Investigation of Social Media Digital Footprints on Tourist Destinations

I.M.C.Desaman

Investigation of Social Media Digital Footprints on Tourist Destinations

I.M.C.Desaman

Index No: 14000212

Supervisor: Dr. H.A.Caldera

January 2019

Submitted in partial fulfillment of the requirements of the B.Sc in Computer Science Final Year Project (SCS4124)

Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: I.M.C.Desaman

……

Signature of Candidate Date:

This is to certify that this dissertation is based on the work of

Mr. I.M.C.Desaman

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Supervisor's Name: Dr. H.A.Caldera

……

Signature of Supervisor Date:

Abstract

Social media is one of the most active areas on the internet today with millions of active users. It has become a data source for lots of data-oriented social studies since users share their experiences, feelings, thoughts, activities etc. through their social media accounts. Social media digital footprints are the trails which are left by travelers in social media, such as feeds, check-ins, photos etc. Sentiment analysis can be used to mine the opinion from those feeds and also geographical analytics can be done using geotagged feeds in order to find points of interests.

In this dissertation, a novel method is proposed in order to rate the tourist destinations using mainly two approaches which are namely sentiment analysis approach and tourist density-based approach. In sentiment analysis approach the feeds which are created by tourists within a tourist destination are extracted. And find the tourist's opinion on that particular destination using aspect-based opinion mining. In tourist density-based approach, geotagged feeds on a particular tourist destination are extracted and map those feed on a geographical map. Then tourist densities of the census tracts are examined in order to obtain the points of interests. The most importantly, to identify the feeds which are created only by tourist, not by the residents or organizations, the tourist's spatiotemporal sequence is examined.

In the evaluation, two types of evaluations are done for two approaches, the standard methodology of using the confusion matrix is done for evaluation of the opinion mining. And a comparative study is conducted to evaluate the visiting patterns. For this approach, the obtained regions of attraction in every location are tested. To evaluate the regions of attraction extraction, heatmaps are generated using the same dataset and peak points are compared with the regions of attraction.

Preface

A novel methodology to consider the social media feeds in rating tourist destinations is presented in this dissertation. Here two novel approaches are presented in this study. Several studies on aspect-based sentiment analysis on user reviews are conducted in the literature. Using this approach on social media feeds to obtain visitors opinion on a tourist destination was solely my own idea, because social media feeds contains more linguistic value and they contain purely user's opinion. Using a density-based approach to obtain points of interests is the most common approach used in urban analytics literature. So, in this study, this approach is also used to identify which places are mostly visited by the tourists within the year. So, here the aim is to combine these two approaches in order to create effective recommendation rather than using static reviews and ratings done by tourists. In identifying non-tourists, here used an approach which is based on the spatiotemporal sequence of tourist's feeds. This approach is used with a photo trail in the study conducted by Zheng et al. [1] to identify non-tourist travel paths. In this study, aspect-level sentiment analysis is used in to identify tourist interests. To identify the aspects, several grammatical dependency rules are developed and according to those rules, aspect terms are extracted. This aspect extraction process is based on dependency parsing which is discussed in a study by Marie-Catherine et al. [2].

Acknowledgement

I would like to express my sincere gratitude to my research supervisor, Dr. H.A.Caldera, senior lecturer of University of Colombo School of Computing for providing continuous guiding and the supervision throughout the research.

I would also like to extend my gratitude toward Dr. M.D.J.S.Goonetillake, senior lecturer of University of Colombo School of Computing and Dr. T.A.Weerasinghe, lecturer of University of Colombo School of Computing for feedback and the evaluation throughout the research. And also, I would like to acknowledge the assistance provided by Dr. H.E.M.H.B.Ekanayake as the final year computer science project coordinator.

I appreciate the feedback and motivation provided by my friends. I would like to dedicate this dissertation to my loving family members who were always being with me and strengthen me towards my achievements. I take this moment to appreciate all the people who helped me to bring this research to a success.

Finally, I appreciate all your guidance, supervision, motivation without all your help this would be a miracle. Thank you for our support.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ML	Machine Learning
RoA	Region of Attractions
PoI	Point of Interest
RBF	Radial Basis Function
POS	Part of Speech
RAKE	Rapid Automatic Keyword Extraction
OLS	Ordinary Least Squares
Tf-idf	Term frequency inverse document frequency
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
NER	Named Entity Recognition
API	Application Programming Interface
LCH	Leacock Chodorow Similarity
WUP	Wu Palmer Similarity

Chapter 1 - Introduction

Social media is one of the most active and widely spreading data source on the Internet today. Social media consist of millions of data containing the user's daily activities. So, it has been chosen as the data source for many studies. Social media has long-running past from creating Usenet by Tom Truscott and Jim Ellis from Duke University in 1979, which allowed internet users to post public messages. Over since this period Bruce Ableson and Susan Abelson founded Open Diary which brought online diary writers into one community, in 2003 MySpace was founded and, in 2004 Facebook was founded. Today it has become a long journey and social media has evolved into a virtual world which is a computer-based simulated environment inhabited by 3D avatars such as Linden Lab's Second life [3]. Social media contains user's data profiles and connections or interactions between those user profiles. Those are the main components of social media. This vast social media society is structured on those two main components. To denote users, social media uses nodes and to denote connections or interaction it uses edges [4].

In tourism studies, researchers use different kinds of data sources. Visitor surveys, online reviews and transportation statistics are the static data sources because they are infrequently updated and they are highly biased due to the nature that they require volunteer input from online users and require a considerable amount of laborious effort for data acquirement. Dynamic data sources are real-time data sources which are updated real-time. Social media data can be considered as dynamic because social media digital footprints collected in a recent time period which gives real-time insights than static data sources. Those dynamic data can be collected in big sizes in real-time, they come from the natural activities of the users and those data are frequently updated thus provide more reasonable data for the studies [5]. A digital footprint is information on the internet about a particular person or a group as a result of their online activities. Those digital footprints can be collected from photo sharing communities, Social media check-ins, Social media feed geo-locations, Bank transactions, Hotel bills, Internet access using WIFI, Places visitors logged into Internet etc. [6]. When collecting big sizes of data, it is considered three main characteristics, Volume (Size of the data set), Variety (Different formats of

data; structured, semi-structured, unstructured) and Velocity (How fast the data set is evolving, which means, created in real-time or not) [7].

People who used to travel, eventually face some problems, such as when they visit someplace, it may not be what they think of or it may not be a good season to visit there. For example, in some places, there are some seasons which are the climate is worst. And also, people go for hikes and get lost without knowing the area. It is not practical that everybody contacts people who live in that area for precautions before planning a trip. There are lots of resources on the Internet about the tourist destinations but those are infrequently updated and contains static data or people who are planning on a trip do not bother to surf through all over the Internet. So, rather than going through those difficulties it is better if there is a method to get a recommendation from the people who recently visited there. It is more reasonable and reliable as well. So, this study is to find such an effective solution targeting travelers.

In this study, the social media feeds which are created by tourists are extracted in a periodical manner. Then the tourist opinion over a particular destination and tourist density in a particular period are examined. Even for the same destination tourist's opinion and a number of tourists might differ within different periods of the year because of the several reasons such as climate, festival seasons etc. Since this study results a real-time recommendation which takes user-generated social media feeds in real-time and analyze them, this will end up with a more effective solution than the dynamic data sources such as online reviews, ratings and statistics.

1.1 Background to the Research

Urban analytics is the most common analysis followed by lots of researchers in the tourism field. Studying the tourist behavior, their visiting patterns etc. are used in tourism studies which are the main usage of urban analytics. When extracting relevant feeds from the social media there are effective parameters such as hashtags, keywords, geographic boundaries/ coordinates which are frequently used by social media users [5].

Sentiment analysis is a popular approach which is used in mining opinion from a text which is known as the polarity. In the literature, this technique is used by researchers to obtain polarity of the user reviews extracted from review websites [8]. Generally, sentiment analysis is done using two techniques which are namely machine learning

(ML) based approach and lexical based approach. Again, the lexicon-based approach again divided into two approaches which are dictionary-based approach and corpus-based approach [9]. This flow is presented in

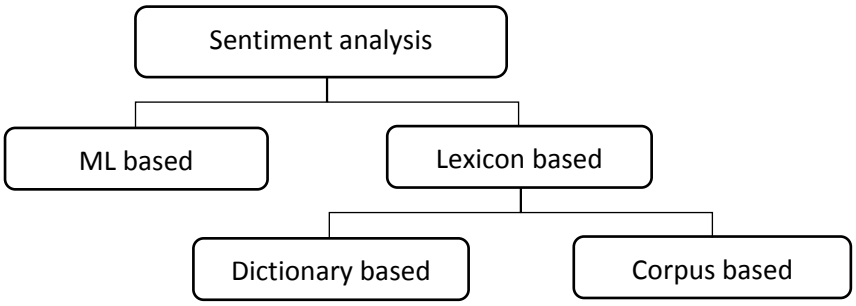


Figure 1-1: Approaches of sentiment analysis

Since this study is focused on finding a recommendation on a tourist destination. It needs to be found the opinion and interests of only the visitors or tourists, not the residents of that area. Otherwise, the study will be biased. All the geotagged feed which is extracted from social media are not created only by the tourists. There are feeds which are created by the residents, organizations etc. Since this study is directly extracting the feeds from social media. The extracted feeds needed to be filtered in order to get the feeds which are created only by tourists. So, one of the challenges in this study is removing the feeds which are created by the non-tourists. To tackle that challenge, this study is taken spatiotemporal sequence-based approach over census tracts. Which is taken a feeds trail correspond to the spatiotemporal sequence and assumed that the feeds created by true tourists spread over large spatial extent within the trail [1].

In urban analytics, it is common practice that most of the studies used a density-based approach in finding human behavior. Which is mainly based on digital footprints using people's online activities. So, it can be found that where they were in a particular time in a day using this evidence. Then the human densities over a particular location can be obtained using the geographical representation of these data. This is the general usage to find the Region of Attractions (RoA) or Point of Interest (PoI) [1]. When rating a particular tourist destination, it's popularity which also can be called as an attraction, is an important factor. So, in this study, the second approach is finding the tourist densities over a particular destination. In this approach, the geo-tagged tourist feeds are geographically represented with the period of the year which they were generated. This geographical representation which is a density map is then divided into census tracts and

the tourist densities and RoA over the period of the year for a particular census tract are obtained.

Combining these two approaches, which are opinion-based approach and tourist density-based approach, rating and a recommendation is presented in this dissertation. In this recommendation, the period of the year which is the feeds are extracted is taken into the account because it is an important perspective which needed to be evaluated in a recommendation.

1.2 Research Problem and Research Questions

Using social media data as a data source for the tourism studies is more effective rather than other data sources such as visitor surveys, transportation statistics and online reviews which are highly sparse because they require volunteer inputs from the users and infrequently updated [5]. Social media data can also be linguistically analyzed in order to get user opinion. But however, unlike reviews, these data are noisy and unstructured. So, the gathered data needed to be structured. Collected data contains lots of information which cost lots of storage and also there will be massive analysis problems [7]. Only the feeds which are created by visitors needed to be filtered otherwise opinion would be biased. So, the feeds created by non-tourists needed to be removed [1]. Then the tourist's opinions over a destination needed to be extracted from the collected feeds and then that opinion is used to create the rating. Need to examine tourist visiting patterns over different periods of the year because recommendation need to be real-time and tourist attractions over a particular destination might depend on the seasonal climatic changes and other factors. For this purpose, geo-tagged feeds and their timestamps are used. In order to provide an effective recommendation as the final outcome, these two methods of tourism studies which creates a rating using social media opinion and examining tourist visiting patterns, are combined and practically applicable solution need to be implemented.

To address the research problem this study is focused on the question of, what is the appropriate methodology to generate a recommendation for tourist destinations by analyzing the digital footprints in social media?

- What is the appropriate method to structure the complex dataset obtained from social media according to traditional date lengths and formats?
- Which is the accepted mechanism to determine whether the feeds are created after visiting the location?
- What is the appropriate mechanism to mine the opinion of the social media feeds created by the users?
- What is the mechanism to rate the opinion over a particular location?
- Which method can be used to obtain visiting patterns of particular location using a density-based approach and identify the RoAs (Region of Attraction)?
- What is the approach to combine social media feed opinion-based rating and tourist visiting patterns to create stories and generate a recommendation?

1.3 Justification for the research

Tourism is one of the main industries in most countries. So, in every country, they try to give the most valuable service to the tourists. Hotel accommodation, tourist's destination maintenance etc. must be in high quality to attract the tourists. The quality of the tourism industry depends on the satisfaction of the tourists. In tourism, recommendation systems are mostly used for business purposes. Those are shown really effective in recommending interests of their users. Most of those systems use prior knowledge about users to identify the position of interests, common user preferences and user interests to create a recommendation [10]. These kinds of recommendations need prior knowledge about the users, otherwise, prediction won't be effective. Nowadays people used to maintain blogs and websites to keep their trip records. They write articles which are visible to other people as well. Those are created according to the writer's opinion; they try to avoid negative impressions in their articles. Visitor reviews, tourism statistics and visitor surveys are statics data sources. Most of the time those data sources are outdated with the time being. So, tourists cannot always rely on those articles. Those are practical issues with the existing recommendation systems. Social media consist of lots of travel records which are created by social media users. Those are more dynamic, created in real time, contains the user's pure opinion and contains location data. So, the motivation for this study is the reliability of using this kind of data source in making a recommendation. There are studies which analyzed tourist digital footprints in social media and examined the tourist behaviors using visiting patterns [1], [5], [6]. Social

media feeds are linguistically rich. Using that value into a recommendation would add not only the tourist's behavior but also the tourist's opinion into the account. To find the tourist's opinion, opinion mining is used. Opinion mining on social media feeds is not straightforward. Because unlike reviews they are always not in the formal language. In social media feeds users use slang words, acronyms, hashtags, emoticons etc. as shown in *Figure 1-2*. This is the main challenges when extracting opinionated text from the

> OMG what a journey… Amazing weather but woorrst location eveeer. Travelling with friends!!! Heading back to home. ☺ ☺ ☹ ☹
>
> #vacation #travel #texas #trippingwithfriends

Figure 1-2: Sample feed

feeds. It requires high effort with preprocessing and subjectivity detection [4]. Another challenge is there are feeds which are created by both tourists and non-tourists. So, a technique needs to be found to identify non-tourists using feeds spatiotemporal sequence [1]. In tourist feeds, opinion is expressed on different aspects. As shown in *Figure 1-2* "amazing weather but the worst location", the weather and the locations are aspects. So, the aspects which hold the opinion is extracted because to create rating it needs to identify the opinion on different aspects. So, user opinion on those aspects will be identified through opinion mining [11].

Therefore, this research is focused on creating an effective recommendation using social media data which takes a combined approach of identifying visitor's opinion and visiting patterns over tourist destinations.

1.4 Methodology

In this study, the feeds are extracted from social media. Then the preprocessing is done. And several steps are followed in order to filter the feeds. Then the main approaches of opinion mining and identifying the tourist visiting patterns are followed. The recommendation is based on this the results of the above two approaches. In the following sections, the dataset of this of this study and the followed approach is briefly described.

1.4.1 Dataset

For this study geotagged feeds from social media are collected. The feeds can be filtered according to several parameters through the search API. Twitter feeds contain more linguistic features and the geolocation data which are important factors in this study. These data are extracted according to their time stamp, in order to collect recent feeds. Collected data objects contain user data, locations, texts etc. as shown in Figure 1-3. This


```
{
  "created_at": "Tue Jan 07
16:57:03 +0000 2014",
  "id": 420599977061396500,
  "id_str":
"420599977061396480",
  "text": "This is the snow I've
been waiting for! Holiday valley
later anyone?",
  "truncated": false,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "user_mentions": [],
    "urls": []
  },
  "source": "<a
href=\"http://twitter.com/downl
oad/iphone\"
rel=\"nofollow\">Twitter for
iPhone</a>",
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str":
null,
  "in_reply_to_user_id": null,
"in_reply_to_user_id_str": null,
"in_reply_to_screen_name": null,
"user": {
    "id": 425468862,
    "id_str": "425468862",
    "name": "yim■o",
    "screen_name": "AyoJimb0",
    ……}
 "geo": {
    "type": "Point",
    "coordinates": [
      42.72795971,
      -78.94105861
    ]
  },
"coordinates": {
    "type": "Point",
    "coordinates": [
      -78.94105861,
      42.72795971
    ]
  },
"place": {
    "id": "94965b2c45386f87",
    "url":
"https://api.twitter.com/1.1/ge
o/id/94965b2c45386f87.json",
    "place_type": "admin",
    "name": "New York",
    "full_name": "New York,
USA",
    "country_code": "US",
    "country": "United States",
    "contained_within": [],
    "bounding_box": {
      "type": "Polygon",
      "coordinates": [
        ……
      ]
    },
    "attributes": {}
  },
  "contributors": null,
  "is_quote_status": false,
  "retweet_count": 0,
}
```


Figure 1-3: Part of extracted feed in JSON format

dataset is unstructured and also there are lots of parameters which are not relevant to this study. So those data need to be carefully filtered. Those feeds contain location and timestamp which are important in examining visiting patterns. So, in this dataset preprocessing steps need to be done to filter out irrelevant data which are not related to this study. The location and timestamp data of feeds are shown in the *Table 1.1*.

Table 1.1: Sample dataset of geotagged feeds

Created time	**User ID**	**Feed**	**Geo-location**
Tue Jan 07 22:40:35 +0000 2014	326382156	Mother Nature please give me some more time to memorize Spark Notes on the Odyssey 😭	[42.94996304, 42.94996304]
Tue Jan 07 21:49:02 +0000 2014	357798883	I just wanna thank God and Mother Nature for helping me to avoid taking this Bio test😭😭😭	[42.74354998, 42.74354998]

Tue Jan 07 16:57:03 +0000 2014	425468862	This is the snow I've been waiting for! Holiday valley later anyone?	[42.72795971, 42.72795971]
Tue Jan 07 03:30:01 +0000 2014	194157721	MAYBE THIS IS A SIGN THAT MOTHER NATURE WANTS TO SEE THE #BUFFALOBLIZZARD BECOME AN @MLS CLUB IN 2014!!! #justsaying	[43.15013776, 43.15013776]
Tue Jan 07 00:54:10 +0000 2014	119161349	Perks of working at hotel, if I get stranded tomorrow, I have a place to sleep 👯	[42.87556381, 42.87556381]
Mon Jan 06 22:24:06 +0000 2014	597795303	Mother nature ain't got shit on my jeep	[42.8475388, 42.8475388]
Mon Jan 06 07:16:30 +0000 2014	278226840	On a Paul Rudd binge. Love that guy #wanderlust #thisis40	[42.96656351, 42.96656351]
Mon Jan 06 04:21:54 +0000 2014	33138618	I don't want to go to sleep. Does any one want to go on an adventure or something with me?	[43.00558983, 43.00558983]
Sat Jan 04 20:33:09 +0000 2014	78197212	I want to travel the world someday	[42.9560489, 42.9560489]
Fri Jan 03 02:42:28 +0000 2014	1323179102	It will be a national holiday the day Jennifer Lawrence and Josh Hutcherson get married. Mark my words, it will be a glorious day.	[42.89614503, 42.89614503]
Thu Jan 02 15:57:11 +0000 2014	142955084	anyone want to crash a hotel, pretend we got a room and go swimming/hot tubbing for a couple hours?	[42.90662542, 42.90662542]
Thu Jan 02 15:18:33 +0000 2014	498836477	always a good start when you're not even packed and you leave for vacation tomorrow ...	[42.87920872, 42.87920872]
Tue Jan 07 20:38:22 +0000 2014	478249634	Travel ban in EA but Bar Bill is open 😮 I need a ski-dooo!	[42.77109636, 42.77109636]

1.4.2 Approach

As the first step, the feeds are extracted from the social media. Here, mainly focus on Twitter because Twitter search API provides easy extraction functionalities of tweets with all required data. Then the feeds are filtered according to their subjectivity, the feeds which are not related to the domain are removed. Only the geotagged feeds are filtered because this study focuses on both geographical and linguistic data. Next

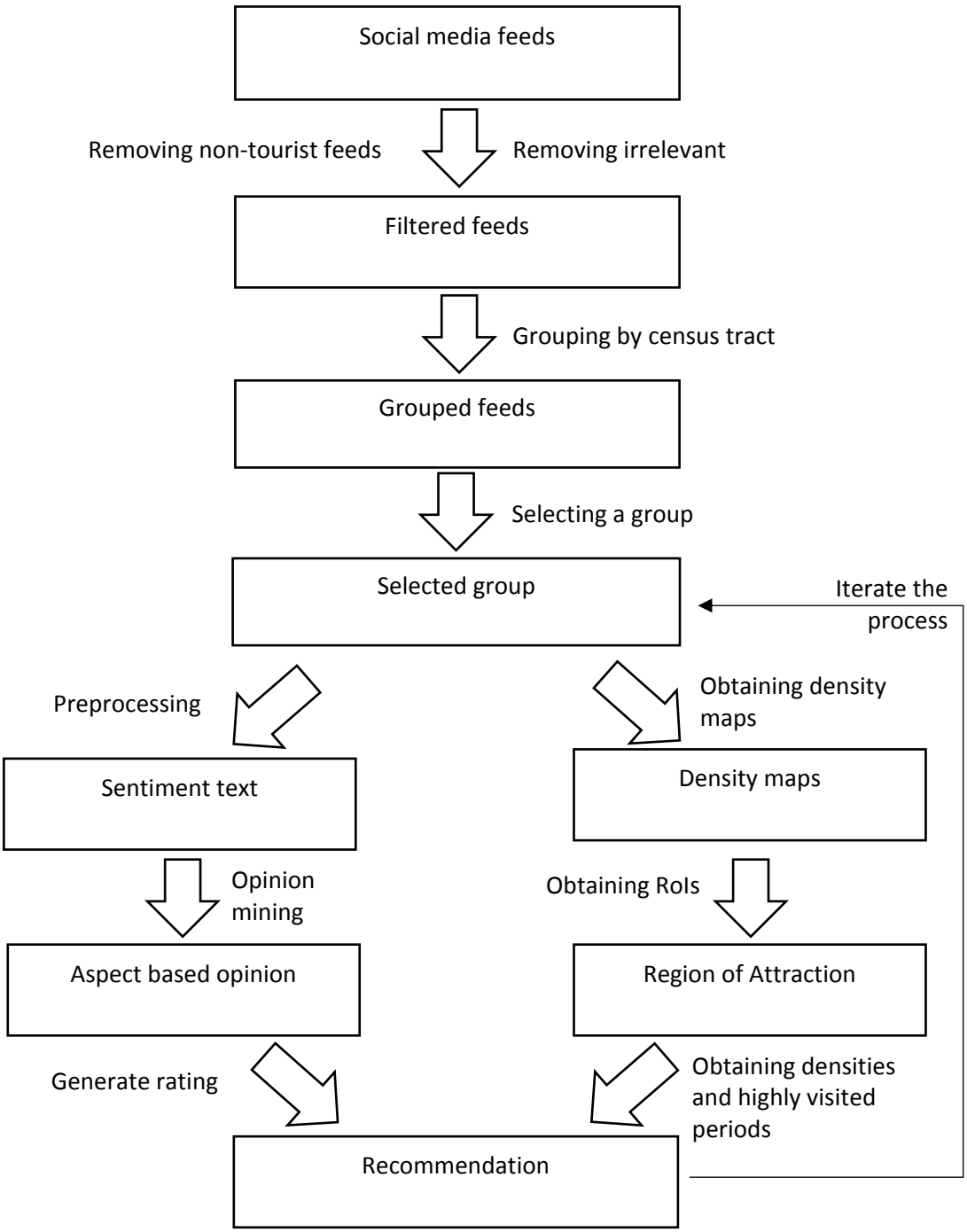


Figure 1-4: Research methodology

identifying the non-tourist feeds and removing those feeds is done using spatiotemporal sequence, in order to reduce biasing the data. Then the filtered feeds are grouped according to their geolocation. Here, the grouping is done in three levels. First, the census tract of the feed is extracted using its geolocation and they are grouped according to the census tract. For the next level the census tracts are grouped according to the city where they are situated. And in the next level, the cities are grouped according to the country. The recommendation is generated according to these three levels. The user-generated feeds in social media are noisy and unstructured. So, as the next step, the preprocessing is done on the feed's text and the opinionated sentiment text is extracted. Using the sentiment text aspect-based opinion mining is done for each and every tweet of each and every location and rating is done according to the given opinion. Also, the density maps are generated using the geolocations of the feeds and RoAs are extracted. So, now the rating over the opinion and the RoAs over a particular census tract is extracted. Then the stories are generated using the tourist opinion over aspects, RoAs and highly visited periods. The high-level diagram of the proposed methodology is given in the *Figure 1-4*.

1.5 Outline of the Dissertation

This dissertation is structured as follows. Chapter 2 - describes the existing approaches, theories and related work in the domain of sentiment analysis and urban analysis. The proposed research design and the approaches, taken to address the problems are described in Chapter 3 - . Chapter 4 - demonstrate the implementation details. Chapter 5 - provides the evaluation model and the results of this study. Chapter 6 - presents the findings of the study. In this Chapter, all the research problems discussed in the first chapter are addressed and the conclusions are given. Also, the areas which need to be further researched and future work are discussed.

1.6 Definitions

More effective recommendation of tourist destinations is proposed in this study. The data source for this study is social media. It is assumed that social media feeds which are created by tourist are more important, reliable and effective in finding visitor opinion and their visiting patterns in order to create a recommendation. This study is proposed

to find a more real-time solution than the existing recommendation systems since this uses social media feeds which are dynamically generated by users rather using static data sources.

User-generated social media feeds are collected and then to generate the recommendation two approaches are combined. Those are creating rating by visitor opinion and obtain tourist visiting patterns using geo-tagged feeds. So, a combination of these approaches; Rating on the destination, Period and Places to visit will be obtained. The main aim of this study is to focus more on visitor's opinion because they are more realistic and richer of their pure satisfaction. And most of the recommendation systems only consider the tourist visiting patterns and PoIs [5]. So, the aim of this study is to add value to social media opinion in tourism studies and experiment its participation in recommending tourist demands.

In this study, literature which is related to the domain are reviewed to identify different approaches, tools and technologies used to do a linguistic study on social media user feed and extract the opinion, and also to identify different approaches, tools and technologies used to obtain visiting patterns using geo-locations of the social media feeds. Finally, finding of this study might add value to the literature of sentiment analysis and urban analytics domain.

1.7 Delimitations of Scope

Since there are restrictions on some social media, some private user data cannot be extracted from the API endpoints. So, some important data will be lost. Some feeds might not be created in that exact location. So, geo-location might give inaccurate results. Thus, extra filtering needed to be done by finding whether the location names which are located in the geotagged area are presented within the text or hashtags. This process also filtered out some feeds since they do not have location names in the text or hashtag even, they are created in the exact location. The same issue arises with the timestamp, the feeds might not be created within the exact time which the location is visited. To tackle this issue the tweets are filtered in a monthly basis (not in a daily basis). All the travelers are not active in social media or they are not always sharing about their travels. Due to this reason, some of the useful opinions will be lost. The opinion might vary according to the visitor's perspective. In social media, the user has the freedom to

post any data that comes to their mind and they can tag any location in their feeds. However, in this study, all the user's opinion is taken into the account in order to make a recommendation. So, an assumption is made that the majority of the users are honest.

1.8 Summary

In this Chapter introduction to this study is given. The background of this study and its justifications are given. The research problem and the research question which are identified to tackle that problem are discussed within this Chapter. In order to answer that research questions, a hypothesis is built and methodology is briefly discussed. The approaches which are taken in this study and solutions for the problems which are faced when taking those approaches are also discussed in this Chapter. Then the outline of this dissertation and the definitions for the study are given. Also, the limitations of the scope are discussed. This Chapter has laid the foundation for the rest of the Chapters in the dissertation.

Chapter 2 - Literature Review

2.1 Theories

Bigdata is a broad term for a data set that is large in size or complex. There are several advantages of tourism big data [7]. Those are, Reliability (Big data are based on user's real actions, not on surveys. So, this data is more reliable since they do not require user effort to input their data. They are not biased since they are based on evidence), New information flows (Useful to analyze the consumer demands for different tourism products and services. The data can be collected from data sources such as social media and open public data) and Real-time data and nowcasting (Uses real-time data. Do not use official data sources).

There are several steps before capturing big data [7]. Those are, Objective (Benefits of collecting data, Visualizing big data (Finding methods which data could be effectively collected from different sources) and Structuring big data (Arranging data according to traditional data length and format).

Big data contains lots of information which creates storage and analysis issues. So, the problem is how to use this large dataset in tourism forecasting. These data need to be structured in order to reduce these issues. There are two methods of selecting and shrinking a large amount of structured data; the factor model and the LASSO method [7].

Social networks can be divided into two groups such as which are designed for social interactions (Facebook), or which are designed for a different service such as content sharing (Flicker) [4]. There are different players on social media; influential players and decile players. The opinions/reviews of influential players are usually getting into account in the analysis. The opinion of those players influences the opinion of others. Clustering techniques of data mining can be used to model opinion of affected nodes and unaffected nodes. Users with the same opinion are mapped under the same node. In opinion mining Support Vector Machine, Naive Bayes and Maximum Entropy were majorly considered in most of the studies which are done to mine opinion from user reviews [12].

When reading the recent literature, there can be found several recent studies related to several approaches of this study. This study mainly carries on according to two main approaches, those are sentiment analysis and urban analysis. Apart from this, in filtering out the non-tourists this study is focused on feeds spatiotemporal sequence-based approach. So, the related literature can be categorized into those subcomponents. The sentiment analysis and the urban analysis is described in the following sections.

2.1.1 Sentiment analysis

In earliest approaches of predicting semantic orientation, adjectives are used. Hatzivassiloglou et al. [13] attempt to predict the orientation of adjectives by analyzing pairs of adjectives (conjoined by and, or, but, either-or, or neither-nor) extracted from a large unlabeled document set. Usually "and" conjoins two adjectives of the same orientation, while "but" conjoins two adjectives of opposite orientation. They extract all conjunctions and adjectives. Conjunctions of adjectives are extracted into train and test sets. The classifier is based on the LOG-linear regression model, which classifies pairs of adjectives either as having the same or as having a different orientation. And they use the train set to train the classifier. Then the classifier is applied to the test set. Then using a clustering algorithm divided the adjectives into two subsets which place as many words of the same orientation as possible into the same subset. Using the intuition of positive adjectives tend to be used more frequently than negative ones the cluster contains a higher average frequency of terms determined as positive.

Liu et al. [14] identified three component of opinion mining; Opinion holder (who gives the opinion), Object (on what the opinion is expressed about) and Opinion (polarity or sentiment on the object). Topic information retrieval is pre-step of opinion mining. The opinion mining is then done on subjective data. There are three types of opinion mining; document level, sentence level and aspect level. In document-level opinion mining, the whole document is considered as one object and it is assumed that each document focuses on a single object and presented by one opinion holder. Then the opinion for the whole document is calculated. In sentence level opinion mining a single sentence is considered. In sentence level opinion mining first determines whether the sentence is subjective or objective. After that extract the opinion from the subjective sentence. Aspect-based opinion mining is used when the document contains opinions about more than one aspect. So, first, it identifies and extracts object aspects from the document and

opinion is determined for each aspect. And tend to produce aspect-based opinion summary using multiple reviews [15].

In the literature, there can be found two main approaches for sentiment analysis [9]. Those are machine learning and lexicon approaches.

a) Machine learning approach

Use classifiers to classify sentiments according to their polarity. They use classification techniques such as Naive Bayes, Maximum Entropy and Support Vector Machines. These classifiers are trained by labeled dataset using supervised learning [9].

b) Lexicon based approach

This relies on a sentiment lexicon, a collection of known and pre-opinionated opinion terms. They use sentiment dictionaries with opinion words and give the polarities to the terms in the text which is needed to be determined the opinion. They again divide the lexicon-based approach into two; Dictionary-based approaches and Corpus-based approaches as shown in *Table 2.1* [9].

Table 2.1: Types of lexicon-based approaches

Dictionary-based approach	**Corpus-based approach**
This uses a predefined dictionary of positive and negative words score the opinion using word counts, indices and frequency etc. In this approach at the beginning, it collects a small set of known opinion words and grows the set by searching in lexical dictionaries (e.g. WordNet) for synonyms and antonyms [1]. Opinion words share the same sentiment as their synonyms while antonyms share the opposite. Qiu et al. [16] tries to find the sentiment sentence in web forum context. They used this method to find the sentiment of adjectives.	The corpus-based method can produce opinion words in relatively high accuracy. This method uses ML too. This method needs a very large set of training data [17]. Here also it starts with very small seed set which includes opinion words which contain the opinion universally irrespective to their context. Then using the words in the seed set it grows the set by choosing words using conjunctions with seed words or their co-occurrence. So, in here the polarity of the words depends on their context or domain in the text [18].

2.1.2 Urban analytics

Urban analytics is the most common analysis followed by lots of researchers. Event detection, venue recommendation for city-scale events, characterizing mobility patterns, characterizing environment and human behavior in cities etc. are the most popular recent

studies in urban analytics. Tourism studies are the main component of urban analytics. Studying the tourist behavior, their visiting patterns are used in tourism studies. For these studies' social media check-ins, geo-locations and users' activities in the internet are vastly used. Most of the studies in these field follow this density-based approach. In this approach, tourist densities are obtained from the user activities. Related work in urban analytics is explained in Section 2.2.2 [5].

2.2 Related work

This study touches two main areas. So, related work can be found in these two main domains. These domains are Sentiment analysis and Tourist density analysis. Within this topic, related work are divided into two sub topics and discussed them separately.

2.2.1 Sentiment analysis

There are several related works can be found in the sentiment analysis domain. These related works are based on two approaches. Those are machine learning approach and lexical based approach as mentioned in Section 2.1.1. The related works on these two approaches are presented in the following sub-points.

a) Machine learning approaches

The study conducted by Pang et al. [8] was based for lots of research studies. In this study, they introduced ML to perform sentiment analysis. They used Naive Bayes, Support Vector Machines and Maximum Entropy on the movie review domain and report accuracies between 77% and 83% depending on the feature set, which included unigrams, bigrams and part-of-speech tagged unigrams. In another study, Pang et al. [19] used subjectivity extract model. They labeled the sentences in the document as either subjective or objective, discarding the later. Then apply a standard machine-learning classifier to the resulting extract to prevent considering irrelevant. They explore subjectivity extract using minimum-cuts. In [20] which also conducted by Pang et al. they attempt to infer the implied numerical rating, such as "three stars" or "four stars". They used One-vs-all, Regression and Matric labeling algorithms for their evaluation.

Applying ML techniques for favourability analyze is discussed by Lane et al. [21]. Favourability analysis is very closely related to sentiment analysis. They have conducted

their study using multiple classifiers. They have used Naïve Bayes, RBF (Radial Basis Function) networks, Bayesian networks, Decision trees. In general, two issues that affect ML approaches are the selection of features and the presence of imbalanced data. They have experimented for two tasks. Pseudo-subjectivity (detecting the presence or absence of favourability.), Pseudo-sentiment (distinguishing between documents with generally positive and negative favourability). The most notable difference between the two tasks, pseudo-subjectivity and pseudo-sentiment, is that the best classifier for the sentiment task was Naive Bayes in every case, whereas the best classifier varies with dataset and feature set for the pseudo subjectivity task. They used five types of features; Unigrams, Bigrams, Trigrams, Entity words and Dependencies.

A study conducted by Esuli et al. [22] also proposed a novel method for determining the orientation of terms. The method relies on the application of semi-supervised learning for the task of classifying whether the terms are either Positive or Negative. The semi-supervised learning is applied to term representations obtained from glosses of a freely available machine-readable dictionary.

a) Lexicon based approaches

Kim et al. [23] researchers selected the sentences containing the opinion words and calculate the polarity of the sentence. First, they create a seed word set and the set is grown using the WordNet using the synonyms and antonyms of the word set. Using the adjectives of words, they score the polarity of opinion words. They identify opinion holder for a sentence. Usually, this will be an organization or a person. Then identify sentiment region which near to the opinion holder. So, when classifying the sentence this sentiment region is used.

Adjectives are used as opinion words by Hu et al. [24] and they used WordNet to predict the polarity and grow the seed set. When predicting the sentence orientation for each opinion word inside the sentence they have calculated the orientation using the constructed seed list. If a sentence contains negation words such as “No”, “Not”, “Yet” etc. the opinion orientation is marked as opposite to its original opinion. In their work, they extracted frequent aspects which present in reviews and extract the opinion on those aspects.

Wang et al. [25] propose a system called TwiInsight to identify the insight from Twitter data. They mainly focused on two facts; Topic extraction, Extract opinion on a specific topic. Topic extraction algorithms they used; Skyttle, Rapid Automatic Keyword Extraction (RAKE), GATE Twitter Part-Of-Speech Tagger. Opinion Extraction algorithms; Stanford CoreNLP-based Algorithm - Uses sentiment dictionary, Haven OnDemand-based Algorithm - Based on positive and negative phases, Monkeylearn-based Algorithm - Use ML to extract relevant data from text.

Agarwal et al. [26] built a model to classify tweets for two classifications; positive, negative and three classifications; positive, negative and neutral. They have experimented with three types of models; unigram model, feature-based model and tree kernel-based model. In their experiment three kernel-based model outer performed other two. In their preprocessing stage, they assigned the emoticons with their polarity using labeled emoticon dictionary and replaced all the acronyms by their translations using dictionary which has translations for the acronyms. They took a dictionary form affected language and grow it using the WordNet and gave polarity scores. So, they took these prior polarities of words and their part of speech tags as the features. They have shown that using these features that feature based model and tree kernel-based model outer performs the Unigram-based model.

Walha et al. [27] adapt opinion analysis which uses the lexicon sentiment analysis method. They have given a sentiment score for user comments using a lexical database composed of emoticons and opinion word dictionaries. The emoticon decorated texts can give insight about the polarity of text. Hogenboom et al. [28] created a framework to automate sentiment analysis. They extract emoticons from the text and determine their sentiment and give the sentiment to the affected text.

2.2.2 Feeds density analysis according to the location

Geo-tagged photos are used to analyzed tourists travel patterns by Zheng et al. [1]. They constructed a photo trail of a spatiotemporal sequence of a photographer, concatenating timestamps on a daily basis. They took RoA by a density-based approach using the number of geotagged photos. They generated a density map using feed's geolocation. They classify the photographer as tourist or non-tourist using the method that photos of a true tourist to be spread over a large spatial extent within the tour.

Salas-Olmedo et al. [6] compare three data sources; Sightseeing (photo-sharing services), Consumption (Foursquare check-ins), Being connected (Twitter). Here the method of analyzing tourist activity data is different from previous work. They have preprocessed the dataset to be counted the number of tourists in each location according to each data source. They have followed this methodology to avoid multiple counting and make the results comparable. The information from different data sources is integrated through cluster analysis and spatial correlation analysis to characterize the areas of tourist concentration according to the type of activity. They have found that some census tracts show a high density of photographs since it offers a good vantage point for taking photographs while some are low because of restrictions of taking photographs. They have also found that a positive correlation is medium in Twitter-Foursquare data sources while it is low in Panoramio-Twitter and Panoramio-Foursquare data sources. There are several steps followed in this research; Number of tourists per census tract (To find the number of single tourists in each census tract for each data source. The raw data was converted to single tourists using joint spatial aggregation), Tourists density by census tract (The number of tourists in each census tract depends on concentration of tourists and size of census tract. The larger census tract tends to be shown a greater number of tourists. So, to avoid this problem tourist density is taken), Rescaling of the data (They have detected a different number of tourists in each data source. So, in order to eliminate the effect of the range and make the density more comparable they have rescaled the data, Descriptive analysis is helpful to compare the degree of concentration of tourists), Density maps and descriptive statistics (Density maps provides an initial visual overview of the density distribution of the tourists), OLS analysis (The coefficient of determination reveals the common part of the variation between each pair of data sources and the differences between each pair of sources can be analyzed using the maps of standardized residuals), Cluster analysis (Cluster analysis was used to integrate the information from three data sources and characterize the census tract according to the tourist activities.), Spatial autocorrelation analysis (Spatial autocorrelation techniques do not consider each location in an isolated way, but the relation to the locations in its environment. Looks for the relation of each activity between census tracts for each data source and results were combined to determine the specialization of each census tract). The cluster was calculated using the k-means algorithm and suing rescaled density values [6].

Dhiratara et al. [5] have compared Instagram data with official tourism statistics and online review data. Their target is to show that social media can provide different reflections of the tourism from other data sources and to identify how social media data provide real-time insights about visiting patterns of tourists in different tourist locations in big events. When filtering relevant SM data there are effective parameters such as Hashtags, Keywords, Geographic boundaries/ Coordinates. Selection of parameters is crucial and need to be carefully designed to avoid biasing interpretation of the results. They have found that social media characterizes tourist locations differently, featuring more landmark locations. In this research, they have developed an alternative method using Instagram's search function due to the change of Instagram's platform policy for collecting data through API endpoint. They used location tags and the timestamp of the Instagram feed to get tourist population of each location. To compare data from different data sources they used parametric rank correlation namely Kendall's 7 rank correlation based on pairwise agreements between tourist locations [5].

2.3 Summary

This study is focused on proposing an effective technique to create a recommendation over tourist destinations by following two main approaches. The two approached followed here are opinion mining on social media feed and obtain visiting patterns using social media feeds geo locations. So, in order to lay a background to this approach, it is referred to two domains in the literature. Those two domains are Sentiment analysis and Urban analytics. For sentiment analysis, there can be found two approaches in the literature those are ML Based approach and Lexicon based approach. The lexicon-based approach again can be done in two different techniques; Dictionary-based approach and Corpus-based approach. To obtain visiting patterns Urban analytics-based literature followed a density-based approach by finding RoAs and tourist flows. The challenging part is determining non-tourist feeds. To address that issue the related works created a spatiotemporal sequence-based approach which relies on the fact that true tourist locations spread within a large spatial extent within their trail. The related works on the above approaches are discussed in this Chapter.

Chapter 3 - Design

3.1 Introduction

The research design caters for three challenges. Those are Opinion mining on social media feeds, obtaining tourist visiting patterns, and generating the recommendation. Subjectivity detection, filtering non-tourists, preprocessing and verifying the feeds

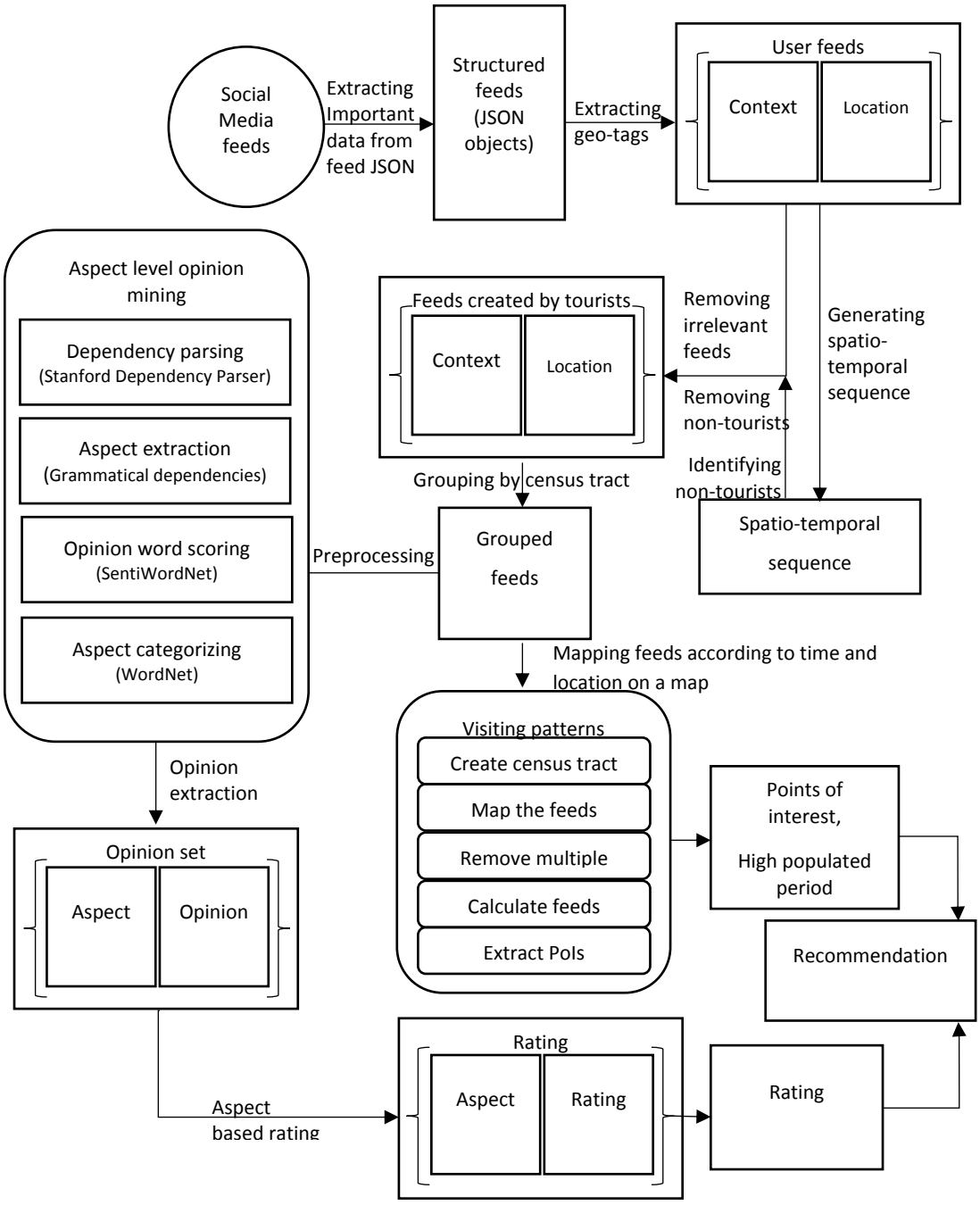


Figure 3-1: Flow of the research design

geotags and timestamps needed to be addressed within the design. For those tasks, the proposed design is presented in the following sections. The overall project design is presented in *Figure 3-1*.

3.2 Extracting data

There need to be effective parameters to extract the useful information from a large dataset. So, Hashtags, Keywords, Geographic boundaries/ Coordinates are effective parameters that can be used. Selection of parameters is crucial and need to be carefully designed [5]. Using the twitter search API by providing search queries feeds can be extracted. In this study several travel and tourism related keywords used in order to extract the tweets. The keywords used in this study are "travel", "holiday", "vacation", "traveling", "TTOT", "wanderlust", "RTW", "backpacking", "adventure", "hotel", "nature" [29]. When the search query is run for defined keywords, API finds for the feeds which contain the defined keywords. But search API gives the real-time feeds which are created in recent time, which means in order to collect all the feeds generated within a month, the API needs to be run continuously. Feeds cannot be filtered according to the time through the API. In search API the requests per an IP are limited in order to reduce the traffic. By considering these reasons, in the system implemented in this study, the IDs of the tweets which are generated within a day are extracted. And then by giving the tweet ID, the feed JSON is extracted using the search API. If a recommendation is generated for a particular location, generalizing the process over other locations is straightforward. So, in this study, the tweets which are created since 2014 January to 2018 December around the location of Niagara Falls are extracted. Extracted feeds are preprocessed in order to remove irrelevant feeds and find the opinionate text. Then the census tract for each feed geolocation is found and grouped the feeds according to the census tract.

3.3 Identifying non-tourists

Feeds created by tourists can be differentiated from those created by residents based on the time period and location which some users created the feeds [5]. There are two methods to identify the non-tourists in the system. In the first method, user location is extracted from the feed and if the user location is situated close to the feed's geo-location

then the user is identified as a resident and all user feeds are removed from the particular census tract. In the second method, using feeds geo-location in its census tract, the spatiotemporal sequence is created for each user over the timestamp. If a particular user has a large number of tweets within the same area within the year, he is identified as a non-tourist, because only a non-tourist such as resident or an organization would create multiple feeds during the year within same census tract using the same user account [1]. A true tourist's feeds would lie within large spatial extent during some period of the year (during his vacation in that area). Then the feeds identified as non-tourists are removed. The steps described above are summarized in the following points.

1) Get all feeds in the census tract.
2) Group the feeds by the user.
3) Identifying non-tourist
    a. If the user location is situated within the census tract, identified as non-tourist.
    b. If the user's feeds timestamps are distributed more often within the same year, identified as non-tourist.
4) Removing all the feeds within the census tract which are created by the identified non-tourists.

3.4 Eliminating irrelevant feeds

The feeds which are not related to the domain need to be removed. Otherwise, there is no purpose of using feeds which are not about tourism to create a recommendation. When a feed is related to a particular domain, it is assumed that the user tends to use words which are related to that domain. Using that assumption feeds are identified whether they are related to the domain or not. So, in this study existing review dataset which is related to the tourism is taken and find for the most frequently used words within that dataset [30]. The word frequencies are obtained using Tf-idf vectorizer.

Then these high frequent words are extracted as keywords and those word are used to filter out feeds extracted for this study. To do so, nouns and verbs which have been used in a particular social media feed are taken, for each of that word, synsets are taken using

WordNet. For each word in synset if that word present in keywords then the feed is identified as relevant word as shown in the *Eq(1)*.

$$\forall x(\in E)[(synset(x) \cap K \neq \emptyset) \rightarrow q] \quad \text{Eq(1)}$$

$where,$
$E = Entity\ words$
$K = Keywords$
$q \rightarrow Feed\ containing\ E\ is\ relavent\ to\ the\ domain$

Another problem is removing the feeds which aren't created at the exact location. Because of this reason, the feed's geo-location might inaccurate with the exact destination's location which might reduce the reliability of the results. To eliminate this issue, Named Entities and the Hashtags in the text are used. If a location name contains within the Named Entity or a Hashtag, the geo-location of that location is taken. If the feed's geo-location is not located within close proximity (defined by a threshold) to the previously extracted location, then the feed is identified as inaccurate feed. So, the feed is removed from the dataset.

3.4.1 Term frequency inverse document frequency (Tf-idf) vectorizer

Tf-idf vectorizer gives a value to the words in a document of a corpus according to the importance of the word. Term frequency is the number of times a term occurs within a document. Usually, terms like "the" occur more frequently. So, they get high weights. Because of this reason important but non-common terms which represent the characteristics of the documents get low weights. Using the inverse document frequency this issue is fixed. Using the term frequency inverse document frequency weight of the rare but important terms get high weights and common terms get low weights [31].

3.4.2 WordNet

WordNet is a freely available lexical database which contains synsets. The current version of the WordNet is WordNet 3.0. Synsets are the synonyms grouped and linked together according to their lexical relationships of the meaning. The synonyms are grouped according to specific senses of the words. In WordNet words which are found in close proximity are semantically related. The synsets are encoded according to several relations in wordnet. In super-subordinate relations, hypernyms are the synsets that are

more general and the hyponyms are the synsets that are more specific. Hyponyms have an “is-a” relationship to their hypernyms. Also, there are “is-made-of” and “comprises” relationships. holonyms are things that the item is contained in. Meronyms are components or substances that make up the item. Using WordNet, we can get similarities among synsets [32]. WordNet represents these synsets in a hierarchical structure. In WordNet similarities among synsets can be calculated. Using these measures, it can be calculated how similar two words. There are several similarity measures can be calculated, Path similarity (compute the number of edges from of the shortest path from one sense to another sense), Leacock Chodorow Similarity/LCH (uses the negative log algorithm to the measure of path similarity), Wu Palmer Similarity/WUP (similar to the LCH but assigns weights to the edges based on the distance in the hierarchy) [33].

3.5 Preprocessing

The social media data is noisy. So, it contains lots of unnecessary data. This study is based on aspect level sentiment analysis. So, the text in the feeds needs to be preprocessed to extract the aspect and their opinion. In the preprocessing, several steps are followed in order to obtain the sentiment text. First of all, unnecessary characters are removed from the text, such as unidecode characters, escape characters, HTML escape characters and HTML tags. Then the contractions in the text are fixed. The characters which are not letters, numerical characters or punctuation marks are removed. Those are the basic text preprocessing steps carried out in this study. Now there are several social media specific entities in the feed text. Those are URLs, usernames, hashtags, emoticons, repeated characters and slang words/acronyms etc. Even though hashtags and emoticons are also important features when extracting the opinion, since this study focus on aspect-based sentiment analysis, it is hard to identify their targeted aspects. So, these entities are removed from the text. Also, the punctuation marks are not important in mining the opinion, but in this study word and sentence boundaries needed to be identified in order to extract the aspects. So, punctuation marks are not removed in the preprocessing steps. Repeated character, acronyms/slang words are different usage of the language. They needed to be taken into consideration when mining the opinion. So, they must be converted into a readable format. By repeated characters it means, the words such as “woooorst”, “cooool”, “booooringg” etc. Those kinds of words do not make any sense. So, they are replaced with the correct form. Ex: “worst”, “cool”,

"boring". The acronyms are words such as "LOL", "OMG" etc. those words needed to be replaced with the meanings. Ex: "Laughing Out Loud", "Oh My God". This is done using a pre-compiled dictionary of popular slang words in social media [26]. The preprocessing steps which have been discussed above can be summarized as follows.

1) Removing unidecode characters, HTML escape characters, HTML tags and escape characters.
2) Fixing contractions.
3) Removing URLs, usernames, hashtags, emoticons.
4) Fixing repeated characters.

3.6 Opinion mining

After these pre-processing steps, the sentient text is extracted and fed to the opinion mining model. There are three opinion mining techniques identified in the literature; document level, aspect level and sentence level [15]. As discussed in Chapter 1 - , even though social media feed contains a limited number of characters, they might contain several opinions on different aspects as seen in *Figure 1-2*. In document level and sentence level sentiment analysis whole content in a document or a sentence is given a single polarity according to identified features in the text. But aspect level sentiment

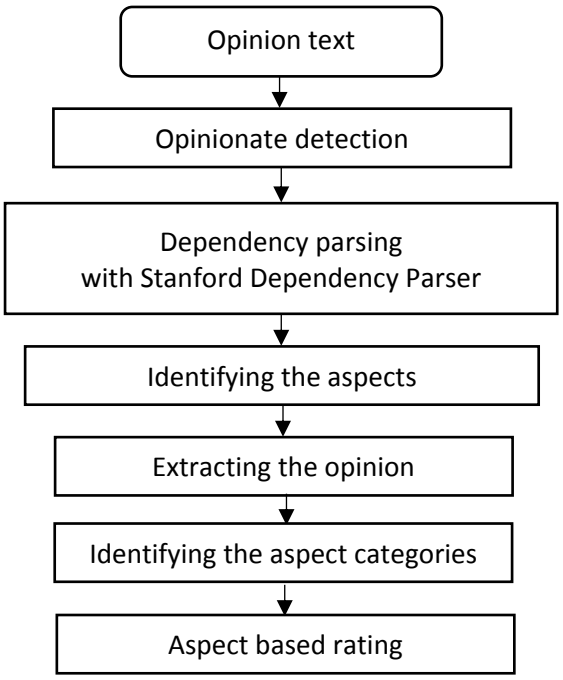


Figure 3-2: Aspect-based opinion mining

level is different from those two models. In aspect level opinion mining the aspects addressed in each sentence are extracted and opinion on each aspect is obtained. So, this method is chosen as the most appropriate model for this study.

Now the feeds are grouped according to the census tract, filtering and preprocessing is done. From each census tract, each and every feed is taken. First, it needed to identify the aspects presented in the text. In order to do so, the grammatical representations among the terms in the sentiment text are used. To identify aspects several grammatical dependency rules are defined and the terms which follow those rules are identified as the aspects [11] [24]. The terms which holds the relationship according to the rule are extracted as the opinion words for that aspect as shown in *Figure 3-2*. This process is presented in the *Eq(2)*. When extracting grammatical relationships Stanford dependency parser is applied to extract the parse tree with the dependencies and the relationships [2].

$$\{x, y \in W \mid (r(x, y) \in R) \rightarrow (x \in A \,\&\, y \in O_x)\} \qquad \text{Eq(2)}$$

$where,$
$R = Rules$
$W = Words\ in\ the\ text$
$A = Aspects$
$O_x = Opinion\ words\ of\ aspect\ x$

The grammatical rules which are used for aspect and opinion word extraction can be defined as following steps.

1) If the term has a clausal complement with the internal subject or a clausal complement with the external subject and the term has a sentiment, the complement is an aspect. If the complement has a nominal subject or a nominal modifier, that noun is an aspect. The term is an opinion word of the aspect.
2) If a verb has an adverbial modifier and the verb has a direct object, nominal modifier or nominal subject the noun is an aspect, if the adverb has a sentiment then the adverb is an opinion word and if the verb has a sentiment then the verb is an opinion word.
3) If a noun has an adjectival modifier and the modifier has a sentiment, then the noun is an aspect and the adjective is an opinion word.
4) If a term has a relationship with auxiliary or copular verb and the terms have a sentiment and nominal subject or a modifier, then the noun is an aspect and the term is the opinion word for the aspect.
5) If an extracted opinion word has a negation then the opinion word is negated.

Another possible way of extracting the aspects is identifying the frequent terms within the feeds. Association Rule mining is used in this approach, where all the nouns within feeds in a census tract are taken as noun transactions. Then identify the frequent association rules in order to identify frequent noun combinations [11]. But this method requires a high amount of transactions which need more feeds. But in this study, the aspects are extracted on a monthly basis. So, the feeds for a census tract are limited. So, this approach is not applicable in this situation.

To extract the opinion from the opinion words extracted in the previous step, the lexicon-based approach is applied using the SentiWordNet. According to the opinion word's sentiment score, overall sentiment score for the aspect is given as shown in *Eq(3)*.

$$\{x, y, z \in A \mid \forall x \exists y \forall z[(sim\_score(x, y) < sim\_score(x, z)) \rightarrow (\{x, y\} \subseteq C^1)]\} \quad \text{Eq(3)}$$

$where,$
$A = Aspects$
$C^1 = Cluster$

Now the extracted aspects are clustered into groups in order to categorize the aspect. The average aspect scores within each category are assigned to the category. When clustering the aspects, the similarities among the aspects are considered. More similar aspects are clustered together as shown in the *Eq(4)*.

$$S_a = \sum_{i=0}^{n} \left(pos_{score(O_i)} - neg_{score(O_i)}\right) \quad \text{Eq(4)}$$

$where,$
$S_a = Sentiment\ score\ of\ aspect\ a$
$O_i = i^{th}\ opinion\ word\ of\ aspect\ a$

3.6.1 Stanford dependency parser

Stanford dependency parser is a tool which provides the grammatical dependencies between words in a sentence. The grammatical relations provided by the parser are binary relationships, which holds between the governor and the dependent. There are approximately 50 grammatical relations in the representation. In the results, the dependency is represented as *relation_name(governor, dependent)*. There are five types of dependency representations available in the parser. These five types are Basic-dependencies (each word in the sentence are dependent on exactly another term), Collapsed dependencies (prepositions, conjuncts and relative clauses are collapsed),

Collapsed dependencies with propagation of conjunct dependencies (propagation of the dependencies can be extracted from the conjunctions), Collapsed dependencies (dependencies which do not preserve the tree structure are omitted), Non-collapsed dependencies (includes basic dependencies as well as the extra ones, without any collapsing or propagation of conjuncts) [2].

3.6.2 SentiWordNet

SentiWordNet is a sentiment dictionary which based on WordNet. The current version of SentiWordNet 3.0 is based on WordNet 3.0. SentiWordNet gives a positive, negative and objective score for the synsets in the WordNet. For each and every synset summation of the positive, negative and objective scores are equal to 1. These scores are given b automatically annotating the WordNet synsets according to their degree of positivity, negativity and neutrality. In SentiWordNet 1.0 the automatically annotating was done by weak-supervision, semi-supervised learning algorithm. But in the SentiWordNet 3.0 authors have included random-walk step to refine the scores. This lexical resource is publicly available and can be used in sentiment classification and opinion mining applications [34].

3.7 Obtaining the tourist visiting patterns

Geotags and the timestamps are obtained from the feeds. Then they are mapped on the map in order to get a visual representation. The multiple feeds which are created by the same user are removed to avoid multiple counting. Feed densities needed to be taken in order to identify the tourist flows [6]. RoAs are taken for each and every census tract. This process is presented in *Figure 3-3*. As mentioned in Chapter 1 - , density representations are done in three levels, country, city and census tract level. When taking the RoA the tourist density distribution is considered. The point which have the highest

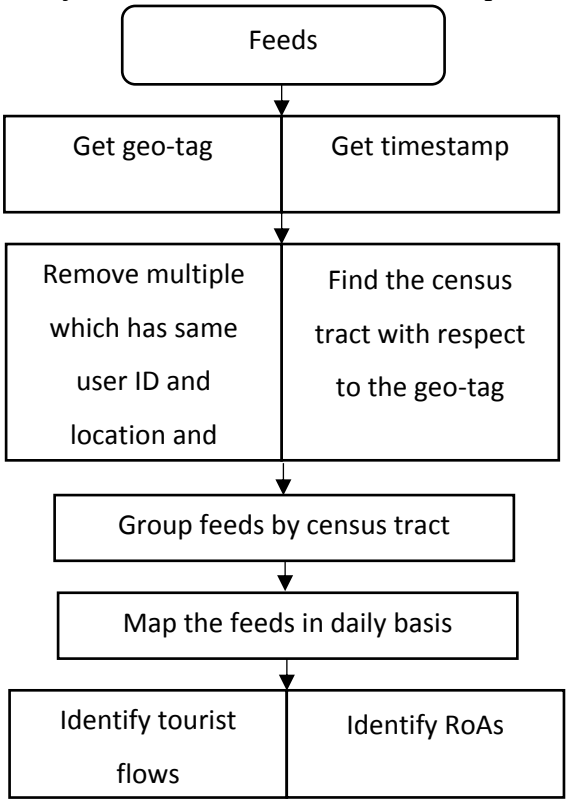


Figure 3-3: Obtaining the tourist visiting patterns

number of tourists within the closest proximity are taken as the RoA. First of all, the points are grouped according to the closest proximity, then the area which have the highest number of tourists is taken as the RoA as shown in the *Eq(5)*.

$$\{x, y \in P \,\&\, A^1 \in A \mid \forall x \exists y[(dintance(x, y) < T) \rightarrow (\{x, y\} \subseteq A^1)]\} \qquad \text{Eq(5)}$$

$$RoA = Maximum\ sized\ set\ in\ A$$

$where,$

$P = Points$

$A = Closest\ proximities$

3.8 Generating the recommendation

Now, the opinion scores for the identified aspect categories are obtained. Also, the RoAs within the census tracts is extracted. The feeds are grouped by the timestamp. In here, the tweets are grouped in monthly basis. Because there is a possibility where some feeds are not created in the exact time which the place is visited. Using the above details, stories can be generated about a particular census tract. The highest scored aspect category is recommended as the aspect which is most of the tourists are interested in that particular area. The feed densities within each and every period are taken and the most visited period is obtained. The RoA is suggested as the most popular point within a census tract. The recommendation is generated for the three levels which have been mentioned early, country, city and census tract. The process is generalized from the census tract level to higher levels. Finally, the recommendation is generated with the visual information with most popular area, visitor's opinion on popular aspects on that particular location and the highly visited period of a particular location. So, by this process visitor's opinion and the tourist densities can be specified according to the seasons within the year.

3.9 Summary

This Chapter provides a detailed description of the research design including the subcomponents of the system. The research design consists of 3 main components, which are opinion mining and obtaining visiting patterns. Finally, these 2 components combined to generate the recommendation. Extracting dataset and preparation of the dataset for the two components is descriptively provided within this Chapter. These preparation steps include identifying non-tourists, eliminating irrelevant feeds and preprocessing.

Chapter 4 - Implementation

4.1 Introduction

This Chapter describes the implementation details of the proposed solution in this study. The implementations of the proposed design in Chapter 3 - are addressed here. In the first part of this Chapter, software tools used in this study are described. Then implementation details of data preparation, removing irrelevant, non-tourist detection are presented. Then the implementation details of the main components of the study, which are opinion mining and obtaining density maps are described and in the latter part generation of the recommendation is presented.

4.2 Software Tools

Analysis part of this study is implemented using python 3.0. Twitter Search API is used for data collection. Packages from python nltk library are used in preprocessing the dataset. In eliminating irrelevant feeds, TfidfVectorizer from scikit-learn library is used to identify frequent keywords. Python geopandas and geopy packages are used in geography-based operations. Then, Stanford Dependency parser is used to extract grammatical dependencies from a given text. In scoring the opinion words, SentiWordNet lexical dictionary is used. Similarity scoring between words is carried out using WordNet. To present the obtained results, a web application is implemented using Angular 6. Angular modules are used for the implementation of the web application. Google Maps are used for the geographical representation of the feeds. Representation of the opinion scores in charts is done using the primeng charts in Angular.

4.3 Collecting the dataset

For the dataset, the Twitter feeds are collected through Twitter Search API. As mentioned in Chapter 3 - , the tweet can be searched using the search queries by providing the keywords to be filtered. The keywords used for the filter are “travel”, “holiday”, “vacation”, “traveling”, “TTOT”, “wanderlust”, “RTW”, “backpacking”, “adventure”, “hotel”, “nature” [29]. To eliminate the request limitation first the tweets

are searched for these keywords for a particular period and then the tweets are extracted. using the tweet ID via search API as shown in *Figure 4-1*.

```
filters = [“travel”, “holiday”, “vacation”, “traveling”, “TTOT”, “wanderlust”, “RTW”, “backpacking”, “adventure”, “hotel”, “nature”]
tweets = [ ]
tweet_ids = search_tweets(since=yesterday, until=today, query=filters)
for tweet_id in tweet_ids:
        tweet = get_tweet(tweet_id)
        if (tweet is not retweet) and (tweet is not reply) and (tweet has geo_tag):
                tweets.append(tweet)
```

Figure 4-1: Extracting tweets

4.4 Grouping the feeds according to the census tract

The preprocessed feeds are grouped according to their census tract. To this task, the geotag is taken from the tweet JSON. The census tract of the corresponding location

```
grouped_feeds = { }
for feed in feeds:
        address = get_address(latitude= latitude, longitude= longitude)
        census_tract = address[“post_code”]
        city = address[“city”]
        country = address[“country”]
        key = cenus_tract+”_”+city+”_”+country
        if not  key in grouped_feeds:
                grouped_feeds[key] = [ ]
        grouped_feeds[cenus_tract+”_”+city+”_”+country].append(feed)
```

Figure 4-2: Grouping the feeds by census tract

coordinates is extracted through GeoPy API and the feeds which are having same census tract are grouped together. Through this API the address of the corresponding location can be obtained. So, the locations can be grouped according to the census tract, city and

country as demonstrated in *Figure 4-2*. First, all the filtered feeds are taken, then by finding the census tract, city and the country of each feed, they are grouped.

4.5 Identifying the non-tourists

There are two methods if the user's location is in close proximity to the census tract location then that user is a non-tourist. Or all the feeds created by each user are taken and group them on a daily basis. The users who have more feed distribution around the same location during the whole year are identified as non-tourists and they are removed from the dataset. This implementation is shown in *Figure 4-3*. In this implementation for each group of census tract, all the feeds are taken. Then by looping through each feed, the feeds are grouped according to the user in order to identify and remove non-tourists. For each tweet the user's location is compared with the feed location if these two locations are close in distance then the user is identified as a non-tourist. In other timestamps of all the feeds of a particular user are taken and then these timestamps are grouped by the year and non-tourists are identified by time difference among tweets within the year.

```
for census_tract, census_feeds in grouped_feeds.items()
        user_feeds = { }
        for feed in census_feeds:
                user = feed[“user”]
                user_id = user[“user_id”]
                if not  user_id  in grouped_feeds:
                        user_feeds [user_id] = {“user”: user, “feeds”: []}
                user_feeds [user_id][“feeds”].append(feed)
        for key, feeds in user_feeds.items():
                user_location = feeds[“user”][“user_location”]
                timestamps = {}
                for feed in feeds[“feeds”]:
                        location = feed[“location”]
                        date = feed[“datetime”]
                        if not date[“year”] in timestamps:
                                timestamps[date[“year”]] = []
                        timestamps[date[“year”]].append(date)
                        if is_closer(user_location, location):
                                delete user_feeds[key]
                                break
                if year, dates in timestams.items():
                        dates = sort(dates)
                        date_diff = dates[-1] – dates[0]
                        if date_diff > 2months and date_diff < 1year:
                                delete user_feeds[key]
                                break
```

Figure 4-3: Identifying non-tourists

4.6 Eliminating irrelevant feeds

For this step, the word similarities are used. As mentioned in Chapter 3 - the terms within the feed's text are compared with the terms in the predefined keywords as shown in *Figure 4-4*. If a feed does not hold any of the similarity with the any of the keywords then the feed is identified as irrelevant. In order to remove the feeds which are not created in the exact location, the Named Entities and the Hashtags within the feeds text are used as mentioned in Chapter 3 - . If that entity is a location and it is closer to the feed's location, then the feed is taken as a relevant feed. Otherwise, it is irrelevant and it is removed as shown in *Figure 4-4*. In this implementation for each tweet all the entity words are taken, then for each and every entity word synset is taken and synset is compared with the defined keyword sets. If there's no similar word exist for any of the

```
for feed in feeds:
        relevant = false
        text = feed["text"]
        hashtags = feed["hashtags"]
        entity_words = get_nouns_and_verbs(text) + hashtags
        for word in entity_words:
                synset = get_synset(word)
                if synset ∩ keywords ≠ ∅:
                        relevant = true
                        break
        feed_location = feed["location"]
        named_entities = get_namd_entities(text) + hashtags
        for entity in named_entities:
                location = get_location(entity)
                if is_closer(feed_location, loction):
                        relevant = true
                        break
        if not relevant:
                delete feed
```

Figure 4-4: Removing irrelevant

entity word then feed is removed. In the other step the feeds Named Entities are taken and from those entities, locations are extracted and by their distance to the feed location irrelevant feeds are identified.

4.7 Preprocessing

In preprocessing unnecessary characters from the text are removed and clear text is obtained. Also removing unwanted entities, fixing contractions, replacing repeats and slang words/acronyms is done. To remove the unnecessary characters, unwanted entities, and to replace repeats simple regular expression replace is done. To replace slang words/acronyms, they are matched with a precompiled acronym dictionary.

4.8 Opinion mining

To extract aspect, the text which is taken from the preprocessing stage and it is parsed through the Stanford dependency parser [2] in order to obtain the relationships of the terms within the sentences. Then the aspects are extracted from the text. In identifying the aspects several rules are applied the terms which satisfy those rules are taken as the aspects. And the term which holds that rule with the aspect is taken the opinion word for that aspect as shown in *Figure 4-5*. If an opinion word has a negative modifier then the score is negated.

Then the identified aspects are clustered according to the similarity. By this method, the similar aspects are likely to cluttered in same category as shown in Figure 4-6. The similarity matrix is taken for all the aspect and then according for each aspect most similar aspect is taken and then they are appended to the same group. The similarities are taken using the WordNet synset similarities.

```
aspects = { }
if open_clausal_complement(word, term) or clausal_complement(word, term) and has_sentiment(word):
        if aspect = nominal_subject(term) or aspect = nominal_modifier(term):
                aspects[aspect] = [is_neg(term)? neg(term): term]
If is_verb(word) and (adverb = adverbial_ modifier(word)):
        If aspect = direct_object(word) or aspect = nominal_subject(term) or aspect = nominal_modifier(term):
                if has_sentiment(adverb):
                        aspects[aspect] = [is_neg(adverb)? neg(adverb): adverb]
                if has_sentiment(word):
                        aspects[aspect] = [is_neg(verb)? neg(verb): verb]
If is_noun(word) and (adjective = adjectival_ modifier(word)):
                if has_sentiment(adjective):
                        aspects[word] = [is_neg(adjective)? neg(adjective): adjective]
if aux(word, term) or cop(word, term) and has_sentiment(word):
        if aspect = nominal_subject(word) or aspect = nominal_modifier(word):
                aspects[aspect] = [is_neg(word)? neg(word): word]
```

Figure 4-5: Aspect extraction

Then for each aspect in aspect category sentiment score is taken and the score is averaged in order to take overall sentiment score for each category as shown in the Figure 4-7 for the negated opinion words score is negated.

```
sim_matrix = get_similarity_matrix(aspects)
categories = [ ]
deleted_indexes = [ ]
function get_cetegories(i)
        deleted_indexes.append(i)
        items = [ ]
        if i not in deleted_indexes:
                min_index = get_index_of_min(sim_matrx[i])
                items = [aspects[min_index]]
                for item in items:
                        items = items + get_cetegories(min_index)
        return items
for i in 0:len(aspects):
        if i not in deleted_indexes:
                category = [aspects[i]] + get_cetegories(i)
                categories.append(category)
```

Figure 4-7: Aspect categorization

```
category_scores = { }
for category in categories:
        category_score = 0
        for aspect in aspects:
                aspect_score = 0
                for word in aspect[“opinion_words”]:
                        score = get_sentiment_score(word)
                        aspect_score += neg(word)? (score*(-1)): score
                category_score += aspect_score/len(aspect[“opinion_words”])
        category_scores[category] = category_score/len(aspects)
```

Figure 4-6: Scoring aspects

4.9 Obtaining tourist densities and RoAs

This process is implemented in the web app. The grouped feeds according to the census tract are taken. The tourist densities are taken using the tourist counts of each location. To identify the tourist densities according to the period the feeds are grouped according to the season as mentioned in Chapter 3 - . So highly visited periods and densities can be obtained. For each census tract RoAs are identified. When identifying the RoAs the points within the close proximity are clustered together. The cluster which has the highest tourist count is taken as the RoA in that particular census tract as demonstrated in *Figure 4-8*. In here, first for each point closest point are identified using a defined threshold. Then for each and every point, their closest points are clustered in to one cluster.

```
distance_matrix = get_distance_matrix(points)
point_groups = [ ]
deleted_points = [ ]
function get_groups(i)
        deleted_points.append(i)
        items = [ ]
        min_indexes = [ ]
        if i not in deleted_points
                for j in 0:len(points):
                        if get_distance(points[i],points[j])<T:
                                items = items □ [point]
                                min_indexes = indexes □ [j]
                for index in min_indexes:
                        items = items □ get_ groups(index)
        return items
for i in points:
        if i not in deleted_points:
                point_groups.append([points[i]] + get_ groups(i))
```

Figure 4-8: Getting RoAs in a census tract

4.10 Generating the recommendation

Since all the required data are now extracted generating the recommendation is straightforward. Now the previous implementations provide the tourist opinion over as aspects and RoAs within that census tract. So, stories are generated using this information. Then that story is presented as the recommendation as mentioned in Chapter 3 - . For visual representation, the web application is used and the results are presented as charts and maps. The aspect rating over the period is visualized using charts. The RoAs are visualized using Google maps. To visualize the populated areas heat maps are used. So, overall recommendation is a representation of these visualized charts, population maps and heat maps. Using the opinion scores and the tourist densities.

4.11 Summary

In this Chapter, overall system implementation is given. In the software tools Section, software tools and the functions are discussed. In the later Sections all the implementation details described with the flow proposed in Chapter 3 - . Data collection, data filtering, preprocessing, opinion mining, obtaining tourist densities and RoAs and finally generating the recommendation is presented with pseudo-code segments within these sections.

Chapter 5 - Results and Evaluation

5.1 Introduction

The results and the evaluation model are elaborated within this Chapter. In the rest of the Sections, the proposed evaluation model and obtained results will be presented. There is no overall evaluation model to test the whole system propped within this study. There are mainly two components needed to be evaluated. First, in opinion mining model. And the second is comparing the correctness of the results obtained from extracting locations which have high interests of tourists. As mentioned in Chapter 3 - , this study uses the tweets which are created since 2014 January to 2018 December around the location of Niagara Falls as the dataset for the evaluation. Even though past feeds are extracted, the dataset is limited. The dataset contains 9886 feeds. It is because when searching for past feeds, Twitter gives results of a limited number of feeds. So, these limited number of feeds represents the whole population.

5.2 Evaluation model

Evaluation of this study is done for two main approaches. To evaluate the accuracy of the opinion mining model confusion matrix measures are taken. The confusion matrix is shown in *Figure 5-1*. A manually annotated sentiment datasets in different domains which are extracted from Kaggle data repositories are used for the evaluation [35] [36] [37]. These datasets have tweets and they are annotated according to the sentiment of the overall feed. The annotated dataset is processed through the proposed opinion mining model. This opinion mining model gives sentiment over the aspects. So, to get overall sentiment of the feed, the aspect scores are averaged. Then the obtained results are compared using the confusion matrix. The precision, Recall and accuracy measures are presented in *Eq(6) - Eq(9)*

$$Precision = \frac{TP}{TP + FP} \quad \text{Eq(6)}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{Eq(7)}$$

$$FMeasure = 2 \times \frac{Precision \times Recall}{Rrecision + Recall} \quad \text{Eq(8)}$$

Eq(9)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

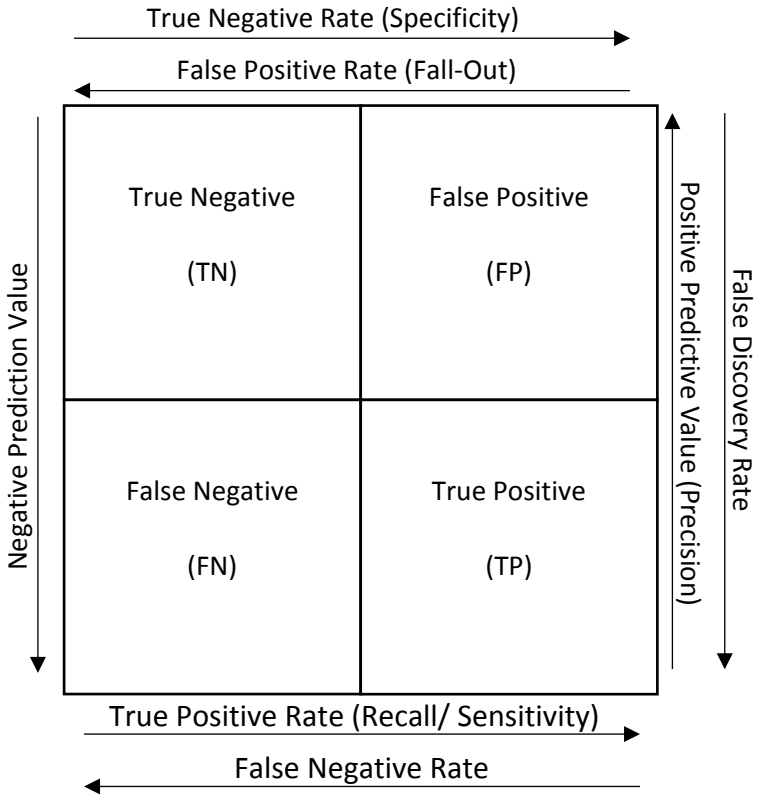


Figure 5-1: Confusion matrix representation

In the system RoAs are taken for each and every location using the feeds. So, to evaluate this model the heatmaps and the algorithmically identified RoA points are compared. The heatmaps are generated using the feeds of the particular area. If the heatmaps shows peak location within the map, then it has to be identified as a point which have highest density.

5.3 Results

Using the described data extraction mechanism twitter feeds around Niagara Falls from 2014 January to 2018 December are extracted. The feeds within the dataset contain the parameters such as feed ID, text, hashtags, geo-location, user data etc. which are considered in this study. The data are extracted in JSON format as shown in Section 1.4.1. Then in order to remove the unnecessary feeds the data are filtered and preprocessed. After these steps, the sentiment text is obtained as shown in *Table 5.1*.

Table 5.1: Obtained sentiment text after filtering and preprocessing the feed text

Feed Text	**Preprocessed Text**
At our holiday party. #BuffaloNY @ Canadian/USA Border (Buffalo. NY). https://t.co/8qcUjRcs3y	At our holiday party . canadianusa border buffalo . ny .
A little outer harbor adventure this evening #buffalo @ Buffalo Lighthouse https://t.co/6IL5aHkZkp	A little outer harbor adventure this evening Buffalo lighthouse
ft. me, alli, and manal on our vacation to Niagara Falls @ Niagara... https://t.co/fv17vOisum	ft. me alli and manal on our vacation to Niagara Falls Niagara .
The hotel I am staying in has HBO and I just got to watch the new episode of GOT. Thank god.	The hotel I am staying in has HBO and I just got to watch the new episode of GOT . thank god .
Travel RN - Med Surg / Tele /... - Supplemental Health Care: (#Buffalo, NY) http://t.co/IRTwxRCfbv #Healthcare http://t.co/5ljE3az1Sy	Travel right now Med Surg Tele . Supplemental Health Care NY

The preprocessed texts are parsed using the Stanford dependency parser and for each word within the text, grammatical dependencies, POS tag and NER tags are obtained. Then using the defined rules aspects are extracted and then scoring is done. Sample results obtained from parser, Aspect extraction module and Opinion mining module are shown in *Table 5.2*.

Table 5.2: Parse tree for feed text, extracted aspects and opinion scores

Text 1	
Text	What a beautiful view on afternoon jog. The lake is nice and calm
Parse results	

[{
'what': {'pos_tag': 'WP', 'lemma': 'what', 'ner': 'O', 'dep': 'view', 'punct': '.'},
'a': {'pos_tag': 'DT', 'lemma': 'a', 'ner': 'O'},
'beautiful': {'pos_tag': 'JJ', 'lemma': 'beautiful', 'ner': 'O'},
'view': {'pos_tag': 'NN', 'lemma': 'view', 'ner': 'O', 'det': 'a', 'amod': 'beautiful', 'nmod': 'jog'},
'on': {'pos_tag': 'IN', 'lemma': 'on', 'ner': 'O'},
'afternoon': {'pos_tag': 'NN', 'lemma': 'afternoon', 'ner': 'TIME'},
'jog': {'pos_tag': 'NN', 'lemma': 'jog', 'ner': 'O', 'case': 'on', 'compound': 'afternoon'}, '.':
{'pos_tag': '.', 'lemma': '.', 'ner': 'O'}
}, {
'the': {'pos_tag': 'DT', 'lemma': 'the', 'ner': 'O'},
'lake': {'pos_tag': 'NN', 'lemma': 'lake', 'ner': 'O', 'det': 'the'},
'is': {'pos_tag': 'VBZ', 'lemma': 'be', 'ner': 'O'},
'nice': {'pos_tag': 'JJ', 'lemma': 'nice', 'ner': 'O', 'nsubj': 'lake', 'cop': 'is', 'cc': 'and', 'conj':
'calm'}, 'and': {'pos_tag': 'CC', 'lemma': 'and', 'ner': 'O'},	

'calm': {'pos_tag': 'JJ', 'lemma': 'calm', 'ner': 'O'}
}]

Aspects and opinion words	Opinion score

[{
'view':
[{'word': 'beautiful', 'pos_tag': 'JJ'}]
},{
'lake':
[{'word': 'nice', 'pos_tag': 'JJ'}, {'word': 'calm', 'pos_tag': 'JJ'}]
}]
	{
	'view': 0.6875,
	'lake': 0.2875
	}

Text 2

Text	The room was clean and comfortable but the view was not nice.

Parse results

[{
'the': {'pos_tag': 'DT', 'lemma': 'the', 'ner': 'O'},
'room': {'pos_tag': 'NN', 'lemma': 'room', 'ner': 'O', 'det': 'the'},
'was': {'pos_tag': 'VBD', 'lemma': 'be', 'ner': 'O'},
'clean': {'pos_tag': 'JJ', 'lemma': 'clean', 'ner': 'O', 'nsubj': 'room', 'cop': 'was', 'cc': 'but', 'conj': 'nice', 'punct': '.'},
'and': {'pos_tag': 'CC', 'lemma': 'and', 'ner': 'O'},
'comfortable': {'pos_tag': 'JJ', 'lemma': 'comfortable', 'ner': 'O'},
'but': {'pos_tag': 'CC', 'lemma': 'but', 'ner': 'O'},
'view': {'pos_tag': 'NN', 'lemma': 'view', 'ner': 'O', 'det': 'the'},
'not': {'pos_tag': 'RB', 'lemma': 'not', 'ner': 'O'},
'nice': {'pos_tag': 'JJ', 'lemma': 'nice', 'ner': 'O', 'nsubj': 'view', 'cop': 'was', 'neg': 'not'},
'.': {'pos_tag': '.', 'lemma': '.', 'ner': 'O'}
}]

Aspects and opinion words	Opinion score

[{
'room':
[{'word': 'clean', 'pos_tag': 'JJ'}, {'word': 'nice', 'pos_tag': 'JJ'}],
'view':
[{'word': 'nice', 'pos_tag': 'JJ', 'neg': True}]
}]
	{
	'view': 0.6875,
	'lake': 0.2875
	}

Text 3

Text	I like swimming in this pool, but it is not clean Today.

Parse results

[{
'i': {'pos_tag': 'PRP', 'lemma': 'i', 'ner': 'O'},
'like': {'pos_tag': 'VBP', 'lemma': 'like', 'ner': 'O', 'nsubj': 'i', 'xcomp': 'swimming', 'punct': '.', 'cc': 'but', 'conj': 'clean'},
'swimming': {'pos_tag': 'VBG', 'lemma': 'swim', 'ner': 'O', 'nmod': 'pool'},
'in': {'pos_tag': 'IN', 'lemma': 'in', 'ner': 'O'},
'this': {'pos_tag': 'DT', 'lemma': 'this', 'ner': 'O'},
'pool': {'pos_tag': 'NN', 'lemma': 'pool', 'ner': 'O', 'case': 'in', 'det': 'this'}, ',': {'pos_tag': ',', 'lemma': ',', 'ner': 'O'},
'but': {'pos_tag': 'CC', 'lemma': 'but', 'ner': 'O'},

'it': {'pos_tag': 'PRP', 'lemma': 'it', 'ner': 'O'}, 'is': {'pos_tag': 'VBZ', 'lemma': 'be', 'ner': 'O'}, 'not': {'pos_tag': 'RB', 'lemma': 'not', 'ner': 'O'}, 'clean': {'pos_tag': 'JJ', 'lemma': 'clean', 'ner': 'O', 'nsubj': 'it', 'cop': 'is', 'neg': 'not', 'nmod:tmod': 'today'}, 'today': {'pos_tag': 'NN', 'lemma': 'today', 'ner': 'O'}, '.': {'pos_tag': '.', 'lemma': '.', 'ner': 'O'} }]	
Aspects and opinion words	**Opinion score**

[{
'pool':
[{'word': 'like', 'pos_tag': 'VBP'}, {'word': 'clean', 'pos_tag': 'JJ',
'neg': True}],
'today':
[{'word': 'clean', 'pos_tag': 'JJ', 'neg': True}]
}]
	{
	'pool': 0.106250,
	'today': -0.1875
	}

Likewise, when taking the aspects of a census tract all the feeds within that census tract are considered. All the aspects extracted from the feeds and they are clustered according to the similarity. In *Table 5.3* it can be seen three aspect words are identified within the feeds. These aspects are identified from the tweets extracted within the period of January of 2014 to December of 2018 as shown in *Figure 5-2*. The aspects are hotel, morning, vacation, holiday and she. When it comes to only the period of the year 2015, the aspects are reduced to hotel, holiday and she as shown in *Table 5.4* and *Figure 5-3*. So, it can be clearly seen that the model identified "she" as an aspect. Since this model uses grammatical relations and the used dataset is social media feeds where language is use differently and also data might contain grammatically unstructured sentences, this behavior can be expected.

Table 5.3: Aspect words and categories, Census Tract 14215, around the Niagara Falls, Jan 2014 - Dec 2018

Aspect category	**Aspect words**	**Score**
Category 1 – hotel	hotel	0.010416666666666666
Category 2 – morning	morning	0.19474969474969475
Category 3 – vacation	vacation	-0.026636904761904764
	holiday	0.10416666666666667
Category 4 – she	she	-0.10227272727272728

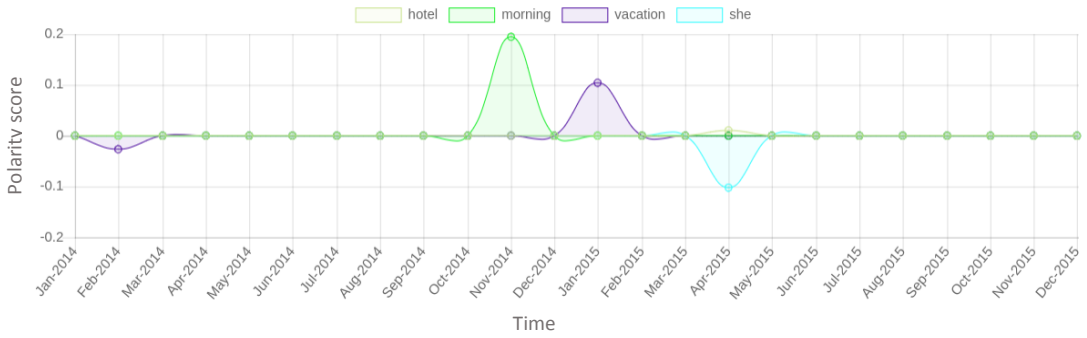


Figure 5-2: Aspect polarity distribution, Jan 2014 - Dec 2018, Census Tract 14215, around the Niagara Falls

Table 5.4: Aspect words and categories, Census Tract 14215, around the Niagara Falls, year 2015

Aspect category	**Aspect words**	**Score**
Category 1 – hotel	hotel	0.010416666666666666
Category 2 – holiday	holiday	0.10416666666666667
Category 3 – she	she	-0.10227272727272728

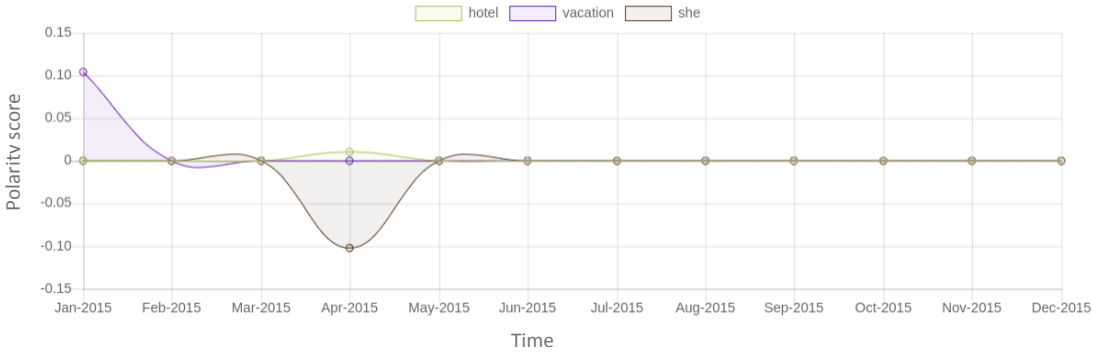


Figure 5-3: Aspect polarity distribution, year 2015, Census Tract 14215, around the Niagara Falls

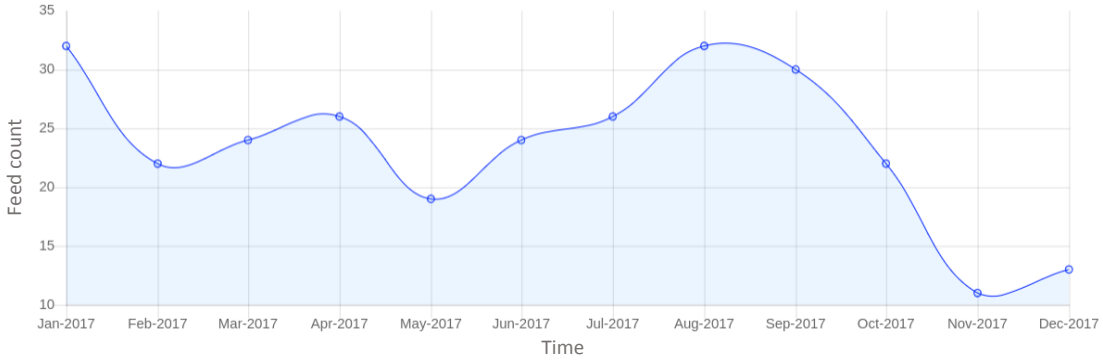
Using the measures of confusion matrix, the opinion mining model is evaluated. For the evaluation three annotated sentiment datasets in different domains are used. They are extracted from the Kaggle data repositories. Those are Sentiment140 dataset, First GOP Debate Twitter Sentiment dataset and Twitter US Airline Sentiment dataset. In the results shown in *Table 5.5*, it can be seen that the accuracies became low with the annotated datasets in the different domains. The main reason for the issue is ambiguity of the social media feeds. Language usage in the social media feeds and usage of

sarcasm, unpredictable termination of the words is the main cause to mislead the grammatical parser. Considering those issues future implications to improve this model are given in the Section 6.5

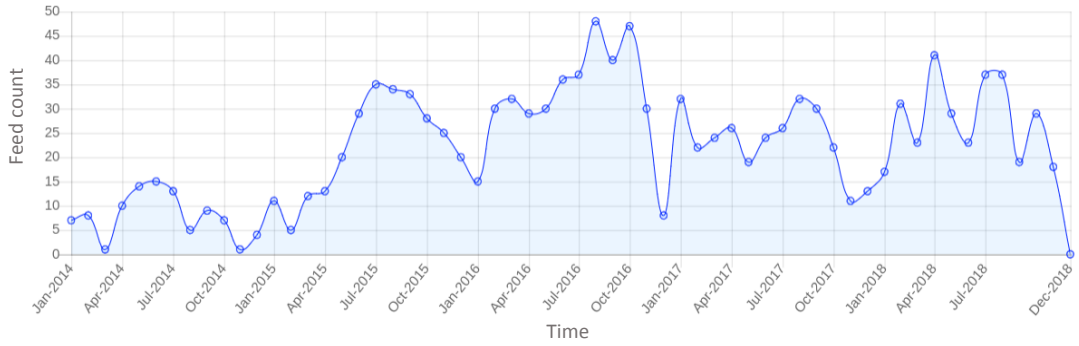
Table 5.5: Comparison with the manually annotated data

Dataset	Sentiment	Precision	Recall	F Measure	Accuracy
Sentiment140 dataset	**Negative**	0.589	0.576	0.582	0.535
	Neutral	1.000	0.026	0.051	
	Positive	0.498	0.755	0.600	
First GOP Debate Twitter Sentiment dataset	**Negative**	0.692	0.490	0.574	0.427
	Neutral	0.166	0.006	0.013	
	Positive	0.216	0.712	0.332	
Twitter US Airline Sentiment dataset	**Negative**	0.786	0.547	0.642	0.478
	Neutral	0.164	0.012	0.022	
	Positive	0.204	0.760	0.322	

The tourist densities are taken in three levels country vise, city vis and census tract vise. Since in this dataset only tweets from the area of the Niagara Falls are extracted only tourist count for that particular area is interpreted. Since the Niagara Falls situated in the border of the USA and Canada results are shown for these two countries. The tweet counts and opinions can be filtered by the season. *Figure 5-4* shows the country level feed distribution around the Niagara Falls in Canada border. In *Figure 5-5*, it demonstrates the city level feed density distribution in the city of St. Catherine. And in *Figure 5-6* the census tract level feed density distribution is shown for the census tract (postal code) L2M 4T5 in the St. Catherine. In all the Figures (a) interprets the density distributions in the year 2017 while (b) interprets the overall tweet densities from January of 2014 to December of 2018. Likewise, this interpretation is generated for all the feeds geotags. So, for each country, there are multiple cities and for each city, there are multiple census tracts. This density distribution is represented for each of these



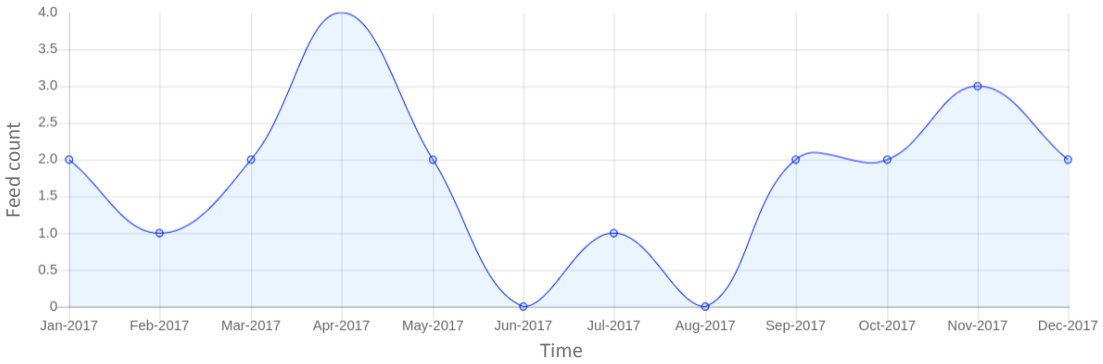
(a) Feeds count distribution of all the areas within Canada around Niagara Falls, year 2017



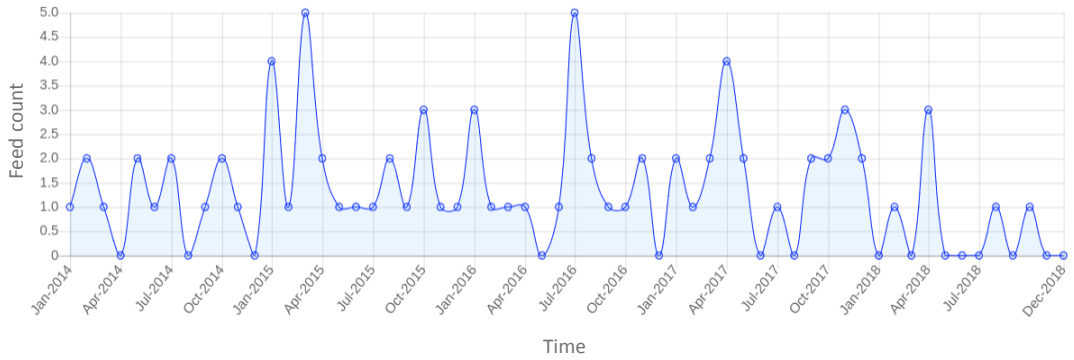
(b) Feeds count distribution of all the areas within Canada around Niagara Falls, Jan 2014 - Dec 2018

Figure 5-4: Country level feed count distribution for the area around Niagara Falls

objects. In all these distributions the only the area near the Niagara Falls are considered because the dataset extraction is done only for the area around the Niagara Falls.

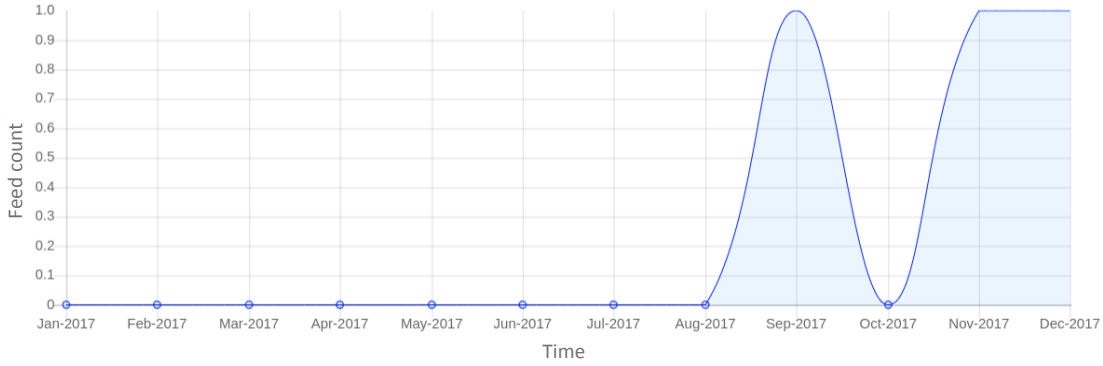


(a) Feeds count distribution of the areas within St. Catherine around Niagara Falls, year 2017

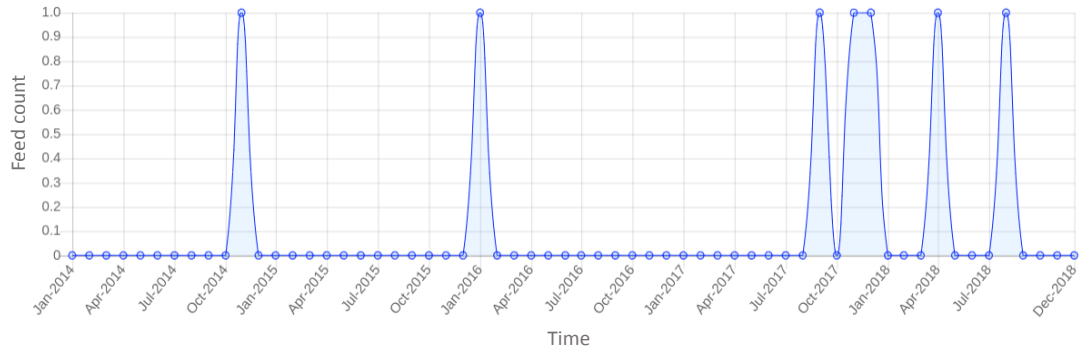


(b) Feeds count distribution of the areas within St. Catherine around Niagara Falls, Jan 2014 - Dec 2018

Figure 5-5: City level feed count distribution for the area around Niagara Falls in St. Catherine



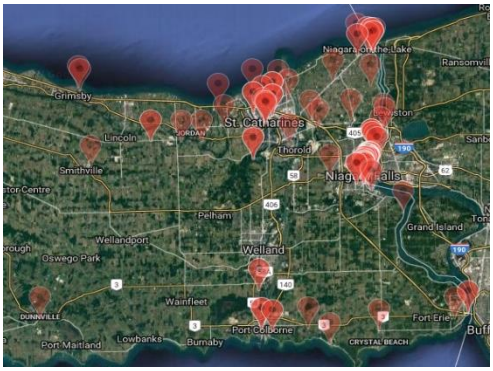
(a) Feeds count distribution of the areas within Census tract L2M 4T5 around Niagara Falls, year 2017



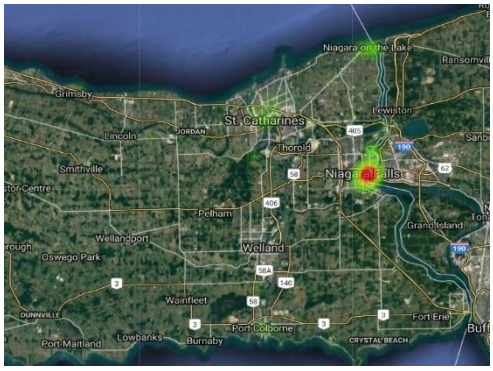
(b) Feeds count distribution of the areas within Census tract L2M 4T5 around Niagara Falls, Jan 2014 - Dec 2018

Figure 5-6: Census tract level feed count distribution, Census tract L2M 4T5, St. Catherine

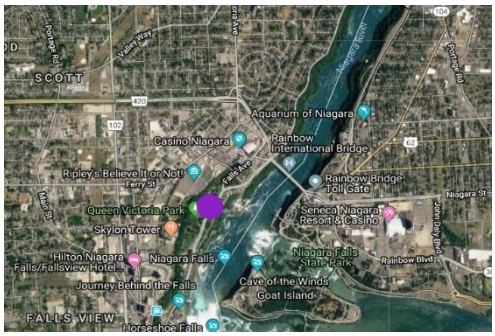
Heat maps are generated for each location using the density distribution. By the heat maps, the highest visited point are represented. According to the heat maps, the highest peaks can be identified. The algorithm described in Section 4.9 is used to identify the highest populated points. Using these two methods the density variation for each location is represented. This approach is also demonstrated for above mentioned three levels. So, the populated points are varying according to these levels. In *Figure 5-7* - *Figure 5-12* shows the density maps respectively in country, city and census tract levels. In each Figure, (a) shows the mapped markers of all the feeds identified within that particular object, (b) shows the heat map of the population of feeds in that object and the (c) shows the algorithmically identified RoAs within the particular area. For all levels, these figures demonstrate filtered feeds from January of 2014 to December of 2018 and the feeds within the year 2017 respectively.



(a) Markers of the feed locations



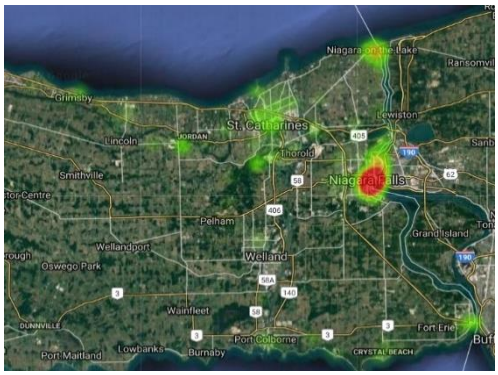
(b) Heatmap of population



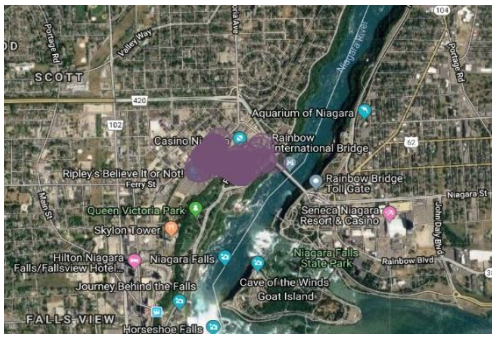
(c) Identified RoAs

Figure 5-7: Country level density maps around Niagara Falls within Canada, year 2017

(a) Markers of the feed locations

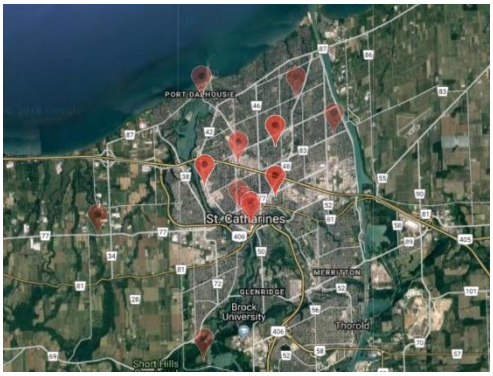


(b) Heatmap of population

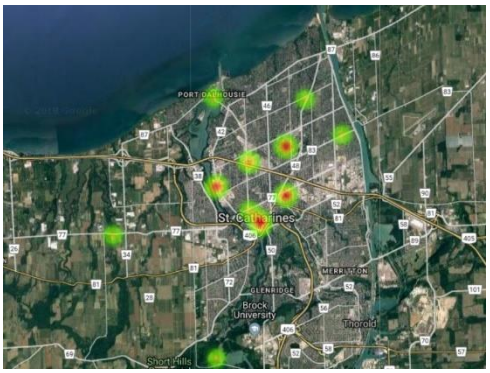


(c) Identified RoAs

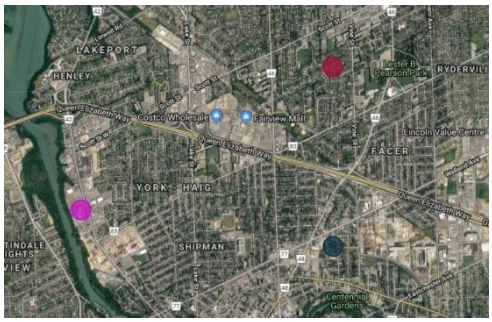
Figure 5-8: Country level density maps around Niagara Falls within Canada, Jan 2014 - Dec 2018



(a) Markers of the feed locations

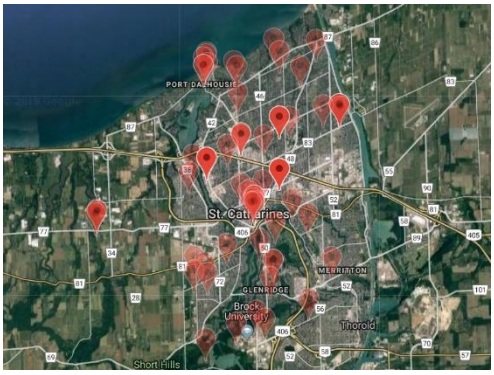


(b) Heatmap of population

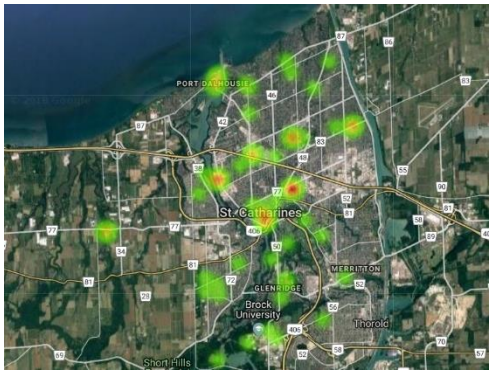


(c) Identified RoAs

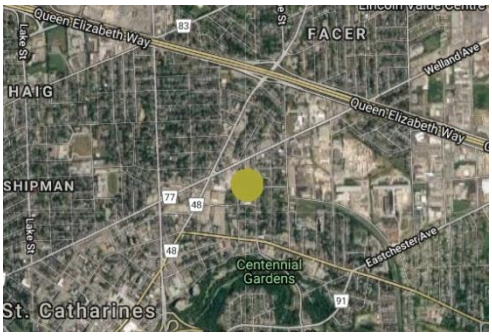
Figure 5-9: City level density maps for the area around Niagara Falls, St. Catherine, year 2017



(a) Markers of the feed locations

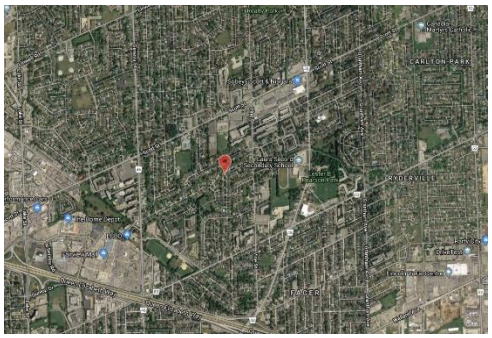


(b) Heatmap of population

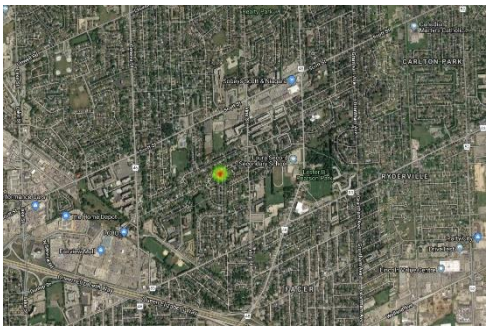


(c) Identified RoAs

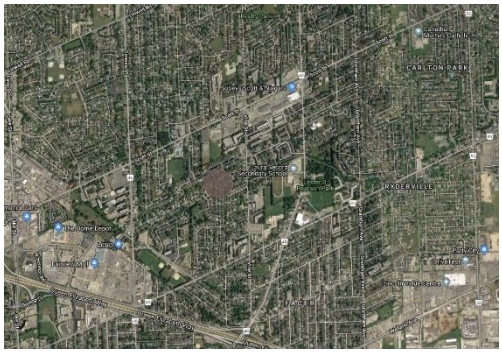
Figure 5-10: City level density maps for the area around Niagara Falls, St. Catherine, Jan 2014 - Dec 2018



(a) Markers of the feed locations

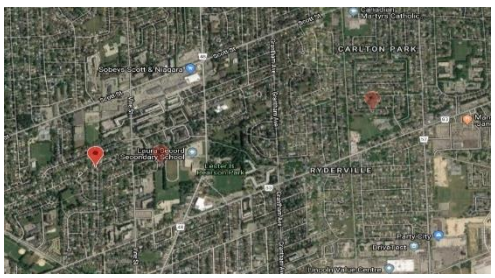


(b) Heatmap of population

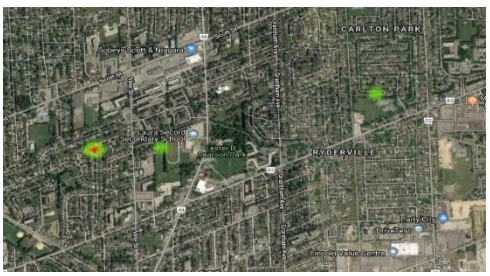


(c) Identified RoAs

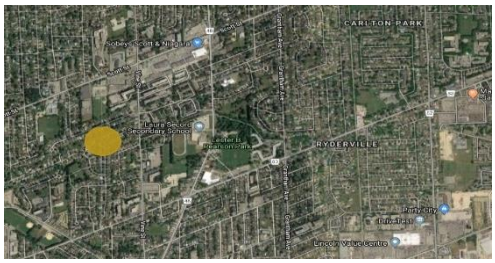
Figure 5-11: Density maps of the areas within Census tract L2M 4T5 around Niagara Falls, year 2017



(a) Markers of the feed



(b) Heatmap of population



(c) Identified RoAs

Figure 5-12: Density maps of the areas within Census tract L2M 4T5 around Niagara Falls, Jan 2014 - Dec 2018

The recommendation is customized using data extracted in the above phases. Generating recommendation is straightforward. It presents the visualized data in readable format. The recommendation is generated for all the three levels (country, city, census tract) independently. The recommendation texts are demonstrated for these three levels respectively in *Figure 5-13* - *Figure 5-15*. In *Figure 5-15* it can be seen for the census tract L2E 3L4, there is no highest rated aspect in the most visited period of January of 2018. It occurs when there's no aspect present in the feeds created within that particular period or all the aspects have negative ratings within that period.

Raiting : **1/5**
Mostly visited period : **Apr-2018**
No. of Visitors : **41 visitors**
Rating : **rating 1/5**
Highest rated aspect : **winter**
Rating : **4/5**
Highest rated period : **Jul-2018**
No. of Visitors : **37 visitors**
Rating : **rating 1/5**
Highest rated aspect : **hotel**
Rating : **4/5**

Mostly visited city : **Niagara falls** No. of Visitors : **247 visitors**
Rating : **2/5**

Mostly visited area : **[latitude 43.09261 and longitude -79.07524]**
No. of Visitors : **96**

Figure 5-13: Recommendation text for the area around Niagara Falls in Canada, year 2018

Raiting : **2/5**
Mostly visited period : **Apr-2018**
No. of Visitors : **35 visitors**
Rating : **rating 1/5**
Highest rated aspect : **winter**
Rating : **4/5**
Highest rated period : **Jul-2018**
No. of Visitors : **26 visitors**
Rating : **rating 1/5**
Highest rated aspect : **holiday**
Rating : **3/5**

Mostly visited census tract : **L2E 3L4** No. of Visitors : **97 visitors**
Rating : **2/5**

Mostly visited area : **[latitude 43.09261 and longitude -79.07524]**
No. of Visitors : **96**

Figure 5-14: Recommendation text for the area around Niagara Falls, city of Niagara Falls, year 2018

Raiting : **2/5**
Mostly visited period : **Jun-2018**
No. of Visitors : **15 visitors**
Rating : **rating -1/5**
Highest rated aspect :
Rating : **0/5**
Highest rated period : **Sep-2018**
No. of Visitors : **10 visitors**
Rating : **rating 2/5**
Highest rated aspect : **falls place**
Rating : **4/5**

Mostly visited area : **[latitude 43.09261 and longitude -79.07524]**
No. of Visitors : **96**

Figure 5-15: Recommendation text for the area around Niagara Falls, Census Tract L2E 3L4, year 2018

5.4 Summary

This Chapter shows the proposed evaluation model and the obtained from the system. This showcases the dataset and the results obtained from preprocessing, opinion mining and density maps approached. At the beginning of this Chapter, the dataset is briefly described. Then it can be seen the preprocessing results in order to obtain the opinion text. Then the parse results are shown with grammatical dependencies. Also identified aspects and their polarity scores are presented. Next, the location-based results are shown for a sample location. In aspect extraction over a location, the aspects are identifying from the tweets. They are ordered by the created time, so that opinion can be filtered by time. In this opinion model, some words are categorized as aspects but it is clear that they are not the aspects related to this domain. Then the density maps are presented with heatmaps, RoA maps and marker maps. It can be seen that algorithmically identified RoAs are similar to the visual heat locations of the heatmap. Finally, the recommendation is presented. Readable representation of the visualized representation is shown in the results.

Chapter 6 - Conclusions

6.1 Introduction

This Chapter tries to address the identified research question with the proposed design. In the first section, the research question is reviewed. Within this Chapter, the research design and the results are compared in order to present how well the approach taken in this research design addressed the identified research problem. The hypothesis is taken into discussion. In the latter sections, limitations and the further implications for the research are discussed.

6.2 Conclusions about research questions (aims/objectives)

The main aim of this research is to implement a combined approach of sentiment analysis and urban analytics using social media data in order to generate a recommendation on tourist destinations. The main objective of this study is to focus more on visitor's opinion because they are more realistic and richer of their pure satisfaction because existing recommendation systems focus more on static data sources such as reviews and statistics. From current work, it is examined that social media contains more opinion rich data. They can be used to find the opinion over a tourist destination and the geographical data about the destination and to obtain the visiting patterns. A major challenge in this study was identifying the non-tourist. Then the appropriate methods for opinion mining and obtaining the visiting patterns needed to be identified. In the research design, another problem raised that how to identify the feeds which are created after visiting the location or how to identify the feed geolocation is consistent with the exact tourist destinations location. And finally, an approach to generate a recommendation needed to be identified. This research design is based on above identified research questions. The appropriate approaches to tackle these problems are presented in the design. For the data extraction, Twitter Search API is used and the twitter feeds which are created from January 2014 to December 2018 geo located around Niagara Falls are extracted. The results of this study are presented from this dataset. Using this dataset, it is tested whether the results generated by this system are closer to the actual. When identifying the non-tourists, spatio-temporal sequence-based

approach is taken. Where it considers the particular users feed distribution and density within the census tract area for a particular year is examined. To evaluate the feeds location NERs are used. Locations are extracted from the text and they are compared with the feed location. After those filtering steps, opinion mining approach and density maps-based approach is taken. In opinion mining aspect-based sentiment analysis is tested for this domain in this study. Here the lexicon-based approach is taken for identifying aspects. Through the grammatical dependencies, aspects are identified. The main drawback in this method is, the accuracy of the model depends on the correctness of the dependency parser. The user-generated feeds within social media might not always grammatically correct. Grammatical dependencies depend on the grammar of the feed. And also, usage of the words and punctuations mislead the parser and might generate inaccurate parse results. So, identified aspects tend to be incorrect in situations like this. It is the main drawback of the approach taken in this study. So, in the extracted results in Chapter 5 - , it can be seen some of the invalid aspects are extracted. In the evaluation model of testing the accuracy of opinion mining using an annotated dataset shown decrease of the accuracies. Main problem to this issue is inconsistency of the parse results. Since language usage of this domain is not persistent with the actual grammatical rule, this kind of issue arise. Proposed fixes for this issue are given in the further implications section. This shows the employed model need to be more adapted.

In identifying the visiting patterns, RoAs are taken using a defined algorithm. This extracted RoAs are compared with the heat maps generated using the densities of the feeds. When extracting the past feed from Twitter, the number of feeds which can be extracted for a month is limited. For the extracted feeds the densities tend to generate accurate results. The RoAs are taken using these densities. The recommendation is a combination of these all approaches. The resulting recommendation can be generated with the RoAs and the tourist interests obtained from the opinion mining module.

6.3 Conclusions about research problem

The main problem with existing recommendation systems is they are infrequently updated and used static data sources like surveys, reviews and statistics. Is there a possible solution to address this issue. It is using dynamic data sources which are updated real-time. In this study, that problem is addressed using the domain of social media. But

when considering these data sources, main problems are language usage in social media. Since this study takes the aspect-based opinion mining model, the grammatical relationships needed to be extracted from the text. The basic preprocessing steps need to be aligned to address this language. According to the obtained results there are some issues when extracting the aspects since this vulnerability. For other problems like non-tourist detection, above proposed solutions are used in this study. Taking the RoAs systemically is another task addressed in this study. For that, the feeds are grouped according to the targeted destination. So, that only the feeds within that particular area are obtained and the points are grouped according to the closest proximity and the highest density is taken as the RoA. Using the above approaches, a solution is found for the overall problem of generating the recommendation from the dynamic feeds.

6.4 Limitations

When understanding the language usage of the social media feeds it requires bit effort. There are some words which are language specific even though the feed is in the English language there might be terms which the user used in the tongue language. That kind of terms are hard to identify by the parser and when applying rules, they generate inaccurate results. Another problem is it is assumed that the majorly of the users provide their honest opinion in the feed. It is not an issue a smaller number of users not being honest because average opinion is taken. The geotag coordinates of the tweet cannot have guaranteed as the exact location that the tourist was being because when tagging the location, the location coordinates are suggested by the third-party applications or the google map coordinates or the mobile GPS location. If there are any issues with those services there might be misaligned coordinates. Also, the user might not know the exact location that needed to be tagged. So, this kind of issues affects the accuracy if the result.

6.5 Implications for further research

As mentioned early this study focuses on generating a recommendation using dynamic social media feeds which are generated by the user. In this study, the recommendation is generated for a particular location. But in social media, some studies use collaborative filtering and content-based filtering methods in generating recommendations for products, hotels etc. This approach can be integrated with this research study when

generating a user-based recommendation. In this study, the location-based data is used to analyse the user's interest on the location. To generate the user-based recommendation the user's data from the social media can be collected and then identify the user's behavior. Then the location can be recommended to that user according to his behavior.

In this system, for aspect-based opinion mining, grammatical rules are used. But as shown in Section 5.3 there is a drawback in the results because of the inconsistent language usage in this domain. These grammatical inconsistencies like invalid grammars, invalid termination symbols or usage of sarcasm needed to be previously addressed before dependency parsing. If those are correctly identified, this approach can be fine-tuned in order to get better accuracies. And also, for the opinion mining, as identified in Chapter 2 - , there are two approaches. Machine learning approach and lexicon-based approach. In this study, a lexicon bases approach is employed. But there are studies which use the ML based approach. For ML approach, the main drawback identified in the literature is domain dependency, because the ML models learn from the corpus and the corpus is domain dependent. But for this problem for opinion mining ML based opinion mining model can be tested and state of art comparison can be done with this result. Another implication is using a ML model only to identify the aspect (not for whole opinion mining model) and the scores can be generated using the SentiWordNet.

References

[1] Y. Zheng, Z. Zha and T. Chua, "Mining Travel Patterns from Geotagged Photos," *ACM Transactions on Intelligent Systems and Technology,* vol. 3, no. 3, pp. 1-18, 2012.

[2] Marie-Catherine de Marneffe, B. MacCartney and C. D. Manning, "Generating Typed Dependency Parses from Phrase Structure Parses," in *Proceedings of LREC 2006*, 2006.

[3] A. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons,* vol. 23, no. 1, pp. 59-68, 2010.

[4] C. Aggarwal, "An Introduction to Social Network Data Analytics," in *Social Network Data Analytics*, 2011, pp. 1-15.

[5] A. Dhiratara, J. Yang, A. Bozzon and Geert-Jan Houben, "Social Media Data Analytics for Tourism - A Preliminary Study," in *Proceedings of KDWeb*, 2016.

[6] M. Salas-Olmedo, B. Moya-Gómez, J. García-Palomares and J. Gutiérrez, "Tourists' digital footprint in cities: Comparing Big Data sources," *Tourism Management,* vol. 66, pp. 13-25, 2018.

[7] H. Song and H. Liu, "Predicting Tourist Demand Using Big Data," in *Analytics in Smart Tourism Design*, 2016, pp. 13-29.

[8] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[9] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal,* vol. 5, no. 4, pp. 1093-1113, 2014.

[10] F. Khoshnood, M, Mahdavi and M. Kiani sarkaleh, "Designing a Recommender System Based on Social Networks and Location Based Services," *International Journal of Managing Information Technology,* vol. 4, no. 4, pp. 41-47, 2012.

[11] N. Zainuddin, A. Selamat and R. Ibrahim, "Improving Twitter Aspect-Based Sentiment Analysis Using Hybrid Approach," *Intelligent Information and Database Systems,* pp. 151-160, 2016.

[12] M. Adedoyin-Olowe, M. Medhat Gaber and F. Stahl, "A Survey of Data Mining Techniques for Social Media Analysis," *Journal of Data Mining and Digital Humanities,* vol. 7895, pp. 135-145, 2013.

[13] V. Hatzivassiloglou and K. McKeown, "Predicting the Semantic Orientation of Adjectives," in *Proceedings of the 8th Conf. on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, Association for Computational Linguistics, 1997.

[14] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2011.

[15] M. Missen, M. Boughanem and G. Cabanac, "Opinion mining: reviewed from word to document level," *Social Network Analysis and Mining,* vol. 3, no. 1, pp. 107-125, 2012.

[16] G. Qiu, X. He, F. Zhang, Y. Shi, J. Bu and C. Chen, "DASA: Dissatisfaction-Oriented Advertising Based on Sentiment Analysis," *Expert Systems with Application Journal Elsevier,* vol. 37, no. 9, p. 6182–6191, 2010.

[17] J. Jiao and Y. Zhou, "Sentiment Polarity Analysis based Multi Dictionary," *Physics Procedia,* vol. 22, pp. 590-596, 2011.

[18] D. R. Rice and C. Zorn, "Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies," in *Proceedings of NDATAD*, 2013.

[19] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, 2005.

[20] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 2004.

[21] P. Lane, D. Clarke and P. Hender, "On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data," *Decision Support Systems,* vol. 53, no. 4, pp. 712-718, 2012.

[22] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, 2005.

[23] S. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, 2004.

[24] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, 2004.

[25] Z. Wang, G. Bai, S. Chowdhury, Q. Xu and Z. L. Seow, "TwiInsight: Discovering Topics and Sentiments from Social Media Datasets," 2017.

[26] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment analysis of Twitter data," in *Proceedings of Workshop on Languages in Social Media*, Portland, Oregon, 2011.

[27] A. Walha, F. Ghozzi and F. Gargouri, "ETL Transformation Algorithm for Facebook Opinion Data," in *Proceedings of the 11th International Conference on Web Information Systems and Technologies*, 2015.

[28] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong and U. Kaymak, "Exploiting Emoticons in Sentiment Analysis," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC '13*, 2013.

[29] E. Luxton, "Let's Be Travel Bloggers," [Online]. Available: https://letsbetravelbloggers.com/social-media/twitter/hashtags. [Accessed 02 January 2019].

[30] M. Braunhofer and F. Ricci, "TripAdvisor Dataset," ResearchGate, [Online]. Available: https://www.researchgate.net/publication/308968574_TripAdvisor_Dataset. [Accessed 02 January 2019].

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825--2830, 2011.

[32] G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM,* vol. 38, no. 11, pp. 39-41, 1995.

[33] D. Jurafsky, J. H. Martin, Speech and Language Processing 2nd Edition.

[34] S. Baccianella, A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in *Proceedings of Language Resources and Evaluation (LREC)*, 2010.

[35] W. Gao, "Sentiment140 | Kaggle," Kaggle, [Online]. Available: https://www.kaggle.com/pirateshadow/sentiment140-test. [Accessed 05 11 2018].

[36] "First GOP Debate Twitter Sentiment | Kaggle," Kaggle, [Online]. Available: https://www.kaggle.com/crowdflower/first-gop-debate-twitter-sentiment. [Accessed 05 11 2018].

[37] "Twitter US Airline Sentiment | Kaggle," [Online]. Available: https://www.kaggle.com/crowdflower/twitter-airline-sentiment. [Accessed 05 11 2018].

Appendix A: Diagrams

From *Figure A-1* - *Figure A-6* demonstrate the user interfaces of the implemented web app in three levels.

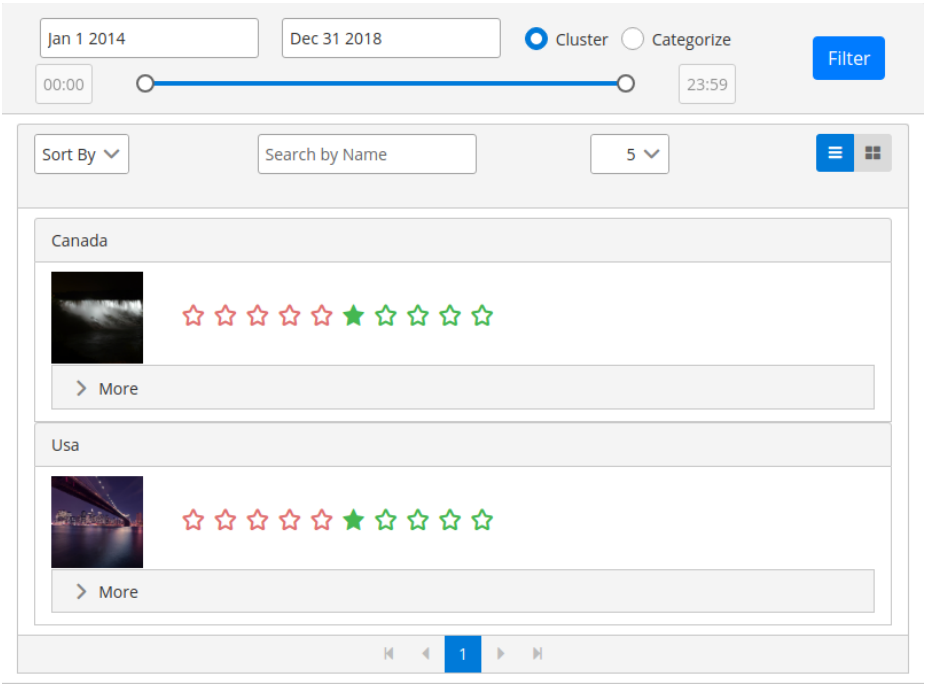


Figure A-1: Visualization of the Country level rating

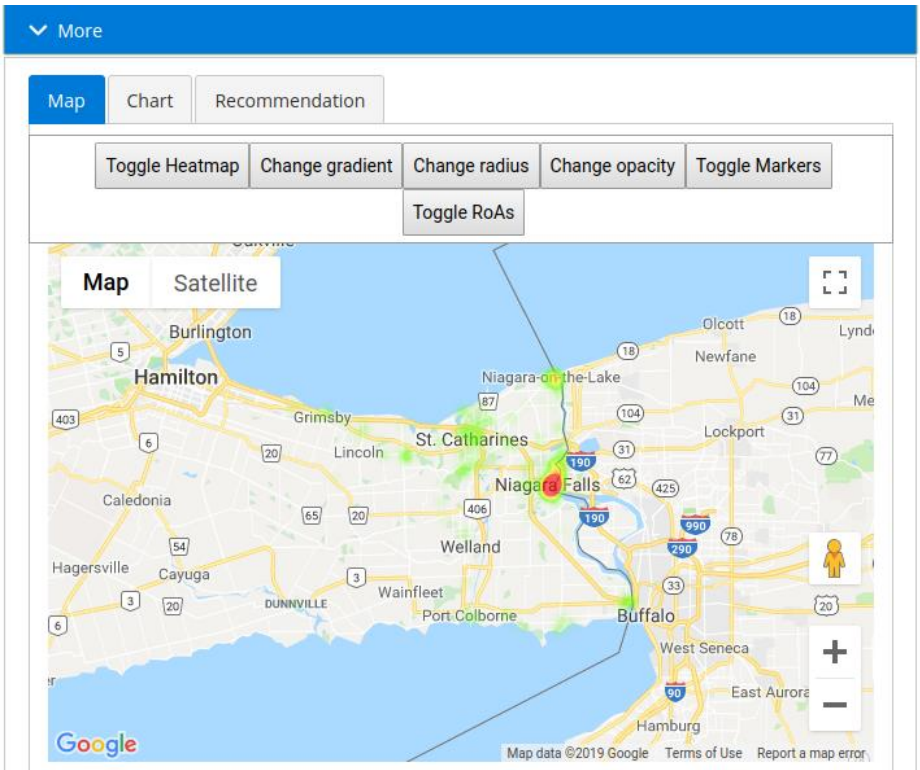


Figure A-2:Visulization of the maps, charts and the recommendation

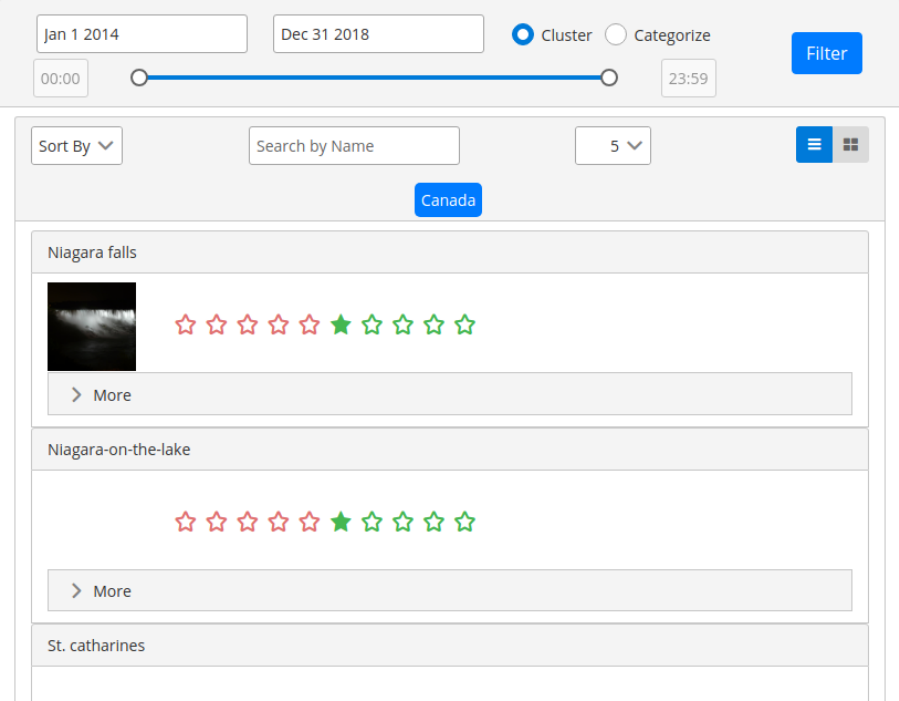


Figure A-3: Visualization of the City level rating

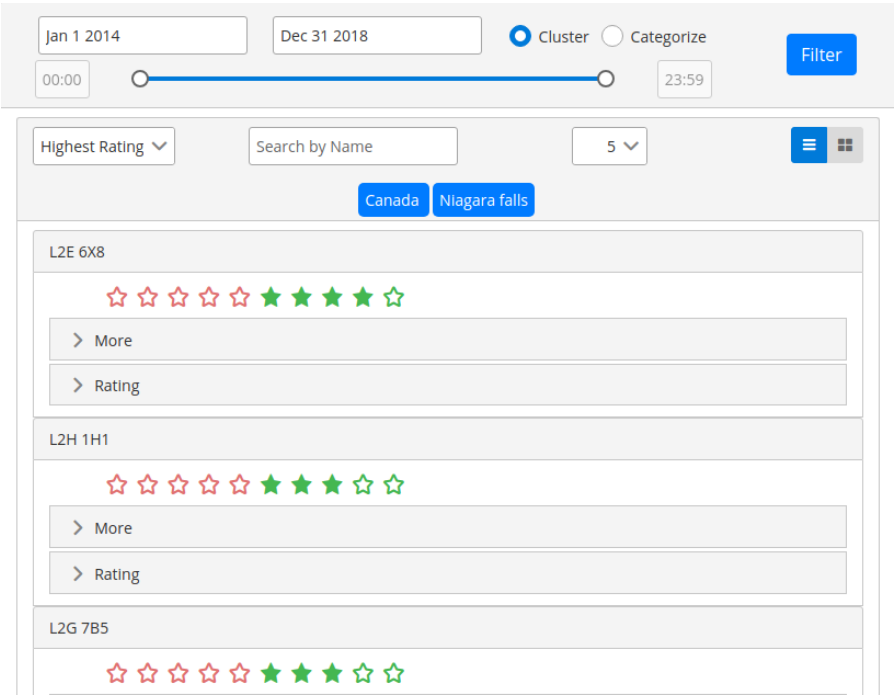


Figure A-4: Visualization of the Census Tract level rating

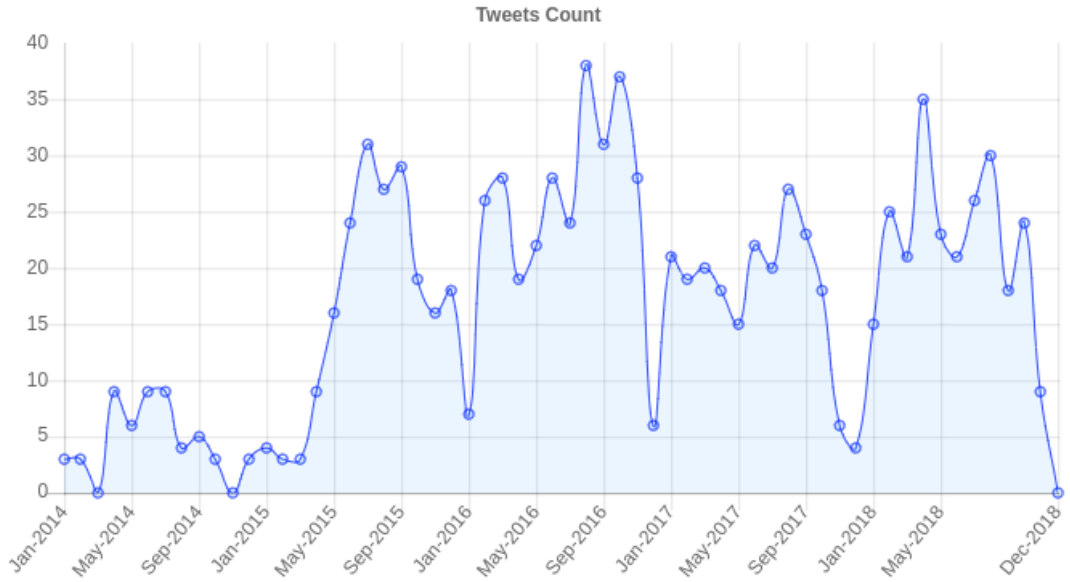


Figure A-5: Feed count distribution in the City of Niagara Falls

Figure A-6: Aspect polarity distributions

Figure A-7 shows the identified RoAs with the heat map, and *Figure A-8* shows the identified marker positions and the heat map.

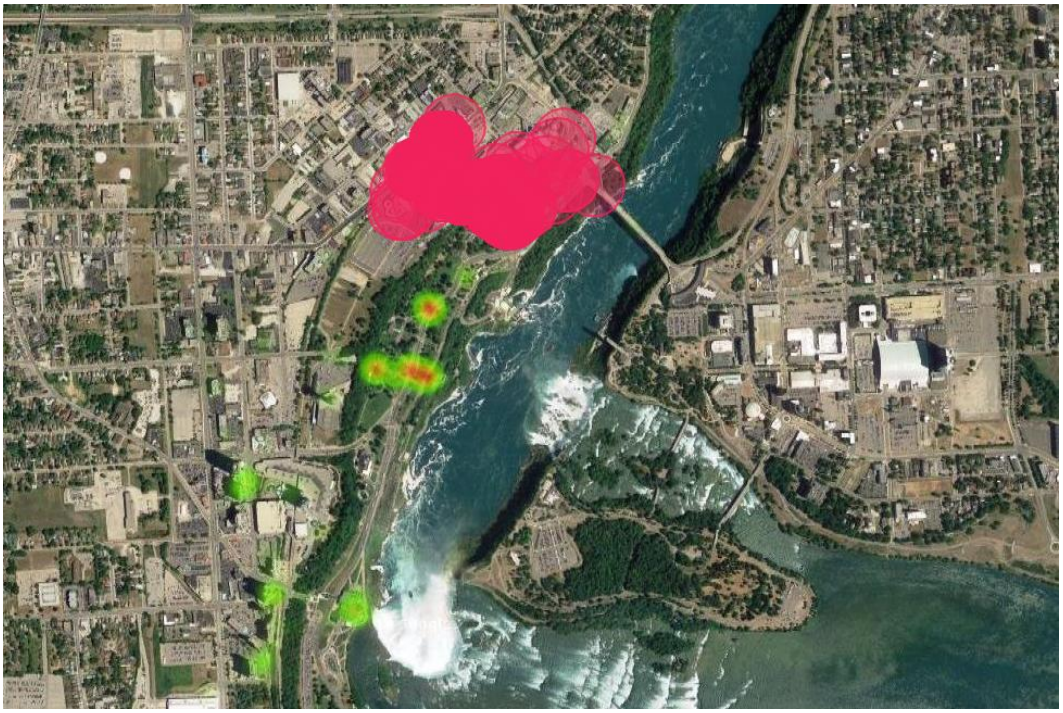


Figure A-7: Identified RoAs within the City of Niagara Falls

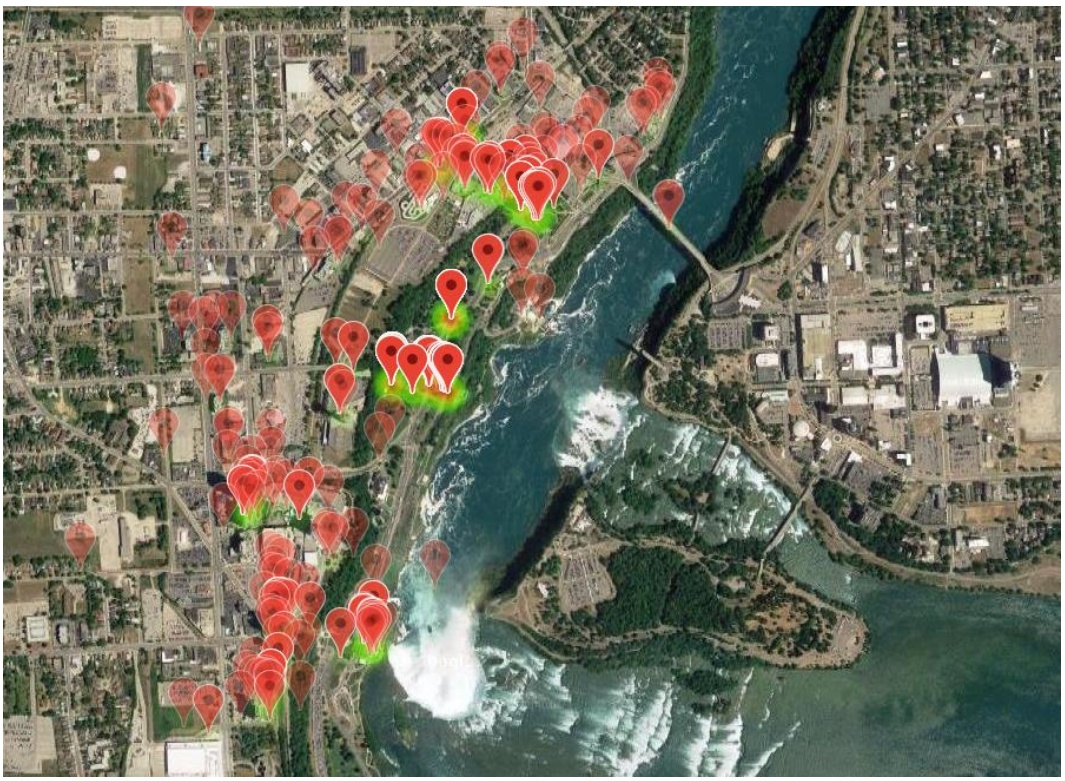


Figure A-8: Marker positions of the feeds

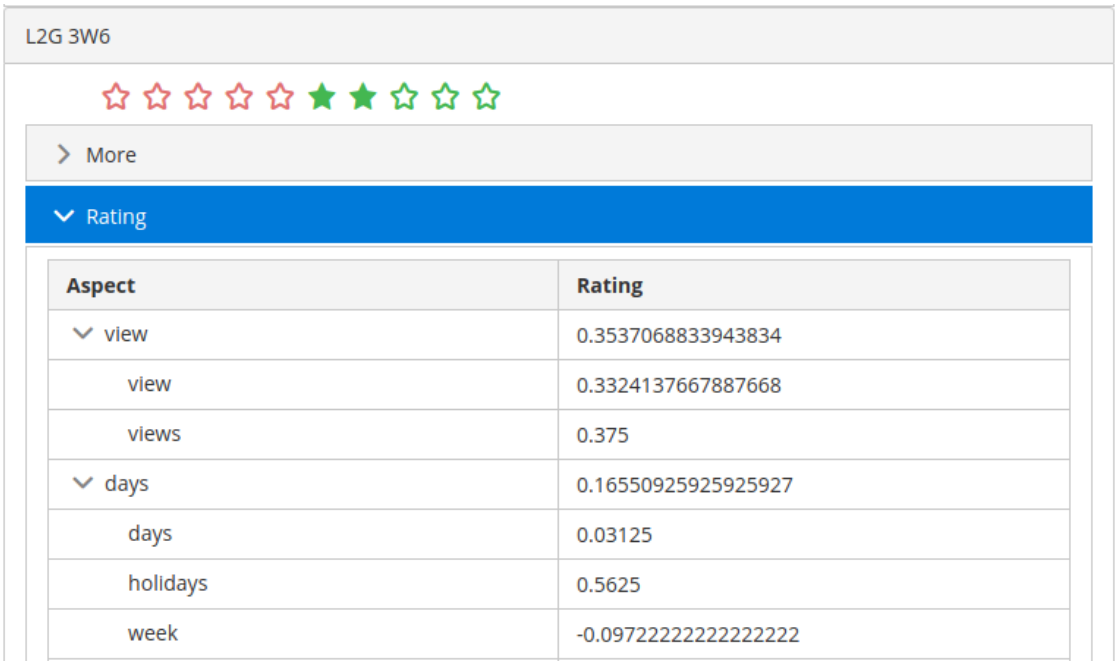


Figure A-9: Aspect ratings for the L2G 3W6 Census Tract

Aspect rating are demonstrated as in *Figure A-9* in the web app

Appendix B: Code Listings

Python implementation of the clustering aspects is given below. First it takes the aspects and then assign related aspect to each aspect according to the similarity in the getAspectClusters function. Then the aspects are grouped according to the relationships. In the clusterAspects function.

```
def getAspectClusters(self, aspectlist):
        aspectsynsets = wordnetdic.getSynsets(aspectlist)
        removeindex = 0
        for aspect in aspectlist:
                if not self.isAspect(aspectsynsets[removeindex]):
                        del aspectlist[removeindex]
                        del aspectsynsets[removeindex]
                        removeindex -= 1
                removeindex += 1
        aspectgroup = {}
        for (i,a1) in enumerate(aspectlist):
                aspectgroup[a1] = {'group':a1, 'score':0}
                synsets1 = aspectsynsets[i]
                for (j,a2) in enumerate(aspectlist):
                        if a1 != a2:
                                synsets2 = aspectsynsets[j]
                                vals = [val if val else 0
                                for val in [wordnetdic.getWupSimilarity(s1,s2)
                                for s1,s2 in product(synsets1,synsets2)]]
                                if(len(vals)>0):
                                        maxscore = max(vals)
                                        if aspectgroup[a1]['score'] < maxscore:
                                                aspectgroup[a1]['score'] = maxscore
                                                aspectgroup[a1]['group'] = a2
        return aspectgroup
def clusterAspects(self, aspects):
        categories = {}
        if "hashtags" in aspects:
                categories["hashtags"] = aspects["hashtags"]
                del aspects["hashtags"]
        aspectlist = [aspect for aspect in aspects if aspect != "hashtags"]
        if len(aspectlist)==0:
                return categories
        if len(aspectlist)==1:
                categories[aspectlist[0]] = aspects
                return categories
        aspectgroup = self.getAspectClusters(aspectlist)
        def createGroup(key):
                del aspectgroup[key]
                items = []
                for key1,value1 in aspectgroup.items():
                        if key == value1['group']:
                                items.append(key1)
                for item in items:
                        items = items + createGroup(item)
                return items
        for aspect in aspectlist:
                if aspect in aspectgroup:
                        categories[aspect] = {aspect: aspects[aspect]}
                        for aspectword in createGroup(aspect):
                                categories[aspect][aspectword] = aspects[aspectword]
        return categories
```

Implementation of the extracting point according to their distance is implemented in following getRIs function. The following is the typescript implementation for the web app.

```
getRIs(){
  var pointgroup = {};
  for(var idx1=0; idx1<this.data.length; idx1++){
    var point1 = this.data[idx1];
    pointgroup[JSON.stringify(point1)] = {'group':[]}
    for(var idx2=0; idx2<this.data.length; idx2++){
      var point2 = this.data[idx2];
      if(idx1!=idx2){
        var dis = this.getGeoDistance([point1.latitude,point1.longitude],
          [point2.latitude,point2.longitude]);
        if(dis < minRIDistance){
          pointgroup[JSON.stringify(point1)]['group'].push(point2);
        }
      }
    }
  }
  var maxCount = 0;
  var groups = [];
  for(var idx=0; idx<this.data.length; idx++){
    var point = this.data[idx];
    var key = JSON.stringify(point);
    if(key in pointgroup){
      var group = this.createGroup(pointgroup, point);
      groups.push(group);
      if(group.length > 1 && group.length > maxCount){
        maxCount = group.length;
      }
    }
  }
  for(var idx1=0; idx1<groups.length; idx1++){
    group = groups[idx1];
    if(maxCount == group.length){
      var latlnggroup = [];
        var color = this.getRandomColor();
        for(var idx2=0; idx2<group.length; idx2++){
          point = group[idx2];
          latlnggroup.push({lat:point.latitude,lng:point.longitude});
          this.circes.push(new google.maps.Circle({
            strokeColor: color,
            strokeOpacity: 0.8,
            strokeWeight: 2,
            fillColor: color,
            fillOpacity: 0.35,
            center: new google.maps.LatLng(point.latitude, point.longitude),
            radius: minRIDistance
          }));
        }
        this.RIs.push(latlnggroup);
        this.colors.push(color);
      }
    }
  }
```