Media Engineering and Technology Faculty
German University in Cairo

Spatio-Temporal Urban Sentiment Analysis

Bachelor Thesis

Author:	Rana Emad Abdelmoniem
Supervisors:	Assoc. Prof. Mervat Abuelkheir
	Prof. Ahmed Hassan Mohamed Yousef

Submission Date: 01 August, 2021

Media Engineering and Technology Faculty
German University in Cairo

Spatio-Temporal Urban Sentiment Analysis

Bachelor Thesis

Author: Rana Emad Abdelmoniem
Supervisors: Assoc. Prof. Mervat Abuelkheir
Prof. Ahmed Hassan Mohamed Yousef

Submission Date: 01 August, 2021

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowlegement has been made in the text to all other material used

Rana Emad Abdelmoniem
01 August, 2021

Acknowledgments

I would like to express my gratitude to my supervisor, Mervat Abuelkheir, for guiding me and supporting my ideas and efforts through out the project. I would like to thank my colleagues for offering deep insights into my research. I wish to show my appreciation for my family and friends for their advice, support and patience for me during the entire duration of the project.

Abstract

The perceptions of people towards neighborhoods reveal insights into their satisfaction with their living environment, its aspects and their quality of life. Recently, a number of websites and social media platforms are designed and used by people to find and share their opinions about neighborhoods to find suitable places to live. Such online neighborhood review data provide novel opportunities for studying the perceptions and sentiment of people toward their neighborhoods.

Urban sentiment analysis provides perception into how citizens feel about their environment and therefore, how it can be improved or used to enhance the living situations of the residents. In this context, the main objective of the proposed work is to tackle and identify the citizens' sentiment, track and analyze the results, locate obstacles and limitations and examine different tools and algorithms needed for performing real time urban sentiment analysis on the residents opinions of their cities and neighborhoods. In this paper we analyze the reviews shared on social media platforms about their neighborhoods and cities. We use Targeted sentiment analysis to identify the sentiment conveyed about a certain neighborhood - the target.

The analysis is based on a dataset of online reviews about some cities and neighborhoods in England; relying on two well-known classification algorithms, logistic regression and stochastic gradient descent. We then apply the same analysis techniques on another dataset of reviews of urban cities and neighborhoods that is collected from Twitter, relying on Vader and TextBlob, text processing and sentiment classification libraries. We apply NER on the data to detect the target location. The SGD model performed better than the LR model reaching accuracy of 85 percent and Vader performed better than TextBlob on the twitter corpus reaching accuracy of 75 percent.

For the scope of this research, only 7 locations were analyzed in details from both datasets, but the general approach can be followed on any other location to analyze the residents' sentiment towards the location. For these locations we extracted the aspects and the sentiment towards that aspect to draw a general conclusion of how the residents feel living in these cities.

Contents

Chapter 1

Introduction

Sentiment analysis or opinion mining is a method of detecting patterns of emotions and opinions on specific topics using data mining applied to unstructured data, such as tweets. This research field has an increasingly interest to both the scientific and the business communities. Conducting sentiment analysis on urban areas can provide perception on the residents and how they feel about the quality of life where they live by analyzing their sentiments towards different aspects of the urban life. These insights can be used to improve or enhance the living conditions of the citizens therefore, the main objective of the proposed work is to track and analyze the sentiments of residents of a neighborhood over a specified period to identify the issues affecting their lives and how they are dealt with over time.

Social media platforms has led to a widespread of unstructured data where there is a rapid growth in the number of users contributing. This fact has brought new opportunities and challenges to how the information is retrieved and analyzed. The opinions shared by the users about different issues are made in spontaneous manner in real time. This provided us with accurate data about the sentiment of the citizens given a specific period as well as, introducing difficulties as to how to process and handle the data to better extract accurate sentiment from text. More novel tools are introduced to handle some of the challenges as scrapping tools and tools for analyzing continuous data streams.

Many scientific proposals took place the last few years aiming to study and tackle various sentiment analysis challenges as well as analysis of spatio-temporal data. However, not many that address spatial and temporal aspects of sentiment related to urban areas enabling a better understanding of the mood of people when using social media, which would aid in investigating the behaviour of the people in their cities leading to finding explanation and ways to deal with unhealthy behavioural patterns of the people.

Concerning the spatial dimension, the considered geographical location of a tweet can be the location where the social media user delivered the message, or the user's home location, or even a location eventually mentioned in the message. Concerning the temporal dimension, we are interested in possible opinion changes over time. Related work has considered the spatial dimension based only on geocoded messages. The main problem with such kind of approach is due to the fact that there are still few information sources which provide accurate geocoded messages. Targeted sentiment analysis investigates the classification of opinion polarities towards certain target entity mentions in the tweet.

The remainder of this paper is organized as follows. Chapter 2 provides a literature review on related work. Chapter 3 presents the core ideas of the models tested in this work for extracting the target place and analyzing the sentiments in relation to that target over time. Chapter 4 presents the details of the datasets, compares the performances of multiple experiment parameters and the analysis of the results. Chapter 5 concludes all our work and results of the study and propose potential future directions that can be pursued to improve the study.

Chapter 2

Background

In this chapter, we briefly discuss different approaches related to sentiment analysis papers of spatio-temporal nature.

In Reference [9] the authors conducted a sentiment analysis approach on Twitter Microtexts, analyzing the Portuguese tweets related to FIFA's Confederations Cup during the period between 12 April and 12 August 2013 - approximately 2 months before the beginning and 2 months after the end of the competition, where temporal data were collected on daily basis. Naive Bayes and SVM classifiers approaches were used aiming to replace Part-Of-Speech Taggers in the identification of opinionated tweets. Both approaches works with two classifiers - one to detect whether a tweet presents opinionated content and the other to classify the subjective polarity of the message as either positive or negative. To perform spatial analysis, GIR technique was applied. 2 approaches were conducted; the first analyzes sentiment expressed in emoticons; the second used manual labelling, where 1,500 collected tweets were randomly chosen and separated for manual labeling of the sentiment polarity. Sentiment polarity was validated using 10-folds cross validation with all labelled tweets. The accuracy for the SVM classifier was 0.800 whilst the Naive Bayes classifier was 0.777, the F-Measure was also better using the SVM classifier. The sentiment polarity throughout the time period was predominantly positive, while the spatial distribution of the sentiment showed that the sentiment polarity varies through the regions. One of the limitations was that the targeted tweets were related to the 2013 FIFA Confederations Cup, but some of the opinions expressed in the messages may refer to other entities. This problem could be tackled using Named Entity Recognition techniques.

The study Spatial and Temporal Sentiment Analysis of Twitter data, reference [15], focused on spatio-temporal variation of georeferenced Tweets' sentiment polarity, to understand whether sentiment polarity on Twitter exhibits specific time - location patterns. Tweets posted in campus were divided into 6 spatial zones and 4 time zones. Tweets were assigned polarity using Starlight Data Engineer then were further processed in Feature Manipulation Engine to be converted into spatial and temporal feature classes, where GIS (Geographic Information System) was used to map the spatial and temporal patterns of the tweets. Results show that the highest percentage of positive Tweets occurred in the social science area (42.1 percent), while science and engineering and dormitory areas had the highest percentage of negative postings. The spatial data then indicates that students or staff in the science and engineering areas could feel slightly more negative compared with students or staff in the social science areas. The temporal data shows that the largest percentage of negative feeling occur at the beginning of semester (21.6 percent). This contradicts the hypothesis which was that the highest negative percentage would be during the finals. The smallest number of negative opinions took place after examination. Certain temporal patterns of Twitter polarity only occurred at specific locations which strengthen the hypothesis that Twitter sentiments are time-location specific, where they might depend on the activities conducted at certain locations. The output of Starlight Data engineer was not perfectly accurate, where emoticon could not be analyzed and sarcasm and irony tweets polarity were not accurate. Some tweets, implied emotions but sentiment words were not explicitly written, were misclassified.

Similar study was conducted, on twitter data in 2016, in the study A Systematic Spatial and Temporal Sentiment Analysis on Geo-Tweets [10], which explored the sentiment of the local users using Geo- tweets extracted from January to December 2016, where they were analyzed on monthly, daily and hourly basis where hourly summary of Geo-tweets sentiments calculated every four hours. Local Indicators of Spatial Association (LISA) was used to determine the spatial clusters, Latent Dirichlet Allocation (LDA) was used to classify positive sentiments into different topics and Valence Aware Dictionary and sEntiment Reasoner (VADER), a rule-based tool specially tuned for sentiment analysis of weblogs and social media text, was used for sentiment analysis. Results showed that some patterns were found relating the nature of Twitter content and the characteristics of places and users. Weekend events and friend and family gatherings are the time that users prefer to post positive tweets. In the western part of the US, users love to post photos on Twitter more than in other parts of the US. Users post more between January and April, Geo- tweets peaks on Saturday, more negative tweets are posted on the weekend. Positive sentiment tweets count goes down from Saturday to Sunday. The counties near San Diego in California and Miami in Florida have significant positive sentiment number. These places are famous for its miles of white-sand beaches and amazing weather while, counties that are located at middle west part of country, such as Montana, North Dakota have showed that negative sentiments is larger than positive sentiments. This all reveals that spatial and temporal patterns do exist. Some of the limitations, which weakened the study, were the cleaned dataset still includes noisy data and research mainly focuses on the positive sentiment analysis.

The 2016 US Elections were examined to discover sentiment on Twitter [13], towards either the democratic or the republican party at US counties over any arbitrary temporal intervals, using a large collection of geotagged tweets from a period of 6 months leading up to the US Presidential Election in 2016. After Geo-Tweets are collected they are passed through classifier that filters political tweets only, then finds the political alignment. LDA is used to find politics related words then Word2Vec is used. Both classification phases - political and non political then affiliation - are done using linear classifiers - SVM and

Logistic Regression. Different models have been tested, such as Multinomial Naive Bayes, LSTM and FastText. FastText was the best performing model with an accuracy of 84.4 percent with a possibility of reaching 95 percent or above if emoticons were included; where the next best model was the LSTM based model with accuracy of 82 percent. Less tweets were received from non-urban areas which leads to lower performance. Although the sentiment map was closely accurate at describing the popularity of the candidate, it did not guarantee an actual win to the voting process. Moreover, Twitter is mainly popular among the younger generations which in my personal opinion would reflect a more accurate sentiment analysis for future events where the main participants would be those younger generations, while not as accurate for ones that envolves the more senior generations, unless twitter became popular among most/ all generations.

The most recent study [12], examines how sentiment analysis could be used for risk assesments and disaster detection for a particular location at different time intervals. The proposed algorithm - Risk Assessment Sentiment Analysis was done using LSTM and then validated using SVM, Naive Bayes, max. Entropy, LR, random forest, XGBoost, stochastic gradient descent and CNN in 2-folds: the first was Binary, the second was multi class (for 3 classes). Results showed the the proposed algorithm RASA performed better at both scenarios, where the increase in the binary class was 1 percent higher than XGBoost - the next best performance, and 30 percent increase in the multi class on average , compared to all over techniques which revealed that the model is better suited for multi class scenarios. Some of the limitations of this study were: It was not across all domains, where the accuracy is more probable to decrease if other domains were to be considered, emoticons are not included in the analysis, sarcastic tweets are not handled well and the study focuses on English tweets only.

The most relatable studies conducted were [Marzieh Saeidi et al. 2016] [14] performed on Sentihood dataset and [Yingjie Hu et al. 2019] [11] on online neighborhood reviews for understanding the perceptions of people toward their living environments. The former explores targeted aspect-based sentiment analysis and introduces the SentiHood dataset, which is extracted from question answering platform of Yahoo!, where urban neighborhoods in England are discussed by the users; developing strong baseline models using Logistic regression and LSTM for future benchmarking. Aspect based sentiment analysis can recognise the positive and negative opinions expressed towards a service aspect but not the target entity of the opinion. Targeted aspect based sentiment analsysis on the other hand extracts the target and its respective aspects with relevant sentiments. Results showed that the best performing model is logistic regression when location masking and POS tagging is applied with accuracy of 91.6 percent. LSTM was shown to be underperforming with accuracy of 87.2 percent, probably due to the lack of training data. The out-performance of logistic regression over LSTM calls attention to the advantage of features engineering when the availability of data is limited. The study does not handle the temporal dimension only the spatial dimension.

The latter study analyzes an online neighborhood review data, extracting the semantics - the topics - and the sentiments - the emotions. Latent dirichlet allocation (LDA) and and multi-grain LDA (MG-LDA) are explored in the paper. LDA is discovered to be the better performing model, discovering more topics accuratly that are not discovered by MG-LDA. MG-LDA was under-performing as well, discovering topics that have mixed themes. This showed that LDA is more effective in identifying semantic topics from neighborhood review data with an explanation that people tend to care more about the same core aspects of neighborhoods, as location, safety and convenience. Naive Bayes and lexicon-based approaches were explored as well, where naive bayes approach performed better. It is stated that the naive bayes, lexicon-based and LDA approaches perform relatively well with higher agreements for some aspects but not so well for others. This performance difference suggests that there exist varied difficulties in deriving correct ratings for different aspects. The study as well does not tackle but suggests exploring temporal changes of the perceptions of the residents towards the neighborhoods to use the insights to evaluate effectiveness of urban planning policies by understanding if the policy improves the satisfactions of the residents when implemented.

Chapter 3

Methodology

3.1 Datasets and Data Collection

The main data source, on which this study was conducted, was an existing location-tagged dataset - SentiHood Dataset [2]. SentiHood dataset is extracted from question answering platform of Yahoo!. It has been used in several studies before as it contains qualitative reviews of urban neighbor-hoods in England. The dataset consists of the text, which contains the review of a user, the ID and the opinion, which consists of the sentiment, the aspect and the target location for which the aspect and sentiment are conveyed. The drawback of using SentiHood, as the main source of data, was that it was small in size. Sentihood dataset consists of 5215 total entries and only 3530 imbalanced enteries after cleaning.

The other data source chosen, for collecting reviews and opinionated text about urban areas was Twitter, as it has proven to have a great deal of varying data that is very expressive. The drawback is the amount of noise and unrelated data that needs to be filtered.

The go-to tool for Twitter data collection is Tweepy - Twitter API. The API introduced a challenge as it limits the user to 5000 tweets at largest per month for standard users. The alternate tool chosen is TWINT - Twitter Intelligence Tool [1]. TWINT is a Twitter web scrapping tool that offers unlimited tweets retrieval with the same effectiveness and features as Tweepy. Both, Arabic tweets and English tweets can be collected, defining the language constraint extracting only the needed features.

Collecting accurate, related data can be achieved by defining the location and/or by using keyword search.

3.1.1 Geotags Approach

To collect tweets that addresses a specific place, the longitude, latitude and radius of search is defined in TWINT. Defining the radius produces a drawback, since areas shapes are not circular therefore, some tweets are ommited for it not being within the radius of search area, while other tweets, that are not related to the location specified, is retrieved. Additionally, the circles of some areas might overlap producing redundant data.

3.1.2 Searching using keywords

This approach is precise and straightforward. We use keywords relevant to the scope of the project as the name of a city e.g. "Toronto" or an aspect e.g. "prices", "traffic". Some of the limitations that needed to be tackled for this approach is that even though the required keyword is present in the tweet, it might not be relevant such as when the keyword holds several meanings.

3.2 Data Pre-processing

Twitter data contains a large amount of noise. Since the text is essentially informal, many challenges must be taken into account in order to perform the sentiment analysis on tweets including grammatical errors, slang, and repeated characters hence, the collected tweet corpus goes through various stages of text pre-processing to be able to process and analyze only the insightful content. Regardless of the collection method, all tweets are binded into a file to go through the pre-processing stage, which on some occasions might vary slightly according to the language or the dialect but, there exist some pre-processing steps that text goes through nevertheless, as Normalization, Lemmatization or Stemming and Removing Stop-words.

Before the pre-processing, we clean the dataset from any unnecessary or redundant data. We first specify the languange, English in our case, then we remove any duplicated tweets and remove columns with no valuable information, such as username or replies count and finaly, remove any rows where exists no valuable information, such as tweets that only contains a photo.

The first Pre-processing stage is Normalization. This stage consists of:

- Removing URLs and HTTPs
- Removing handles, retweets or users mentioned
- Transform abbreviations into their original form e.g. aren't changed to are not
- Removing Punctuation, numerics, emoticons and special characters
- Splitting joined words and transforming words into their standard form e.g. VerryyyHapppy change to very happy

The next stage in pre-processing is lemmatization or stemming. The goal of both stemming and lemmatization is to reduce inflectional forms of a word to a common base form. However, Stemming refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time. While lemmatization uses vocabulary and morphological analysis of words to return the base or dictionary form of the word (the Lemma). In the scope of our project, we use lemmatization using Gensim Library. Gensim was used for lemmatization in this project, as it returns the word to its root form as well as, lower cases the word and remove Stopwords - a set of commonly used words in a language that carry little or no useful information. Lastly, Part-Of-Speech tagging (POS) is applied to the clean tweet to improve the accuracy of the model.

3.3 Target Extraction

The target entity is the location entity mentioned in the tweet for which the opinion - aspect and sentiment - is expressed. We were able to identify the location entities mentioned in a tweet using Named Entity Recognition (NER) from SpaCy Library [7], whether the location represents the target entity or other locations that may infer to the aspect e.g. Westminister Abbey could refer to aspect of tourism, where the named entities are returned with their respective tags - Person, Geographical Entity, Organization, etc. We then filter the Geopolitical Entities (GPE) found in the tweet to extract out target location then, proceed to extract the opinion expressed for it.

3.4 Sentiment Classification

Now that we have our clean tweets and relevant data, we can proceed to classifying the tweets sentiments. In our annotation, we chose to go with binary class sentiment classification - positive as (1) and negative as (0) sentiment labels. This is because most of out data comes across aspects with polarity and neutral classes would not provide much insight to the study.

In previous studies, the classifiers that showed the best performances in previous research papers are Logistic Regression, Stochastic Gradient Descent (SGD), VADER for unlabeled data and LSTM. LSTM generally performs better as it is able to extract some context for the data provided and therefore produces more accurate sentiment. However, the models to be explored for this research are Logistic Regression, SGD and VADER. It is to be noted that the main model to be explored is the Logistic Regression model for the SentiHood dataset and the reason for that is the Logistic regression model out-performed the LSTM model in previous experiments which is probably due to lack of training data which is similar to the current situation of this research. VADER is also to be explored for the twitter data as it presented the best performance for unlabeled data in previous studies.

For the machine learning models used for the annotated dataset SentiHood - logistic regression and SGD models - the independent data shall be passed as numeric vectors as the model does not understand text hence, we use feature extraction techniques to prepare our SentiHood data for sentiment classification. We experiment with Gensim's Word2Vec [3], it uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence, and We also use SciKit-Learn's CountVectorizer [4], to fit all the words in our tweets in a bag of words so, we can compare results for optimal accuracy and performance.

We try passing the vectorized corpus as is to the machine learning models once, and we pass the corpus after fixing the imbalance another, getting 2598 positive tweets and 1862 negative tweets, to compare performances. We combat the class imbalance in our SentiHood datset by trying to re-sample our dataset. Re-sampling can be done by adding copies of instances from the under-represented class called over-sampling (or more formally sampling with replacement), or by deleting instances from the over-represented class, under-sampling. Considering the size of out dataset, we chose over-sampling.

For training the machine learning models on SentiHood, we SciKit-Learn as well [5] [6]. We normalize the sentiment feature for each tweet between 0 and 1 rather than the original 'positive' and 'negative' to fit the machine learning models parameters and for later aggregation functions as averaging and summing. We experiment with the machine learning algorithms, to reach optimum levels of accuracy and performance, by trying POS tagging and location masking. For our twitter corpus, we use VADER [8]. We experiment later with other models as well, such as TextBlob and NRCLex. No feature extraction is needed for this model, as it is built-in the process so, we just make sure the dataset goes through the same pre-processing and data cleaning stages.

Chapter 4

Experiments and Analysis

In this chapter, we discuss the analysis and results obtained from the experiments done based on our methodology. For every section, we show the results obtained for both of our datasets - SentiHood and Twitter Corpus.

4.1 Datasets and Data Collection

For the datasets used, as mentioned in the methodology, SentiHood dataset is a previously annotated dataset reviewing 47 boroughs in England. It was imbalanced so, we had to apply over-sampling, copying enteries of the under-represented class to increase its frequency. Results showed that the sentiment classification models performed better with the over-sampled dataset (balanced) with an increase in the accuracy of 6 percent.

Our Twitter corpus size originally was a little north of 76,000 entries. After discarding the non-english, duplicated and entries of no valuable information, the size was reduced to around 39,000 entries for 18 foreign urban cities. This was the data collected using the 2 methods mentioned before - Geotagging and Keyword Search - where geotagged tweets collected were around only 1100 from over 39,000 which is around only 2.8 percent of the entire corpus and most of it was New York. This is presumed to be due to privacy concerns, as most users do not feel comfortable allowing location tagging. This is why the keyword search approach was our main source for the twitter data. It proved more reliable, as the size of the data retrieved was reasonable, the data itself already contained the location needed for analysis, unlike geotagged tweet where the location tagged is not always mentioned in the tweet, and the tweets predominantly contained sentiment or opinion about that location mentioned. The cities searched for building our corpus were chosen based on their popularity for urban planning, size of their urban population and most importantly the users activity in those cities.

Tracking, analyzing and visualizing the data of 18 active cities and their aspects would have been challenging and not very beneficial. Alternatively, we selected the 3 most active cities -Toronto, Vancouver and Birmingham- and 1 city -Budapest- with medium activity but was divided on a longer time span. The final twitter corpus size is around 13,500 reviewing 4 cities.

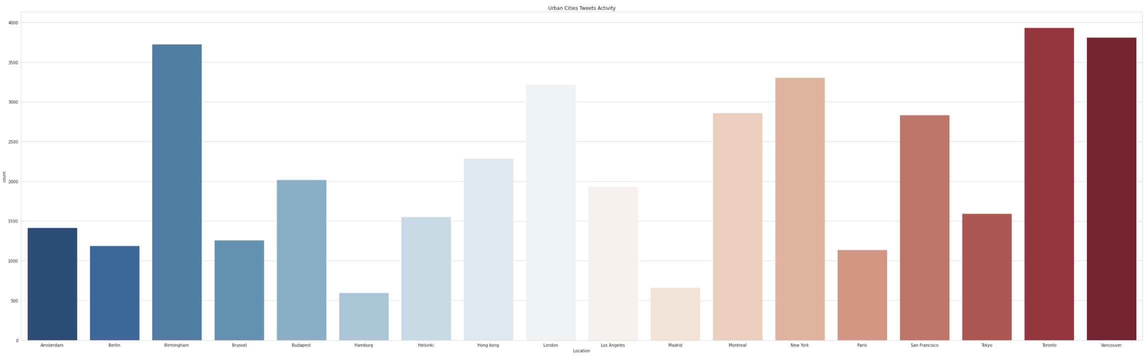

Figure 4.1: Activity of urban cities of the twitter corpus

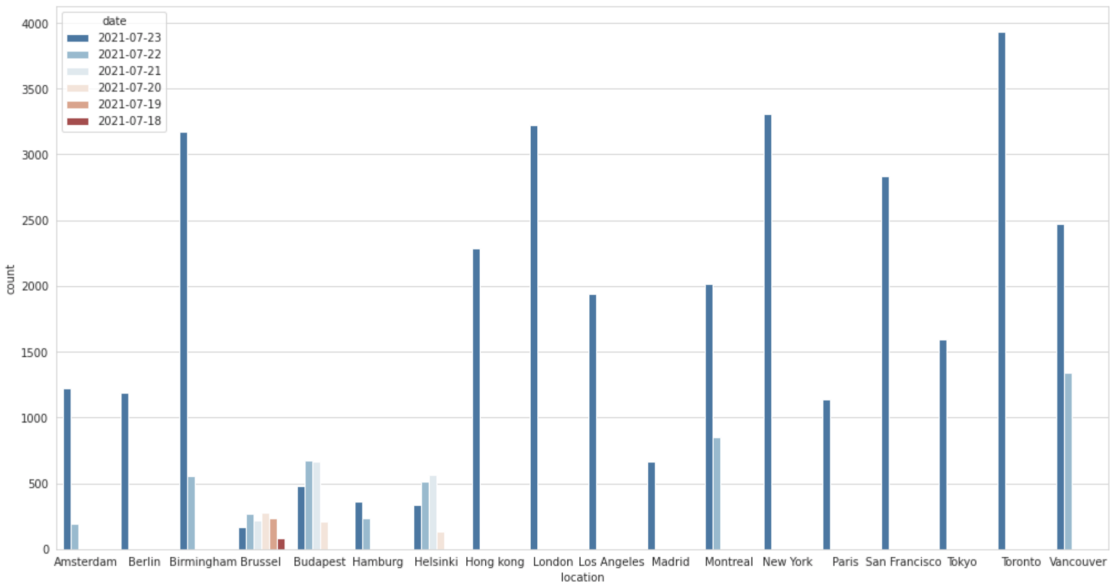

Figure 4.2: Activity of urban cities per tweet date

4.2 Target Extraction

NER using SpaCy is used to extract the named entities in the text. We then filter only the location related tags as GPE to extract the locations mentioned in the tweet. Initially, for the Sentihood dataset, the model was not performing well so, it had to be trained on our data. After the model was trained on England boroughs related text, the accuracy increased to reach 93 percent.

Running the same SpaCy model on the twitter corpus worked smoothly without any training, as the locations mentioned were popular urban cities that fall under the tag Geopolitical, on which the model is previously pretrained.

4.3 Sentiment Classification

The spatio-temporal sentiment analysis consists of 3 parts:

- Analyzing the performances of the sentiment classifiers.
- Temporal Sentiment Analysis: associates both positive and negative sentiments along a time period.
- Spatial Sentiment Visualization: provides sentiment association with locations which consider sentiment polarity in each geographical region.

4.3.1 SentiHood Experiments and Analysis

Starting with SentiHood dataset results, we first started experimenting with the logistic regression model to reach the optimal accuracy. Results obtained from applying the model directly to the vectorized data showed that the model was not performing well, reaching only 60 percentile. Applying POS tagging seemed to have solved this problem. We then explore the different feature extraction models, starting with Word2Vec using Gensim then, CountVectorizer using Scikit-Learn.

Table 4.1 presents the summarized results achieved for sentiment detection for SentiHood showing the weighted average results over all the metrics. The accuracy for the Logistic Regression model when CountVectorizer was used exceeded 80 percent whilst the Word2Vec model reached about 75 percent, showing that CountVectorizer performance was better for out dataset. The CountVectorizer accompanied by location masking performed better across all metric as well. For that, we use location masking along side CountVectorizer through out the rest of the study.

Feature Extraction Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***
LR - Mask - Word2Vec	0.749	0.750	0.750	0.750
LR - No Mask - Word2Vec	0.742	0.750	0.740	0.740
LR - Mask - CountVectorizer	**0.848**	**0.850**	**0.850**	**0.850**
LR - No Mask - CountVectorizer	0.832	0.830	0.830	0.830

Table 4.1: Logistic Regression Model Results

From the table, It is inferred that applying POS tagging, location masking and using CountVectorizer make the highest performing technique. Now we focus on comparing the models, fixing the other variables. Table 4.2 shows the results of the weighted average of all the performance metrics. We can see that the results for the 3 approaches are very close nevertheless, the Stochastic Gradient Descent model performce better than the best Logistic Regression results with and without applying location masking. The best performing approach is using SGD along side POS, location masking and CountVectorizer with accuracy reaching 0.852.

Classifier and Technique	***Accuracy***	***Precision***	***Recall***	***F1-Score***
LR - Mask - CountVectorizer	0.848	0.850	0.850	0.850
SGD - Mask - CountVectorizer	**0.852**	**0.860**	0.850	0.850
SGD - No Mask - CountVectorizer	0.849	0.860	0.850	0.850

Table 4.2: Results of best LR models and SGD models

After building the sentiment classifier and validating the results, the next step was to obtain the general semantic orientation of the sentiments expressed in the tweets regarding the boroughs of England.

It is important to note that Sentihood dataset did not contain the time of posting variable so, for the sake of this study, we simulated the date for every tweet using the time variable from part of the twitter corpus.

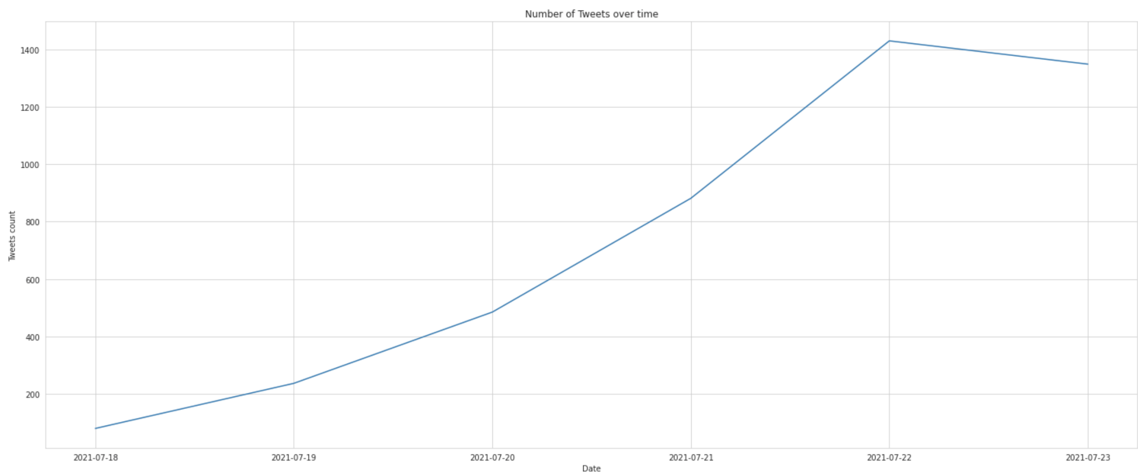

Figure 4.3: Tweets posted per day

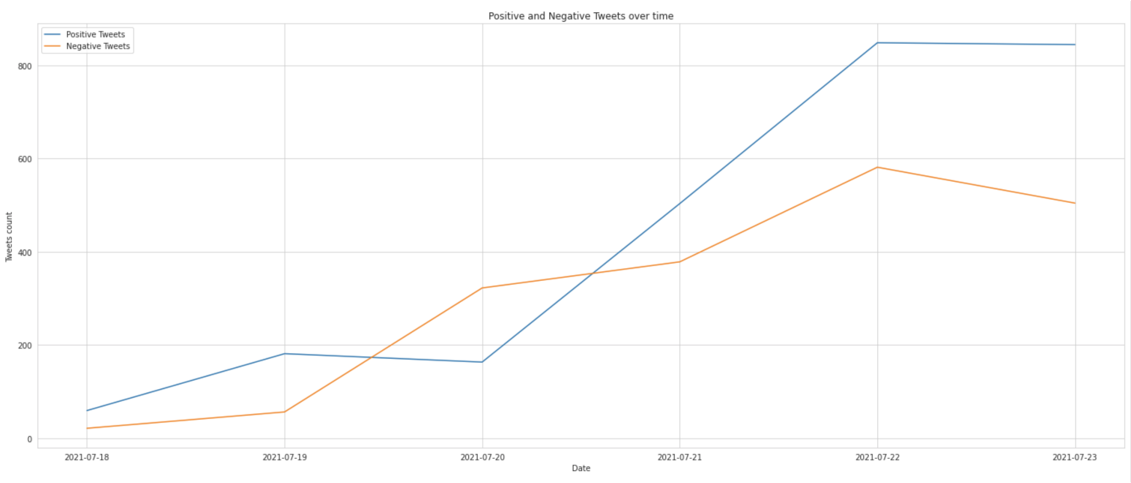

Figure 4.4: Number of Positive Tweets vs Number of Negative Tweets

From figures 4.3 and 4.4, we can see that the number of tweets reached its peak on 22nd of July, with both the positive and negative tweets reaching their highest counts. It is also shown that the positive tweets were leading most of the time except on the 20th of July, where the negative tweets were higher.

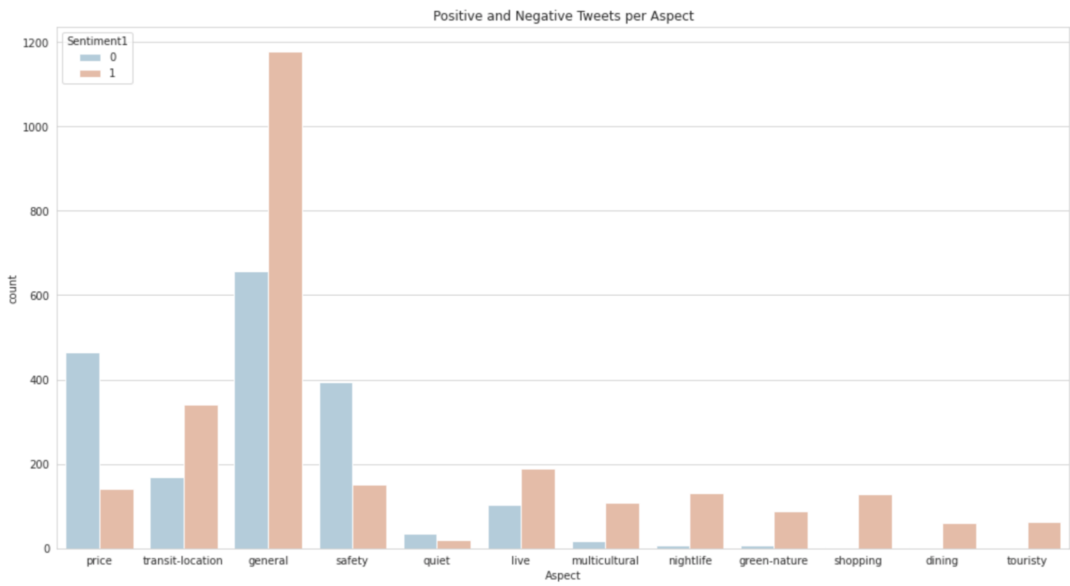

Figure 4.5: Number of positive and negative tweets for every aspect extracted

From figure 4.5, It is deduced that the users most of the time express their opinion directly towards the location in general e.g. "I have friends living in Suffolk and love it" or "Don't go to Somerset I lived there for 3 years its horrible", where the users do not provide any reason or mention the aspect producing their sentiment. In other respects, the aspects, associated with the highest number of negative tweets, were Price followed by Safety. This agrees with the hypothesis, as England is very notorious for its high-priced living and considering this dataset was collected in 2016, safety is another aspect expected to be negatively perceived, as that year was dominated by the UK's vote to leave the European Union and the subsequent political fallout, raising concerns about stability and safety. As for the aspects associated with high positive tweets, transit-location and livability were the highest mentioned. This again agrees with the hypothesis, as England is well known for its high living standards - although pricey - providing a lot of facilities as high quality education, housing and of course, transportation system that facilitates and alleviate the transit aspect of the country.

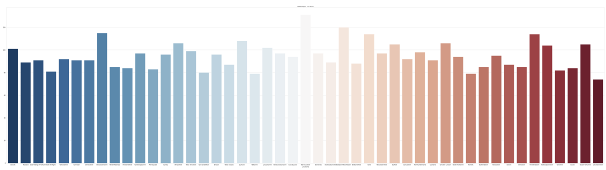

Figure 4.6: Number of tweets per location

Since we are interested in providing a spatial and temporal sentiment analysis, we are interested in understanding the sentiment distribution on both temporal dimensions as shown above - when the sentiment may vary as the time passes - and spatial dimension, where it can be possible to identify geographic regions with their negative or positive sentiment. In our approach, we first try to plot the activity and then sentiment for each location. Trying to plot the number of tweets posted for every location proved profitless, as can be seen in figure 4.5, as it is hard to analyze and track the count for 47 boroughs.

Instead, we are going to analyze the activity of only 3 locations:

- The most active location: Warwickshire - The location with the highest number of positive tweets: Kent - The location with the highest number of negative tweets: Gloucestershire

It is quite noticable, from figure 4.7, that the most active location, Warwickshire, almost has an overall neutral sentiment, where the difference between the positive and the negative tweets count is diminutive. Whilst the other 2 locations overall sentiment is positive for the location with the highest number of positive tweets, Kent, and negative for the location with the highest number of negative tweets.

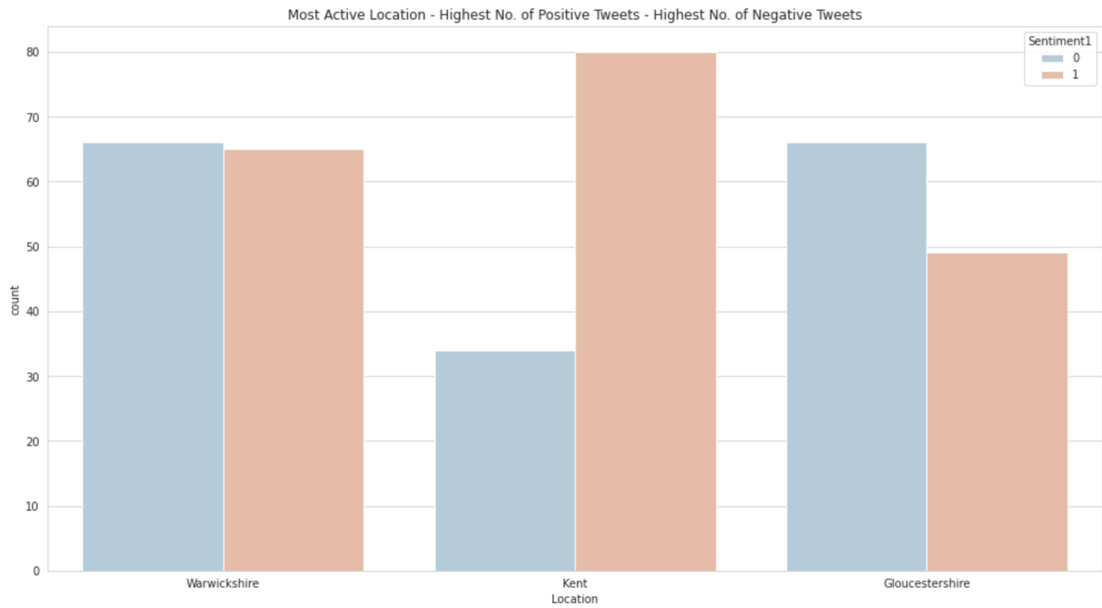

Figure 4.7: Number of Positive Tweets vs Number of Negative Tweets for Warwickshire, Kent and Gloucestershire

From figures 4.8, 4.9 and 4.10, We observe that through out the 5 time periods for Warwickshire, the positive tweets count and the negative tweets count take turn to lead, resulting in an overall neutral sentiment over time and ending on the 23rd of July with a diminutive difference.

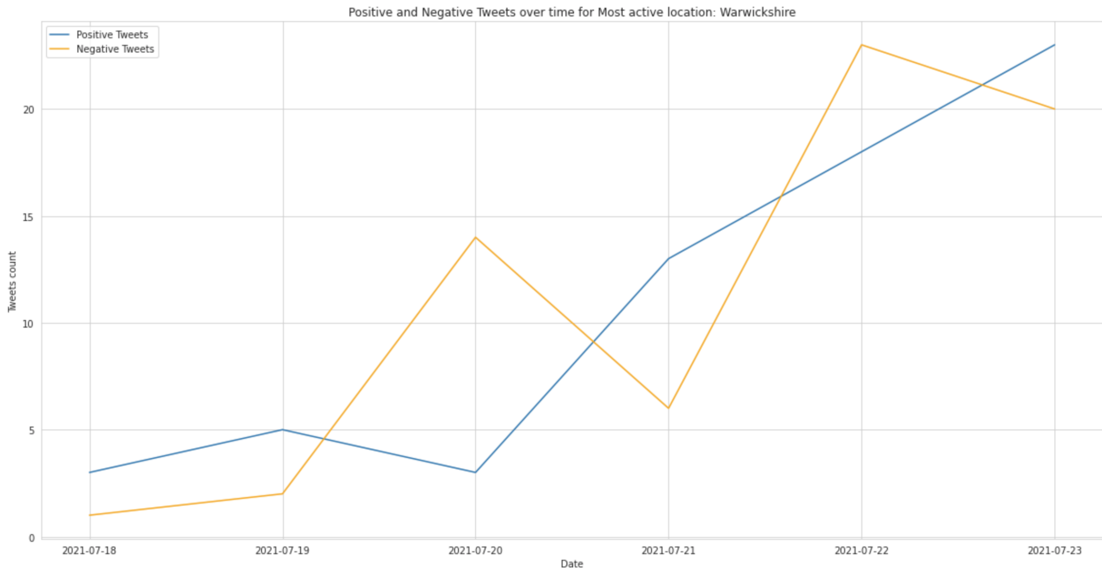

Figure 4.8: Warwickshire sentiments counts per day

As for Kent, the number of positive tweets is higher continually except for a very small period of time on the 21st of July, where the difference was still minor; ending with a considerable gap between the number of positive and negative tweets, with the number of positive tweets is much higher.

Meanwhile for Gloucestershire, the negative sentiment is dominating the graph through out almost the entire time period, with a considerable gap between the sentiments.

The 3 location start off the time period with low activity (number of tweets) ending with a significant increase in the activity.

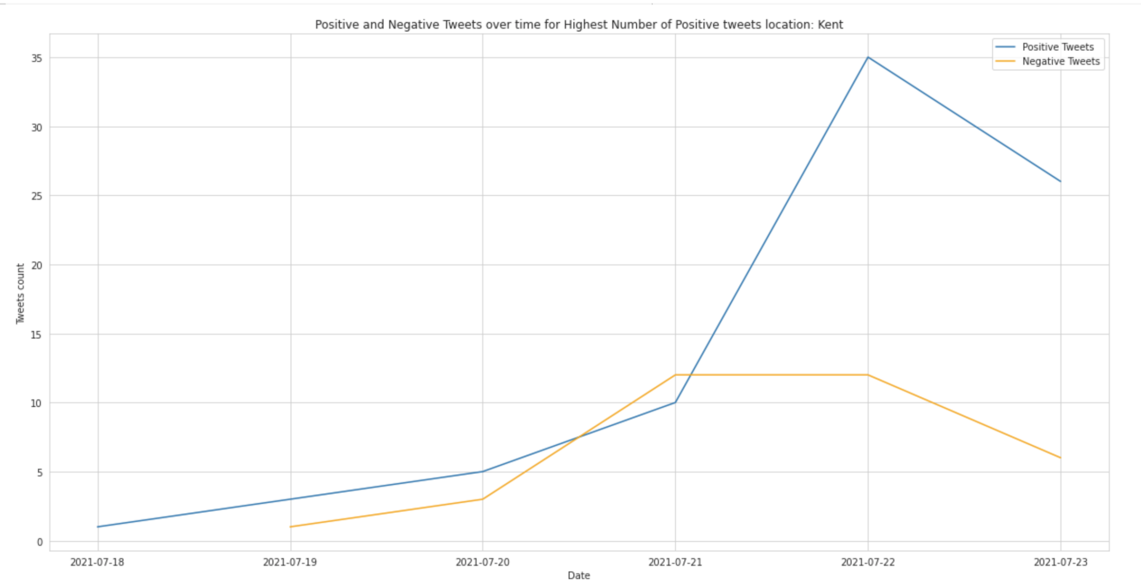

Figure 4.9: Kent sentiments counts per day

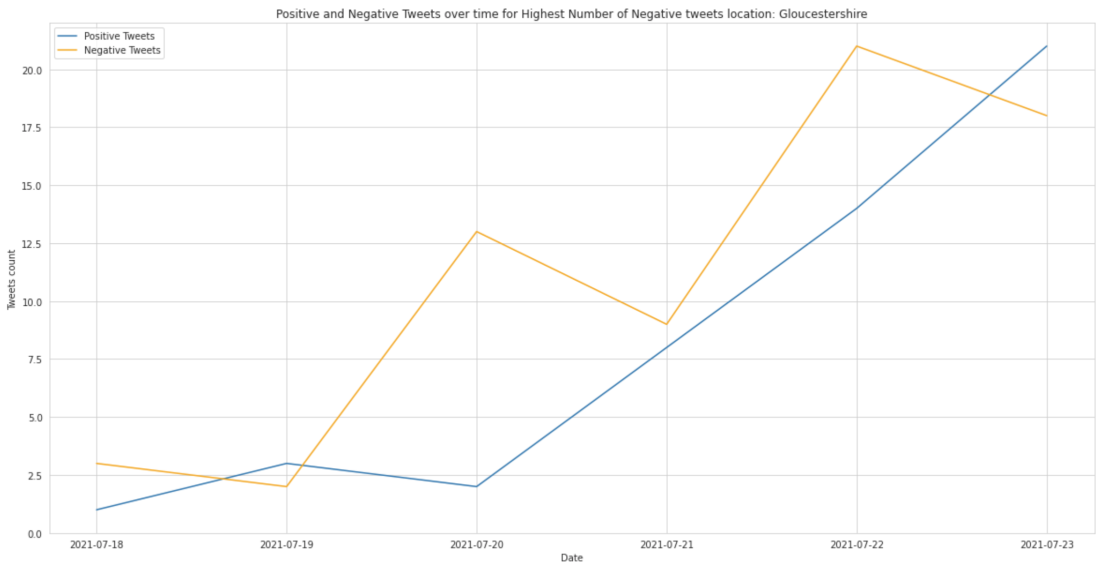

Figure 4.10: Gloucestershire sentiments counts per day

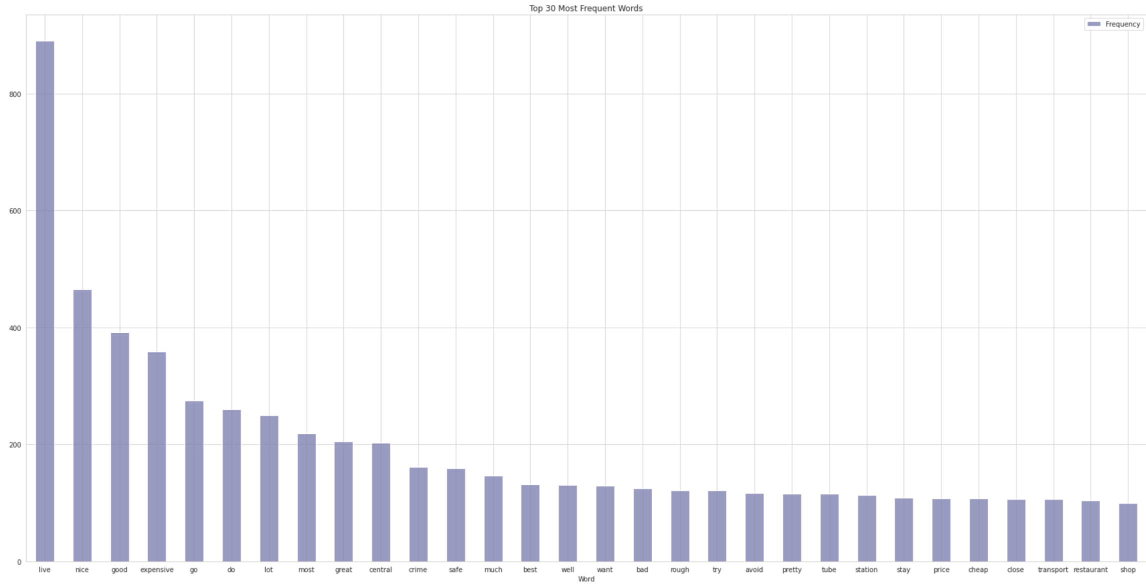

Figure 4.11: Top 30 most frequent words

Figure 4.12: Wordcloud of the most frequent words

The most recurring words in the dataset conveyed the most mentioned aspects e.g. live relating to the aspect of live or livability, expensive relating to Price; where other frequent words express the positive sentiment e.g. nice, good, which is expected, as the original dataset had 2598 positive tweets compared with 931 only negative tweets. Furthermore, after balancing the dataset, the positive tweets count was still higher so, it is logical that the more frequent words represent positive sentiment.

4.3.2 Twitter Corpus Experiments and Analysis

We followed a similar approach with minor differences for the twitter corpus. A sample of 340 tweets were randomly picked out and manually labeled as either negative class (0) or positive class (1). This method was used to compare the classifiers' results. Two class sentiment classification was applied to fit our labeled data.

Table 4.3 presents the weighted average results for the performance metrics. It is observed that Vader, with the neutral class counted with the positive class, performed the best. As a result, for the rest of the experiment Vader was our elected model.

Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***
TextBlob - neutral perceived as negative	0.66	0.66	0.66	0.66
TextBlob - neutral perceived as Positive	0.72	0.74	0.72	0.69
Vader - neutral perceived as negative	0.68	0.68	0.68	0.68
Vader - neutral perceived as Positive	**0.75**	**0.75**	**0.75**	**0.73**

Table 4.3: TextBlob and Vader results on labeled sample

For the rest of the corpus, we were using Vader for the sentiment classification and the feature extraction process was disregarded; we only applied POS tagging and location masking. Three class sentiment classification was applied to the rest of the corpus since, unlike SentiHood, the twitter corpus was not previously annotated and included great amount of neutral tweets that would be later ignored since, we only care for the aspects that the users perceive well or perceive badly to be reported.

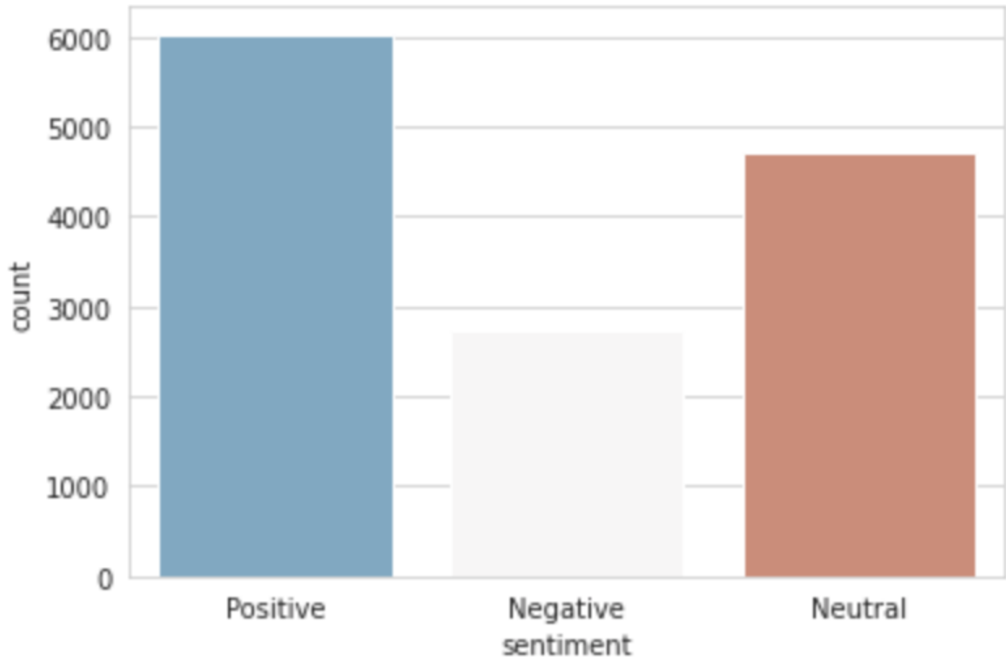

Figure 4.13: Sentiment polarities count

Figure 4.13 shows the result of the Vader model applied to the twitter corpus. Around half the tweets were classified as positive whereas the negative class represents less than quarter the corpus.

The rest of the spatial and temporal analysis process is practically identical to the SentiHood analysis process. We start of presenting the temporal sentiment analysis then the spatial sentiment analysis.

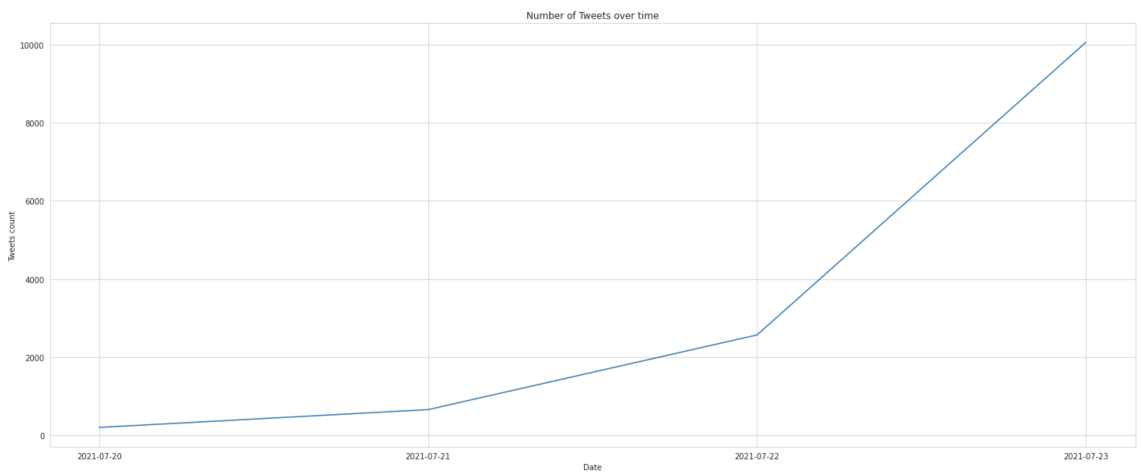

Figure 4.14: Tweets posted over time

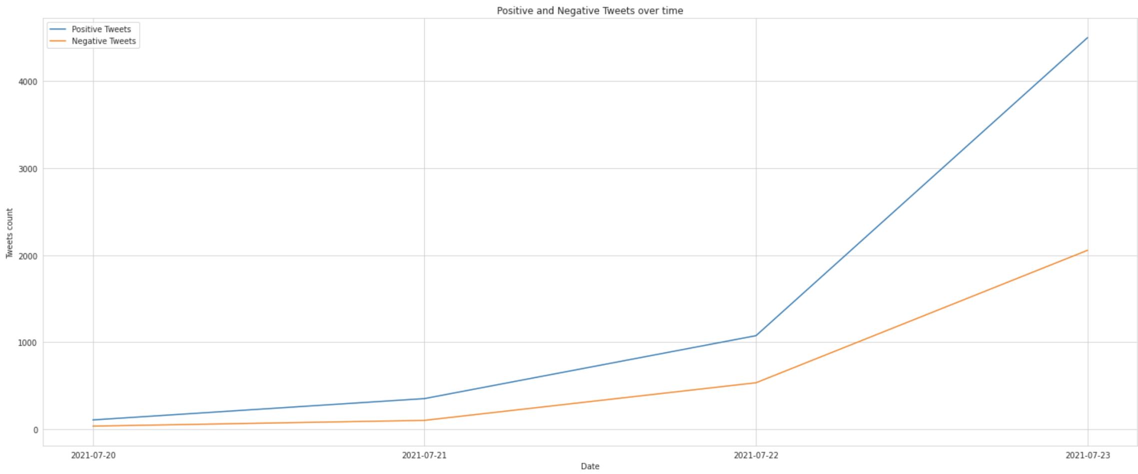

Figure 4.15: Number of Positive Tweets vs Number of Negative Tweets

From figures 4.14 and 4.15, It can be observed that at the beginning of the time period, the tweets count was minor, increasing significantly by the end of the time period. The positive tweets was leading the entire time, growing the gap between the positive and the negative tweets count as time goes by. By the end of the time frame provided, both the positive and negative tweets reach their peak.

As mentioned in the Data Collection Section at the beginning of the chapter, tracking the 18 urban cities was macroscopic for the scope of this research. The cities that were explored are: Toronto, Vancouver, Birmingham and Budapest.

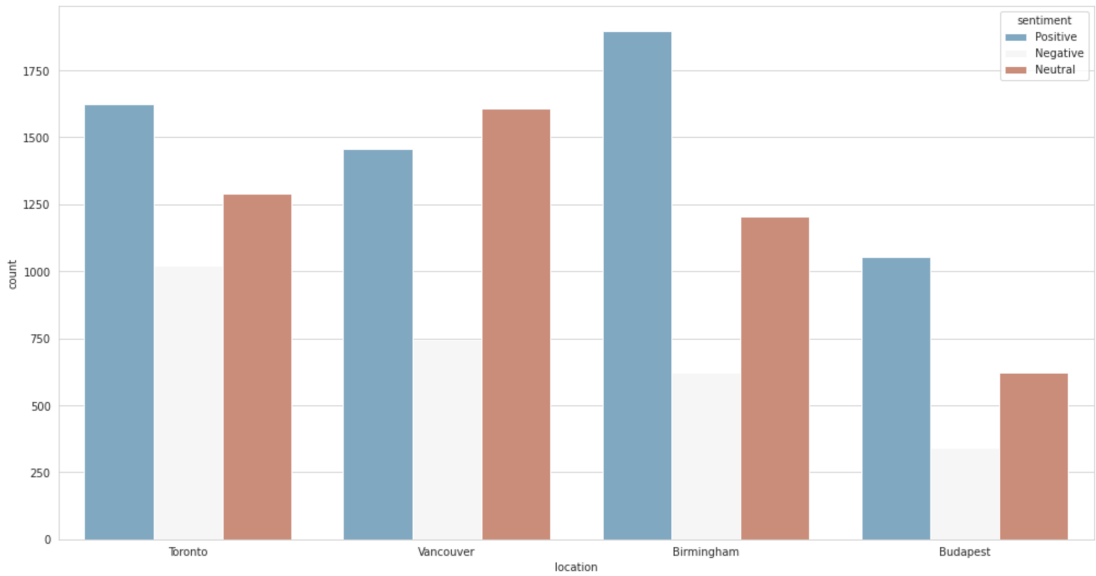

Figure 4.16: Sentiment polarities for Toronto, Vancouver, Birmingham and Budapest

It is quite noticable, from figure 4.16, that Birmingham has the highest number of positive tweets and the largest gap between the sentiment polarities of the tweets. Vancouver's highest number of tweets is neutral and the overall prevalent sentiment is positive. This agrees with the results presented in figure 4.15.

From figures 4.17, 4.18, 4.19 and 4.20, We observe that through out the entire graph, the positive sentiment in Toronto and Birmingham is always ruling, resulting in an overall positive sentiment. The overall prevalent sentiment in Vancouver and Budapest is positive sentiment as well but, for Vancouver we can see that the negative sentiment count increases significantly to reach the positive tweets count before it drops. For Budapest on the other hand, the positive and negative tweets count meet at several point and the count differs greatly over time, dropping then rising several times, to end with a high rise in the positive tweets and a low drop for the negative ones.

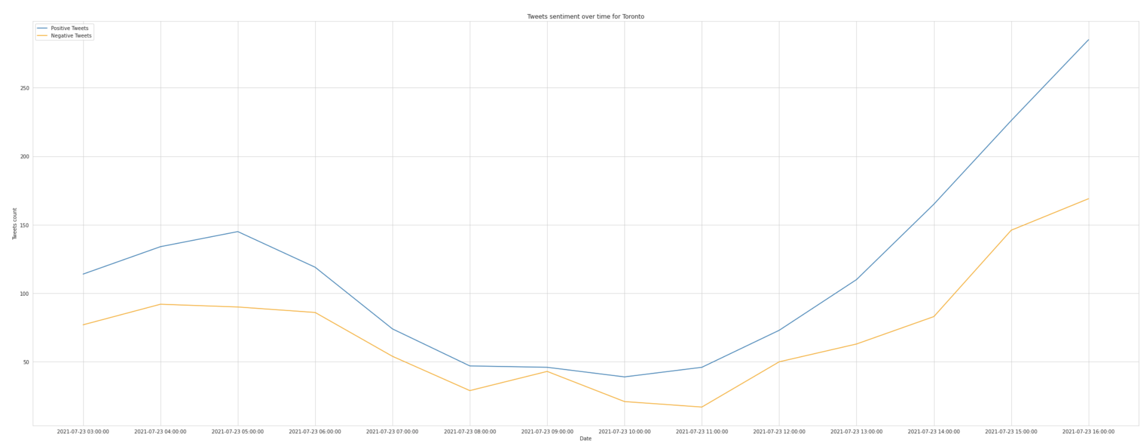

Figure 4.17: Toronto sentiments count over time

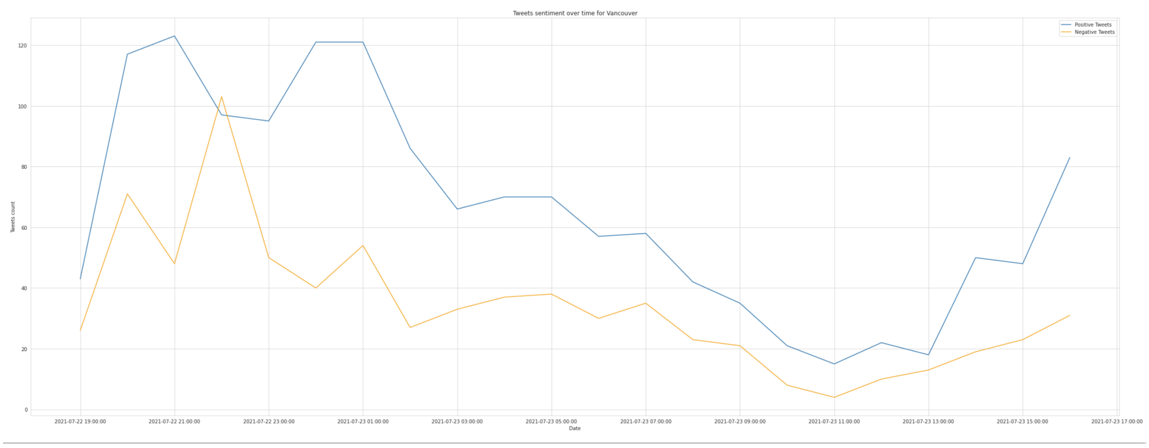

Figure 4.18: Vancouver sentiments count over time

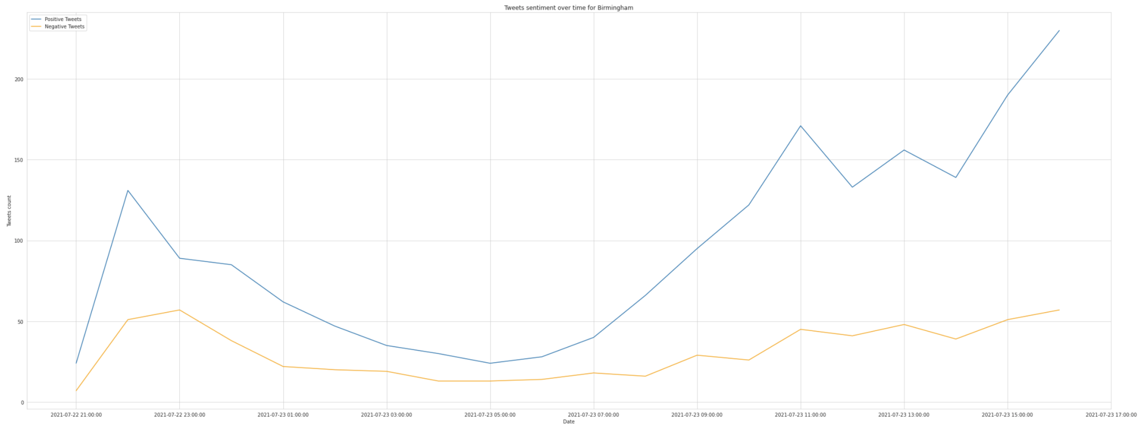

Figure 4.19: Birmingham sentiments count over time

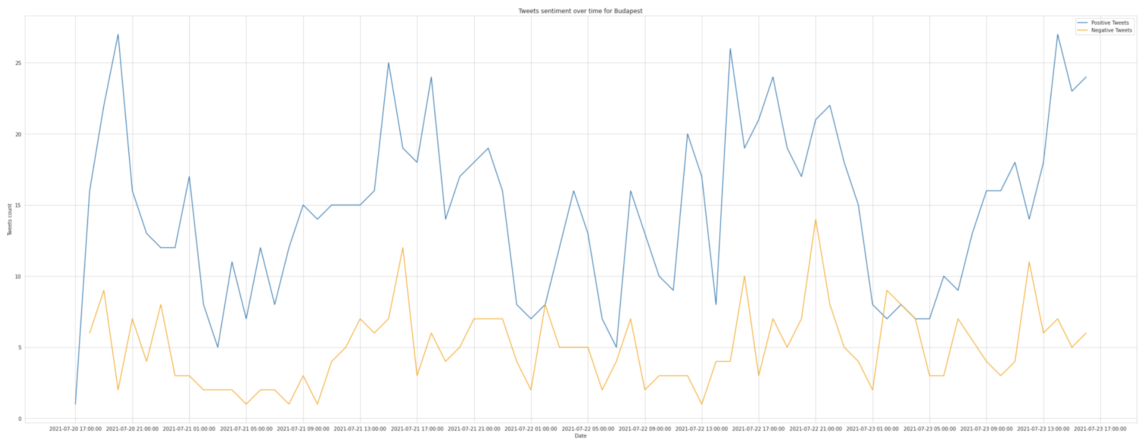

Figure 4.20: Budapest sentiments count over time

The time frame for the tweets is: 1 full highly active day for Toronto, 2 days for Birmingham and Vancouver and 4 evenly active days in Budapest. They all take place during the period between 19th of July and 24th of July, which is the heart of the summer vacation for most people. The 3 cities - Toronto, Vancouver and Budapest - are considered to be popular areas for tourism and family vacations.

Enormous amount of students study in both Toronto and Vancouver, resulting in a lot of family vacations taking place in both cities. They both also host some of the most popular festivals e.g. Vancouver Mission Folk Music Festival and other several Jazz concerts as Groovin' in the Park in Toronto. The consistant posting of tweets over the time of the day resulted by the posts of the family vacation, whereas this spike that can be seen in the Toronto graph before it dropped at around 5 in the morning, and in Vancouver starting at 11 pm, is caused by the music festivals as most festivals start at 12 AM. The spike that we see stating at 12 PM is caused by an event, that is unrelated to our study as most its tweets are neutral, where the Toronto football team took place, causing the spike in tweets count. As for the negative tweets in Toronto and Vanvouver, they were caused by residents calling out the encampment residents and park evictions, that were caused by the city's worsened ongoing affordable-housing crisis by displacing people who are badly in need of shelter, as well as the brutality of the police force attacking the homeless encampment. The rise in the tweets count before 11 PM is not caused by a particular event. Our explanation for that would be that this is the time when most people in general are awake and online.

The positive tweets in Birmingham are general expressions of admiration for the city as "Oh Birmingham how Ive missed you" whereas, the negative tweets are mostly about covid-19 and the vaccine e.g. "It's the unvaccinated folks that are letting us down." " It's Time to Start Blaming the Unvaccinated Folks for COVID Surge" Birmingham birmingham pheonix cricket match at 15 caused the spike, that again is unrelated to our study as we chose to ignore the neutral tweets

As for Budapest, most of the positive tweets were about good hotels and touristic attractions. This agrees with the timing of the summer vacation, and since Budapest is the touristic city it is, that result was expected. The negative tweets were caused by the Pride march protest against anti-gay laws. We can see that the activity of the tweets in Budapest forms a pattern, where the counts rise at mid-day (5 pm) then drops at night (1 to 5 am). The reason for that might be that most residents and tourists both are most active during the day during the summer.

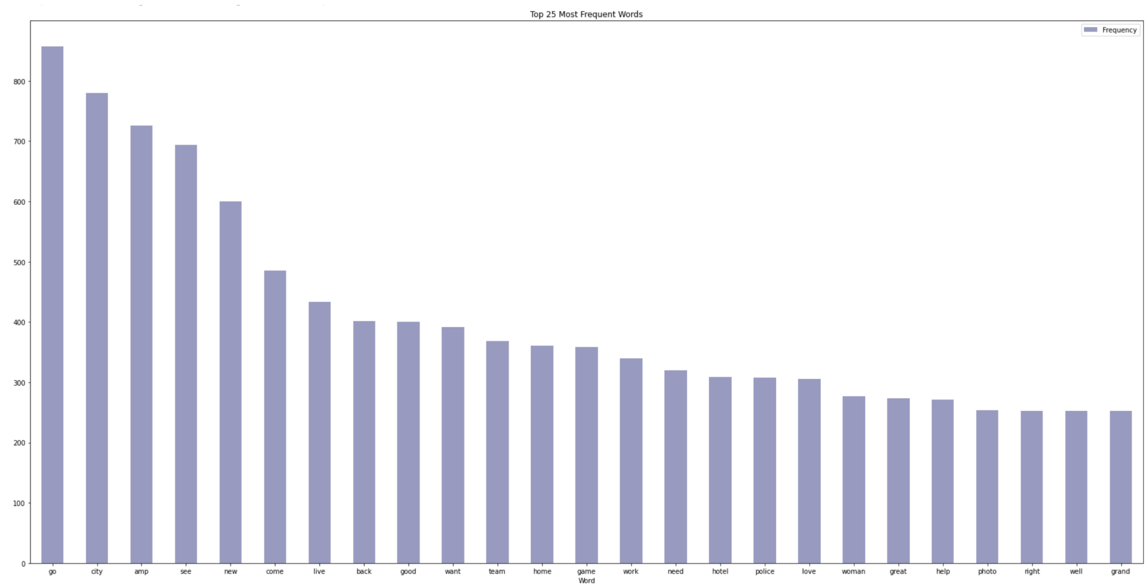

Figure 4.21: Top 25 most frequent words

Figure 4.22: Wordcloud of most frequent words

From figure 4.21 and 4.22, we can clearly observe that the most dominant words are activity related as go, see, come, live which is expected as most tweets for these cities express sentiment in general for the city without aspects e.g. "Ppl be like I love my city and the city is Toronto" (ppl is slang for people) and "I've heard Toronto is dope, would love to go sometime.", " Woohoo! Got tickets for Toronto!! See you in January". The tweets express sentiment for the ongoing events taking place and excitement for visiting and going to these cities, which is why the most frequent words are related to movement.

Most of the tweets for the twitter corpus express sentiment towards doing an activity in the city rather than an aspect like price or safety. This is expected as those cities are very popular among tourists and as family vacation spots so, it is expected to express sentiments related to activities or movement rather than a stable aspect that is perceived by a resident.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The purpose of this study was to analyze the tweets and opinions posted by users of social media platforms, as Twitter or Yahoo, to find out how people perceive their living enviroment.

In this paper, We investigated and implemented various techniques on data mining and text pre-processing to track and analyze their performances and accuracy in extracting opinions from tweets for urban analysis, given a certain period of time. We explore 2 different datasets, a previously annotated dataset -SentiHood- and a twitter dataset, that is collected from recent twitter posts, about several urban foreign cities. We also explore the targeted sentiment analysis in this paper. Two approaches were explored for SentiHood, logistic regression and stochastic gradient descent, improving their results using POS tagging to better the context as well as location masking. The SGD approach provided the optimum results with accuracy of 85 percent when grouped with CountVectorizer, that proved to perform better than Word2Vec model. NER was used for detecting the location tagged in the tweet.

Results showed that the residents of Warwickshire are the most active on social media, expressing both positive and negative sentiment towards their borough regarding different aspects. The residents of Kent were the most satisfied with their borough unlike Gloucestershire, where most residents were discontent with it. Most of the users would express their general opinion or feeling towards the area without specifying the aspect. This made it more challenging to pinpoint exactly what needed to be changed or enhanced to improve residents opinions about the place. For the scope of this research, only these 3 locations were explored, but the general approach can be followed on any other location and people's sentiment towards it can be analyzed using the aforementioned methods.

Vader and TextBlob text processing libraries were explored for the analysis of the twitter corpus. Vader classified the sentiment with higher accuracy than TextBlob with accuracy reaching 75 percent.

The results for the twitter corpus showed that people in Toronto, Vancouver, Birmingham and Budapest are mainly interested in activities that can be performed with family and friends, which coincides with the summer vacation activities and general spirit. It was observed that residents in Toronto and Vancouver are concerned about the affordable-housing crisis and the police brutality while handling the homelessness problem which raises concerns for safety. On the other hand, the residents in Birmingham are mainly concerned with their health with a covid-19 surge starting once more in the city. Budapest showed the most stable pattern for posting on twitter, where people are more active during the day and much less active at night, posting about the beauty of the city and its touristic attractions. That is due to the touristic nature of the city coinciding with the summer vacation, increasing the number of visitors and their opinions for the city.

Some of the limitations of this study were, not handling emoticons or sarcasm in tweets well. Conducting this research on other language would improve the insights collected as not all users in a place post in English. This study was initially going to be conducted on Egyptian cities, however, even though Egyptian tweets were available, they did express sentiments towards neighborhoods or cities rather than express sentiment in general; therefore, extracting insights from Egyptian tweets proved to be a challenge, that lead us to working with more popular and touristic urban cities. The availability of larger datasets would have improved the insights generated for this study, as more opinions would have been analyzed and better deep learning models as LSTM could have been used. One of the most important limitations to be considered is that Twitter is popular among the younger generations. This excludes the more senior generations and some of the important opinions as when it comes to place of living, almost everybody has an opinion as to how to improve their living situation.

5.2 Future Work

Further work will explore techniques to deal with emoticons and sarcasm to improve sentiment prediction and exploring different platforms for less noicy and more specific, urban related opinions. Last but not least, to gather more detailed, accurate and live data, we are encouraged to explore detecting sentiment and aspect from audio data, rather than text from social media, by collecting real-time audio of the urban neighborhoods to be analyzed.

Appendix A

Lists

List of Figures

List of Tables

Bibliography

[1] Twint: Twitter Scraping Without Twitter's API.

[2] SentiHood Dataset.

[3] Gensim Word2Vec Model.

[4] ScitKit-Learn CountVectorizer.

[5] SciKit-Learn Logistic Regression.

[6] SciKit-Learn Stochastic Gradient Descent.

[7] Named Entity Recognition (NER) using spaCy.

[8] Sentiment Analysis using Vader.

[9] André Luiz Firmino Alves, Cláudio de Souza Baptista, Anderson Almeida Firmino, Maxwell Guimarães de Oliveira, and Anselmo Cardoso de Paiva. A spatial and temporal sentiment analysis approach applied to Twitter microtexts. *Journal of Information and Data Management*, 6(2):118–118, 2015.

[10] Tao Hu, Bing She, Lian Duan, Han Yue, and Julaine Clunis. A systematic spatial and temporal sentiment analysis on Geo-Tweets. *IEEE Access*, 8:8658–8667, 2019.

[11] Yingjie Hu, Chengbin Deng, and Zhou Zhou. A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments. *Annals of the American Association of Geographers*, 109(4):1052–1073, 2019.

[12] M Parimala, RM Swarna Priya, M Praveen Kumar Reddy, Chiranji Lal Chowdhary, Ravi Kumar Poluru, and Suleman Khan. Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach. *Software: Practice and Experience*, 51(3):550–570, 2021.

[13] Debjyoti Paul, Feifei Li, Murali Krishna Teja, Xin Yu, and Richie Frost. Compass: Spatio temporal sentiment analysis of US election what twitter says! In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1585–1594, 2017.

[14] Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. *arXiv preprint arXiv:1610.03771*, 2016.

[15] Zhiwen Song and Jianhong Cecilia Xia. Spatial and temporal sentiment analysis of twitter data. *European handbook of crowdsourced geographic information*, 205, 2016.