A dissertation submitted to the **University of Greenwich**
in partial fulfilment of the requirements for the Degree of

Master of Science
in
Data Science

Identification of tourist locations from geo-tagged social media posts

Name: Thushani Nipunika Kalamullage

Student ID: 001178245

Supervisor: Dr. Jia Wang
Submission Date: September, 2022
Word count: 10,036

IDENTIFICATION OF TOURIST LOCATIONS FROM GEO-TAGGED SOCIAL MEDIA POSTS

Thushani Kalamullage

Liberal Arts and Sciences, University of Greenwich, 30 Park Row, Greenwich, UK.

(Submitted 16 September 2022)

Abstract: At the present time, social media platforms revealed user behaviour and attitudes from their posts. The Twitter platform also plays a major role in the social media world. This project study is focusing extracting tourist attractions based on the London area using geo-enabled Twitter data. This project study proposed to get the dataset from Kaggle and presented the data pre-processing techniques, text cleaning techniques, frequency analysis techniques, sentiment analysis techniques, and Geographic Information System techniques based on spatial analysis. The proposed techniques were applied to peak and off-peak hours Twitter data in the London area and the outcome of this project is based on the 26th of July 2018 Twitter data. The expected outcome of this research is to extract tourist attraction locations and the visualized tourist intensity map. Apart from the intensity layer, tourist feelings about the tourist attraction are also considered. All the results will be embedded into a website.

Keywords: social media, Geotagged tweets, tourism, Natural Language Processing, Sentiment Analysis, Hot Spot analysis

Acknowledgements

I would like to express my gratitude to my supervisor Dr Jia Wang for guiding me and introducing this opportunity to do this project topic. I'm grateful for your great assistance throughout the project.

I am also particularly grateful for the extraordinary support of my husband who stands at my side both in good and bad moments.

I would like to take this chance to thank my colleagues who accompanied me during the research period.

Thank you.

Table of Contents

List of Figures

List of Tables

1 Introduction

The word itself social media plays a very important role in society, and more than half of the world population at least uses one of the popular social media platforms [3]. As per the numbers, it was 59% of the world population. Since it was embedded in people's lifestyles, those social media profiles themselves showcase the attitude and the social and physical behaviours of the profile owners. This is an inevitable truth, and the average person uses social media for 2.47 hours per day in 2022 with respect to most of the social media platforms featuring geotag posts for the users, such as users being able to tag the current location while posting [35]. This gives the chance to analyse those posts and the geotags and predict the behaviour of the users. In this proposed research project, we are expecting to use these data to create a tourist intensity layer of London. Tourist data is underrepresented in the location planning models because tourists' behaviour did not represent in the census. Most tourists are roaming around the tourist attractions. This project aims to identify tourist attractions through social media posts such as Twitter and develop a tourist intensity layer for London, based on the tourism-based tweets.

Among all the famous social media platforms, Twitter is selected in our scope due to the research-friendly data availability. Most of the social media analyst projects were conducted using Twitter since it was well established in the research areas like this project. On the other hand, with the limited timeline that needs to be considered to develop this project, we need to reduce the project scope to the Twitter platform only. There were two possible ways that can collect the Twitter data, and they were as follows,

- Developing a script that can extract Twitter data from the platform.
- Using already extracted data from research sites such as Kaggle and GitHub.

the first approach was not very successful due to the Twitter regulations that they have regarding university students. Hence a second approach was planned to use in the data extraction stage of the project. As per the Proposed steps for the project, the data set needs to be unbiased to external factors like locations, weather conditions, etc. Geolocation also needs to be included in the data set. If the considered data set does not include a significant amount of the geolocation, the whole data extraction part needs to be repeated until proper unbiased data extraction. We were able to find this Twitter dataset in London on the 26th of July 2018.

After the proper data extraction from the Twitter platform, Project is proposed to work on the data set cleaning to avoid unrelated data. This project works expecting to pre-process the dataset considering the London area and avoid duplicate and missing values. The above steps were implemented to avoid the effect of unnecessary data on the outcome. As a data cleaning technique tokenization and lemmatization are also used in text processing.

After the text cleaning and the pre-processing phase, the project is focused on frequency analysis of the data set. Frequency analysis will be processed during peak and off-peak times and will extract the most frequent words of the tweets in the data set. As a final outcome of the frequency analysis, a histogram graph will extract based on the result.
In the later phase of the project, the sentimental analysis will be used to extract the attitudes towards the research project. The final outcome of the sentimental analysis will also be visualised using a London map including positive, neutral and negative locations in the data set. afterwards, the final set of the data will be visualised on the london map according to the peak time and the off-peak time as well. Above mention products will be the final outcome

of the project and all products will be implemented on a small website that users can easily explore.

There are many industries bound with tourism. Accommodation, food and beverage, transport, travel agencies, and cultural activities organisers are a few of them. The project outcome is expecting to use this tourist density map for the above areas. This will help tourism-related industries can make their business decisions very easily and effectively. Even this project is able to address personalised advertisements using the extracted density map of London. According to the user's attitudes and preferences. Gathered data will represent the personal interest of the tourist attractions. They can improve their business by providing the facilities to visitors such as increasing the supply according to demand, especially in busy areas, opening new branches in high-density areas, arranging exhibitions, fairs and many more things. This might help to develop the identified tourist areas and much more positive impact on the community as well. This project will be benefited tourists and the tourism-related industry to find tourist attractions and improve their businesses.

Apart from that this will be very advantageous in taking regulatory decisions in special circumstances. Such as the COVID19 pandemic. If we had such a density map, we were able to introduce covid regulations with special attention so that tourism would be less affected. Not only in the pandemic situation this project can be implemented in many more areas as well.

Apart from the mentioned areas, this project study represents the other attractive places that tourists were been in the day. For example, areas that have the most attractive restaurants and pubs. These were not considered landmarks or historical locations but tourist attractions for other factors such as food and nightlife. The present data set in this project didn’t include the night-time data for tweets. This will be a part of the future work on the project that needs to be addressed in the future. As one of the outcomes shopping related areas also will be extracted as a tourist attractions in London. In the future extended level, this needs to be classified separately and should be able to represent in a separate map. This helps to get more understanding of advertising and business promotions for entrepreneurs.

As another outcome, this project will provide immensely helpful information to promote the accommodations in these tourist attractions. The local community of these areas will directly be benefitted from this information, and it will encourage the development of tourist services as well. The final visualization will be expected to host in the public domain, so users can easily access the project outcome.

1.1 Research goals

The main research goal of this thesis is to identify the tourist location using Twitter data by evaluating the tourist flow in London. Apart from the main research goal, the following tasks had to be achieved.

- Literature analysis, processing, and spatial analysis of data gathered from Twitter.
- Development of methodologies for spatial analysis with natural language processing techniques using Twitter data for analysing the tourist behaviour
- Evaluation and visualisation of the tourist flow as an intensity layer on top of the London map

London is the capital of the UK, and it has approx. 9 million of the population. Due to regional scale research, the selected area is the biggest city in the UK, which has 1,572 km² [18].

1.2 Research Question

Research questions were determined during the evaluation of the acquired geo-tagged data. These were focused on frequency analysis and sentiment analysis of the geo-tagged Twitter data. Every research question is to identify the flow of tourists. All the results of this project were visualised and evaluated in comparison to tourism data from local councils in London. After filtering out the tourism-related tweet, sentiment analyses can be used. Then evaluated the tourist expression of the tweet that they published. And make an analysis of the spatial distribution of the positive, negative and neutral attitudes.

Additionally, it must be emphasised that a wide range of applications are available to analyse tourism flows, ranging from straightforward spatial distributions to more complex ones such as detecting movement paths of the tourists. The next chapter of the literature review introduces a few of the alternatives. However, the most important research questions were considered due to the scale of our master's thesis.

The investigation facts that focused on spatial analysis with sentiment analysis are shown below.

- Filtering the tweets related to the tourism
- Identifying the sentiment orientation of each individual tweet.
- Identifying the most visited tourist locations.
- Generation of a Heat Map for peak and off-peak times
- Analysis of spatial sentiment orientation in London.

2 Analysis of the Twitter Dataset

2.1 Legal, Social, Ethical and Professional issues

This project did not have any social, or ethical issues. As per the project plan, any copyright materials such as images, data etc not be included in this project. In the nature of the project, there are no health and safety issues. Because of the non-usage of confidential data, this research does not breach any condition according to the "Data Protection Act 2018" in the UK. In this research, I'm only considering publicly available data and geo location and tweets are a major consideration.

2.2 Literature review

The different fields of topics had to be included in this section. For the economy of the country, tourism plays a major role. By 2025, 40.4 million visitors may visit London annually and 25.7 million are international visitors of it [4]. Social media gives enormous support to many research fields, and it contains a huge amount of travel-related data. During this literature review, this project has focused on applications and their methodology. Other existing works related to tourism and other research fields regarding spatial analysis and other methodologies were studied in this project.

After launching Twitter in 2006, a lot of research is researched on the usage of Twitter data [1]. An extensive study on Twitter data analysis is the work "Twitter Data Acquisition and Analysis: Methodology and Best Practice" by Stephen Dann [7]. There are many fields were studied on Twitter data. The "The Routledge Handbook of Tourism Geographies" presented the broad literature on spatial tourism analysis [29].

Consideration of the methodologies that were used that used in other researchers; sentiment analysis was mentioned in a lot of research papers as the natural language methodology to identify the feeling of the text. The sentiment analysis was used in the "Tourism, travel and tweets: algorithmic text analysis methodologies in tourism" research paper [5]. As well as the article used "Towards Citizen-Contributed Urban Planning Through Sentiment Analysis of Twitter Data" used sentiment analysis, spatiotemporal analysis, and topic extraction [19]. This research gives a deep insight into sentiment analysis. The "Inferring international and internal migration patterns from Twitter Data" research by Zagheni, Grimella and Weber developed an estimator that can estimate relative changes in trends when including new Twitter data [38]. Hansi Senaratne, Arne Broring, Tobias Schreck, Dominic Lehle published the "Moving on Twitter: Using episodic hot spot and drift analysis to detect and characterise spatial trajectories" research paper that detects trajectories using hot spot analysis. Kernel Density Estimation (KDE) was used to detect hotspot clusters of Twitter activities [33]. There is another research that used the geo-tagged Twitter data is "Twitter as Sentinel in Emergency Situations: Lessons from the Boston marathon explosions" (Cassa et al., 2013). This research analyses the tweets in emergency situations and their early recognition [2]. One of the use cases of geo-tagged Twitter data is identifying the earthquake. The "Earthquake shakes Twitter Users: Real-time event detection by social sensors" (Sakaki, Okazaki and Matsuo, 2010) monitored tweets in real-time to detect and notify the earthquakes. Semantic analysis was used in this research and Kalman filtering, and particle filtering was applied to estimate the location. This research identifies earthquakes much faster than the official announcement from the emergency agency [30]. The "Identification of tourist hot spots based on social networks: A comparative analysis of

European metropolises using photo-sharing services and GIS" (García-Palomares, Gutiérrez and Mínguez, 2015) research was demonstrate the identification and analysing of the tourist attraction in London, Berlin, Barcelona, Athens, Madrid, Paris, Rome, and Rotterdam [8]. This research used spatial statistical techniques in a GIS. The "Using social media to identify tourism attractiveness in six Italian cities" (Giglio et al., 2019) research is based on the identification of the attractiveness of different tourist locations in Italy using the Flicker social media platform [10]. Geo-tagged photos were used to analyse the data and results were shown on the map to identify the annual trends. Mathematical and Machine Learning approaches were used for analysing the data. In the "Travel Recommendation Using Geo-Tagged Photos in Social Media for Tourist" (Memon et al., 2014) research done by Imran Memon, Ling Chen, Abdul Majid, Mingqi Lv, Ibrar Hussain and Gencai Chen proposed a method to identify the travel locations for tourist according to their preferences. In this research Flicker was used to collect photos and semantic analysis was used to implement this tourist recommendation system [24]. The sentiment-based approach was used in the "Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter" (Padilla et al., 2018) by Jose J. Padilla, Hamdi Kavak, Christopher J. Lynch, Ross J. Gore, and Saikou Y. Diallo to determine the temporal and spatiotemporal effect on tourist emotions [26]. The "Understanding Tourist Destination Choices from Geo-tagged Tweets" (Hasnat and Hasan, 2018) presented the analysis of tourist destinations according to the collected geo-enabled tweets [13]. The conditional Random Field (CRF) model was used to predict the next destination of the tourist. The "Predicting crime using Twitter and kernel density estimation" (Gerber, 2014) represented the crime prediction using spatiotemporally tagged tweets [9]. In this research, Twitter-specific linguistic analysis and statistical topic modelling were used to predict crime in the United States.

Geo-tagged Twitter data can be used for identifying tourist behaviour. For analysing the spatial distribution, Hotspot Analysis and Kernel Density Estimation methods were used [32]. Sentiment dictionary, grammatical rule, and sentiment score methods were used for exploring the spatiotemporal pattern of the sentiment changes in tourist flow [17]. The spatial distribution and Latent Dirichlet Allocation (LDA) methods were used as topic modelling in geotagged Tweets classification research [21,39]. 'Spatiotemporal sentiment variation analysis of geotagged COVID-19 tweets from India using a hybrid deep learning model' research analysed the spatiotemporal pattern of sentiments. This is for Covid-19-related geotagged tweets and this research used a hybrid deep learning model with bidirectional long-term short memory (BiLSTM) and convolutional neural network (CNN) [20]. The 'Spatial and Temporal Patterns Of Geo-Tagged Tweets' research explores spatial and temporal patterns of geo-tagged tweets and examines human mobility patterns [15]. Apart from that, word2vec, fastText, GloVe, and TF-IDF word embedding models use for creating the feature space. As well as to get classification results Naïve Bayes support vector machines, and a convolutional neural network can be used. Those are the overview of the possible studies that are considered in the literature research. Those techniques and methods may be beneficial to investigate the spatial distribution of tourism in London. Therefore this research will expose new opportunities in Twitter data usage with spatial scale.

3 Methodology

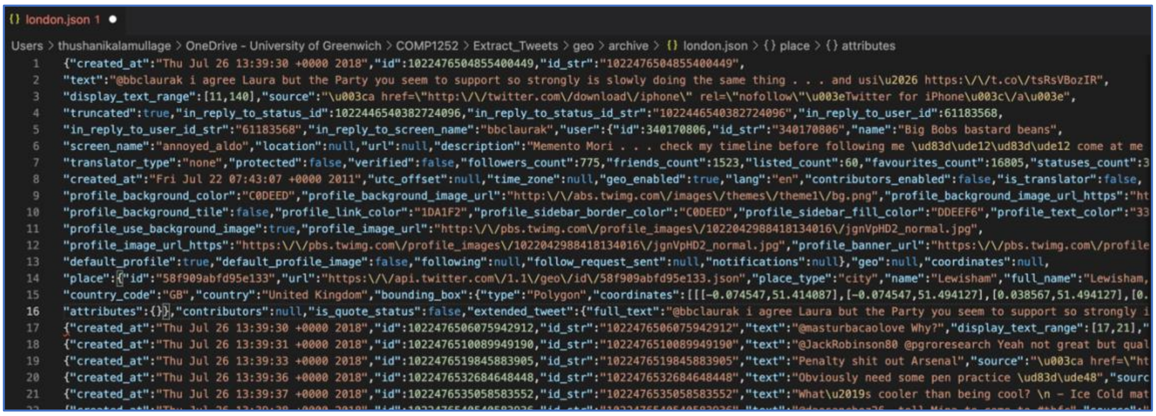
One of the main objectives of this master's project is the development of sentiment analysis and frequency analysis for the tourism industry. This chapter is introduced, the methodology for each step that is used in this research. Python language is used for the development of this project and Arc GIS Notebook was used as an Integrated development environment (known as IDE). The first few sub-sections introduced the data acquisition, processing steps, and filtering steps. Then discuss the natural language processing methodology that is used in this research. Finally, discuss the visualisation methodology. As per the pre-research, the most suitable methodologies were selected after a closer look and evaluation.

3.1 Data acquisition

There are three approaches to acquiring Twitter data. The first one is collecting Twitter data from the official REST API. This method was not used in this project due to its official restrictions on achieving historical data [34]. The second one is scrapping the Twitter data using the ‘Twitterscraper’ Python package. This method enables scraping tweets based on the given location, but this method did not return the exact geo-coordinates of the location. The final one is to find a publicly available Twitter dataset that relates to tourism. There are many available Twitter datasets related to tourism on Kaggle and GitHub but finding the dataset with geo-coordinates is critical. The next sub-section describes the nature of the dataset.

3.2 Dataset Preparation

The dataset was collected in JSON format from Kaggle [31]. The JSON format is not giving a better picture of the dataset than the dataset converted into the comma-separated file (CSV) format. This contains more than 100 columns. Figure 1 illustrates a sample dataset taken from Kaggle.


{} london.json 1 ●
Users > thushanikalamullage > OneDrive - University of Greenwich > COMP1252 > Extract_Tweets > geo > archive > {} london.json > {} place > {} attributes

```
{"created_at":"Thu Jul 26 13:39:30 +0000 2018","id":1022476504855400449,"id_str":"1022476504855400449",
"text":"@bbclaurak i agree Laura but the Party you seem to support so strongly is slowly doing the same thing . . . and usi\u2026 https:\/\/t.co\/tsRsVBozIR",
"display_text_range":[11,140],"source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e",
"truncated":true,"in_reply_to_status_id":1022446540382724096,"in_reply_to_status_id_str":"1022446540382724096","in_reply_to_user_id":61183568,
"in_reply_to_user_id_str":"61183568","in_reply_to_screen_name":"bbclaurak","user":{"id":340170806,"id_str":"340170806","name":"Big Bobs bastard beans",
"screen_name":"annoyed_aldo","location":null,"url":null,"description":"Memento Mori . . . check my timeline before following me \ud83d\ude12\ud83d\ude12 come at me
"translator_type":"none","protected":false,"verified":false,"followers_count":775,"friends_count":1523,"listed_count":60,"favourites_count":16805,"statuses_count":3
"created_at":"Fri Jul 22 07:43:07 +0000 2011","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,
"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"ht
"profile_background_tile":false,"profile_link_color":"1DA1F2","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"33
"profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/1022042988418134016\/jgnVpHD2_normal.jpg",
"profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/1022042988418134016\/jgnVpHD2_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile
"default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,
"place":{"id":"58f909abfd95e133","url":"https:\/\/api.twitter.com\/1.1\/geo\/id\/58f909abfd95e133.json","place_type":"city","name":"Lewisham","full_name":"Lewisham,
"country_code":"GB","country":"United Kingdom","bounding_box":{"type":"Polygon","coordinates":[[[-0.074547,51.414087],[-0.074547,51.494127],[0.038567,51.494127],[0.
"attributes":{}},"contributors":null,"is_quote_status":false,"extended_tweet":{"full_text":"@bbclaurak i agree Laura but the Party you seem to support so strongly i
{"created_at":"Thu Jul 26 13:39:30 +0000 2018","id":1022476506075942912,"id_str":"1022476506075942912","text":"@masturbacaolove Why?","display_text_range":[17,21],"
{"created_at":"Thu Jul 26 13:39:31 +0000 2018","id":1022476510089949190,"id_str":"1022476510089949190","text":"@JackRobinson80 @pgroresearch Yeah not great but qual
{"created_at":"Thu Jul 26 13:39:33 +0000 2018","id":1022476519845883905,"id_str":"1022476519845883905","text":"Penalty shit out Arsenal","source":"\u003ca href=\"ht
{"created_at":"Thu Jul 26 13:39:36 +0000 2018","id":1022476532684648448,"id_str":"1022476532684648448","text":"Obviously need some pen practice \ud83d\ude48","sourc
{"created_at":"Thu Jul 26 13:39:37 +0000 2018","id":1022476535058583552,"id_str":"1022476535058583552","text":"What\u2019s cooler than being cool? \n - Ice Cold mat
```

Figure 1 - Sample JSON dataset from Kaggle

According to the higher number of columns, ‘created_at', 'user.id', 'extended_tweet.full_text', 'text', 'place.name', 'geo.coordinates', 'coordinates.coordinates', and 'place.bounding_box.coordinates' columns were selected from the whole dataset. Using geo. Coordinates column and create new columns for longitude and latitude.
The below table describes the columns of the final dataset.

Column	Description	Data type
created_at	date and timestamp data	Datetime
user.id	Twitter used id	Integer
text	Tweet data	String
extended_tweet.full_text	Tweet data with more details for longer tweets	String
place.name	Name of the town	String
geo.coordinates	Latitude and longitude	String array
coordinates.coordinates	Latitude and longitude	String array
place.bounding_box.coordinates	Set of coordinates of the area	Array of String array
LATITUDE	Longitude coordinate	Double
LONGITUDE	Latitude coordinate	Double

Table 1 - Description of the dataset columns

This dataset is suitable for this research, and it contains 261 tweets with geolocation data. Poplar, Sutton, Harrow, Eltham, London, Camden Town, Islington, Grays, Guildford, Hounslow, Lambeth, East Ham, South East, Dartford, Camberwell, Tottenham, Kensington, Paddington, City of London, Slough, Brent, Watford, Lewisham, Richmond, Wandsworth, Barnet, Bromley, Hammersmith, Croydon, Hackney, Romford, Greenwich, Walthamstow, Hillingdon, Waltham Abbey, Merton, Kingston upon Thames, Southall, Stratford, and Ealing are the places that cover in this dataset.

The dataset is based on London with covers the above cities with travel tweets from 26/07/2018 at 13:39:37 to 26/07/2018 at 17:03:05. This dataset covers the peak and off-peak times in London [23]. According to this dataset, 13:39:37 to 15:59:00 can be taken as the off-peak time Twitter dataset and 16:00:00 to 17:03:05 taken as the peak time Twitter dataset.

3.3 Data Pre-processing

The project development is started with the data pre-processing step. Firstly, loaded the CSV file data as a dataframe using the ‘pandas’ library. The below code shows how to import the CSV file as a dataframe in python. The raw_df is the data frame with tweet data.

```
raw_df = pd.read_csv('london.csv', index_col=0)
```

The dataset has been checked to identify unnecessary data and the following steps were followed to identify unnecessary data and get a better idea of the dataset.

3.3.1 Data Cleaning

The scope of This project is to identify the tourist attractions in London. Because of that, all the tweets contained outside London were removed using geo-coordinates. All tweets in the dataset were converted into lowercase to avoid case sensitivity. The ‘str.lower()’ method was used to convert text into lowercase and python code is shown below.

```
raw_df['text'] = raw_df['text'].str.lower()
```

There are some columns that have the same information, and it does not give value addition. According to those unnecessary duplicated columns were removed from the dataset. Using the ‘drop’ function unnecessary columns were removed help of the below code.

```
raw_df = raw_df.drop(['extended_tweet.full_text'],axis=1)
```

After cleaning the dataset number of rows dataset was reduced to 228 and no of cities was limited to 34.

3.3.2 Data Quality Assessment and Exploratory Data Analysis

In this section, consider the data types mismatches and duplicates in the dataset.
The 'text' is the main input attribute to the model and 'geo.coordinates' is the target of this project. When checking the attribute's type in the dataset, the 'user.id' attribute's type was corrected from 'float64' to 'int64'.
Using this info() method gives information about the dataset and according to that result, datatype mismatches were identified.

```
print(raw_df.info())
```

Then change the datatype of the column into an integer type using the below code.

```
raw_df['user.id']=raw_df['user.id'].astype(int)
```

Then duplicated values and missing values were checked in the Twitter data and removed from the dataset. To identify the duplicates of the dataset, the below code was used.

```
raw_df.duplicated().sum()
```

To get a better idea of the dataset, the below codes were used.

```
raw_df.describe()
raw_df.groupby('place.name').describe()
```

To identify the missing values, the below code was used. The Exploratory Data Analysis is for studying the dataset. In this part, the statical information and the shape of the dataset were considered.

```
raw_df.isnull().sum()
```

3.3.3 Text Cleaning and Text Analysis

Text cleaning was applied to the columns that have contains text data. According to the Twitter dataset, the ‘text’ column contains text data. There are many short-form words such as 'won't', and 'can't'. Firstly, those short-form words were converted into long forms. To do this, a python function was created, and conversion happens according to the below code. The ‘re’ represent the regex python library and the sentence is the tweet.

```
sentence = re.sub(r"won't", "will not", sentence)
```

Use the ‘Spacy’ library and apply tokenisation and lemmatisation. In this step, tokenised words were converted into their root format and the text data was normalised. And also, lemmatization helps to reduce vocabulary size. After that use the same library to remove the stop words and punctuations. To do this ‘en_core_web_lg’ model was used with the ‘spaCy’ library. The ‘en_core_web_lg’ is the largest English model of ‘spaCy’. The set codes were implemented to achieve these steps and relevant python codes were shown below.

```
nlp = spacy.load('en_core_web_lg')

def textcleaning(column):
         tk = decontracted(column)
        col2 = [w.lemma_ for w in nlp(tk)if not (w.is_stop) or (w.is_punct)]
        return (" ".join(col2))
```

After that, HTML tags, special characters, web tags, web URLs, numbers, emails, and unnecessary white spaces were removed from the dataset using regex. Because those data do not give considerable meaning to the text.

```
df['text']=df['text'].replace(r'http\S+', '', regex=True)
```

The above code shows a sample of using the regex to eliminate web URLs in the dataset. There were more codes like this to eliminate unnecessary characters and phrases in the tweets. After that, the Twitter dataset was split into two sets according to the peak and off-peak times in London [23]. According to the exploratory data analysis of the text data, the following results were got.

- Minimum number of tokens: 16
- Maximum number of tokens: 103

The number of tokens per tweet is between sixteen to one hundred and three tokens. To get this result, the tokenisation was reapplied to the text data using the ‘nltk’ library.

3.4 Sentiment Analysis

Sentiment analysis is also called opinion mining or emotion AI and it is a subset of text mining. This technique shows the sentiment orientation of the tweets by categorising them into classes such as neutral, positive, and negative [12]. This method is referred to as the polarity format. Apart from that, there is a valence-based form, and it reveals the intensity of the sentiments. As an example, in polarity form, ‘excellent’ and ‘good’ words are categorised as positive and in valence-based form is identified the word ‘excellent’ is more positive than the word ‘good’ [12]. Identifying the sentiment orientation is not an easy task because this method has to deal with rejection, ridicule, the ambiguity of the text, context etc. But after better determination, this method can be applied to many fields such as marketing fields, sociological fields etc. Moreover, sentiment analysis is popular in product review analysis and social media-related analysis. The entire text was considered in this research apart from the characteristics of the text. The sentiment analysis approach has more similarities to the text classification because sentiment analysis categorised given text based on the three sentiment categories. The lexical approach and machine learning approach are the two well-established approaches. The lexical approach needs one or more lexicon that contains words with positive and negative labels. According to the existing individual word in the text, the whole text is categorised as positive, negative and neutral. But to achieve a better result, a comprehensive lexicon is needed with high quality. The lexical approach cannot determine the context of the sentence as well as it is required a lexicon for languages. The context of the sentences is important because the same word can be given a different idea or sentiment. To avoid those issues in the lexical method, have to combine them with the machine learning modules and use this approach as a hybrid approach [25].

3.4.1 VADER Methodology

In order to analyse the text, there are two approaches. One approach is lexicon-based and another one hybrid approach. In this research, a hybrid approach was considered and as a hybrid approach, lexicon and rule-based approaches were used. This lexicon and rule-based approach is known as VADER. In 2014, Hutto and Gilbert developed this VADER method [16]. In addition, this method is capable of mining mixed languages since the non-English text is also translated through the machine-translation web service [19]. But this feature does not apply to this research because the Twitter dataset only contains tweets that are English. VADER is denoted by Valence Aware Dictionary for Sentiment Reasoning, and this is a lexical and rule-based model. The 'SentimentIntensityAnalyzer' class with the 'vaderSentiment' python library were used to develop the VADER sentiment analysis.

text	Compound	Positive	Negative	Neutral	val
"cool coolice cold matcha lattesimple n delish..."	0.3182	0.277	0.000	0.723	Positive
"surreyrugcare cheam"	0.0000	0.000	0.000	1.000	Neutral
"well cask sucker ryedrink wryneck rye thamess..."	-0.3182	0.200	0.324	0.476	Negative
"house spangle swimwearexclusive print apparel..."	0.0000	0.000	0.000	1.000	Neutral
"byelondon soonsolongfarewelladieu parting swe..."	0.4588	0.375	0.000	0.625	Positive
"heat continueampyes need headband stop quiff ..."	-0.7184	0.000	0.462	0.538	Negative

Table 2 – Intensity values of sentiment analysis with VADER

VADER is act as a valence-based approach in sentiment analysis because it considers the sentiment and its intensity. Table 2 shows the intensity levels and ranking for five tweets. According to the compound, texts were classified into classes. The lower compound text was classified as a negative category and the Higher compound text was classified as a positive category. And other texts were categorised into the neutral class. The VADER method categorised each sentence into negative, neutral, and positive classes. This is based on the compound score. The compound score is the summation of all lexicon ratings and is standardised between -1 and 1 [36]. The compound score is near 1 means the tweet tends to be more positive and it is closed at -1, which means the tweet is more negative than other tweets. For the neutral tweets, the compound score is going 0.

This VADER method is very useful in social media-related sentiment analysis because in social media text contains emotions, abbreviations etc. Moreover, 'omg' (oh my god) like abbreviations are included in the lexicon. To identify the intensity of positivity and negativity those abbreviations were helped.

This project can market the tourism industry as a product to identify the most tourist attraction areas. Visitors publish their feeling about the tourist attractions, and sentiment analysis is helped to extract their feedback and feelings. According to the sentiment analysis can improve the offer or plan for the development of the tourism in future.

3.4.2 Development

The final dataset contains 200 tweets with 31 cites and the sentiment analysis was developed using python language using the VADER method. These are the codes that were used for sentiment analysis using VADER methods.

```
sid_obj = SentimentIntensityAnalyzer()
compound = sid_obj.polarity_scores(sentence)["compound"]
pos = sid_obj.polarity_scores(sentence)["pos"]
neu = sid_obj.polarity_scores(sentence)["neu"]
neg = sid_obj.polarity_scores(sentence)["neg"]
```

These codes calculate the numerical values for compound, positive, neutral, and negative intensities for a given sentence. According to the VADER method, 68 were categorised as positive and 17 as negative. Sentiment analysis with the VADER method classified the tweets according to the compound score, which is more biased to 1 categorised as positive, more biased to -1 categorised as negative and 0 compound score categorised as neutral. This project considers the below conditions to categorise the tweets into positive, negative, and neutral.

```
if compound >= 0.05 :
   sentiment = 'Positive'

 elif compound <= - 0.05 :
   sentiment = 'Negative'

 else :
   sentiment = 'Neutral'
```

Sentiment	Count
Neutral	115
Positive	68
Negative	17

Table 3 - Tweet counts per sentiment

Most of the tweets were categorised as neutral according to Table 3. With the nature of the project, more neutral sentiment tweets were expected. Many visitors do not express their feelings while tweeting about where are they travelling or what are they doing. Table 4 shows categorised tweets using sentiment analysis with the VADER algorithm.

Tweet	Sentiment
“cool coolice cold matcha lattesimple n delish..”.	Positive
“byelondon soonsolongfarewelladieu parting swe...”	
“uniform arrrivego look smart campdayplus”	
“well cask sucker ryedrink wryneck rye thamess..”.	Negative
“waste time look loselife mean travel backwards”	
“heat continueampyes need headband stop quiff ...”	
“surreyrugcare cheam”	Neutral

“unveil new logonew logo alertnew logo look li...”	
“house spangle swimwearexclusive print apparel...”	

Table 4 - Different sentiment representations for the tweet data

As well as using the scatter plot can visualise the sentiment analysis results. In this thesis, the k-means clustering approach was used. The colours indicate the compound sentiment. Figure 2 shows the sentiment analysis results using a scatter plot with six clusters. According to the compound value, the cluster was determined. Apart from this scatter plot, this research also created plots to visualise the feelings of visitors about the places in the London area during the peak and off-peak times.

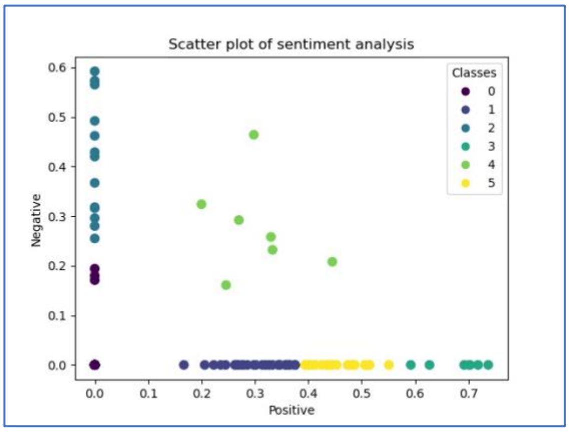


Figure 2 - Scatter plot for sentiment analysis

3.5 Frequency Analysis

Term frequency analysis can be applied to identify tourism-related keywords in the dataset. Then the tweets were examined by their behaviour to identify the patterns. Due to our expert knowledge of the topic, the highest relevance to tourism keywords was manually determined from all the tweets. Identification of words that appear in documents frequently is a very useful way to identify the topics most discussed on social media. This research considers only tourism-related topics and uses frequency analysis to identify the most frequent words and reveals the locations that got a high number of tweets. Term frequency analysis was implemented by using python language. For the development ‘of spaCy, and ‘CountVectorizer’ python libraries were used and for visualisation ‘seaborn’ library was used. Word clouds, bigrams, and histograms were used to visualise the frequency analysis output.

In the field of text mining, word clouds are considered a very visually appealing and straightforward text analysis approach. Then word clouds are one of the best ways to visualize text data. Based on term-frequency words that appear with more frequency were visually emphasised using larger fonts and the colour. Word clouds are easy to read and understand. Moreover, word clouds are informative. It is equivalent to static text summarization technically, but this is still a useful tool for test analysis tasks [14]. In this research ‘WordCloud’ python library was used for generating the word clouds.

To visualise the word clouds the below set of codes was used.

```
stops = nlp.Defaults.stop_words
wordcloud = WordCloud(stopwords=stops).generate(' '.join(df['text'].tolist()))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Other than the word clouds, bigrams and histograms were also used to identify the most frequent tourist-related keywords. Figures 3 and 4 show the word clouds of the tweets that publish for peak and off-peak times.

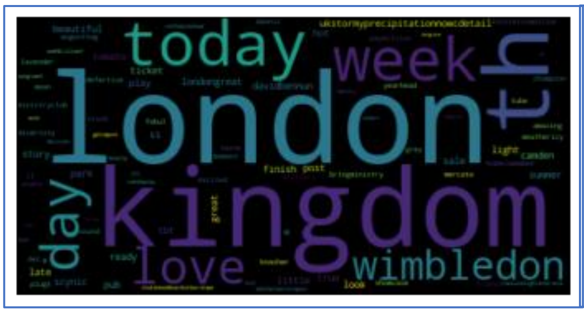


Figure 3 – Word cloud for peak time data

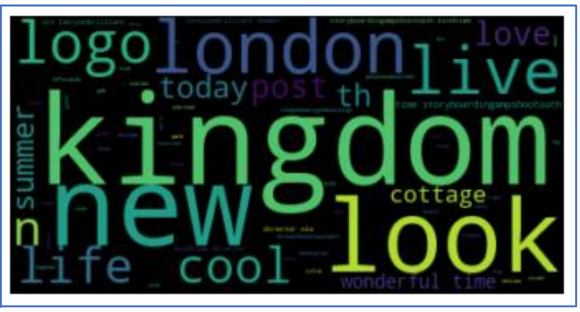


Figure 4 - Word cloud for off-peak time data

According to the above graphs, 'london' and 'kingdom' are the most frequent keywords in the peak time. As per in the off-peak time, 'kingdom' and 'new' are the most frequent keywords.

Moreover, bigrams are also a useful tool in terms of frequency. This method plays a big role in natural language processing. This method is commonly used for simple statistical analysis of text data. In the area of computational linguistics, speech recognition and many other applications benefit from this. The bigram model determines the probability of a word with respect to all previous words [6]. Equation 1 shows the conditional probability prediction.

$$\boldsymbol{P(w_n| w_n^1) \approx P(w_n| w_{n-1})} \tag{1}$$

There were a few python functions to create bags of words and calculate the word frequencies. Those functions are shown below.

```
def tokenize_as_is(x):
  return x

def get_top_ngram(spacy_tokenized_corpus, ngram_range=(2,2), top_n= 50):
    vec = CountVectorizer(ngram_range=ngram_range, max_features=top_n,
    lowercase=False, tokenizer=tokenize_as_is).fit(spacy_tokenized_corpus)
    bag_of_words = vec.transform(spacy_tokenized_corpus)
    vec2 = CountVectorizer(ngram_range=ngram_range, lowercase=False,
                tokenizer=tokenize_as_is).fit(spacy_tokenized_corpus)

    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.items()]
    words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
    return bag_of_words, words_freq
```

Figures 5 and 6 show two bigrams for peak and off-peak times. The below python code was used to visualise a bar plot. The 'sns' is represented by the 'seaborn' python library.

```
sns.barplot(x=ngram_freqs,y=ngram_labels)
```

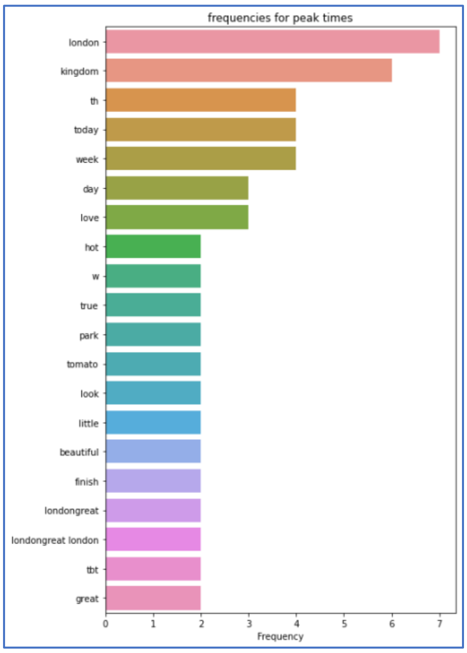


Figure 5 - Bigram for top 20 words during peak times

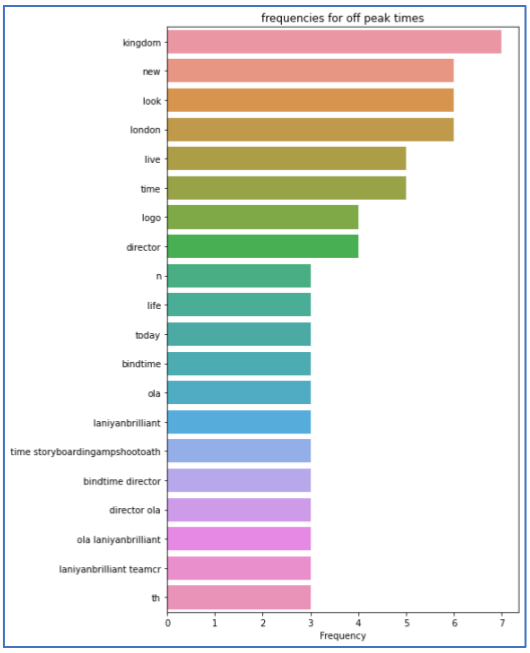


Figure 6 - Bigram for top 20 words during off-peak times

According to those bigrams, the same results appear compared to the world cloud graphs. Above bigram is shown the top 20 most frequent words that tourists used while tweeting during peak and off-peak times in London. Figure 5 and Figure 6 are the final graphs after filtration of the tourist-related keywords. There are a few words that are not relevant to tourism such as words related to jobs, work, apartments, advertisements etc. After figuring out the unrelated word in the dataset, eliminate those data from the Twitter dataset. After this step, the number of rows in the Twitter dataset is reduced to 200 rows.

Another tool used in this project is the histogram. According to the simplicity and versatility histograms are popular in the data science industry. The diagrammatic representation gives a better idea about the dataset. Further, histograms allow viewers to compare data very easily and it capable of working with a wide range of information. Histograms can be used in many different situations to offer an insightful look at frequency distribution.

In this thesis, a histogram was used to illustrate the place that got a higher no of tweets during the peak and off-peak times. The number of tweets with respect to places in London was illustrated using the below histogram. Figure 7 shows two histograms in the same plot considering the peak and off-peak times and targeting 31 places in London.
Theses set of codes is for creating the histogram for peak and off-peak hours in the same plot.

```
fig = plt.figure()
ax = fig.add_subplot(1, 1, 1)
ax.hist(peak['place.name'])
ax.hist(off_peak['place.name'], alpha=0.7)
plt.xticks(rotation = 90)
```

```
ax.set_xlabel('Place')
_ =ax.set_ylabel('Frequency')
_= ax.legend(['Peak', 'Off peak'])
plt.show()
```

Orange colours represent the off-peak time and blue colours represent the peak time tweets. According to this histogram, Hammersmith, Kingston upon Thames, and London get more tweets. This means around London, Hammersmith and Kingston upon Thames cities have more tourist attractions in off-peak times. As well as London and Kingston upon Thames cities published more tweets during peak times compared to other cities in London. Sutton, Kensington, Watford, Lewisham, Croydon, Greenwich and Walthamstow cities did not capture any tweets during the peak times.

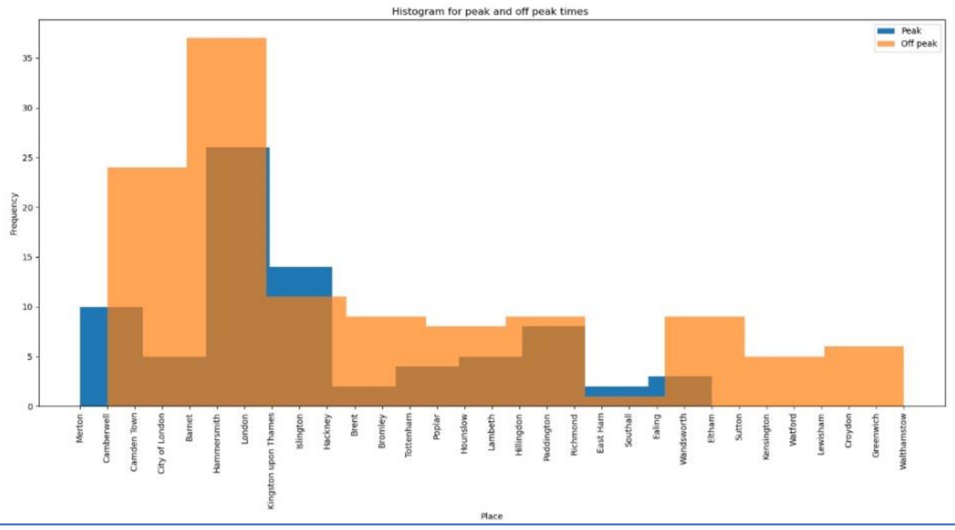


Figure 7 – Histogram for peak and off-peak times according to the 31 towns

3.6 Spatial analysis and Visualisation

Geographics Information System (GIS) is for collecting, analysing, and managing spatially related data [11]. In most of industries used GIS to monitor and identify their problems. Modern GIS offers Apps, Maps and Data to perform analysis and solve problems. With relevance to this thesis, creating a tourism-related map is the most obvious application. This chapter focused on spatial analysis techniques used in this research in order to evaluate the Twitter geodata dataset. Spatial analysis is the study of geographical patterns use of spatial data. According to the location data and its attributes, the special analysis examines the relationships between features and determines spatial patterns and trends. Spatial statistics consider the statistical methods that develop using mathematical computations with help of spatial characteristics. In this research, the Mapping clusters Arc GIS toolkit was used to determine hot spots. Mapping clusters is mainly ide for identifying similar characteristics in groped spatial features. Arc GIS tool allows for modification of source codes according to the analysis. As well as this tool support JavaScript, Java, Python, and many programming languages. Apart from other statistical tools, mapping clusters are the particular interest during this project. According to the first law of geography defined by Waldo Tobler, "Everything is related to everything else, but near things are more related than distant things” has a relation with the mapping clusters. These clusters are able to identify the location and the intensity of existing clusters.

Identification of the spatial distribution in the London area was evaluated in this section. The goal of the project is to statistically analyse and identify the tourist locations based on the Twitter data and make a comparison of the behaviour of the tourist compared to the peak and off-peak times. There are several ways to create maps using Arc GIS software. The first one is uploading the CSV file to the portal. In this way, CSV file data is imported to GIS as a shape file according to the longitude and latitude. Then imported a spatial data map with the shapefile of the London map. Another way is using the programming language. This research follows this process. The tweet dataset went through different kinds of steps like pre-processing, text cleaning, frequency analysis, etc. The final cleaned dataset was considered in this section. The cleaned dataset was converted to a spatial enable dataset. After creating the spatially enable dataframe. One column was added to the dataset which is the ‘SHAPE’ column. This column type is the ‘point’ type, and it automatically generated its values using longitudes and latitudes in the dataset. Sample data of the ‘SHAPE’ column is shown below.

```
{
    “spatialReference”: {
       “wkid”: 4326
    },
    “x”: -0.080028,
    “y”:51.545341
}
```

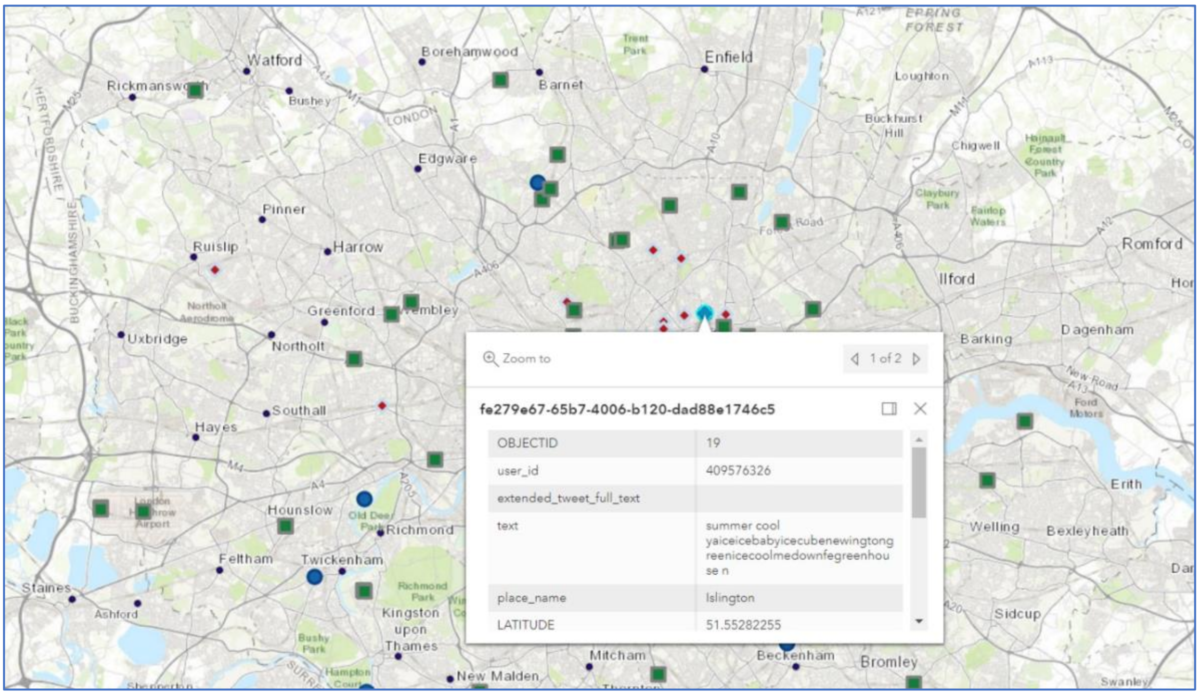
The London Map creates in Arc GIS online tool it was saved in the project repository. Then load the London map by calling its item id. Below python codes showed the spatial data frame creation and how to load the London Map into the notebook.

```
off_peak_sdf = pd.DataFrame.spatial.from_xy(off_peak, 'LONGITUDE','LATITUDE')
gis = GIS('home')
map_item = gis.content.get("774de3b4b9f047dda0851a37efedf789")
londonMap_off_peak = gis.map(map_item)
```

Then visualize the tweets data on top of this London map. Here is the line of code that need for the plotting.

```
off_peak_sdf.spatial.plot(
    map_widget=londonMap_off_peak,
    renderer_type='h',
    symbol_type='simple',
    colorRamp = {
            "type":"algorithmic",
            "fromColor": [133,193,200,0],
            "toColor": [255,255,0,0],
             "algorithm": "esriHSVAlgorithm"
    },
    cstep=50,
    ratio = 0.01,
    marker_size=10,
    height = 40,
    width = 60
)
```

The result of this code is shown in Figure 10. As well as there are many types of figures that can be plotted using this tool. Another plotted graph is shown in figure 8. Figure 8 was plotted using geo coordinates and this graph enables one to view the Twitter data corresponding to the point. It automatically pops up the box containing the Twitter data regards the geo coordination. Consideration of the above facts, this approach gives a very good solution for visualising the Twitter data and analysing the distribution of the London area through spatial analysis.



fe279e67-65b7-4006-b120-dad88e1746c5

OBJECTID	19
user_id	409576326
extended_tweet_full_text	
text	summer cool yaiceicebabyicecubenewingtong reenicecoolmedownfegreenhou se n
place_name	Islington
LATITUDE	51.55282255

Figure 8 – Visualisation of the sentiment analysis results in London map

3.6.1 Hot Spot Analysis

Hot Spot analysis is a very common spatial analysis technique that mainly focused on the identification and visualisation of clustering of spatial fields. Using this technique fulfils the research requirements and is able to get an advantage from it. The Hot Spot analysis statistically determines special clusters of high values as the Hot Spot and low values as cold spots. This happens based on Getis-Ord Gi* statistic and set of weights [27]. Simply, a Hot Spot shows the area that has high occurrence points incidents. It is possible to make a transformation from point observation into area computation using hot spot analysis. In this research, Hot Spot analysis was used for identifying the location that published higher travel-related tweets. According to the ArcGIS tool, there are 3 values returned in this analysis.

Those are,

- Standard deviation: z-score
- Probability: p-value

- Confidence level bin: Gi-Bin

The use of z-score and p-value can determine whether the null hypothesis is rejected or not. The features are shown as clustering when the hypothesis was rejected. The reject null hypothesis needs a higher z-value and lower p-values [28]. Table 5 and Figure 9 shows how to deviate confidence levels of z-score and p-values with standard normal distribution.

z-score (Standard Deviations)	p-value (Probability)	Confidence level
< -1.65 or > +1.65	< 0.10	90%
< -1.96 or > +1.96	< 0.05	95%
< -2.58 or > +2.58	< 0.01	99%

Table 5 - Confidence Levels [28]

The definition of the scores listed below corresponds to the ESRI definition [28].

- Z-score: This refers to standard deviations and the score of +1.8 represent the 1.8 standard deviations.
- P-value: This refers to probability and smaller p-values lead to the rejection of the null hypothesis.
- Gi-Bin: This refers to confidence level and results are visualised according to this.

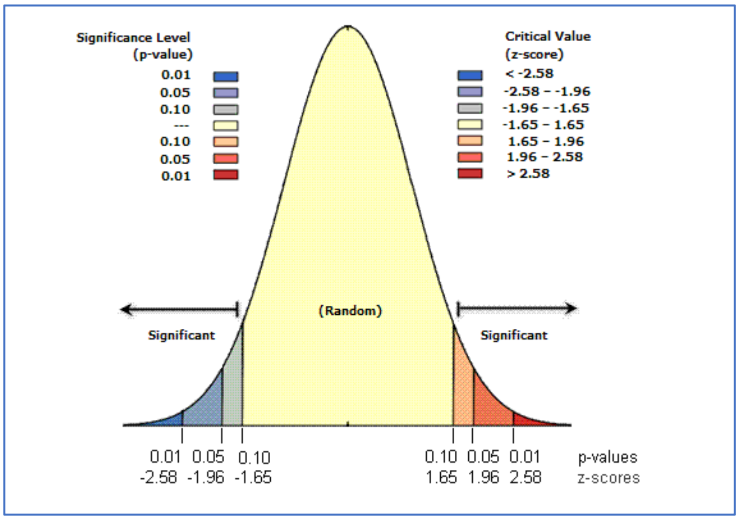


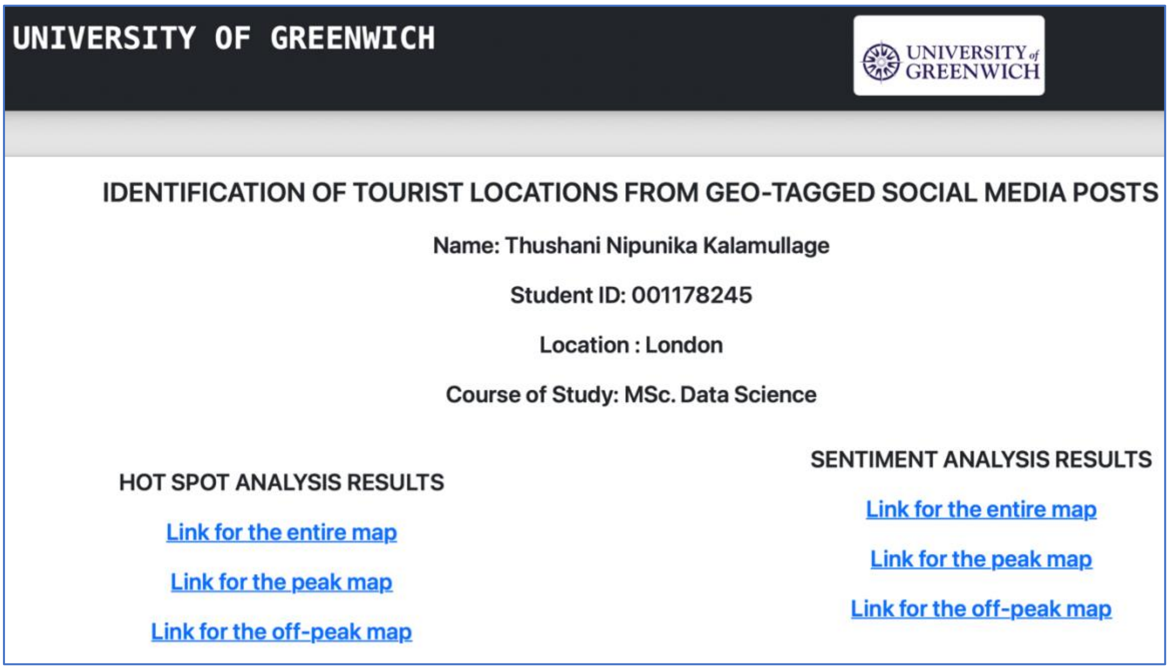
Figure 9 - Standard normal distribution with z-score and p-value [28]

3.6.2 Development

The visualisation development has been carried out using python language. The Arc GIS tool supports python, and it offers an online platform for the developer to implement their analysis. In this project, the Arc GIS notebook was used to develop the visualisation part. For the development purpose 'arcgis' python library was used. Creating Hot Spots for the peak and off-peak times is the main target of this section.

Firstly, spatially enable data frames were created using the 'pandas' library. To do this, the 'LATITUDE' and the 'LONGITUDE' columns were specified in the method.
Then a map item was created using the 'Arc GIS' online tool and saved on my repository. After that, the map for London was loaded to the Arc GIS notebook using the London map item and the 'arcgis' library. Then the heat maps were created for peak and off-peak times using spatially enabled data frames. The London map was taken as the map widget and the colour ramp algorithmic type was used. The map generated using the Arc GIS tool is very useful because this kind of map allows some features like zoom in zoom out and pop up the data when clicked on the map. Figure 10 illustrates the sample map that was generated using the Arc GIS tool for the London area.

Using those heat maps, the final website was created using bootstrap and CSS files. Figure 11 illustrates the final product of this thesis.



UNIVERSITY OF GREENWICH

IDENTIFICATION OF TOURIST LOCATIONS FROM GEO-TAGGED SOCIAL MEDIA POSTS

Name: Thushani Nipunika Kalamullage

Student ID: 001178245

Location : London

Course of Study: MSc. Data Science

HOT SPOT ANALYSIS RESULTS

Link for the entire map

Link for the peak map

Link for the off-peak map

SENTIMENT ANALYSIS RESULTS

Link for the entire map

Link for the peak map

Link for the off-peak map

Figure 10 - Final website for visualisation of the result of Hot Spots analysis and semantic analysis

All the final figures of the extraction were embedded into a small website. The website is on the local host. This web product is developed to demonstrate the figures and as a limitation of this, populated figures can’t be updated dynamically as of now. Since the project figures were visualised on the free plan of the Arc GIS, the dynamic update is not supported for now. Apart from that safari web browser is also having some issues with the website and chrome and edge are preferred for the website. Users are able to explore the below figure on the implemented website.

- spatial distribution by using hotspot figure for peak times
- spatial distribution with using hotspot figure for off-peak time
- spatial distribution by using hotspot figure for whole data
- spatial distribution of the sentiment analysis result for peak times
- spatial distribution of the sentiment analysis result for off-peak time
- spatial distribution of the sentiment analysis result for whole data

There will be populated map files and needful CSS libs in the project web folder. The user only needs to access the index.html file and it’ll open the above view. Users can explore the rest of the visualization from there.

3.6.3 Spatial Distribution

Figure 12 shows the spatial distribution of all tweets across London. Figure 13 shows the spatial distribution during peak times and Figure 14 shows the spatial distribution during off-peak times. Those figures illustrate the hot spot analysis in the London area.

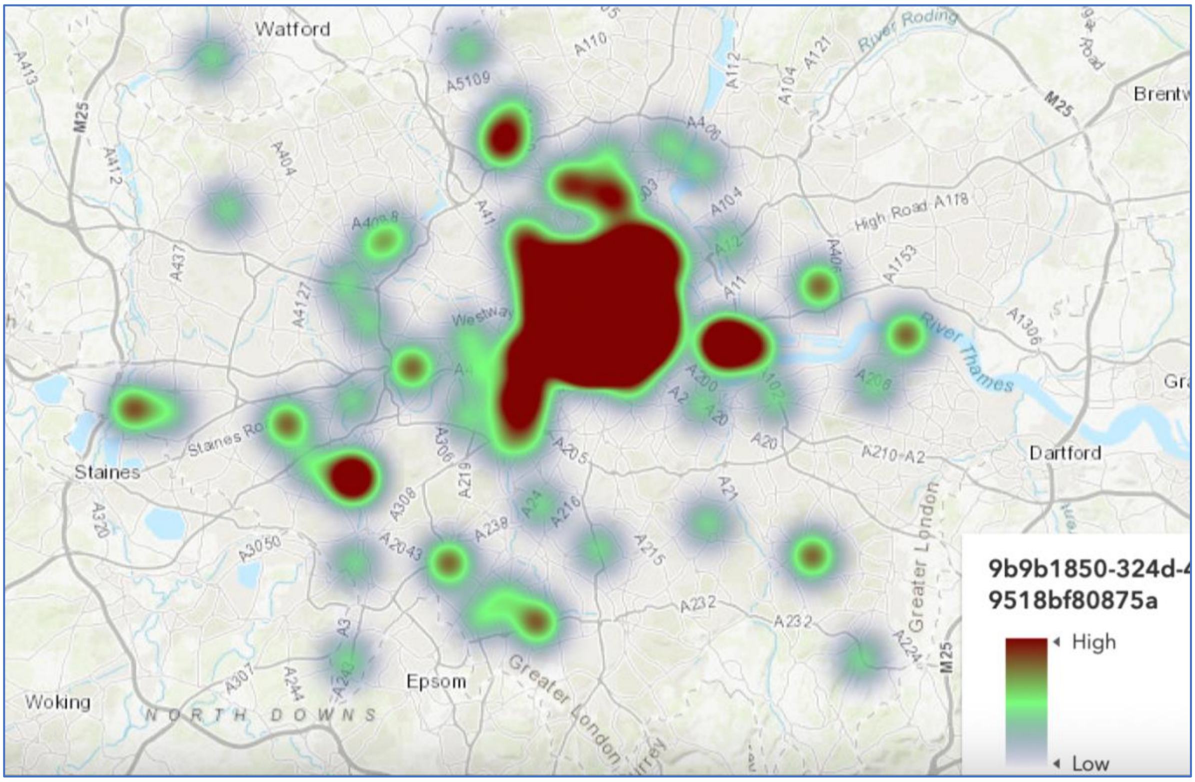


Figure 12 - Spatial distribution with Hot Spot analysis in London

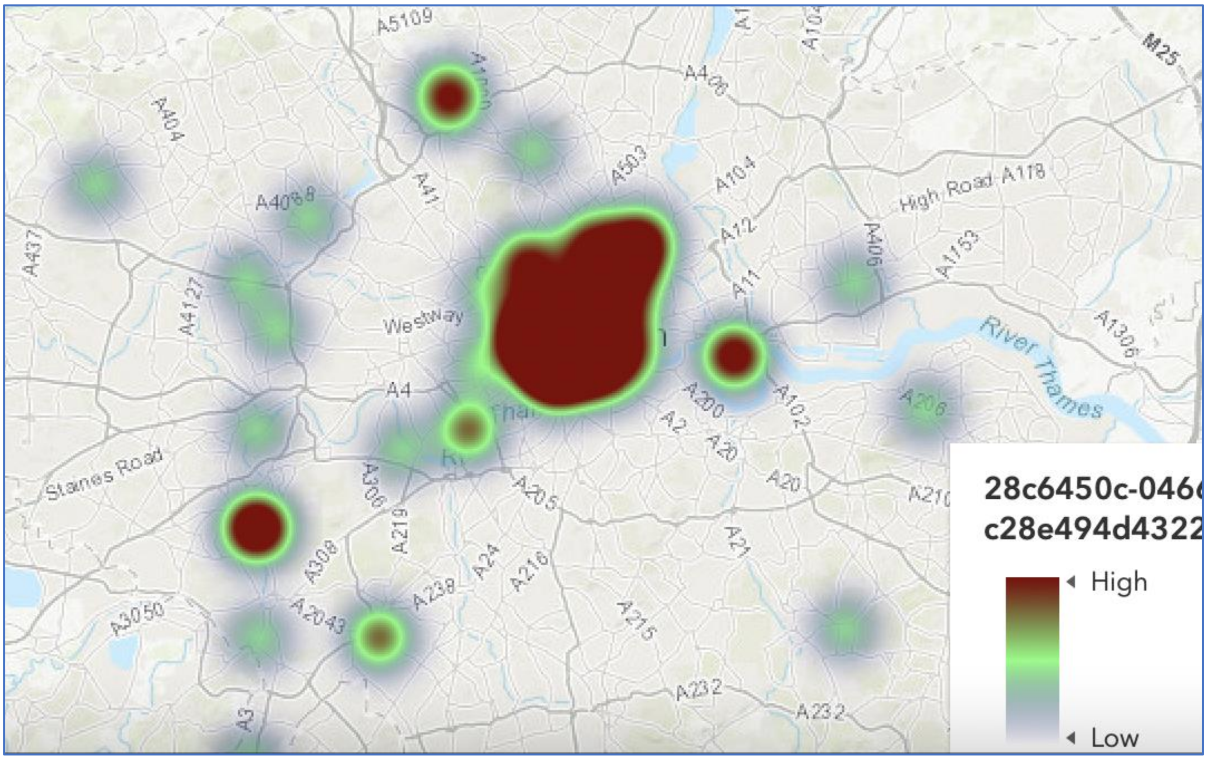


Figure 11 - Spatial distribution with Hot Spot analysis in London during peak times

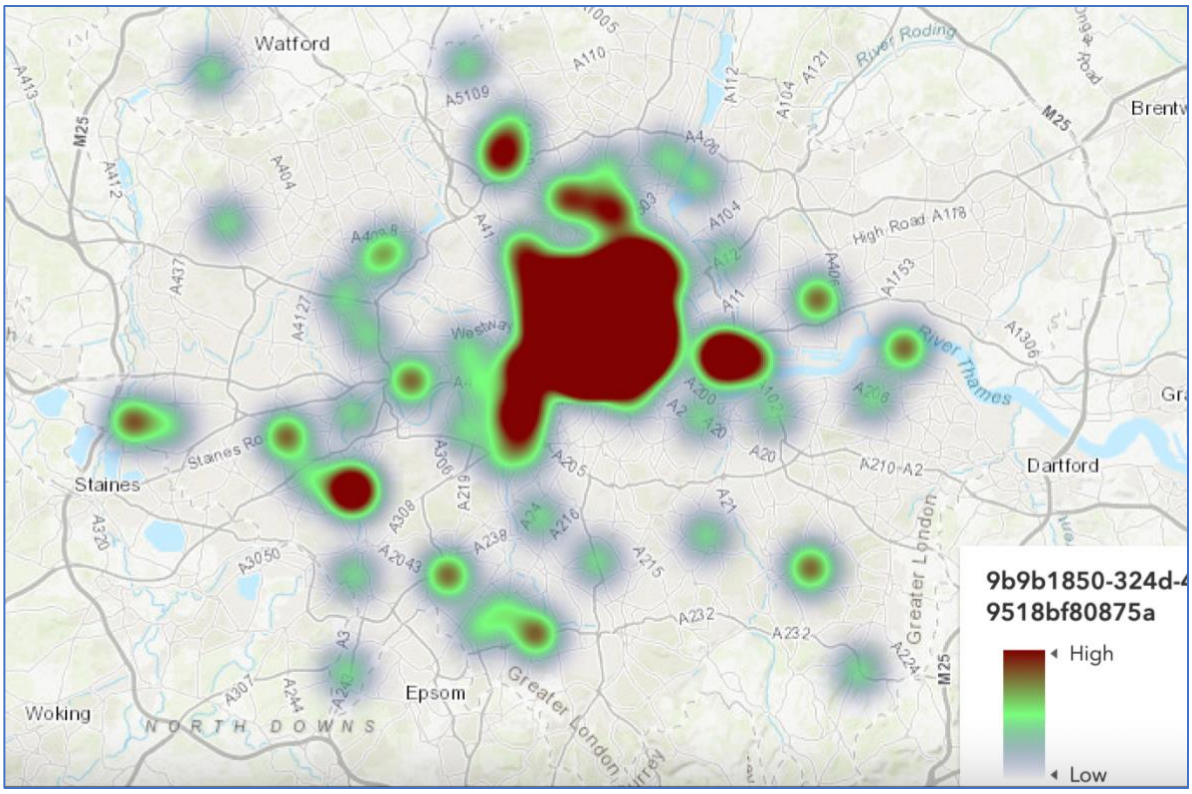


Figure 13 - Spatial distribution with Hot Spot analysis in London during off-peak times

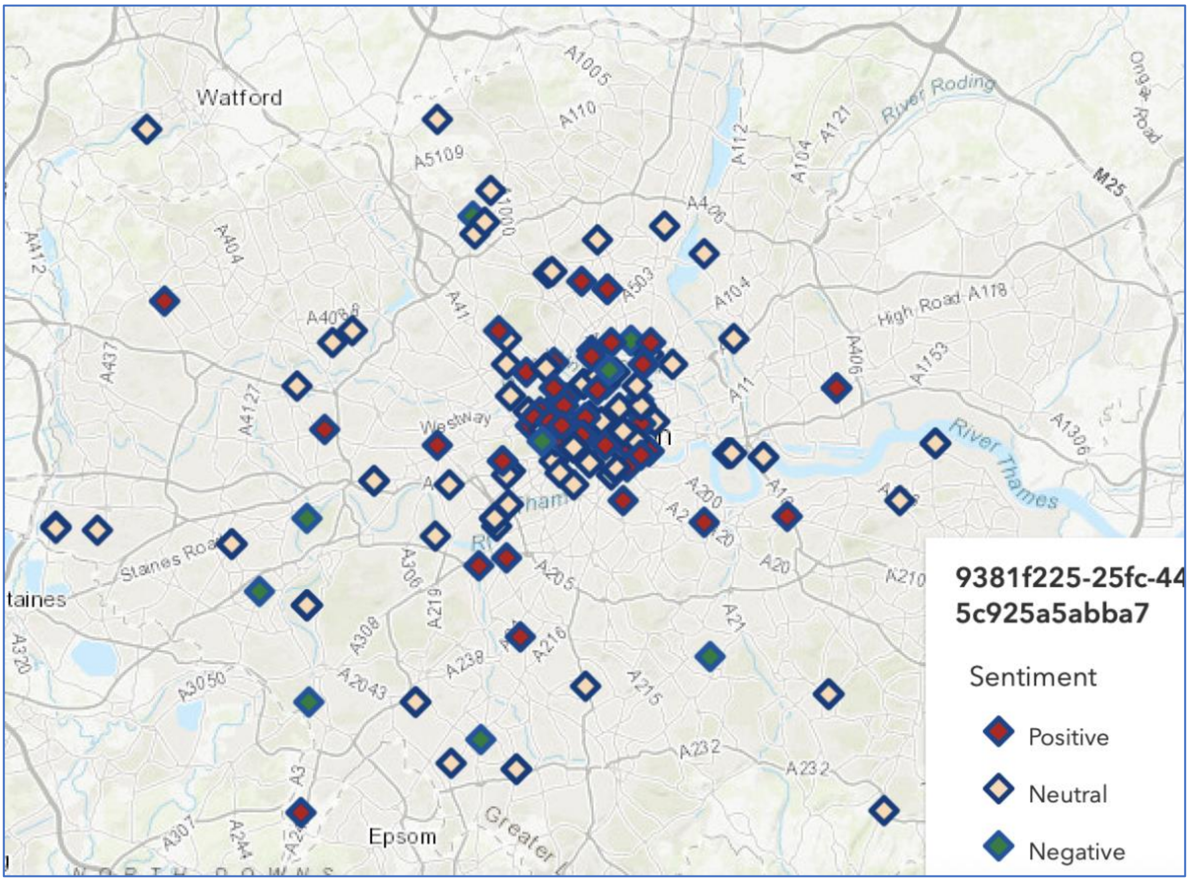


Figure 14 - Sentiment Analysis distribution in London

The darker red colour indicates the places that publish more tweets, and the green colour indicates the low-frequency area according to the number of tweets according to the legend. Figure 15 represents the spatial distribution of the sentiment analysis with the VADER method. Figure 16 and Figure 17 illustrate the sentiment analysis results during peak and off-peak times.

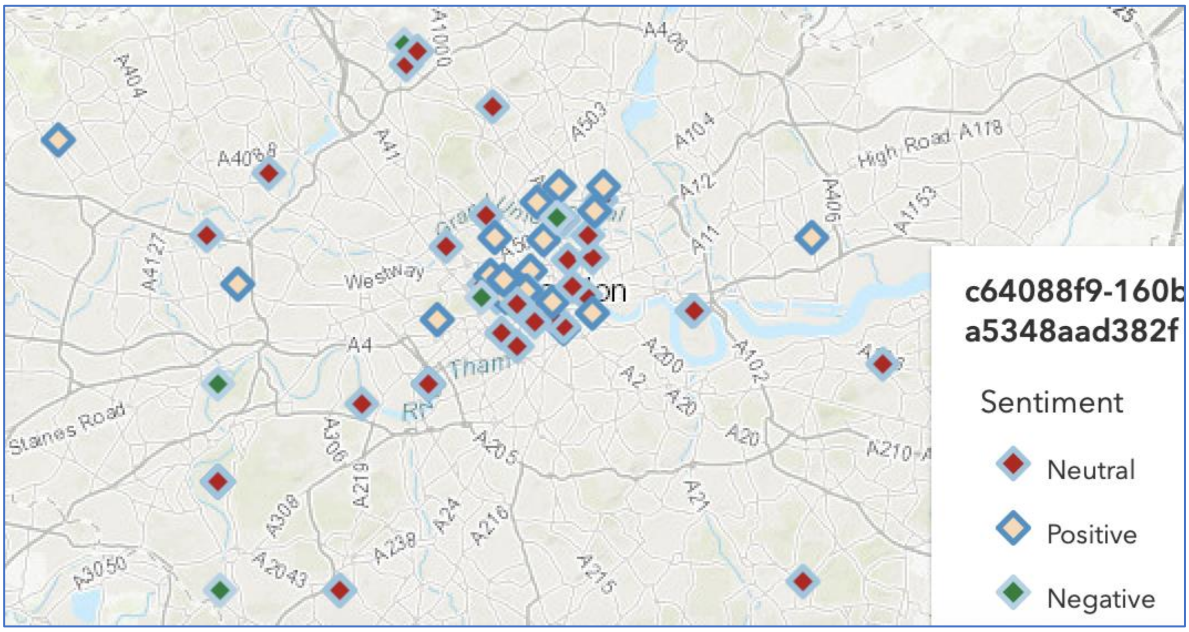


Figure 15 - Sentiment Analysis distribution during the peak hours in London

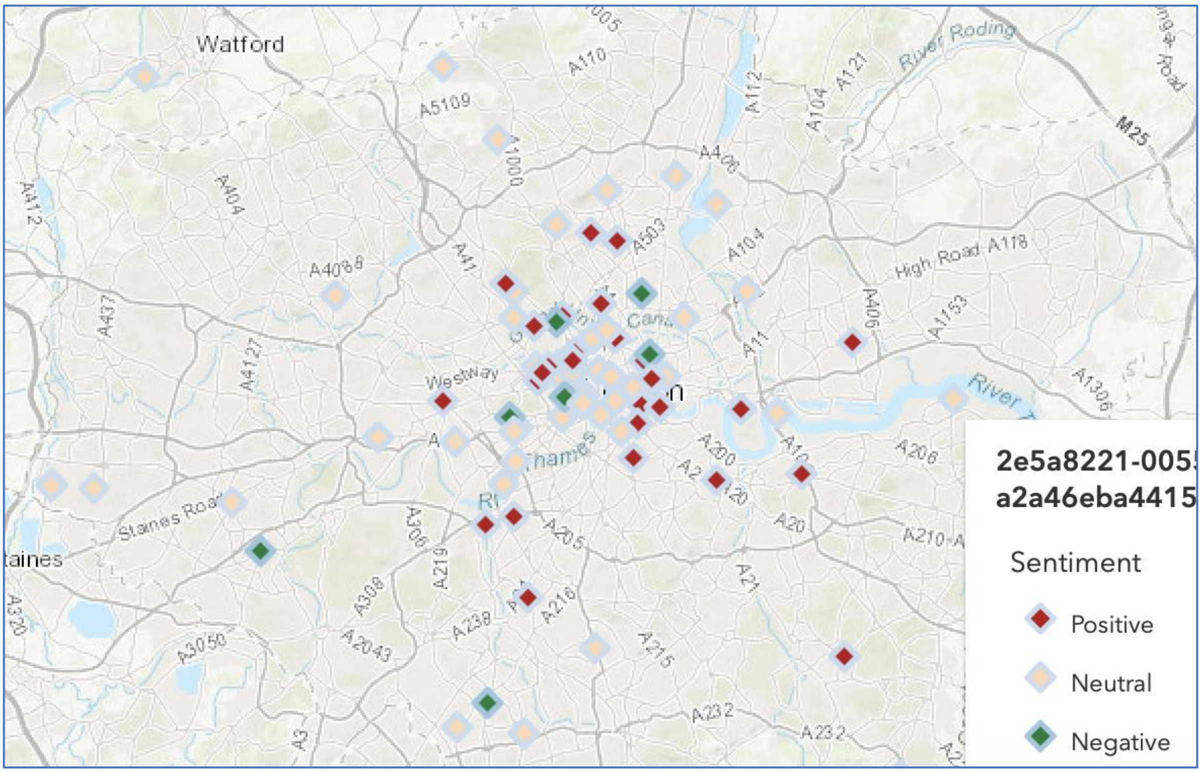


Figure 16 - Sentiment Analysis distribution during the off-peak hours in London

Those six graphs are taken as the outcome of this research to visualise the intensity layer in the London area. The shown legend in those graphs gives better user readability and users can zoom in on the map to reduce the area of the scope. According to the zoomed scope, the sentiment analysis or Hot Spot analysis visualisations were visualised.

According to the Hot Spot analysis graphs, there are more frequent tweets in the centre of London and those areas were coloured darker red. Apart from that, there are a few darker red spots in the graphs. When considering the sentiment analysis spatial graph in peak time, neutral tweets are centralised to the City of London and in off-peak hours neutral tweets have more spread through London.

Table 6 demonstrates the categorised cities in London with tweet counts. 28% of the tweets were published around the city of London. The city of London is the most tourist-attracted location according to Table 6. Apart from that Camden town and Islington cities also have a significant tourist attraction. Due to the limited number of tweets good deviation cannot be seen because 200 tweets in the dataset have to be spread among the 31 cities.

Place Name	Tweet counts
London	56
Camden Town	16
Islington	15
Camberwell	11
Paddington	10
Hackney	8
Poplar	7
Tottenham	7
City of London	7
Barnet	5
Hounslow	5
Hammersmith	5
Richmond	5
Wandsworth	4
Lambeth	4
Sutton	4
Bromley	4
Greenwich	3
Eltham	3
Hillingdon	3
Kensington	3
Kingston upon Thames	2
Brent	2
Merton	2
Walthamstow	2
East Ham	2
Southall	1
Watford	1
Croydon	1
Lewisham	1
Ealing	1

Table 6 - Tweet count of each city in the London area

4 Evaluation

Based on the comparison between the analysis that has developed in this research and the reference data evaluation proceeded. The reference data were collected from the 'Mayor of London' datastores [22]. The reference data contains tourist counts in the London area in 2014. The following Table shows the sample of the reference data. The pre-processing part of this project gives a meaningful dataset, and it gives major support for the applied methodologies.

Boroughs	Total Daytime Population (includes tourists)	Overseas Staying visitors	Domestic Staying Visitors	Day Trip Visitors
City of London	553,103	7,588	16,559	97,572
Westminster	897,293	95,328	5,332	75,282
Camden	495,332	21,053	3,241	36,759
Greenwich	254,966	3,170	1,658	29,782
Croydon	349,228	5,771	1,463	27,175
Southwark	417,029	6,875	1,189	26,383
Bromley	303,344	4,560	2,121	24,767
Islington	328,050	4,667	1,164	24,759
Harrow	307,478	3,064	942	20,882
Wandsworth	306,102	4,738	3,526	20,130
Hackney	252,831	2,600	1,134	18,639
Lambeth	229,311	7,160	2,216	18,145
Barnet	266,496	5,522	1,321	18,066

Table 7 - Sample reference data about tourism in London 2014

According to Table 7, Most tourists are roaming around the City of London. The frequency analysis in this project identified the urelement keywords to tourism. Based on that, only contains tourism-related tweets in the dataset were assumed. Considering Table 6 and Figure 7, most tourism-related tweets were published in the City of London. The second and third most tweets were published in Camden Town and Islington. But according to the reference data, Westminster and Camden got the second and third highest visitor counts in 2014. City of London, Westminster, Camden, Greenwich, Croydon, Southwark, Bromley, Islington Enfield, Newham, and Kensington are the top 10 cities that have more tourist integration according to the reference data. London, Camden Town, Islington, Camberwell, Paddington, Hackney, Poplar, Tottenham, City of London, and Barnet are the most travel-related tweets published. In this case, in the Twitter dataset, most users used 'London' as the location instead of the 'City of London'. Because of that, this Twitter dataset shows the city of London and London as different places. As well as the Twitter dataset contains tweets from a few hours in 2018 and it does not contain any tweets from the city of Westminster.

The sentiment analysis was considered one of the natural language processing methodologies, and it was applied to the Twitter dataset. Implementation of the sentiment analysis was done using the VADER algorithms. For all the comparisons in this project, consider the peak hours and off-peak hours. The evaluation is based on the output of the sentiment analysis which is the three spatial enabled graphs. According to Figure 14, Figure15, and Figure16, most of the positive and neutral tweets were published around the city of London city and negative tweets are very low compared to the positive and neutral tweets. In off-peak times, neutral tweets have more spread over the area compared to peak hours. Positive and neutral tweets in Figure 15 and Figure 16, illustrate that visitors were happy about those attractions or places.

Hot Spot Analysis was used in this project to determine the tourist locations based on published tweets in 2018. As an outcome of the project, those Hot Spots represent tourist intensity layers in London. Three intensity layers were created in this project considering the whole dataset, peak hours, and off-peak hours. According to the hot spots in Figure 12, Figure 13, and Figure 14, darker red colours indicate higher tweet counts and all the three hot spots around the City of London are marked as darker red. That means most visitors visited the city of London during peak and off-peak times.

The limited number of tweets in the dataset is the main obstacle. Considering the analysis carried out in this research, identifying the tourist location using Twitter plays a good job and most of the tourist locations were identified.

5 Discussion and Future work

As per the outcome of the project, we'll be able to extract a few of the products as follows,

- Sentimental analysis graph
- Tourist attraction Intensity layer for the peak time
- Tourist attraction Intensity layer for the off-peak time

Considering the outcome of the sentimental analysis of the Twitter data set that was used in the project, three sets of outcomes were categorized below referring the Table 3.

- Positive – set of the place that has a better emotion of among tourist attraction
- Neutral – set of the place that has a neutral emotion among a tourist attraction
- Negative - set of the place that has very bad feelings of among a tourist attraction

This outcome basically represents the attitude and the feeling that they try to expose in the tweets. Set of the result that represents 0 and 1, were classified as a positive set of locations. These are the set of places that have a very higher probability of being attractive tourist locations in London. Results that represent 0 were classified as neutral and 0 to -1 were classified as the negative set of places that have a very less probability of being a less attractive tourist location in London. Since the data set that was used in the project only covers 26/07/2018 at 13:39:37 to 26/07/2018 at 17:03:05. This was identified as a potential bottleneck of the research. Because the whole data set depends on a single date, there is a chance that the results can be biased. As an example, if there was any tube strike on this date that can be clearly affected by the tourist distribution of that date. Apart from its weather conditions on that date is also a major factor that affects the tourist distribution across London. One of the highest temperatures was reported on 26/07/2018 according to the past weather forecast [37].

As a future work of this project, this analysis needs to be extended to a long period of time including weekdays, weekends and holidays including peak, and off-peak times. As well as weather conditions of each day are also needed to consider. Then and only, the result set can be unbiased and accurate.

the tourist intensity layer with respect to the Twitter data. This can be an initial phase of many future projects such as monitoring the tourist behaviours around London. London is well known as a multicultural tourist attraction regardless the religion and nationality. Every day metropolitan police are giving a higher effort to ensure the safety of the tourist throughout the day. Monitoring the tourist behaviour around London is a very much important factor when it comes to the safety of the people. As a suggested future work, this project can be extended to identify tourist behaviour throughout the day. Then this need to be actively continued by extracting the data and updating the intensity layer.

Apart from safety, this is also can be used during a pandemic situation such as a covid pandemic, Extending the project to update the intensity layer repeatedly can help to identify the tourist distribution and the intensity of the area. This helps to make the most effective regulations and actions to reduce tourist distribution and avoid spreading the covid19. As well as the same system can be used to see the outcome of the regulations that were placed previously. This can be a very effective way of introducing regulations throughout a pandemic period. Not only for the pandemic situation but also for other situations as well, such as the example London Marathon Boston Bombings incident in 2013. Maintaining a frequently updated tourist intensity layer can be helping to identify the before and after the incident. This might be an effective way to help services to the people.

London is also unknown as the busiest city in the world, London also has the most effective and complex underground train system in the world. These were well planned to deliver the most effective service to the Londonist. Even though this service and the other TFL services can be extended to deliver tourist-catered service to the community if we can implement this research project with TFL services. This research is also tailored to deliver the tourist intensity of London, which means they are the place that people visit more and uses the TFL services more. This help to reduce the other services in the area according to the tourist distribution and increases the highly distributed places. As an example, Green Park tube station is getting busiest when there were some functions in Buckingham palace. Updating the intensity layer frequently with the data can help to decentralize the congestion during festival times.

Using the extracted tourist layer with a long period of data can be used to work on effective advertising. Furthermore, this project needs to extract the dining places and etc to cater for the highly accurate intensity layer of the tourist. These won be a tourist attraction layer but also a layer that can be identified as the most tourist socialized place in London. Such a product can be used to organize the advertising requirements for the people who are looking to advertise their business.

On the other hand, this project is able to extend to extract other data such as location and news tweets as well. We were able to run the same set of algorithms that I have used with the dataset, and We can develop a news intensity layer for crime or weather news similar to the tourist intensity layer. This will be a wide variety of outcomes. The importance of having such an intensity layer of crimes helps to identify the areas that need more metropolitan police attention. Furthermore, this can be classified into crimes as well, As an example one intensity layer for the house robberies in the London area and one for the murder cases. Each layer represents the previous crime distribution over the areas. These layers need to be extracted with larger data set covering a few years as well. Because the crime intensity layer won’t be able to plot significant data for a shorter time period. Even for this work, expect to scrape the data from personal profiles as well as the news channels as well. In that case, Twitter will be the best option, other than the rest of the social media platforms. Because Twitter itself focuses on one set of specific tasks. The beneficial side of having such a layer is not only identifying the areas but also helping to place the regulations and other needful actions. Such as, if the house robbery intensity layer reveals the areas that need frequent police patrols and security cameras. On the other hand, the same process can be used to evaluate the effectiveness of the place’s rules and actions. Such as, the crime intensity in these areas should reduce with time if the regulations are effective. Otherwise, it gives the chance to the metropolitan police to reconsider the placed regulations. A conclusion of the areas discussed in this section reveals that this project can be extended to many areas with future works and research projects by using the same algorithm.

6 Conclusions

The way people used geotagged Twitter by tourists in the London area helped analyse the main tourist attraction points in this area in a significant manner. Even with the advantage of the tourist intensity visualization, there wasn't much of the product that can be found in this research area in visualized maps. As per the outcome of this project that proved, the current product was limited due to many factors. Such as limited data available in this scope. The present project study also only used the data set of the 26th of July 2018 due to a lack of data availability. The project studies reveal that there was a significant relationship between the tweet data and the tourist attraction in the London area. As per the data set used in the project, implementation is able to extract 31 locations in London (Table 6). Due to the limited size of the data set that we have used, the number of tweets count that was extracted is low for most of the locations. Ealing, Lewisham, Croydon, Watford and Southall only have one tweet for each location. Thus, all the extracted locations can be considered tourist attractions. A larger number of tweets were extracted from the London city centre. It reveals that the user behaviour of Twitter users was significantly connected to the tourist attractions. As per the second and third most tweets were tagged into Camden town and Islington. That also confirmed the extracted result accuracy when compared with table 7 which is sample reference data about tourism in London in 2014.

The outcome of the frequency analysis also proved that extracted locations can be overlayered with final visualized maps. This outcome was used as input for sentiment analysis. This gave a promising result in the sentiment analysis and the location was mostly aligned with table 7. Sentiment analysis results revealed the attitudes and feelings of Twitter users towards the locations. The statistic referred to in table 7 significantly aligned with the extracted locations. As mentioned in the Sentiment analysis, positive and neutral results were considered the most attractive tourist location. Negative locations can't be considered tourist attractions since the user didn't show any positive attitudes toward them. These areas can be found in the final visualized map as small, faded patches that are located away from London city.

The final map visualization was done using Arc GIS. Red colour areas on the map were the densest areas with tourists and the green faded outline represents the lower intensity of the main attraction. As an example, the Final visualization represents the areas where the tourist was been other than the main attraction. Such as, Considering London city, the map revealed the areas of pubs and other places as well (figure 10). This is because of the advantage of geotagged tweets representing the other attractive areas as well. The final outcome of the project helps the new tourist that is expecting to visit London to choose the most attractive place. On the other hand, it also reviled the busy time of the areas with peak and off-peak intensity layers (figure 12).

These patterns and the extracted areas are very important places to many industries and the government as well. This information was reviling the dimension and the density of the tourist throughout the day. It helps the tourist industry and other areas as well, such as restaurants, and dining areas. If the data set supports the night-time, this will help to extract more information about the tourist behaviour in London. The final website included all 6 outcomes of the project. This research developed intensity layers and emotion classification using geo-tagged Tweets from London. The presented Twitter dataset demonstrates robust pre-processing techniques, text cleaning techniques, frequency analysis techniques, and sentiment analysis techniques. In this study, geo-tagged Tweets got from the London area only and the applied natural language processing approaches. Further to that, future studies could develop on the associations identified in this paper to create predictive tools to achieve their goals likely content of social media posts.

References

[1] Britannica (2019). Twitter | History, Description, & Uses. In: *Encyclopædia Britannica*. [online] Available at: https://www.britannica.com/topic/Twitter [Accessed 15 Jul. 2022].

[2] Cassa, C., Chunara, R., Mandl, K. and Brownstein, J.S. (2013). Twitter as a Sentinel in Emergency Situations: Lessons from the Boston Marathon Explosions. *PLoS Currents*. doi:10.1371/currents.dis.ad70cd1c8bc585e9470046cde334ee4b.

[3] Chaffey, D. (2022). *Global Social Media Research Summary 2021*. [online] Smart Insights. Available at: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ [Accessed 10 Sep. 2022].

[4] Citron, L. (n.d.). A TOURISM VISION FOR LONDON. [online] Available at: https://files.londonandpartners.com/l-and-p/assets/london_tourism_vision_aug_2017.pdf [Accessed 11 May 2022].

[5] Claster, W., Pardo, P., Cooper, M. and Tajeddini, K. (2013). Tourism, travel and tweets: algorithmic text analysis methodologies in tourism. *Middle East J. of Management*, 1(1), p.81. doi:10.1504/mejm.2013.054071.

[6] Daniel, J. and Martin, J. (n.d.). *Speech and Language Processing*. [online] Available at: https://web.stanford.edu/~jurafsky/slp3/3.pdf [Accessed 7 Sep. 2022].

[7] Dann, S. (2015). *Twitter Data Acquisition and Analysis: Methodology and Best Practice*. [online] Maximizing Commerce and Marketing Strategies through Micro-Blogging. Available at: https://www.igi-global.com/chapter/twitter-data-acquisition-and-analysis/131036 [Accessed 10 Jul. 2022].

[8] García-Palomares, J.C., Gutiérrez, J. and Mínguez, C. (2015). Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography*, 63, pp.408–417. doi:10.1016/j.apgeog.2015.08.002.

[9] Gerber, M.S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, pp.115–125. doi:10.1016/j.dss.2014.02.003.

[10] Giglio, S., Bertacchini, F., Bilotta, E. and Pantano, P. (2019). Using social media to identify tourism attractiveness in six Italian cities. *Tourism Management*, [online] 72, pp.306–312. doi:10.1016/j.tourman.2018.12.007.

[11] Goodchild, M.F. (2016). GIS in the Era of Big Data. *Cybergeo: European Journal of Geography*. [online] Available at: https://journals.openedition.org/cybergeo/27647?lang=en [Accessed 1 Sep. 2022].

[12] Gupta, V. and Hewett, R. (2017). *Harnessing the power of hashtags in tweet analytics*. [online] IEEE Xplore. doi:10.1109/BigData.2017.8258194.

[13] Hasnat, M.M. and Hasan, S. (2018). *Understanding Tourist Destination Choices from Geo-tagged Tweets*. [online] IEEE Xplore. doi:10.1109/ITSC.2018.8569237.

[14] Heimerl, F., Lohmann, S., Lange, S. and Ertl, T. (2014). Word Cloud Explorer: Text Analytics Based on Word Clouds. *2014 47th Hawaii International Conference on System Sciences*. doi:10.1109/hicss.2014.231.

[15] Huang, Y., Li, Y. and Shan, J. (2018). Spatial-Temporal Event Detection from Geo-Tagged Tweets. ISPRS International Journal of Geo-Information, 7(4), p.150. doi:10.3390/ijgi7040150.

[16] Hutto, C. and Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, [online] 8(1). Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14550 [Accessed 2 Aug. 2022].

[17] Jiang, W., Xiong, Z., Su, Q., Long, Y., Song, X. and Sun, P. (2021). Using Geotagged Social Media Data to Explore Sentiment Changes in Tourist Flow: A Spatiotemporal Analytical Framework. ISPRS International Journal of Geo-Information, 10(3), p.135. doi:10.3390/ijgi10030135.

[18] Jonn Elledge (2015). *Where are the largest cities in Britain?* [online] City Monitor. Available at: https://citymonitor.ai/environment/where-are-largest-cities-britain-1404 [Accessed 12 Sep. 2022].

[19] Kovacs-Gyori, A., Ristea, A., Havas, C., Resch, B. and Cabrera-Barona, P. (2018). #London2012: Towards Citizen-Contributed Urban Planning Through Sentiment Analysis of Twitter Data. *Urban Planning*, 3(1), p.75. doi:10.17645/up.v3i1.1287.

[20] Kumar, V. (2022). Spatiotemporal sentiment variation analysis of geotagged COVID-19 tweets from India using a hybrid deep learning model. Scientific Reports, 12(1). doi:10.1038/s41598-022-05974-6.

[21] Lansley, G. and Longley, P.A. (2016). The geography of Twitter topics in London. Computers, Environment and Urban Systems, [online] 58(0198-9715), pp.85–96. doi:10.1016/j.compenvurbsys.2016.04.002.

[22] London Datastore. (2009). *Tourism Trips, Borough – London Datastore*. [online] Available at: https://data.london.gov.uk/dataset/tourism-trips-borough [Accessed 14 Sep. 2022].

[23] Matters, T. for L. | E.J. (2018). Londoners encouraged to embrace the 'Wonderful World of Off-Peak' this summer. [online] Transport for London. Available at:

https://tfl.gov.uk/info-for/media/press-releases/2018/august/londoners-encouraged-toembrace-the-wonderful-world-of-off-peak-this-summer [Accessed 3 Jul. 2022].

[24] Memon, I., Chen, L., Majid, A., Lv, M., Hussain, I. and Chen, G. (2014). Travel Recommendation Using Geo-tagged Photos in Social Media for Tourist. *Wireless Personal Communications*, 80(4), pp.1347–1362. doi:10.1007/s11277-014-2082-7.

[25] Nandwani, P. and Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1). doi:10.1007/s13278-021-00776-6.

[26] Padilla, J.J., Kavak, H., Lynch, C.J., Gore, R.J. and Diallo, S.Y. (2018). Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. *PLOS ONE*, 13(6), p.e0198857. doi:10.1371/journal.pone.0198857.

[27] pro.arcgis.com. (n.d.). *An overview of the Spatial Statistics toolbox—ArcGIS Pro | Documentation*. [online] Available at: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/an-overview-of-the-spatial-statistics-toolbox.htm [Accessed 8 Sep. 2022].

[28] pro.arcgis.com. (n.d.). *What is a z-score? What is a p-value?—ArcGIS Pro | Documentation*. [online] Available at: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm [Accessed 10 Sep. 2022].

[29] Routledge, J. (n.d.). *THE ROUTLEDGE HANDBOOK OF TOURISM GEOGRAPHIES*. [online] Available at: https://www.gbv.de/dms/goettingen/645266957.pdf [Accessed 15 Jul. 2022].

[30] Sakaki, T., Okazaki, M. and Matsuo, Y. (2010). Earthquake shakes Twitter users. *Proceedings of the 19th international conference on World wide web - WWW '10*. doi:10.1145/1772690.1772777.

[31] Saraf, S. (n.d.). Twitter_classification. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/shubhamsarafo/twitter-classification?select=london.json [Accessed 2 Jul. 2022].

[32] Scholz, J. and Jeznik, J. (2020). Evaluating Geo-Tagged Twitter Data to Analyze Tourist Flows in Styria, Austria. ISPRS International Journal of Geo-Information, 9(11), p.681. doi:10.3390/ijgi9110681.

[33] Senaratne, H., Bröring, A., Schreck, T. and Lehle, D. (2014). Moving on Twitter. *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks - LBSN '14*. doi:10.1145/2755492.2755497.

[34] Shif Ben Avraham (2017). *What is REST — A Simple Explanation for Beginners, Part 1: Introduction*. [online] Medium. Available at: https://medium.com/extend/what-is-rest-a-simple-explanation-for-beginners-part-1-introduction-b4a072f8740f [Accessed 13 Jun. 2022].

[35] Statista Research Department (2022). *Daily time spent on social networking by internet users worldwide from 2012 to 2022*. [online] Statista. Available at: https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/ [Accessed 10 Sep. 2022].

[36] t-redactyl.io. (2017). *Using VADER to handle sentiment analysis with social media text.* [online] Available at: https://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html [Accessed 1 Aug. 2022].

[37] www.timeanddate.com. (n.d.). *Weather in July 2018 in London, England, United Kingdom*. [online] Available at: https://www.timeanddate.com/weather/uk/london/historic?month=7&year=2018 [Accessed 14 Sep. 2022].

[38] Zagheni, E., Garimella, V.R.K., Weber, I. and State, B. (2014). Inferring international and internal migration patterns from Twitter data. *Proceedings of the 23rd International Conference on World Wide Web*. doi:10.1145/2567948.2576930.

[39] Zhu, R., Lin, D., Jendryke, M., Zuo, C., Ding, L. and Meng, L. (2018). Geo-Tagged Social Media Data-Based Analytical Approach for Perceiving Impacts of Social Events. ISPRS International Journal of Geo-Information, 8(1), p.15. doi:10.3390/ijgi8010015.

Bibliography

[1] Häberle, M., Werner, M. and Zhu, X.X. (2019). Geo-spatial text-mining from Twitter – a feature space analysis with a view toward building classification in urban regions. European Journal of Remote Sensing, 52(sup2), pp.2–11. doi:10.1080/22797254.2019.1586451.

[2] Han, B., Cook, P. and Baldwin, T. (2014). Text-Based Twitter User Geolocation Prediction. Journal of Artificial Intelligence Research, 49, pp.451–500. doi:10.1613/jair.4200.

[3] Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. and Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41(3), pp.260–271. doi:10.1080/15230406.2014.890072.

[4] Oku, K., Hattori, F. and Kawagoe, K. (2015). Tweet-mapping Method for Tourist Spots Based on Now-Tweets and Spot-photos. Procedia Computer Science, 60, pp.1318–1327. doi:10.1016/j.procs.2015.08.202.

[5] www.lexalytics.com. (2020). Machine Learning (ML) for Natural Language Processing (NLP) - Lexalytics. [online] Available at: https://www.lexalytics.com/blog/machine-learning-natural-language-processing/ [Accessed 20 Jul. 2022].