analysis is based on competent machine learning and analysis techniques and we believe that there are no results, techniques or recommendations that discriminate on the grounds of sex, sexual orientation, marital status, nationality, colour, race, ethnic origin, religion, age or disability. Finally, all results and analysis are aimed to assist Westminster City Council's urban planning decisions, which includes aims to boost inclusivity and public wellness.

6.2 Professional competence and integrity

All work has been provided within our professional competence. We have valued alternative opinions on the project from third parties such as my supervisor and relevant members from Westminster City Council. Finally, we have not asked or used any information from any data providers and key stakeholders such as our project supervisor or Westminster City Council that extends beyond the scope of the project.

6.3 Duty to relevant authority

We have ensured that all processes have been carried with due care and diligence based on Westminster City Council's guidance, including no information from them being sent to third parties. No data has been used for personal guidance nor have we misrepresented any results or analysis from the project to take advantage of the lack of relevant knowledge from members of Westminster City Council. All code has been made available to Westminster City Council so that they have full control of any intellectual property generated from this project.

7 CONCLUSION

7.1 Conclusion overview

In conclusion, our findings appropriately answer the research questions answered and that the use of Twitter data does present novel insights that urban planners can use to improve their processes. Through sentiment, topic and spatiotemporal analysis of tweets relevant to the Oxford Street District, we have been able to identify significant patterns within the data and the implications of these has been explored in the discussion sections. In addition, these trends have been in some cases somewhat validated by other datasets such as Google trends, while matching general trends such as the enforcement of lockdown.

7.2 Future developments

To finish, we outline any future developments that we believe would enhance the current analysis undertaken

7.2.1 Expanding beyond Geocoded data

Firstly, the analysis focussed on the use of geocoded data, however other datasets could be extracted relevant to the project based on keyword or hashtag searches. This would prove to be an interesting comparison as Sloan and Morgan (2015) found significant demographic differences between those who do and do not geotag their tweets. However, the negative of this approach is that there is a risk that a significant proportion of tweets captured is not relevant for the Oxford Street District.

Keeping to the use of other Twitter datasets, similar sentiment and topic analysis processes could be used for when tweets are directed at Westminster City Council using the @ tag. This would prove to be useful in understanding the conversations users are directly interacting with Westminster City Council about and how the sentiment of those tweets changes over time.

7.2.2 The use of supervised learning for more accurate topic analysis

Moving away from the use of other Twitter datasets, while unsupervised topic analysis helps build an understanding of broad conversational points, this could be complemented with supervised machine learning solutions to focus more on topics of interests. For example, the issue of public safety and crime could be further examined to understand if this conversational issue is growing or falling in relevance.

In addition, other unsupervised learning techniques could be analysed for topic analysis such as a biterm model. Research on topic modelling for microtexts is still fairly nascent and as a result, there is still significant scope for further analysis in this area. However, to date latent

dirichlet allocation has a consistent track record of being applied to microblog texts successfully, we equally applied this method.
Turning to sentiment analysis, while a lexicon approach has been used for this approach, there is the potential for using supervised learning solutions. This would involve building a classification model based on a large training set. Unfortunately, due to the time requirements for this project this was not pursued and would offer a comparison to the lexicon method employed.

7.3 Distinguishing between user groups

Finally, this project does not distinguish between user groups within the Oxford Street District such as tourists, workers, institutions (e.g. radio stations and shops) and residents. Such distinctions would be able to garner new opportunities from a sentiment, spatial and topic analysis perspective. Previous examples of work splitting out residents and visitors includes classifying users based on their frequency of presence in a certain location such as Abbasi et al (2015) who identified these user types in Sydney for the purposes of city trip analysis.

8 REFERENCES

Abbasi, A., Rashidi, T., Maghrebi, M., & Waller, T. (2015). Utilising location based social media in travel survey methods: Bringing Twitter data into the play. *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 1-9.

Abrahams, A., Coupey, E., Zhong, E., Barkhi, R., & Manasantivongs, P. (2013). Audience targeting by B-to-B advertisement classification: A neural network approach. *Expert Systems with Applications*, 2777-2791.

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science and Engineering*, 72–82.

Arjona Osorio, J., Jiri, H., Radek, S., & Yolanda, G.-R. (2021). Social media semantic perceptions on Madrid Metro system: Using Twitter data to link complaints to space. *Sustainable Cities and Society*, 102530.

Askerov, R., Kwon, W., Song, L. M., Weber, D., Schaer, O., Dadgostari, F., & Adams, S. (2020). Natural Language Processing for Company Financial Communication Style. *2020 Systems and Information Engineering Design Symposium*, 1-6.

Aubrecht, C., Ungar, J., & Freire, S. (2011). Exploring the potential of volunteered geographic information for modeling spatio-temporal characteristics of urban population. *Proceedings of the 7th International Conference on Virtual Cities and Territory*, 57-60.

Bertrand, K., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). Sentiment in New York City: A High Resolution Spatial and Temporal View.

Chen, M., Arribas-Bel, D., & Singleton, A. (2018). Understanding the dynamics of urban areas of interest through volunteered geographic information. *J. Geogr. Syst*.

Cranshaw, J., Schwartz, R., Hong, J., & Sadeh, N. (2012). The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *AAAI Conf. on Weblogs and Social Media*.

Durahim, A., & Coşkun, M. (2015). #iamhappybecause: Gross National Happiness through Twitter analysis and big data. *Technological Forecasting and Social Change*, 92-105.

Edward, A. (2021, June 13). *An Extensive Guide to collecting tweets from Twitter API v2 for academic research using Python 3*. Retrieved from Medium: https://towardsdatascience.com/an-extensive-guide-to-collecting-tweets-from-twitter-api-v2-for-academic-research-using-python-3-518fcb71df2a

Eisenstein, J. (2013). What to do about bad language on the internet. *Proceedings of NAACL-HLT*, 359-369.

Frias-Martinez, V., & Frias-Martinez, E. (2014). Spectral Clustering for Sensing Urban Land Use using Twitter Activity. *Engineering Application of Artificial Intelligence*, 237–245.

Hidayatullah, A., & Muhammad, M. (2017). Road traffic topic modeling on Twitter using latent dirichlet allocation. *2017 International Conference on Sustainable Information Engineering and Technology*, 47-52.

Hoffman, M., Blei, D., & Bach, F. (2010). Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 856-864.

Hollander, J., & Renski, H. (2015). Measuring Urban Attitudes Using Twitter: An Exploratory Study. *Lincoln Institute of Land Policy*.

Kramer, A. (2010). An unobstructuve behavioral model of "Gross National Happiness". *Proceedings of the 28th ACM CHI*, 10-15.

Lansley, G., & Longley, P. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 85-96.

MacQueen, J. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Marta, S., Timothée, G., & Hugues, P. (2015). Twitter data for urban policy making: an analysis on four European cities.

Mimno, D., Wallack, H., Talley, E., Leenders, M., & Mccallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 262-272.

Mitchell, L., Frank, M., Harris, K., Dodds, P., & Danforth, C. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE*.

Nielsen, F. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.

Noulas, A., Mascolo, C., Scellato, S., & Pontil, M. (2011). An Empirical Study of Geographic User Activity Patterns in Foursquare. *Proceedings of the Fifth International AAAI COnference of Weblogs and Social Media*.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). Tracking "Gross Community Happiness" from Tweets. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 965-968.

Quercia, D., Schifanella, R., Aiello, L., & McLean, K. (2015). Smelly Maps: The Digital Life of Urban Smellscapes.

Rossano, S., Daniele, Q., Francesco, A., & Luca, A. (2016). Chatty Maps: Constructing sound maps of urban areas from social media data. *Royal Society Open Science*.

Schwartz, A., Dodds, P., Jarlath, O.-D., Danforth, C., & Ricketts, T. (2019). Visitors to urban greenspace have higher sentiment and lower negativity on Twitter. *People Nat*, 476-485.

Severo, M., Giraud, T., & Pecout, H. (2015). Twitter data for urban policy making: an analysis on four European cities. *Handbook of Twitter for Research*.

Sim, J., Miller, P., & Swarup, S. (2020). Tweeting the High Line Life: A Social Media Lens on Urban Green Spaces. *Sustainability*, 12-21.

Sloan, L., & Jeffrey, M. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS ONE*.

Theodore, D. (2006). Sense of the city: An alternative approach to urbanism. *Journal of Architectural Education*, 69-70.

Tu, W., Cao, J., Yue, Y., Shaw, S., Zhou, M., Wang, Z., . . . Li, Q. (2017). Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science*, 2331 - 2358.

Weiler, A., Grossniklaus, M., & Scholl, M. (2016). Situation monitoring of urban areas using social media data streams. *Information Sytems*, 129-141.

Wessel, G., Karduni, A., Sauda, E., Cho, I., & Dou, W. (2017). Urban Space Explorer: A Visual Analytics System for Urban Planning. *IEEE Computer Graphics and Applications*, 50-60.

Xiaodong, C., MacNaughton, P., Deng, Z., Yin, J., Zhang, X., & Allen, J. (2018). Using Twitter to Better Understand the Spatiotemporal Patterns of Public Sentiment: A Case Study in Massachusetts, USA. *International Journal of Environmental Research and Public Health*.

Xue, J., Junxiang, C., 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*.

Zhang, Z., Ni, M., He, Q., & Gao, J. (2016). Mining Transportation Information from Social Media for Planned and Unplanned Events.

9 APPENDICES

9.1 APPENDIX A

```
# Raw dataset is read in
twitter_df = pd.read_csv("data.csv")

# Basic pre-processing
# Drop line if not English
twitter_df = twitter_df[twitter_df.lang == 'en']
# Drop duplicates
twitter_df = twitter_df.drop_duplicates()
# Drop any NA texts
twitter_df['text'] = twitter_df['tweet'].dropna()
# Convert to lower case
twitter_df['text'] = twitter_df['text'].str.lower()
# Apply preprocessing package - suports cleaning of
# URLs, hashtags, mentions, reserved words, emojis and smileys
twitter_df['text'] = twitter_df['text'].apply(lambda x: p.clean(x))
# Get rid of none-letter characters
twitter_df['text'] = twitter_df['text'].apply(lambda x: re.sub(r"[^a-z\s\(\-:\)\\\/\];='#]", '', x))
# Lemmatization and stemming
# Define lemmatizer and stemmer
lmt = WordNetLemmatizer()
stemmer = SnowballStemmer("english")
# Define function
def lemma_and_stem(strng):
    '''Word stemmer; find the root of the word. E.g. 'dogs' becomes 'dog'''
    strng = strng.lower()
    word = lmt.lemmatize(strng)
    word = stemmer.stem(strng)
    return word
# Apply function
twitter_df['text'] = twitter_df['text'].apply(lemma_and_stem)
# Remove all stopwords and tokenize in one
twitter_df['text'] = twitter_df['text'].apply(lambda x: remove_stopwords(x))
# Drop duplicates on text so only original left
twitter_df = twitter_df.drop_duplicates(subset=['text'])
# Finally tokenise text
# Processing done
```

9.2 APPENDIX B

```
# Now we add two extra columns for sound and smell respectively
# Starting with smell
# Read in smell dictionary
smell_lines = open("lexicons/smell_dictionary_eng.txt", "r")
# Build smell words list
smell_read_in_file = smell_lines.readlines()
temporary_list = []
for element in smell_read_in_file:
    temporary_list.append(element.replace('\t',', '))
smell_list = []
for element in temporary_list:
    smell_list.append(element.strip())
# Split into two lists for overall groupings and individual words
# We just want to see how many relate to smell
smell_words = []
perc_count = 0
for element in smell_list:
    if '%' in element:
        perc_count = perc_count + 1
    if perc_count >= 2:
        smell_words.append(element)
# See how many words have something of interest to the current tweets set
def get_rid_of_perc(data_list):
    new_list = []
    for element in data_list:
        if '%' not in element:
            new_list.append(element)
    return new_list
# Apply to both lists
smell_words = get_rid_of_perc(smell_words)
# Just create list with smell words in it
smell_words_final = []
for element in smell_words:
    element = re.sub(r'([^a-zA-Z ]+?)','', element)
    element = re.sub('\W+', '', element)
    smell_words_final.append(element)
```

```
# Now we do the same process for sound
# Read in sound dictionary
sound_lines = open("lexicons/sound_dictionary_eng.txt", "r")
# Build sound words list
sound_read_in_file = sound_lines.readlines()
temporary_list = []
for element in sound_read_in_file:
    temporary_list.append(element.replace('\t',', '))
sound_list = []
for element in temporary_list:
    sound_list.append(element.strip())
# Split into two lists for overall groupings and individual words
# We just want to see how many relate to sound
sound_words = []
perc_count = 0
for element in sound_list:
    if '%' in element:
        perc_count = perc_count + 1
    if perc_count >= 2:
        sound_words.append(element)
# See how many words have something of interest to the current tweets set
def get_rid_of_perc(data_list):
    new_list = []
    for element in data_list:
        if '%' not in element:
            new_list.append(element)
    return new_list
# Apply to both lists
sound_words = get_rid_of_perc(sound_words)
# Just create list with smell words in it
sound_words_final = []
for element in sound_words:
    element = re.sub(r'([^a-zA-Z ]+?)','', element)
    element = re.sub('\W+', '', element)
    sound_words_final.append(element)
```

```
# Now we add the actual sound and smell columns
# Add sound column
def sound(row):
    text = row['text']
    value = 0
    for element in text:
        if element in sound_words_final:
            value = value + 1
    if value == 0:
        sound = "No"
    else:
        sound = "Yes"
    return sound
twitter_df['sound'] = twitter_df.apply(sound, axis=1)
# Add smell
def smell(row):
    text = row['text']
    value = 0
    for element in text:
        if element in smell_words_final:
            value = value + 1
    if value == 0:
        smell = "No"
    else:
        smell = "Yes"
    return smell
twitter_df['smell'] = twitter_df.apply(smell, axis=1)
```