Identification of local touristic attractions in London, Edinburgh, and Manchester by using geo-tagged Twitter posts

Abstract

In this project, a geospatial analysis combined with NLP implementation is done in order to identify local touristic places for three cities in the UK which are London, Manchester, and Edinburgh. OpenStreetMap data was used to extract amenities and the geographical coordinates of such amenities, and Twitter data was used to perform sentiment analysis on tweets published by users in the three cities mentioned. Numerous research used geotagged tweets to examine crowd behaviour, emotion and sentiment analysis, quality of life assessment, city dynamics, and land use, and this project aims to combine the sentiment analysis with geospatial analysis to examine the touristic behaviour and touristic places in the three cities. The project offers a complete overview on the factors that affect touristic behaviour and assesses the performance of local touristic places as well. The NLP model differentiates the tweets analysed into positive and negative tweets in order to evaluate the feedback provided from the tweets on the amenities in the data. The NLP model achieved suitable results that match the objective of this study with an accuracy of 92% for the sentiment analysis. A dashboard is created to visualise the data acquired to exhibit the insights obtained in this project using Microsoft PowerBI. The project aims to strengthen the engagement with touristic places by providing tourists with useful information for the cities in the study, and at the same time provide urban planners and decision makers with insights that benefit the approach in improving the touristic services to maintain a stable balance in the cities' touristic behaviour. The advice given is a generic advice based on different interests and has no impact on any particular social group, the resources and the product delivered avoid any legal, social, or ethical obstacles.

Keywords: Natural Language Processing; Sentiment Analysis; Twitter; OpenStreetMap; QGIS; PowerBI; Touristic Areas; London; Manchester; Edinburgh

Table of Contents

Table of Figures

Table of Tables

1. Introduction

In recent years, rapid developments in social media services such as Twitter have created a user-friendly and easily accessible space for users to use such platforms interactively, to share day-to-day events and location information. Users publish up-to-date short texts called "Tweet", which they share with associated circumstances and daily activities, with a maximum limit of 280 characters. Tweets consist of two primary parts: a text that can only be 280 characters long and metadata. The metadata for a tweet includes the timestamp, the user's ID, and the location's coordinates.

Users of Twitter can view some other users' tweets by following them. Users can choose to share their tweets with the public profile or specific users. Two popular tweet features that provide information about the locations where tweets are distributed as "coordinates" and "place name". When examining people's spatial behaviour, the geotagged data that was connected to the tweets is useful.

Numerous research used geotagged tweets to examine crowd behaviour, sentiment analysis, quality of life assessment, city dynamics, and land use. In some studies, it is suggested that tweet coordinates are not sufficient and accurate sources, also it might cause a problem for the scientific research where on some occasions users refuse to share their locations in social media platforms and inappropriately tag the location of some places without physically visiting the venue while providing textual information regarding the venue. Contrary to the studies mentioned before, this paper aims to show that accessing coordinate information and user text data of the visited venue is an efficient and reliable source.

In this study, twitter data was obtained from publicly available dataset; This dataset of 169034 lines containing Tweet Id, Date, Hour, User Name, Nickname, Bio, Tweet Content, Favs, RTs, Latitude, Longitude, City, Country, Profile picture, Followers, Following, Tweet Language, and Tweet URL information, in this study the information required has been extracted and reduced to a dataset containing Tweet, Latitude, Longitude, Country, City, and Language information.

OpenStreetMap (OSM) helps facilitate the production of vast amounts of geographical data from community users. OSM is utilized by businesses as well as end users to enable map

applications, location suggestions, and numerous other geospatial services. The OSM dataset for this study could be downloaded on its own, but it was time consuming to switch back and forth between the data source and the development environment, and in addition, only the Point of Interest (POI) data for the cities of Edinburgh, Manchester and London is desired to be obtained. Therefore, The OSM data was directly retrieved to the notebook with the OSMnx Python library, consequently the data was acquired and processed. Two different forms of information about geographic objects are intended to be made available by OSM:

1- geographical boundaries like lines, points, and regions,
2- Tags. The "Key" and "Value" of a tag are used to describe the objects.

Examples of objects include tourist attractions, business locations, and points of interest. While the values describe the individual qualities, the keys give a broad class of features, such as "leisure"="park" and "amenity="restaurant". OpenStreetMap data was used to make general review predictions for POI amenities in London, Manchester, and Edinburgh. OpenStreetMap data was taken separately for each category for London, Manchester, and Edinburgh. The 11 categories are as follows: Pubs, Restaurant, Cafe, Townhall, Post office, Marketplace, Nightclub, Cinema, Social Facility, Bank and Art Centre.

In this study, twitter dataset and OpenStreetMap dataset were combined and an accurate estimation, whether positive or negative, was made for the Points of Interests in the specified cities. In addition, the distance and duration of each venue from the centre was determined.

A general profile about the venue was obtained by predicting positive or negative reviews about the visited location where the text data was obtained from the tweets of the users who tagged the venues. It is also valid data to infer the consumer's interest in the service in order to be able to classify all users' comments with correct accuracy and efficient predictions about the venue. The tweet coordinates, triggered by the customer experience or word of mouth (WOM) and shared the review for the venue based on a daily dialogue, helped the project to be handled from a wider scope, as it would not cause a problem in terms of accuracy if the users did not exactly tag the location at the venue. Since the Twitter dataset does not contain venue names and categories, this information is provided in a single dataset by merging it with the OpenStreetMap dataset. Since one of the goals is to see the tweets of place names according to their point of interest classes, the OSM dataset coordinates and the

tweet dataset coordinates are combined according to the closest spatial location using QGIS features. In addition to making a direct analysis of the places, all the tweets were maintained in the combined data regardless of their distance from the closest amenity in order to analyse whether the regions are also recommended by tourists.

Figure 1 shows the plot graphs of each city in the study along with the area graph, the plot graph can be seen with the black background which shows the boundaries set for each city, and the area graph is shown in blue to highlight the area covered in each city.

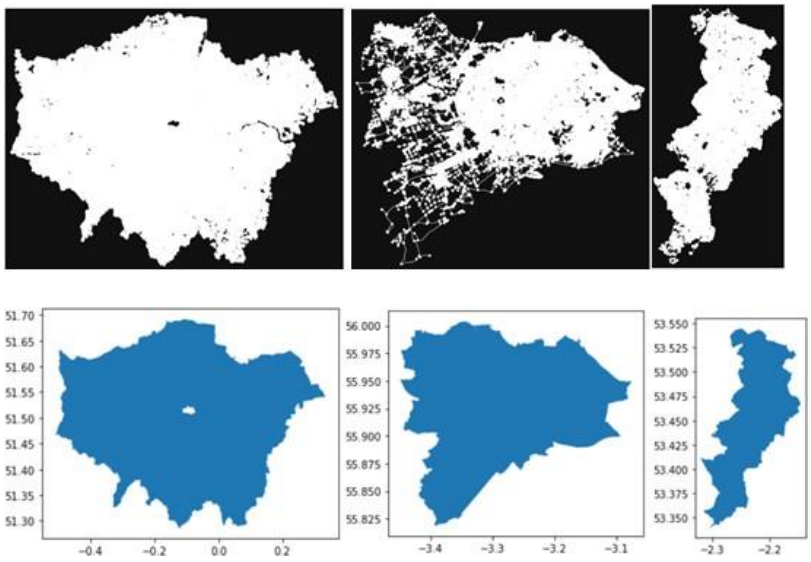

Figure 1 Plot and Area Graphs for London, Manchester, and Edinburgh

In this project, it was aimed to identify local touristic places in London, Manchester, and Edinburgh. Tourism, urban planning, transportation availability, and small businesses, etc. can contribute. To find the best locations for entrepreneurs and business owners, changes in the local customer population might be observed. Analysing cross-visit patterns to guide affiliate marketing and advertising expenditures can be beneficial to create new business opportunities. Changes in how people move and use public spaces can be measured. This can optimize population density models and needs assessments for urban planners and engineers. Apart from that, it can be used, for example, for planning and operations for public transport, maintenance and security of public spaces, and real estate acquisition and development.

Legal and ethical concerns are unaffected by this project's usage of free, publicly accessible resources, and the data collected doesn't contain any private information because it was obtained from a publicly available Twitter dataset. Regarding social difficulties, the advice

given is a generic advice based on different interests and has no impact on any particular social group. The resources and the final product totally avoid these problems, and the project does not call for any individual surveys or questionnaires, thus it is free of any legal, social, or ethical obstacles.

2. Literature Review

Many academic articles that conducted similar studies before were used throughout the studies, the keywords used to find articles are OSM, Twitter API, QGIS, Point of Interest etc. As a result of the research, a more concrete framework was created for the project with the following similar studies. The following papers are studies that have been of great benefit to the project, and more articles have been read for the project being worked on and noted as future references to be used in the future work of the project.

(Zohar, 2021) studied the locations of tweets were extracted using three of the meta data. The spatial link between the acquired sites was then examined using the GeoNames and Open Street Map datasets. It was discovered that there is a strong correlation between the distance and the inferred locations utilizing the text and user-location meta-fields and the distance in their midpoint and the device location. As a result, it might reveal both the precise position of the device and the scene of the occurrence mentioned in the tweet. For inferences involving text and user location, text and user description, and user location, as well as triple location, the centre point was calculated. The distance to the device's position was determined using these central points. The OSM dataset-classified distance distributions and several relationship kinds were presented. The method used to find the distance from the inferred location to the centre point provided with this study is useful for this project.

(Mittal, 2019) presented a study that provides support for the precise location matching of social media data to physical sites. Machine learning and deep learning algorithms were used to properly map physical location utilizing Geo-tweet data and Flickr image data, using geo-tagged data from Twitter and Open Street Map in cities like London as well as picture mapping with Flickr data in cities like London and Kyoto. A preliminary explanation of this work was offered in this paper, this paper provides the distance between the tweet's central location and the geotagged social media item's location (i.e. the spread distance value) by using latitude and longitude of the tweet and the latitude and longitude obtained for the place.

Logistic Regression, and CNN methodologies were used for NLP in this project, the method of finding the distance between the geotagged location and the central used in this paper has been the inspiration for this project.

(Khalilnezhad, 2022) provided a case study for London, in this study, it was aimed to investigate the Twitter features and people's behaviour of using parks in London. Tweets from London residents were collected, in this study, unlike other studies, tweet texts shared for parks were used instead of coordinates, in addition, a separate dataset containing park names was created using OpenStreetMap queries. The average distance travelled by the visitors of each park to visit the park and the number of visitors to the parks were obtained in the study, the reason for this was to evaluate the validity of using the text data of tweets to make environmental and spatial assessments. As a result, the approach using text data instead of coordinates in the project showed that the findings obtained using textual applications are as valid as studies applying tweet coordination. The approach used for textual data in this project is considered to be a viable method for this project as the data structure for both studies are similar.

3. Execution

3.1 Data

In this study, 3 datasets were used, and the data was limited to 3 cities in the United Kingdom which are London, Manchester, and Edinburgh, as the geographical study area examined. In this study, Twitter was used as a source to obtain the necessary information, latitude and longitude coordinates were used to detect tweets posted in the three cities mentioned. The dataset consisting of 160K rows publicly available was cleaned to include only tweets from the cities of London, Edinburgh, and Manchester, which are the focus of the project, and the raw data downloaded was prepared to fit the aim of this project, a dataset of 40438 rows was obtained. The dataset includes tweets, latitude, longitude, country, city, language columns.

OpenStreetMap is used to extract the geographic locations of the features. The OSM dataset for this study could be downloaded on its own, but only the Point of Interest data for the cities of Edinburgh, Manchester and London is desired to be obtained along with the amenity’s general information such as name, amenity etc. Therefore, the OSM data was directly retrieved to the notebook with the OSMnx Python library, the data was acquired and

processed consequently. OSMnx is a Python package that helps download geospatial data from OSM, GDAL and Fiona python packages were utilized to acquire the data from OSMnx. Two different forms of information about geographic objects are intended to be made available by OSM:

1- Geographical boundaries like lines, points, and regions
2- Tags. The "Key" and "Value" of a tag are used to describe the objects

OpenStreetMap data was obtained by specifying "London" as the address where the data should be downloaded with the "Place_name" key. Using the parameter “amenities” a list of OSM amenity categories were passed using the “geometries_from_place” function. The 11 categories are as follows: Pubs, Restaurant, Cafe, Townhall, Post office, Marketplace, Nightclub, Cinema, Social Facility, Bank and Art Centre. Each category of amenity contained many columns, the required columns are amenity, osmid, name, Address:post_code, geometry, add:street , add:city were selected for each category, thus creating separate datasets containing the same columns for each category, these datasets had been appended. A dataset containing the properties of amenities belonging to 11 different amenity categories containing the columns mentioned for London was obtained. After all the steps were performed for the cities of Manchester and Edinburgh, the datasets created for these cities were appended. As a result, a dataset of 21671 rows containing the information of 11 different amenity categories for the 3 cities was obtained.

The OSM dataset coordinates and the twitter dataset coordinates are joined using QGIS features based on the nearest spatial attributes location in order to observe the tweets of place names according to their Point of Interest classifications. Twitter dataset and OpenStreetMap dataset are loaded into QGIS as 2 separate layers. Thus, the Twitter dataset and OpenStreetMap coordinates can determine the name and other characteristics of the closest amenity to the tweet; provided with a new merged dataset created using QGIS's “Join attributes by nearast” feature. The features are combined according to the closest one. For each feature in the first input layer, the closest feature from the second input layer is determined. Twitter dataset was determined as the first input layer and OpenStreetMap dataset was determined as the second input layer, the combined dataset includes the columns of both datasets and the distance between the merged points, the longitude="feature_x and latitude="feature_y" information of the Twitter data selected as the target, and the

longitude="nearest_x" and latitude="nearast_y" information of the OpenStreetMap data selected as the nearest data. The combined dataset provides a more complete overview by connecting the tweets with the amenities according to their geographical coordinates, with this approximate distance analysis, it is predicted that the region where the tweets are posted will help users to form an idea about the region by showing whether it is a positive or a negative region. The merged dataset consists of 41967 data rows. All the tweets were maintained in the combined data regardless of their distance from the closest amenity in order to analyse whether the regions are also recommended by tourists, in addition to performing a direct examination of the locations to determine whether the regions are also favoured by visitors.

3.2 Preprocessing

As mentioned in Chapter 3.1, 3 datasets were studied for the project. Twitter Dataset is mainly based on text data and coordinates containing tweets from the cities of interest, OpenStreetMap dataset is mainly based on amenities and coordinates, it is aimed to match the tweets posted with the amenities in the OSM dataset, and with the help of QGIS, features are measured by determining the closest point to the attributes based on the coordinate columns of the Twitter and OpenStreetMap datasets, and nearest points were joined, and the updated dataset was saved as the last dataset.

Relevant columns were filtered from the publicly available pre-recorded dataset and recreated as Tweet content': 'Text', 'Tweet language (ISO 639-1)': 'Language', 'Place (as appears on Bio)': 'City'} has been named, the language has been filtered to be English only. Null values are removed to provide a more reasonable estimation, the ‘City’ column has been cleaned of all other cities except London, Edinburgh, and Manchester. Phrases such as “Eltham London”, “Manchester England” were replaced with “London”, “Manchester”, “Edinburgh” to form a single united format across the data.

Pre-processing is required to remove irrelevant tweets and reduce the semantic dimensions of noisy data. Therefore, such tweets were deleted from the dataset in the first pre-processing step. In the second step, optimization methods were also applied to ensure that users could provide reliable data for our study. Tweets are usually composed of incomplete expression, a variety of noise and poorly structured sentences because of the frequent presence of

acronyms, irregular grammar, ill-formed words, and non-dictionary terms. Noise and unstructured Twitter data would affect the performance of tweet sentiment classification. Prior to feature selection, a series of pre-processing are performed to tweets to reduce the noise in the micro-blog text. The pre-processing steps are implemented to satisfy the following:

- All non-ASCII and non-English characters be removed from the tweets.
- All URL links be removed as URLs do not contain the sentiment information of the tweet; therefore, it was deleted from tweets accordingly.
- Numbers were cleaned and removed, numbers generally do not contain sentiment information, so it was more efficient to remove the numbers when measuring sentiments and hence all numbers were deleted from tweets in order to refine the tweet content.
- Repetitive characters were cleaned and removed. Removing duplicated data was necessary to conduct a more comprehensive analysis.
- Punctuations were cleaned and removed.
- All characters were set to lower case to avoid distinctions of identical words.
- Stop words such as "the", "above", and "of" have been removed, cleaned up using the classical method of obtaining stop words from a precompiled list.
- Replace emoticons and emojis; The emoticons and emojis are a writer's mood expression in the form of icons in the tweet. the emoticons and emoji were replaced with their origin text form.
- Tokenization.
- Lemmatization.
- Stemming.

3.3 QGIS

Since one of the goals is to classify amenities, tags obtained from OpenStreetMap are assigned, labels obtained from OpenStreetMap are assigned, as one of the goals is to classify amenities. Specifically for London, Manchester, and Edinburgh, by using QGIS' "Join attributes by nearest" feature, tweets were combined with the closest geo-referenced amenity.

All tweets, even if they are far away, were kept in order to be able to approach and analyse the proximity of a point given by the Euclidean distance in the WGS84 coordinate space. The

large distance between tweet coordinates and hub coordinates made the analysis to be more appropriate, thus supporting the analysis that the distance from an amenity to the point where the tweet was posted can lead to investment opportunities for small businesses about the region where the amenity is located. More detailed information has been given in the Visualization chapter. After the join, an attribute table consisting of 40980 rows was obtained.

The OpenStreetMap dataset imported into QGIS was copied to create a more convenient display of point layers and facilitate analysis, each copy was filtered for each amenity. On eleven new layers, the distance between the amenities was determined first.

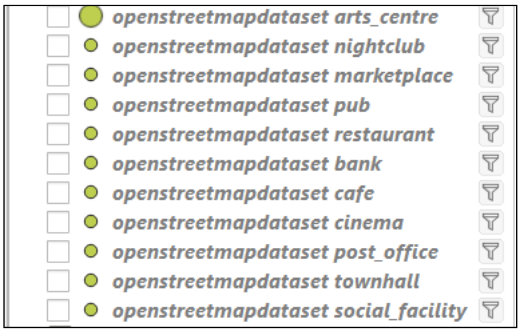

Figure 2 Amenity Layers

Figure 2 shows the eleven layers created for each amenity, the selected layer was displayed as points in the cities of London, Edinburgh and Manchester on the map in the platform.

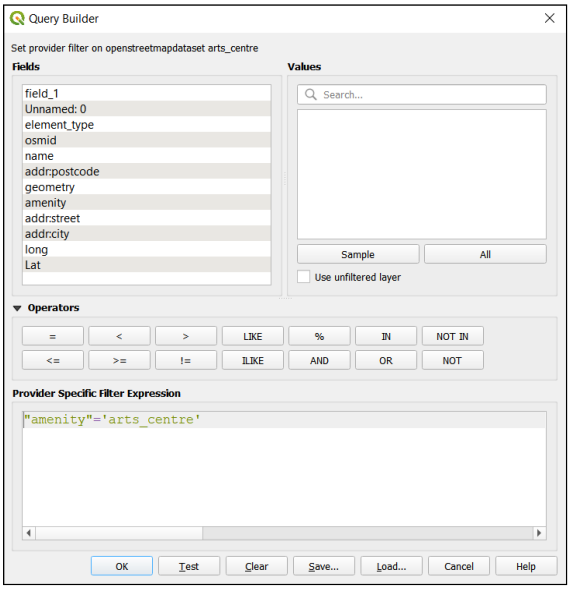

Figure 3 OSM layer filtering

Figure 3 shows how the copied OpenStreetMap layer was filtered for each amenity, in this figure, it can be seen that the filter was applied on the arts_centre amenity and presents all arts_centres in three cities on the map as shown in figure 4 below:

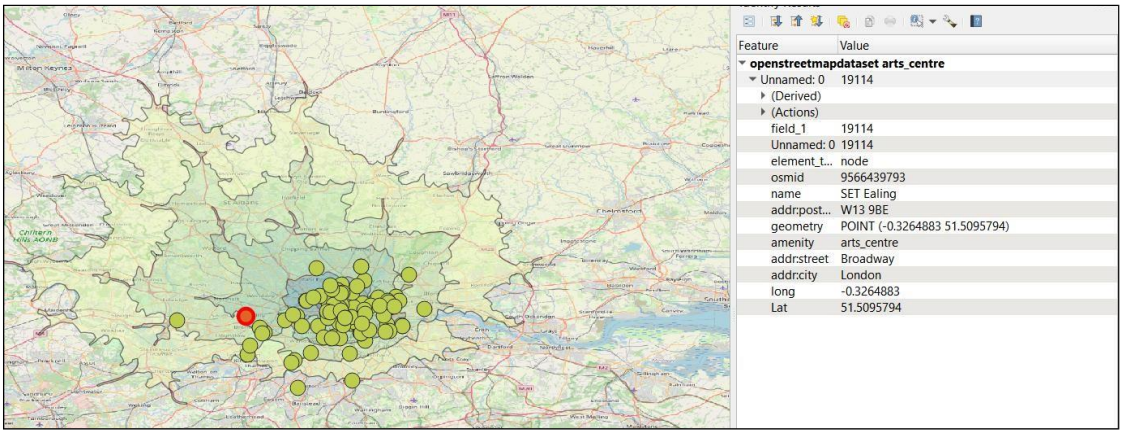

Figure 4 Art Centers in London

Figure 4 shows all the arts_centre points in London on the map. The red point is one of the arts_centre selected for review, as can be seen from the Identity Results tab, important information such as latitude, longitude, address, and name are presented for the selected point.

With the "Nearest neighbour analysis", the distance between the nearest features was determined, the layers previously created for each amenity were selected as input layer and input layer 2 where feature to feature distance was applied between these amenities for the cities of London, Manchester and Edinburgh: post_office to social facility, bank to post office, pub to bank, cafe to pub, 11 layers were created showing the distances and features between restaurant to café, social facility to nightclub, arts_centre to townhall, townhall to restaurant, cinema to arts_centre, nightclub to marketplace, marketplace to cinema. Recommending the closest hub in a different category from the current hub can increase the awareness and popularity of hubs that may not have been discovered by visitors yet, since it is local.

When the point to be examined is clicked, features such as the distance to the nearest amenity, the name of the amenity and the coordinates of both points are provided. The style, colour, symbol, and size of each layer are designed to be easily detectable on the map. As seen in figure 6, figure 7, and figure 8 below; the name, postcode, geometry, amenity, street

address, city, longitude, latitude information of the determined hub, hub name, and hub distance information of the hub to be reached were obtained. Figure 5 shows the eleven hub distance layers created from feature to feature.

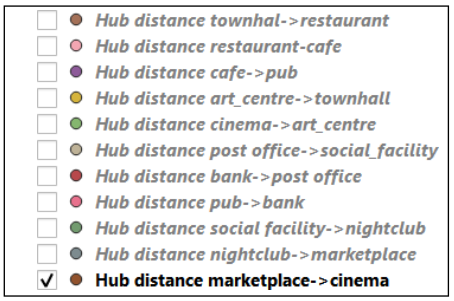

Figure 5 Hub Distance Layers

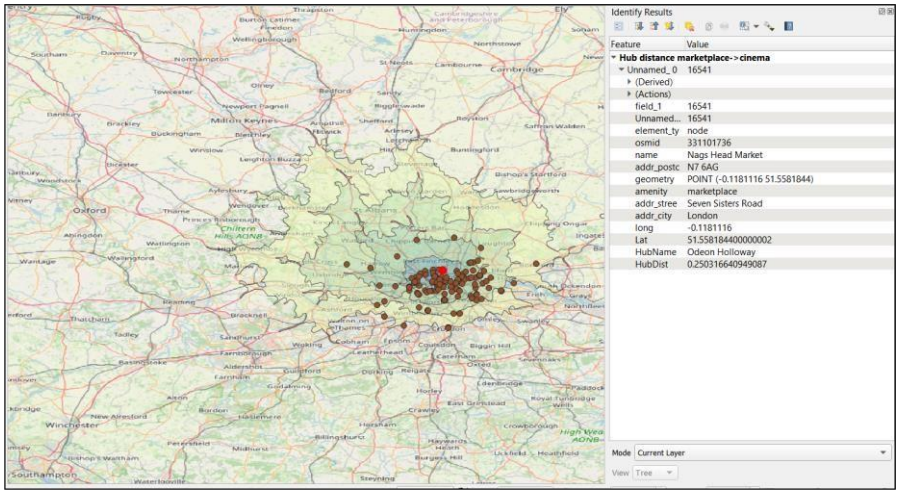

Figure 6 London's Marketplace - Cinema Hub Layer

Figure 6 shows the marketplaces in London with brown points. As can be seen in the figure, the red point is specially selected, it is aimed to determine the closest cinema to this point. The Identity Results tab shows the hub distance from the marketplace to the cinema, in addition, it also provides important information such as geometry information and name of the selected marketplace. For example, the closest cinema to the marketplace named "Nags Head Market" in London is "Odeon Holloway" and the distance is 0.25031 miles. The locations chosen between the hubs were selected according to the user's interest. Desiring to view the nearest cinemas from the Marketplace is a logical combination that can be in the daily plans of the visitors, the nearest cinema suggestion can be another interesting activity after the marketplace tour.

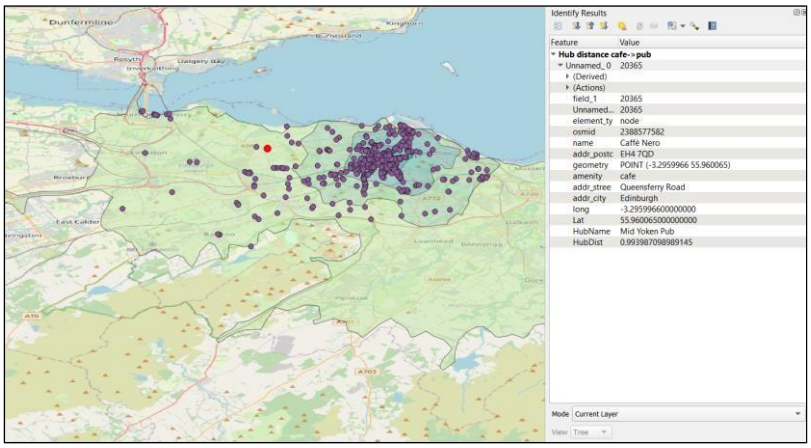

Figure 7 Edinburgh's Cafe - Pub Hub Layer

Figure 7 shows the cafes in Edinburgh with purple points, as can be seen in the figure, the red point is a specially chosen café, it is aimed to determine the closest pub to this point. The Identity Results tab displays the hub distance from the cafe to the pub, in addition, it also provides important information such as geotmery information and name of the selected cafe. For example, the closest pub to the cafe named "Caffe Nero" in Edinburgh is the "Mid Yoken Pub" and the distance is 0.9939 miles.

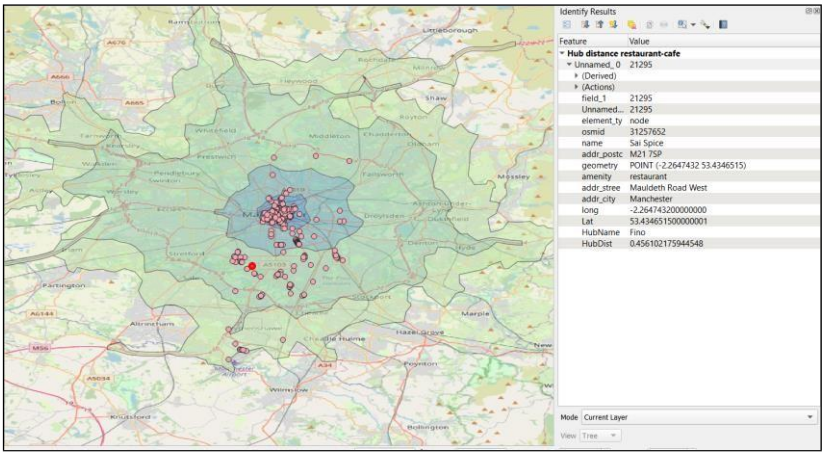

Figure 8 Manchester's Restaurant - Cafe Hub Layer

Figure 8 shows the restaurants in Manchester with pink points, as can be seen in the figure, the red point is a specially selected restaurant, it is aimed to determine the closest cafe to this point. The Identity Results tab shows the hub distance from the restaurant to the café, in addition, it also provides important information such as geotmery information and name of the selected cafe. The closest café to the restaurant named "Sai Spice" in Manchester is the "Fino" café and the distance is 0.4561 miles.

QGIS features are bolstered with the Open Route Services (ORS Tools) plugin, which helps determine the shortest path and fastest path calculations and hence the calculations were executed. Roads are obtained as line features, with information such as the distance and time required to travel from the starting point to the destination being available to achieve useful insights. As seen in Figure 10, the shortest and fastest directions were determined from banks to restaurants in Edinburgh, from arts centers to marketplaces in London, and from pubs to nightclubs in Manchester by car. Osmid was selected as the start id field for the Input start point layer and the name was selected as the end id field of the input end point layer, the reason for this choice was to create a more understandable output for visualization which will be explored in more details in the Visualization section later on.

Figure 9 Amenity Layers by City

Figure 9 shows the six new layers created by filtering according to the cities where the points were intended to be shown after copying the selected amenity layers, in order to show the shortest and fastest paths. Input start point layers were selected in blue color and larger size, red color and smaller size were selected for the end point layer, since it was not desired to cause confusion while detecting directons on the map.

Figure 10 Direction Layers

Figure 10 shows the direction layers created for the shortest paths and fastest paths for each city. Direction layer styles created for fastest paths were chosen as purple line and direction layer styles created for shortest path were chosen as blue dash line. Thus, it was easier to see the difference when comparing the fastest and shortest path on the map.

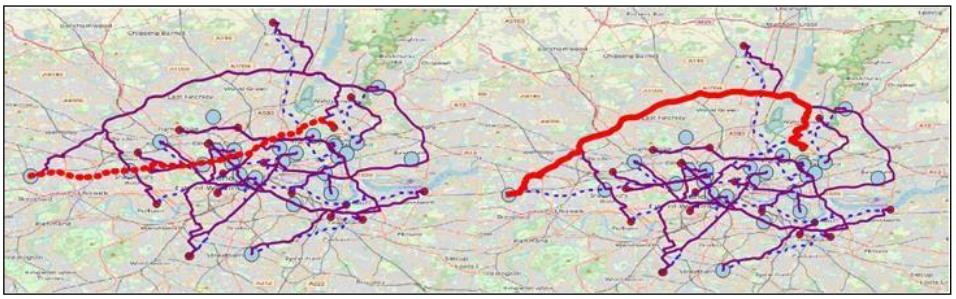

Figure 11 Shortest & Fastest Paths between Art Centers and Marketplaces in London

Figure 11 above and Table 1 below show the shortest and fastest path for the arts center and marketplaces in London. As mentioned before, the purple line on the figure shows the fastest and the blue dash line shows the shortest path, The red line on the map shows the routes that can be reached from the selected arts center to the market places in the shortest way. As can be seen in table 1, the distance from the art center with the ID '9566439793' to the "New Spitalfields Market" is approximately 24.75 km and the duration takes around 1.05 hour which amounts to 63 minutes by car. As for the fastest route, the distance from the fastest road is approximately 35.71 km, and the journey takes around 0.983 hours which amounts to roughly 59 minutes, it can be noticed that the fastest path is longer by almost 10 km but takes approximately 4 minutes less to travel by car. The direction of the destination is important for visitors, the distance and duration of the best way to reach the destination can affect the opinions about the place. As seen in Figure 11, while the shortest route should be the one that can be reached in the shortest time, it was not the case for some paths, therefore an alternative route can be suggested to the visitors showing the fastest route. In addition, traveling long distances even though it takes a short time can have a negative impact on an attractive hub, which should be noted for urban planners and civil engineers.

Table 1 Shortest & Fastest Paths Information

Identify Results

Feature	Value
Directions London arts_centre-market_place shortest from o...	
PROFILE	driving-car
(Derived)	
(Actions)	
DIST_KM	24.745999999999999
DURATIO...	1.054000000000000
PROFILE	driving-car
PREF	shortest
OPTIONS	None
FROM_ID	9566439793
TO_ID	New Spitalfields Market

Identify Results

Feature	Value
Directions London arts_centre-market_place fastest from os...	
PROFILE	driving-car
(Derived)	
(Actions)	
DIST_KM	35.712000000000003
DURATIO...	0.983
PROFILE	driving-car
PREF	fastest
OPTIONS	None
FROM_ID	9566439793
TO_ID	New Spitalfields Market

Finally, to further refine the location analysis, the isochrone option was used to show isochrones. The isochrones of each city were shown to display where one can reach from a

starting point in a given time, durations were assigned for each city as shown in figure 12, the central coordinates of the cities were used as the starting point for all isochrones. As seen in figure 12, it can be estimated approximately how many minutes away from the center of the city to all each isochrone region for each city. The points for each amenity can be assessed using the isochrones provided to estimate the duration between the area of the amenity and the center point of the city.

Figure 12 Isochron durations for each city

Figure 13 below shows the isochrones created on the map for each city, with each isochrone being distinguished in a different color based on the duration needed to arrive at the isochrone from the center point of each city. The figure shows the isochrones for London, Edinburgh, and Manchester respectively.

Figure 13 Isochrones for London, Edinburgh, and Manchester on the Map

The places visualised on the isochrones created for each city are shown in figure 14, the isochrones can be used as an estimate for the duration needed to go from one place to another based on the place's location and position on the isochrones. Figure 14 below shows all the places on the isochrones for London, Edinburgh, and Manchester, respectively. As seen in the Figure, there are many amenities in clusters that overflow out of isochrone in London and Edinburgh, it may be necessary to make an inspection for transportation to the amenities that are more than 55 minutes away from central London, these regions containing dense places

that can attract visitors and develop tourism. Transportation methods and prices may be a subject to be studied as well.

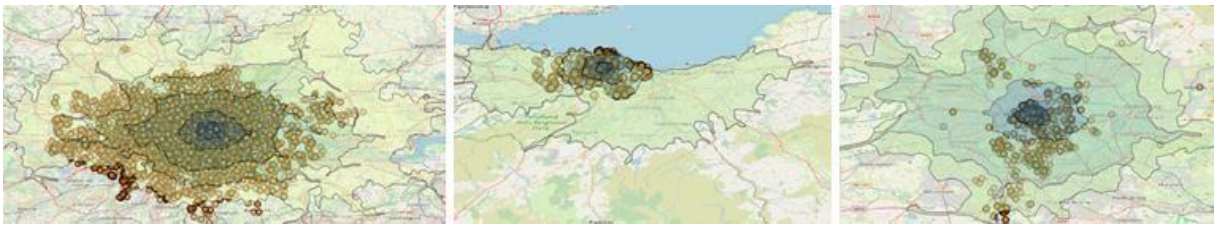

Figure 14 Places Shown on the London, Edinburgh, and Manchester Isochrones

3.4 NLP

3.4.1 Sentiment Analysis

Sentiment analysis involves identifying and categorising the sentiments conveyed in the text source. When analysed, tweets can often produce a significant amount of sentiment data, this data is useful in understanding the opinion of the people about a variety of topics. Lexicon-based classification was used in this study, a classifier is a set of rules by which words are classified as positive or negative along with their corresponding measures of density, generic lexicon methods were thought to provide a good performance since feedbacks for POIs were determined and analysed. The lexicon-based sentiment classifier VADER was used in this study, VADER is an NLTK-based NLP library that comes pre-packaged with sentiment analysis functionality and returns a tweet's polarity score, VADER considers punctuation, capitalization, degree modifiers, conjunctions, pre-tri-gram values when assigning sentiment values: negative, positive, neutral. It was used to get an appropriate result and analysis while classifying tweets.

The polarity score was calculated with Sentiment Analyser and values with scores above 0.05 were assigned as positive and all other values as negative, the neutral label was disregarded as it caused instability in the process, in addition, this approach was appropriate to analyse the positive and negative feedback density of POIs by region, to arrive at the assumption that new possibilities may arise for small businesses in regions with negative texts. The column with the sentiment labels is named "sentiment", as a result, 19457 negative labels and 17378 positive labels were obtained.

Table 2: Sentiment Distribution

```
dataframe.sentiment.value_counts()

negative    19457
positive    17378
Name: sentiment, dtype: int64
```

The factorizer function was used to represent the positive and negative values numerically. Factorizer encodes the object as an enumerated type and categorical variable, non-enumerated labels are presented with a new column called sentiment_n. The labelled "sentiment" column is deliberately retained in its original version, which is important for visualizations as will be seen in the discussion section of this project. Figure 15 shows the number of positive and negative labels, with the value 0 representing the positive sentiments, and the value 1 representing the negative values.

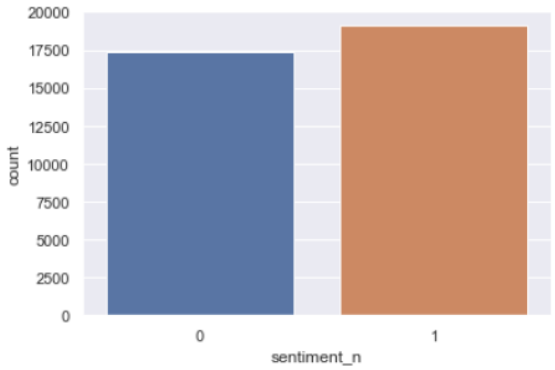

Figure 15 Sentiment Distribution

3.4.2 NLP Data Processing

- Data Preprocessing

As was described in section 3.2, the dataset underwent a number of pre-processing processes, primarily the removal of stop words and emojis, before the model was trained. For easier generalisation, the text is then transformed to be in a lowercase format, punctuation was then cleaned up and eliminated, which minimized the dataset's unwarranted noise. Following that, the repeated characters from the words were also eliminated, along with the URLs since they were of no real use, for better outcomes, the final steps were to execute stemming (which reduces the words to their derived stems) and lemmatization (which returns the derived words to their lemma-like root form). Additionally, the text and sentiment_n columns were selected

for further analysis, positive and negative tweets are separated into data_pos and data_neg features.

- Data Splitting

After Separating input feature (Text_Clean) and label (sentiment_n), the data was split into train and test subsets. The data was divided into 95% training data and 5% test data.

- Transforming the data using the TF-IDF vectorizer and N-grams

Each word in the data was given a different value (weight) using the TF-IDF method. The fundamental principle behind weighting is that words that appear frequently in a document but infrequently elsewhere in the corpus are more significant for that document and are therefore given more weight. Weights were represented by vectors, TF-IDF performed well in improving recall and precision and assisted to get satisfactory accuracy.

Word N-grams features one of the most efficient representation models for sentiment analysis on Twitter and natural language processing. N-Gram constructs vocabulary with multiple words. In this study, it was automatically set for unigram and bigram as N-gram range (1,2), this decision was reached as a result of experimental studies. Bigram and unigram models have demonstrated decent performance for sentiment classification on Twitter data.

- Methodology

In this study, it was aimed to analyse the sentiment of the tweets provided from the merged dataset involving the use of three classifiers. Three different models have been used respectively:

A. Bernoulli Naive Bayes

Bernoulli Naive Bayes classifier is used, when an unwanted word needs to be found or a particular kind of word needs to be labelled in a document. Additionally, it produces binary output in the form of 1-0, True-False, or Yes-No.

B. SVM (Support Vector Machine)

A supervised machine learning model called a support vector machine (SVM) employs classification techniques to solve two-group classification problems. An SVM model can classify new text after being given sets of labelled training data for each category. They offer

two key advantages: greater speed and improved performance with fewer samples. As a result, the approach is excellent for text classification issues, where it's typical to only have access to a dataset with a few thousand tags on each sample.

C. Logistic Regression

Logistic regression is a class-based algorithm that predicts a binary outcome based on a sequence of independent variables. Since the dependent variable's nature is dichotomous, there are only two viable classes. Simply stated, the dependent variable is a binary variable, with data recorded as either 1 (which represents success/yes) or 0 (which represents failure/no).

Figures 16, 17, and 18 show the weighted testing sets recall, accuracy, precision, and f1-score linked with each model. The training dataset is a relatively balanced dataset of data labelled texts, accuracy is used as the performance comparison metric across models.

	precision	recall	f1-score	support
0	0.87	0.79	0.83	869
1	0.83	0.90	0.86	973
accuracy			0.85	1842
macro avg	0.85	0.85	0.85	1842
weighted avg	0.85	0.85	0.85	1842

Figure 16 Bernoulli Naive Bayes Model Performance

	precision	recall	f1-score	support
0	0.93	0.91	0.92	869
1	0.92	0.93	0.93	973
accuracy			0.92	1842
macro avg	0.92	0.92	0.92	1842
weighted avg	0.92	0.92	0.92	1842

Figure 17 SVM (Support Vector Machine) Model Performance

	precision	recall	f1-score	support
0	0.91	0.87	0.89	869
1	0.89	0.92	0.90	973
accuracy			0.90	1842
macro avg	0.90	0.89	0.89	1842
weighted avg	0.90	0.90	0.90	1842

Figure 18 Logistic Regression Model Performance

The SVM (Support Vector Machine) algorithm where each term frequency is given a binary-value had the highest accuracy with an accuracy of 92% and is selected to be the optimum model for this project. Aside from accuracy, the weighted average of precision and recall are 0.92 and 0.92 respectively. The results achieved in this model are also higher when compared to the next two best-performing models by accuracy which are the Bernoulli Naive Bayes and the Logistic Regression models.

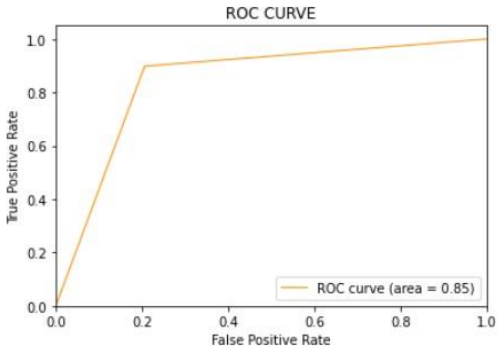

Figure 19 Bernoulli Naive Bayes Model ROC Curve

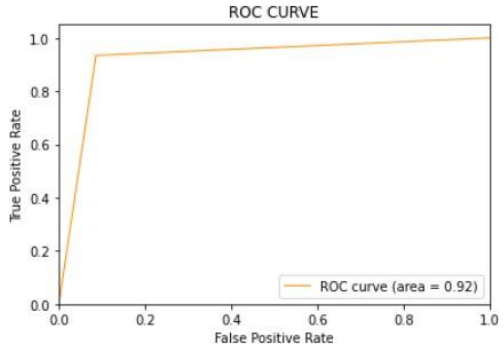

Figure 20 SVM (Support Vector Machine) Model ROC Curve

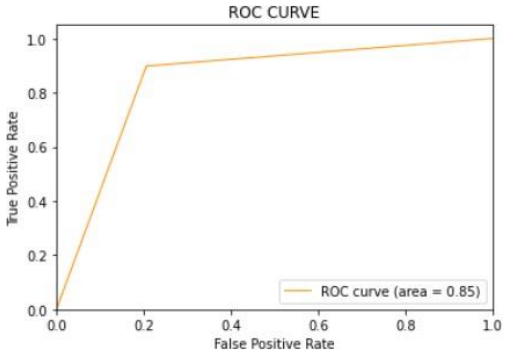

Figure 21 Logistic Regression Model ROC Curve

Figure 19 shows the ROC-AUC Curve for the Bernoulli Naive Bayes model, while figure 20 shows the ROC-AUC Curve for the SVM (Support Vector Machine) model, and figure 21 shows the ROC-AUC Curve for the Logistic Regression model, ROC curves measure the accuracy of the models for the testing set. The ROC curve for Bernoulli Naive Bayes model shows that most elements fall under the curve and with the true positives rate being as high as 0.995, it is clearly shown that the model trains well. The ROC curve for the SVM (Support Vector Machine) model shows the similar metric with the highest rate being around 0.998, and as for the Logistic Regression model, the ROC curve shows approximately the same result when compared to SVM model with slightly lower rate.

Although, the rates are slightly different it can be seen that the SVM model had the highest rate which shows it was the most accurate model when compared to the Logistic Regression model and the Bernoulli Naive Bayes model. The ROC curve clearly shows that the SVM model is well trained and performs well on the testing set.

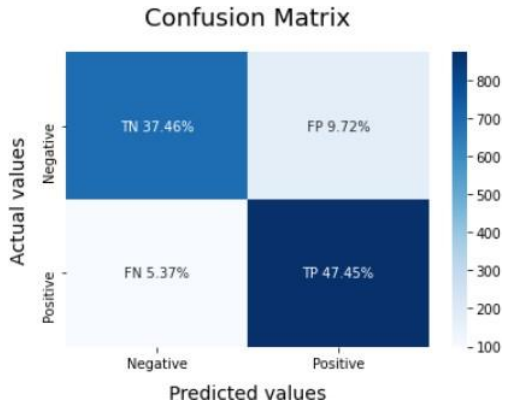

Figure 22 Bernoulli Naive Bayes Model Confusin Matrix

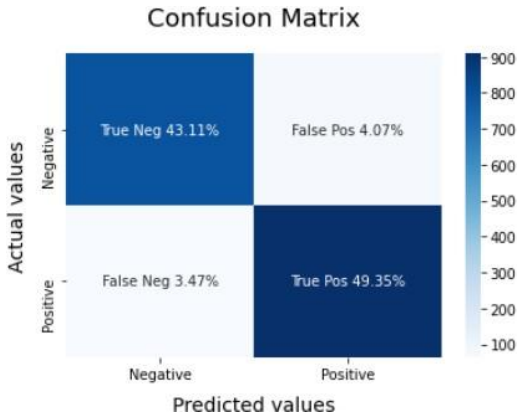

Figure 23 SVM (Support Vector Machine) Model Confusion Matrix

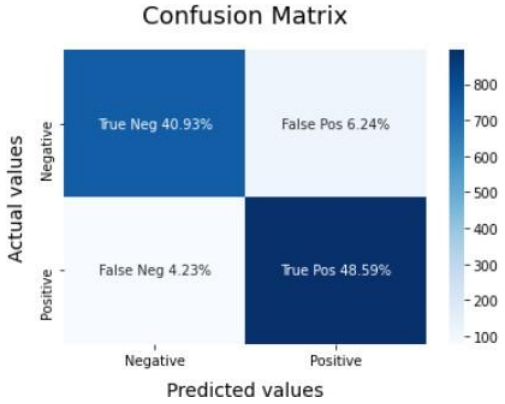

Figure 24 Logistic Regression Model Confusion Matrix

Figures 22, 23, and 24 show the confusion matrices for the correct and predicted labels for each model when run on the test set. In each figure, the upper left box presents the True Negative value, the upper right box presents the False Positive value, the lower left box presents the False Negative value, and the lower right box presents the True Positive value, for each model respectively.

When the above confusion matrix results for the three models are examined, it is seen that the SVM model shown in figure 23, correctly predicted the negative value with a True Negative rate of 43.11% and predicted the negative values as positive with a rate of only 4.07%. While 49.35% of the positive value predictions were predicted correctly, 3.47% were incorrectly predicted. True Negative and True Positive percentages provided much more accurate estimates for SVM model when compared to other models.

3.5 Visualization

The main merged dataset obtained from QGIS for Edinburgh regarding the fastest/shortest paths between banks and restaurant, for London regarding the fastest/shortest paths between

arts centres and marketplaces, and for Manchester regarding the fastest/shortest paths between pubs and nightclubs was analysed via Microsoft's PowerBI tool. The attribute tables, each of which was a layer in QGIS, have been saved as an excel file using the export feature of QGIS. The imported datasets are renamed as London Fastest, London Shortest, Edinburgh Fastest, Edinburgh Shortest, Manchester Fastest, and Manchester Shortest accordingly.

As mentioned before, for all datasets except the main merged dataset in PowerBI , 'Osmid' was selected as the start id field for the input start point layer and name was selected as the end id field of the input end point layer. While visualizing in this way, presenting the names of the destinations instead of the ids made the visualisations and analysis more understandable, relationships have been created between all datasets imported as Excel using PowerBI's relationship management feature.

As seen in figure 25, many to many relationships have been established between the main dataset 'qgismergedata10' and the 'Edinburgh Fastest' dataset, likewise, there are several many to many relationships between the 'qgismergeddata10' and the London fastest, the Manchester Fastest datasets. Other relationships are determined to have many to many relations between the fastest and shortest datasets for each city. There was no need to establish a relationship between qgismergeddata10 and Edinburgh Shortest, London Shortest and Manchester Shortest, as the relationship between the fastest and shortest path datasets automatically provided the necessary relationship for the shortest path datasets with the help of many to many relationships from the fastest paths to the 'qgismergeddata10' dataset.

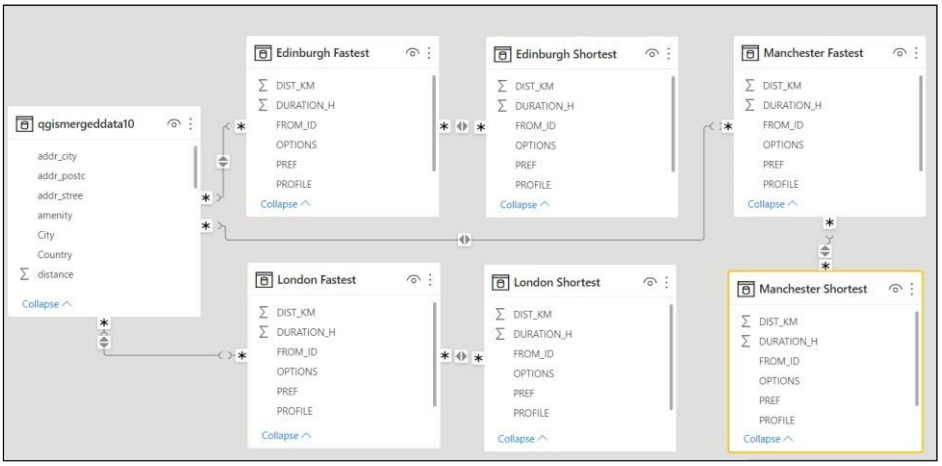

Figure 25 PowerBI Relationship Model

After importing all the datasets to be analyzed, the necessary graphics were selected from the "Visulasations" tab of PowerBI, after the relationships between them were assigned in accordance with the purpose of the analysis to be made. First, the aim is to show the statistics of the main ‘qgisdatamerged10’ dataset, the panel at the top left of figure 26 is a slicer for selecting the three cities Edinburgh, London, and Manchester, which are the focus of the project. The slicer under this panel provides filtering for the amenity, the Clustered Bar Chart shows the number of places in the cities, while the Clustered Column Chart shows the number of tweets by amenities, the Donut Chart in the top right panel shows the number of amenity by city, with London represented in light blue, Manchester in orange, and Edinburgh in dark blue.

PowerBI provides ArcGIS map functionality, which enables to determine the points desired to be displayed on the map from the coordinate information, allowing a deeper geospatial analysis on the data as ArcGIS provides geospatial focused insights integrated into the maps of PowerBI. The map presented in figure 26 is provided by ArcGIS, all the amenties in the three cities are shown on the map. The card panel under the map shows the place names, the number of tweets for this place, the amenity type and the city where the place is located. The card on the right hand side of the map shows the place names and the tweets shared for these places.

As can be seen in the chart named ‘Number of Places by City’, there is a total of 40,000 data points distributed across all eleven amenities, of which 37 thousand are in London, 3 thousand are in Manchester, and 1 thousand points are in Edinburgh. London is the city with the most places compared to the other two cities, the graphic named ‘Number of Tweets by Amenity’ shows how many tweets are shared for each of the eleven amenities, restaurant totalled approximately 16.3 thousand places, while cafés had around 11.3 thousands, pubs measured a total of 7.5 thousand, banks had a total of 2.8 thousand, post office totalled 0.7 thousands, nightclubs had 0.6 thousand places, with social facilities having around 0.6 thousands in total, marketplaces with around 0.4 thousand in total, cinemas 0.3 thousand, arts center 0.3 thousand, and townhalls which had almost 0.1 thousand places in total. The ‘Number of Amenities by City’ graph showed that London and Manchester had 11 amenities, while Edinburgh showed 10 amenities.

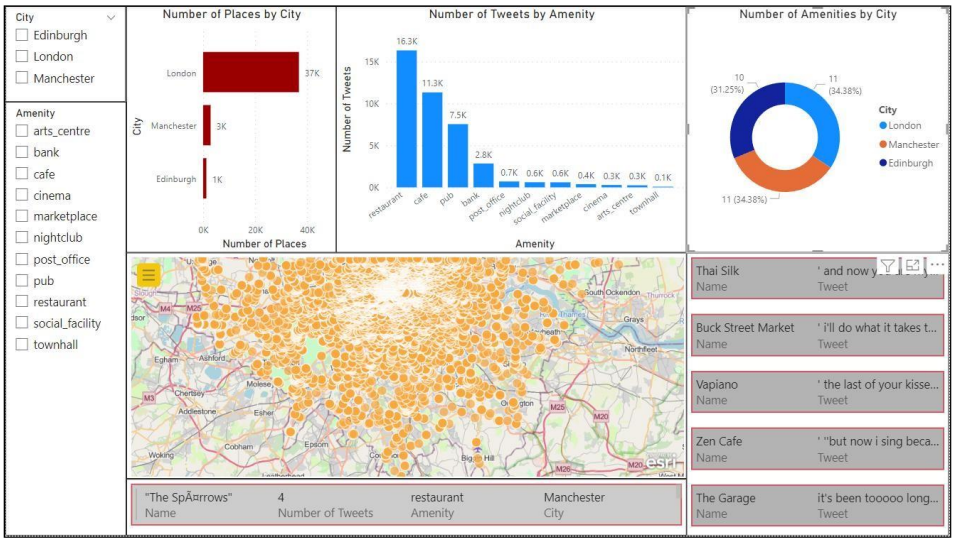

Figure 26 QGIS Merged Data Statistics

Marketplaces in London are displayed in figure 27, there are a total of 363 social facilities in London and a total of 363 tweets were shared from these places. As can be seen in this interactive dashboard, 9 tweets were shared from the "Apple Market", the bottom right corner of the chart shows the tweets from the card for London marketplaces and the marketplace name, the interactivity of scrolling down on both cards shows more marketplaces and relevant information. The total number of tweets and the ability to read the tweets can give visitors an idea about a particular marketplace.

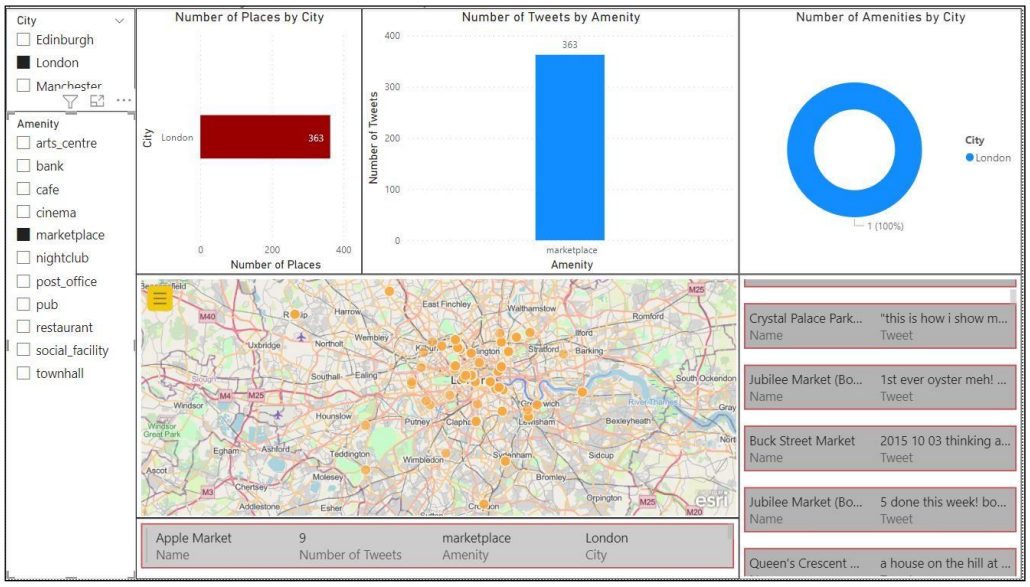

Figure 27 QGIS Merged Data Example

Figure 28 consists of 3 area charts and 3 slicers, the light blue coloured areas represent the shortest path, and the dark blue coloured areas represent the fastest path. Area charts show distance in km on the x-axis and duration in hours on the y-axis, the first area chart presents the distance and duration from the art centres to the marketplaces in London, the area charts in the upper right and lower left show the distance and duration from the pubs to the nightclubs in Manchester, and the distance and duration from the banks to the restaurants in Edinburgh, respectively.

The amenities chosen for the start and destination points have been chosen according to a certain logic, the amenities have been determined to show the direction to destination, which can be the most common visiting point that people can go from the amenity chosen as the starting point. The 3 cards at the bottom right represent the distance and duration of the fastest path from current amenity to the destination, along with the names of the start and end points. For example, the closest marketplace to the "Bow Arts Trust" is the "Nags Head Market" and it can be reached in around 26.4 minutes after travelling almost 9.5 km, by taking the fastest road by car. When the Dashboard is examined from another point, it can be seen that although the fastest path shows a longer direction than the shortest path, the shortest path typically takes more time than the fastest path.

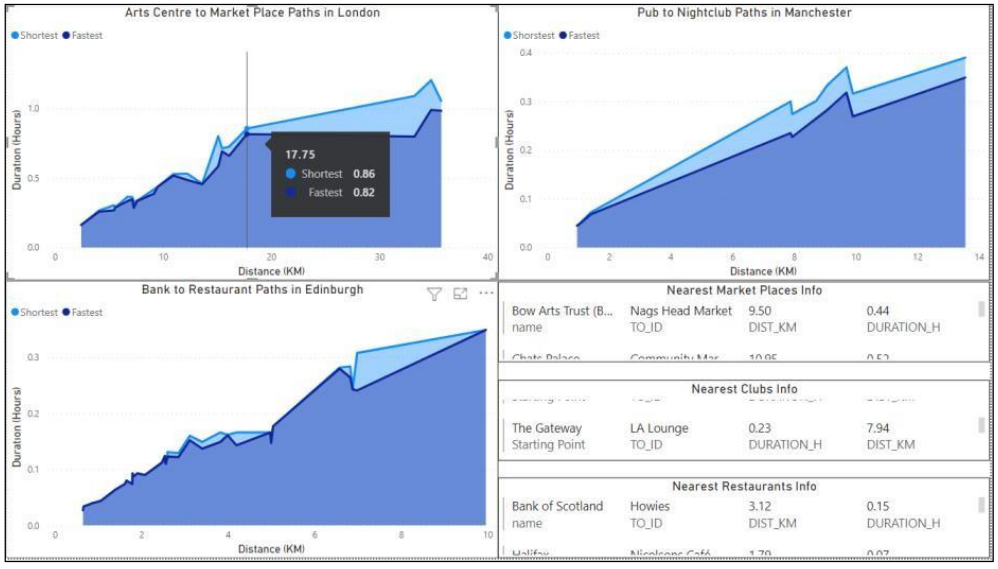

Figure 28 Fastest/Shortest Paths Analysis

In Figure 29, the tweets from each amenity are clustered as positive and negative tweets. The values clustered according to the sentiment reflect the colors of the predominant sentiment, it is aimed to show that the most sought-after areas generally contain positive or negative experiences by visitors according to the predominant values for all amenities, the desired amenity category can be selected by filtering through the panels set on the left side of the page. The dashboard reflects the tweets shared for any amenity selected on the map, and whether the tweets shared for this place are positive or negative.

It is also possible to obtain information about the place by deducing that it is negative based on the visualizations provided. As seen in Figure 29, 3 slicers are used on the left, with the slicer titled "City", the cities to be filtered can be selected in the interactive dashboard, while with the slicer titled "Sentiment", the sentiments to be filtered can be selected in the interactive dashboard. Tweets containing positive or negative emotions, amenities, the total number of positive and negative tweets, and the number of negative and positive tweets according to amenities can be displayed on the dashboard. The amenities filtered with the Amenity slicer can be examined in detail on the dashboard. With the "Number of Places by Sentiment" graphic, tweets containing negative sentiment are assigned a red rank and tweets containing positive sentiment are assigned as blue, the total number of amenities that leave positive and negative impressions on tourists can be determined with this graphic. All the hubs for the amenity category, tweets posted from these hubs, and sentiment belonging to the hub can be displayed in the dashboard as can be seen in figure 29.

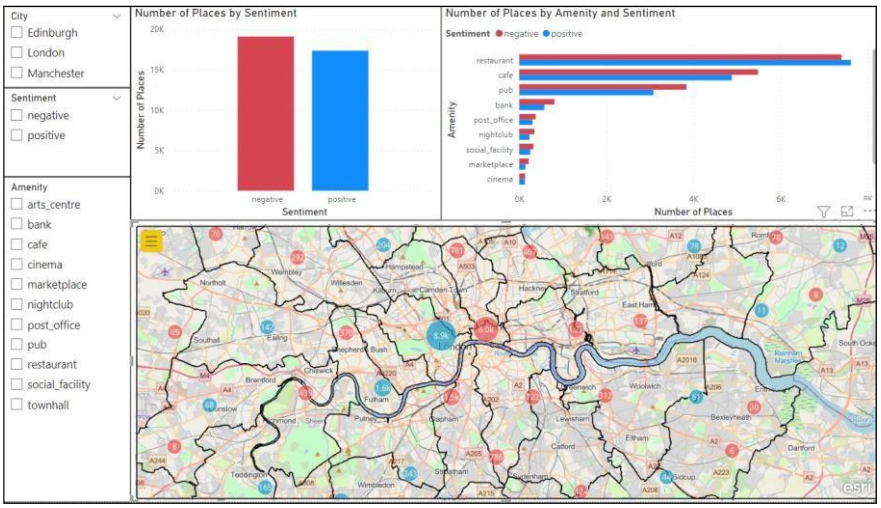

Figure 29 Sentiment Classified Areas Visualization

As seen in Figure 30, pubs that received positive comments by users are displayed on the map. For all pubs in all 3 cities, there are approximately 3K pubs that have been positively evaluated by users, when the map is examined, it is seen that the local pub "Anchor Brewhouse Horselydown" with a view of the "Tower Bridge", with a total of 164 tweets, received mostly positive feedback from the visitors.

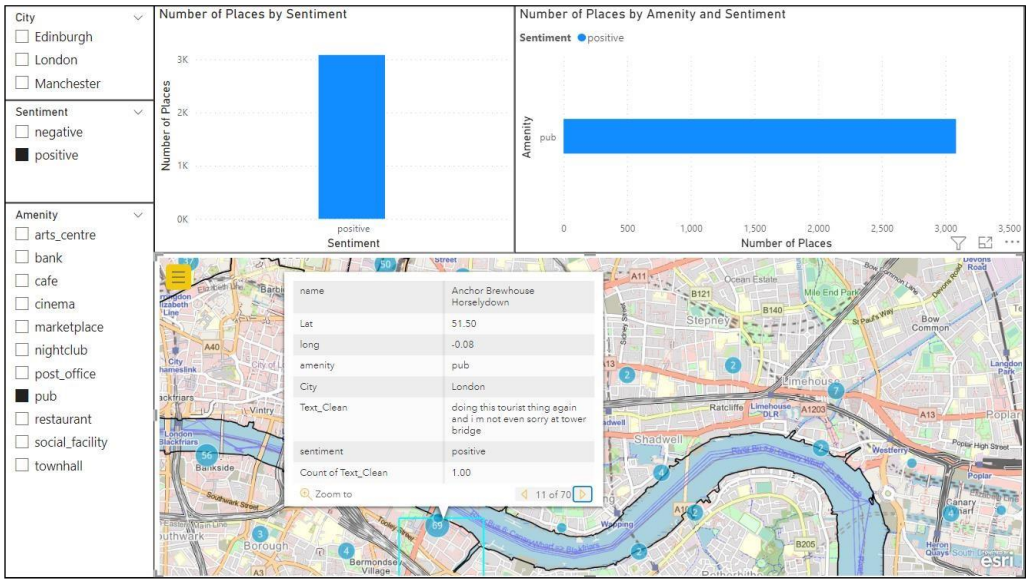

Figure 30 Sentiment Classified Areas Visualization Focused on Positive Impressions for Pubs

4 Data Availability

The Twitter dataset used in this project is publicly available and can be accessed through the following link:

https://github.com/rmsreis/tweet-classifier/tree/master/notebooks/tweet_data

5 Discussion

5.1 NLP

NLP was used to predict the sentiment behind the tweets acquired from the twitter dataset. The lexicon-based sentiment classifier VADER was used in this study, VADER is an NLTK-based NLP library that comes pre-packaged with sentiment analysis functionality and returns a tweet's polarity score. The textual data was prepared by a series of methods consisting of tokenization, stemming, lemmatization, and TF-IDF vectorizing. These methods help reduce the words to their derived stems and return the derived words to their lemma-like root form, the TF-IDF vectorizing was selected based on an experimental systematic approach that

achieved the optimum results for this study. A SVM model was trained and tested for the purpose of this project and an adequate result was achieved, the NLP model achieved an accuracy of 92%.

The output of the NLP part of this project was utilised in the visualisations for this project in order to analyse the positive areas and the negative areas in London, Manchester, and Edinburgh.

5.2 Visualization

In this project, it was aimed to identify local touristic places in London, Manchester and Edinburgh. PowerBI and QGIS tools were used in this study to visualize and analyse the identification of local attractions. Studies have been carried out on QGIS in order to obtain appropriate results with Twitter text data and a dataset containing POI categories obtained from OpenStreetMap with the OSMnx python library. The project has several features that can be of practical use in order to satisfy the objectives set for this project.

Since the main objective of this project is to identify local touristic places in 3 different cities in the UK. QGIS was utilized to implement specific features that would benefit businesses, urban planners, location planners, and even might be used to enhance the transportation system, these features added more depth into the analysis and allowed the project to analyse local touristic places from a wider scope which delivered valuable insights and information that would expand the general touristic experience in these cities by improving the services provided for these places in each city.

The first feature that was added into the project identifies the nearest hubs for a series of specific categories, this feature boosts touristic activities in a particular area by recommending the nearest hub from a specific location in order to raise the visibility and popularity of the hubs that visitors may not have yet visited, this can also benefit the hubs in the surrounding area of that specific location, improving the area’s popularity in general.

The second feature that was utilized using QGIS provides an analytical view on the shortest and fastest paths between a pair of amenities in each city. The feature shows how a trip between 2 places varies in terms of distance in km and duration in hours, the analysis

provides a comparative view on how the trip can be affected by several factors in terms of the travel distance and the journey time. Touristic behaviour could be affected if the services for travelling are not well maintained, and in this analysis, it has been shown that the journeys between the same starting point and destination have lower durations over an increased distance, this may be caused by road congestions and maintenance and other factors as well.

As mentioned before, the shortest route should be the one that can be reached in the shortest time, but this was not the case for all the paths that were examined using this feature, therefore, another route can be suggested to the visitors as an alternative route because traveling long distances even though it takes a short time can have a negative impact on an attractive hub, which should be noted for urban planners and civil engineers.

The third feature in QGIS is the isochrone feature, this feature is particularly useful for urban planners in order to measure the duration needed to reach a certain point in a certain duration. Isochrones help determining the travel time needed when travelling through different locations and help discover which points are reachable within a specific duration.

The dashboard created further explores the analysis done on QGIS, the dashboard is an interactive tool where users can explore the insights of this study and use it accordingly to their preference. The dashboard comprises of 3 pages, the first page aims to analyse the distribution of the number of tweets and the number of places for the amenities in each city, this can be helpful for both individuals and businesses as it provides an overview for each amenity, city, and tweets, the page also includes a map that shows the amenities along with the corresponding tweets for each amenity while highlighting the relevant information regarding the amenities below the map and showing random tweets on the bottom right corner of the page. The page can be used to exract several useful insights for the users.

The second page focuses on analysing the shortest and the fastest paths measured in QGIS. The page offers an in-depth analysis on the measures from which several aspects can be studied in order to improve the services provided by the public and private touristic places which enhances the touristic experience as a whole, hence, fulfilling the objective of this project.

An interesting insight exracted from the second page shows that the paths between pubs and nightclubs in Manchester follow a certain pattern, the linear pattern shows that the duration varies consistently when the distance between the 2 hubs is larger than 1.4 kilometres and that the shortest path always takes longer than the fastest path which might indicate road related issues which might cause tourists to repel from attending the venues discussed.

The third page studies the output of the NLP part of this project as it analysis the regions and amenities according to the sentiment analysis applied on the tweets in the dataset. This page highlights the areas that have a positive impression from the areas that have a negative impression, the amenities in each area were clustured according to the impression predicted by the sentiment analysis, each cluster is distinctly highlighted according to the predominant value of the sentiment. This analysis is extremely useful for identifying the popularity of the regions in each city which helps tourists in forming a general overview about the area of interest, it is also beneficial for businesses, entrepreneurs and decision makers to develop strategies to attract more tourists and enhance the touristic experience in general, for instance, in the areas that have a negative predominant sentiment, small businesses can attract customers and improve the services provided to be a differantial place in the region.

6 Conclusion and Future Work

The sentiment analysis applied in this project delivered accurate and satisfactory results, the process correctly predicted the sentiment of the tweets and classified them into positive and negative impressions. The SVM model achieved an accuracy of 92% where the ROC Curve showed the elements fall under the curve with a rate of approximately 0.998.

The analysis produced in this project was implemented using two main geospatial tools which are QGIS and Microsoft PowerBI. The analysis met the motivation of the project, the closest distance analysis between tourist attraction places and the recommendations could increase the visibility and popularity of the hubs presented. With the detection of the shortest and fastest paths between hubs, it has been estimated that the hubs, which have created a negative impression on the visitors have the impression connected to the journey rather than the hub itself. It might be a subject that urban planners and civil engineers could focus on.

For the cities of London, Manchester and Edinburgh, some good points are mentioned with the isochrones adjusted according to the size of the city, the determination of the travel time to the central point can benefit taking care of transportation problems in order to facilitate the visits for tourist attraction places outside the city.

The project can still be enhanced to be more beneficial in terms of use, such improvements can include creating a wider nearest hubs combination where the user would be able to filter the nearest hub according to the preferred amenity. It can also track the user's activity in order to show the nearest hub recommendations based on the user's activity log.

The fastest and shortest path feature can be further improved as it is restricted to show paths travelled by car, this can be enhanced by adding more transportation methods such as walking and cycling, it can also be applied to show all the paths between all the amenities in each city which can be implemented as a navigation aspect for the project.

The project can be developed to be a mobile application which grants users with an easily accessible tool where it can show the recommended nearest hubs, local places, fastest and shortest directions, and positive and negative areas around the user's location and according to the user's profile.

7 References & Bibliography

Andrejev, S., 2022. *OSMnx: The Fastest Way to Get Data from OpenStreetMap.* [Online] Available at: https://python.plainenglish.io/osmnx-the-fastest-way-to-get-data-from-openstreetmaps-731419d4dc31

Brunila, M., 2017. *Scraping, extracting and mapping geodata from Twitter.* [Online] Available at: mikaelbrunila.fi/2017/03/27/scraping-extracting-mapping-geodata-twitter/ [Accessed 14 04 2022].

Chulong-Li, 2019. *Real-time Twitter Sentiment Analysis for Brand Improvement and Topic Tracking.* [Online] Available at: https://github.com/Chulong-Li/Real-time-Sentiment-Tracking-on-Twitter-for-

Brand-Improvement-and-Trend-Recognition
[Accessed 06 2022].

Cuesta, A., 2014. A framework for massive twitter data extraction and analysis. *Malaysian Journal of Computer Science*, 01, Volume 27(1), pp. 50-67.

GabrielRodriguez66, 2019. *Data-Mining-Final-Project.* [Online] Available at: https://github.com/GabrielRodriguez66/Data-Mining-Final-Project/blob/master/Final_Project.ipynb
[Accessed 08 2022].

github, 2021. *Mapping & Geocoding.* [Online] Available at: https://melaniewalsh.github.io/Intro-Cultural-Analytics/07-Mapping/01-Mapping.html

Gottipati, S. et al., 2021. Analyzing Tweets on New Norm: Work from Home during COVID-19 Outbreak. *IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, Issue IEEE, pp. (pp. 0500-0507).

Häberle, M., Werner, M. & XX, Z., 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*, Volume 52(sup2), pp. 2-11.

Jianqiang, Z., Xiaolin, G. & Xuejun, Z., 2017. *Deep Convolution Neural Networks for Twitter Sentiment Analysis.* s.l., IEEE access, 6, 23253-23260.

Khalilnezhad, S., 2022. Using Twitter as a Means of Understanding the Impact of Distance and Park Size on Park Visiting Behavior (Case Study: London). *Journal of Digital Landscape Architecture*, pp. 146-154.

K, J. D., 2021. *Plot Latitude and Longitude from Pandas DataFrame in Python.* [Online] Available at: https://datascientyst.com/plot-latitude-longitude-pandas-dataframe-python/

Lauer, C., 2021. *Working with OpenStreetMap Data.* [Online] Available at: https://towardsdatascience.com/working-with-openstreetmap-data-37da18d55822
[Accessed 12 04 2022].

Mittal, M. et al., 2019. *Accurate Spatial Mapping of Social Media Data with Physical Locations.* s.l., IEEE.

Motti, Z., 2021. Geolocating tweets via spatial inspection of information inferred from tweet meta-fields. *International Journal of Applied Earth Observation and Geoinformation,* 12, Volume 105, p. 102593.

Müller, M., Salathé, M. & Kummervold, P. E., 2020. COVID-TWITTER-BERT: A NATURAL LANGUAGE PROCESSING MODEL TO ANALYSE COVID-19 CONTENT ON TWITTER. *arXiv preprint arXiv,* 07, 2005(07503), p. 503.

myselfHimanshu, 2019. *Sentiment Analysis : Basic Data Collection.* [Online] Available at: https://gist.github.com/myselfHimanshu/cc6d3a95a644fa2f7f0b65e163df6529
[Accessed 07 2022].

Nathaniel, J., 2021. *Working with OpenStreetMap in Python.* [Online] Available at: https://levelup.gitconnected.com/working-with-openstreetmap-in-python-c49396d98ad4
[Accessed 04 2022].

Open Street Map, 2022. *Retrieving OpenStreetMap data.* [Online] Available at: https://autogis-site.readthedocs.io/en/stable/notebooks/L6/retrieve_osm_data.html

OpenStreetMap, 2022. *OpenStreetMap Wiki.* [Online] Available at: https://wiki.openstreetmap.org/wiki/Main_Page
[Accessed 11 04 2022].

Pascual, F., 2022. *Getting Started with Sentiment Analysis using Python.* [Online] Available at: https://huggingface.co/blog/sentiment-analysis-python [Accessed 08 2022].

Peixoto, F., 2020. *Countries [Latitude & Longitude].* [Online] Available at: https://www.kaggle.com/datasets/franckepeixoto/countries [Accessed 12 04 2022].

Piech, J., 2020. *Tweet-Geolocation-Classifier.* [Online] Available at: https://github.com/rmsreis/tweet-classifier [Accessed 07 2020].

QGIS, 2022. *QGIS Documentation.* [Online] Available at: https://www.qgis.org/en/docs/index.html [Accessed 12 04 2022].

QindanUCL, 2018. *Extract Geodata from Twitter.* [Online] Available at: https://github.com/QindanUCL/Extract-Geodata-Twitter [Accessed 20 04 2022].

Rude, B., 2021. *Extracting geographic location information from twitter.* [Online] Available at: https://brittarude.github.io/blog/2021/08/01/Location-and-geo-information-in-twitter
[Accessed 12 04 2022].

Santhanavanich, J. T., 2020. *How to Download Dataset from OpenStreetMap?.* [Online] Available at: https://towardsdatascience.com/beginner-guide-to-download-the-openstreetmap-gis-data-24bbbba22a38
[Accessed 05 2022].

sharmaroshan, 2019. *Twitter-Sentiment-Analysis.* [Online] Available at: https://github.com/sharmaroshan/Twitter-Sentiment-Analysis [Accessed 08 2022].

Twitter, 2022. [Online] Available at: https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data

Twitter, 2022. *Filtering Tweets by location.* [Online] Available at: https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location [Accessed 10 04 2022].