Harnessing Tweets to get the *Pulse of a City*

by

Esha

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

Twitter is one of the most popular social media applications and is used for a number of reasons. Every day, users share a vast amount of information through tweets that provide location-relevant updates, of events happening in real-time, and to inform other users of upcoming events in a given geographical location. The information in tweets can be used, not only to learn about what is happening in a city, but also to understand users' emotions (e.g., love, fear) and sentiments (e.g., positive, negative) on topics and events as they unfold over time. Such information will be relevant and useful only when the right location is identified for a given set of tweets. Further, considering the volume of data generated on Twitter, both categorization of tweets and visualizations can help users in managing information overload. Categorization of tweets into topic labels can help in identifying broad level categories of topics discussed in a city and filtering unwanted tweets by allowing users to focus on accessing tweets from categories that are of interest to them. Visualization can play a critical role in presenting large and complex data into more easily discerning formats to facilitate comparison on different facets. This research focused on these multiple areas including identification of locations relevant to tweets, visualizations of location-related sentiments and emotions, and categorization of tweets into topic labels.

The identification of tweet-relevant location is a challenging problem as location names are not always explicitly included in most of the tweets. However, location related information is implicitly included with the insertion of user-ids and hashtags in tweets. Thus, the

research aim is to improve identification of tweet-relevant location by harnessing information embedded in user-ids (e.g., @EPLdotCA is the userId of the public libraries in the city of Edmonton) and hashtags (e.g., #yeg is the hashtag for the city of Edmonton). This novel approach, termed *DigiCities*, focused on using this implicit information to identify tweet-relevant locations.

DigiCities are digital equivalents of cities as represented in digital spaces; cities are primarily represented by People, Organizations and Places (POP) in the physical environment, which has digital presence on Twitter as well as through user-ids and hashtags. Digital profiles of cities are created using user-ids and hashtags of people, organizations and places associated with each city and are then used to identify and reinforce city names in tweets. The digital profiles of eight cities from the Province of Alberta in Canada were developed, and a number of classification experiments using different algorithms including k-Nearest Neighbour (kNN), Naïve Bayes (NB) and Sequential Minimal Optimization (SMO) were conducted to evaluate the effectiveness of the proposed approach. The classification accuracy score improved for each algorithm after the implementation of the city profile on Twitter data. Furthermore, tweets from these eight locations were further analyzed to identify users' sentiments and emotions, and associated topics. Multiple visuals of results achieved were developed to compare and contrast sentiments and emotions during different temporal periods at city level.

Acknowledgement

I would like to convey my sincere gratitude for my advisor Dr. Osmar Zaiane for his vital contributions in my Master's Program including this thesis work. His role in my journey through the program is difficult to express in words. His constant support, continuous encouragement and invaluable advise at every step really helped me to complete this thesis work.

I would also like to thank Hamman Samuel for helping me in the data collection phase of this research and giving valuable guidance in the initial phase of this thesis work.

Last but not the least, I would also like to express my special thanks to my family members and friends for their constant support and motivation.

Table of Content

List of Tables

List of Figures

List of Abbreviations

Label	Description
NB	Algorithm used is Naïve Bayes
SMO	Algorithm used is SMO i.e. Sequential Minimal Optimization
kNN	Algorithm used is kNN i.e. k-Nearest Neighbour
B	Dataset used in the experimentation is baseline data i.e., city names are neither appended or replaced
A	Dataset used in the experimentation is the one in which city names were appended (by city names) when terms in the tweets matched the city profile POP Framework elements
R	Dataset used in the experimentation is the one in which city names were replaced (by city names) when terms in the tweets matched the city profile POP Framework elements
SA	Stemming algorithm (SA) means the implemented of stemming algorithm on the dataset. Lovins stemmer as implemented in Weka was used to stem the terms.
WS	Without Stopwords (WS) i.e. dataset from which stopwords were removed. Stopword list as implemented in Weka was used in this research.

Chapter 1

Introduction

1.1 Overview

The popularity of social media applications such as Twitter, Facebook, Instagram, LinkedIn and Pinterest has been increasing over a number of years. These applications have created a large user base and there are over three billion active social media users [114]. For example, social media applications such as Facebook and Instagram have two billion and one billion users respectively [25].

This research focuses on Twitter which is a micro-blogging application launched in 2006 [74]. Originally, Twitter allowed users to post messages of up to 140 characters [49] and recently, the limit was enhanced upto 280 characters [74]. Twitter has 330 million users [25]. According to [7], key statistical facts about Twitter include - among its user base, 34% are females and 66% are males; about, three-fifth of Twitter users are between age 18-49 i.e., 37% are between 18 and 29, and 25% are between 30 and 49 years of age, and; "85% of small and medium business users use Twitter to provide customer service".

The use of Twitter has become ubiquitous – organizations of different sizes including small and medium sized, governments, and individuals use it in a variety of ways and for various reasons. For example: organizations use it to share products information, new product launches, and for customer outreach; governments use it for reasons such as consultation, information dissemination, and engagement with citizens, and; individuals use it to share information, stay in touch with family and friends, seek advice,

learn about the latest trends in the community (or place of interest), and get news. For example, "74% of Twitter users say they use the network to get their news" [7].

The real-time information shared through Twitter can help users, organizations, governments and other stakeholders to understand the pulse of a geographical location (e.g., city) by identifying key topics emerging from conversation and understanding users' emotions (e.g., love, fear) and sentiments (e.g., positive, negative) unfolding over time in a city. We call it the '***Pulse of a City***' because the topics discussed as well as the emotions and sentiments expressed may not only vary from city to city, but also changes with time, possibly in a rhythmical way reminding of the throbbing of an artery. We use this metaphor as we intend our analysis to provide us with the 'heartbeat of a city' with regard to the interest expressed on social media by its citizens. However, such pulse of a city is representing topics, emotions and sentiments as expressed by Twitter users of a city and does not necessarily represent the whole population of a city.

It is critical that the right city related to posted tweets is identified or else the *Pulse of a City* would be misleading and would become irrelevant. Further, considering the volume of tweets posted in a city's context makes it challenging to understand the *Pulse of a City* in its raw form. Also, it is important that appropriate visualization of city-related tweets is done to get a better grasp of that city's pulse. Thus, the proposed research focuses on improving identification of a city, relevant to the set of tweets, and creating appropriate visualizations to create a pulse by identifying topics, emotions and sentiments for a particular city or a geographical location.

1.2 Motivation

Twitter users are active and post over 500 million tweets every day worldwide, that comes to approximately 6,000 tweets per second or over 180 billion tweets per year [7][127]. A large number of tweet postings generates a very large amount of data that is potentially causing information overload for users, but at the same time, it is creating opportunities for researchers to explore and investigate a rich media platform from different perspectives. The seemingly absurd fact about Twitter in particular is that people post about anything and most posts are irrelevant to most people considering it information overload, or even useless information. So the majority of posts are relevant to a minority of people and the majority of users are a priori interested in a minority of posts or in limited topics [69]. However, this sea of seemingly garbage and useless posts is what paradoxically allows us to get this pulse of a location in terms of topics of concern, by distilling this apparently useless information into practical and useful facts.

Tweets have a lot of informational value, and, depending upon the use, and usefulness, could be organized or classified in various ways such as sentiment-based, emotion-based, topic-based, and location-based. A number of researchers such as Papadopoulos et al. [88], Sobkowicz et al. [115] and Xia et al. [136] noted that a variety of meaningful patterns can be extracted from the posted tweets to benefit different types of users i.e., citizens, governments, and organizations. For example, organizations can market products and services to users in a particular city through location-based targeting. Citizens can reach out to government to highlight issues and challenges in their local communities. Further to this, accident reporting via tweets (other than emergency 911 calls) can help law enforcement units in government to provide quick help to citizens in need, as well as allow commuters to decide on whether to use or avoid

that particular route. Businesses like restaurants could also benefit as related sentiments (or reviews) can help users to make informed decisions in choosing their next dining outlet in a city. Moreover, an intense discussion on a particular disease, such as flu, could suggest a potential sign of outbreak [2].

Researchers such as Aiello et al. [4] noted that users of social media, communicate and exchange information on "real-world events and dynamics" (p.1268), and they further stated that with the increase in the number of users and their participation on social media, the "social media streams [would] become accurate sensors of real-world events" (p.1268). User contributions can be harnessed from multiple perspectives including reflection on topics associated with tweets, emotions (e.g., joy, love and fear) [31] and sentiments (e.g., positive or neutral) [3] on topics (or events) in a given geographical location. Likewise, different groups can effectively use analysis of on-going conversations on social media platforms, and connect them with the right locations for various purposes. Tsou [126] argued in favour of developing and designing appropriate techniques to "geo-locate the contents of individual posts and web pages from cyberspace to realspace" (p.1). The development of such techniques would help in enhancing the accuracy of city-relevant tweets. Such increased accuracy would further help in identifying the *Pulse of a City*, including emotions and sentiments.

A large number of tweets are posted each day, and even a very small fraction of them can make tweet data huge in the context of a given city and can cause information overload for users [101][139]. Users post tweets on a wide range of topics but a large number of users are interested in a limited set of topics. It can be a challenging task to find an important and interesting event from Twitter [143] due to sheer volume of tweets and thus categorization of tweets into topic categories is important as it will help users

to find information on topics that are of interest to them. Further, comprehending meanings and patterns from such a dataset, for example, related to topics, sentiments and emotions can be a challenge. The use of appropriate visualization can play a critical role in mitigating such challenges to a large extent [110]. A good visualization can enhance human interaction with data [93], helps in synthesizing and learning about patterns or trends in data on different facets such as temporal and spatial [84], enhances decision making [130] and "increase social awareness and discourse by exposing underlying patterns in data that is submitted by citizens" [129, p.3461]. This research aimed to develop a technique that would help in identifying the appropriate location of a tweet and thus, show the *Pulse of a City* through better visualization.

1.3 Challenges

There are a number of challenges associated with harnessing themes and patterns, mining emotions and sentiments, detecting topic categories and identifying locations from tweets. These challenges are due to limited text lengths (e.g., 140 characters), data sparsity, use of shortcuts (e.g., 'coz' for 'because'), deliberately or inadvertently misspelled words and use of 'out of vocabulary' words (e.g., 'LOL' for 'Laughing Out Loud'), and limited direct, or specific mention of location in a tweet text [44].

In the context of location detection, Chang et al. [22] and Inkpen et al. [44] noted that extracting geographical location from tweets is a very challenging task due to multiple reasons such as location-related data sparsity, particularly having limited information related to a specific city name. When users refer to location, they include varying levels of granularities such as: 'Whyte Ave' or '82 Ave' (the street name or number) in Edmonton; 'NAIT' (Northern Alberta Institute of Technology), a polytechnic institute in Edmonton; or include an incorrect/misspelled place name (e.g.

'St. Alberta' for 'St. Albert'). Cheng et al. [24] in their research randomly selected a sample of one million users using Twitter and found out that in one million sample of tweets "only 26% have listed a user location as granular as a city name (e.g., Los Angeles, CA); the rest are overly general (e.g., California), missing altogether, or [had] nonsensical location information (e.g., Wonderland)" [24, p.759] or even imaginary locations like Narnia, a country in a fantasy novel [66].

An argument can be made that the information, included in the metadata record associated with posted tweets, can be used to identify location relevant to that particular tweet. Researchers such as Watanabe et al. [135] noted that only 0.7 percent of tweets are geo-tagged and the metadata record associated with posted tweets is not good data for location identification of a given tweet. Their initial assessment highlighted that the metadata for each tweet has primarily two types of geolocation information: a) geolocation information provided by users in their profile, and b) the geolocation captured by the application when a user posts the tweet. Neither of these geolocations recorded in metadata may be relevant to location discussed in tweet content. This issue can be examined using a scenario described in the following paragraph.

John (a hypothetical user) has a Twitter account. John resides in St. Albert but has added Edmonton as the user location in his Twitter profile. Currently, John is traveling to Toronto. He is sitting in a restaurant and watching a hockey game on TV played in Calgary and tweets about it – "Just watched an amazing game by Calgary #Flames played @TheSaddledome #YYC". The posted tweet will have two geolocations in metadata – 'Edmonton' from his Twitter profile and 'Toronto' captured due to the location of user at the time of positing of the tweet. Based on this scenario, Calgary is actually relevant (or event-related) location, while the other two geolocations captured

in the metadata record are not relevant to the content of the tweet posted by John. This scenario re-iterates the Wantanabe et al.'s [135] point that the location information in metadata records is not relevant to a tweet's posted content.

1.4 Proposed Research Work and Thesis Statement

In order to contextualize the on-going discussions on social media (Twitter), and to achieve maximum benefit of shared information relevant to a given location, it is important to accurately identify location by minimizing the above noted challenges, including reducing the over-reliance on geolocation information captured in tweet metadata. In a number of cases, a tweet content will have relevant, contextual location information which can be harnessed to identify appropriate location (or city) of an event that a user is referring to in his/her tweet. One of the key assumptions here is that a tweet has some explicit and/or implicit information related to a city associated with an event. The explicit information could be the name of the city itself and the implicit information could be an indirect reference to a city by including names of key people, places and/or organizations that are associated with the city. For example, the tweet by John (as noted in the previous sub-section) talks about an event happening in a specific location by use of specific terms relevant to the city of Calgary i.e., hockey team (#Flames) playing in the hockey arena (@TheSaddledome) in Calgary (Calgary, #yyc – an airport code for Calgary).

Warf and Sui [134] noted that "we are rapidly entering a new age of the metaverse – virtual worlds that serve as digital equivalents to the atom-based physical world" (p.202). Kindberg et al. [56] noted that the information on the Internet portrays our physical world but "there are few systematic linkages to real world entities". The authors argued that "the physical world and the virtual world would both be richer if they were

more closely linked" (p.935). Drawing upon the viewpoints of [56] and [134], it can be argued that a geographical location, such as a particular city, has the potential of being represented in the virtual/digital world by multiple facets. The proposed research does not aim to use geolocation information available in the metadata record associated with individual tweets but proposes a novel approach by creating a linkage between the digital world and the physical world.

DigiCities i.e., the digital avatar of real world cities, is represented by facets such as People, Organizations, and Places (POP) on social media platforms, including on Twitter. This research developed a novel approach, labelled as *'DigiCities'* which harnesses information associated with the different facets of the 'POP' framework. For example: the people facet of the POP framework is represented by the mayor of a city (Don Iveson is the Mayor of Edmonton); the organization facet of the POP framework is represented by the city's public library (including its branches) (Edmonton Public Library in the City of Edmonton), and; the place facet of the POP framework is represented by a popular public park (e.g., Fort Edmonton Park in Edmonton). These elements, when combined together, would provide a real world geographical location (i.e., city) and its digital presence on a social media platform such as Twitter. The POP elements in the real world have names and/or some identifying values or ids. These POP elements also embody digital names or identifications in the digital world. For example: people, organizations and places now have digital names as well on social media sites such as Twitter and they are represented by user-ids (which starts with '@') and hashtags (starting with '#'). The proposed research work primarily focuses on accurately classifying location(s) using information in tweet content (e.g., use of hashtag representing a particular location such as #yeg – an airport code for Edmonton).

Further, in order to gain maximum benefits of shared information relevant to the given location, and also to benefit from the improved location-relevant tweet identification, this research also endeavours to understand, what we labelled, the *Pulse of a City*. The goal of identifying the tweet-based *Pulse of a City* is to gain insight into different emotions and sentiments, and associated topics emerging over a period of time. Considering the volume of tweets posted in the context of a city makes it very challenging to draw meaning and identify pattern associated with the *Pulse of a City*. Thus, this research also aims to develop appropriate visualizations so that the relatively accurate pulse, from the perspective of emotions and sentiments, can be identified and associated with particular topics of interest.

1.5 Thesis Statements

Considering numerous benefits and challenges associated with the identification of location relevant to tweets, the research presents a novel approach of *DigiCities* that will significantly improve the identification of location relevant to tweet which will further help in providing a more accurate reflection of the *Pulse of a City*.

- #1: This thesis proposes a novel approach labelled as '*DigiCities*' which uses a POP Framework to identify location relevant to tweets.
- #2: *DigiCities* i.e., real world cities in digital environment including social media platform such as Twitter is represented by three key facets i.e., People, Organizations, and Places (POP).
- #3: The proposed novel approach is tested by using three classifiers i.e., Naïve Bayes (NB), k-Nearest Neighbour (kNN) and Sequential Minimal Optimization (SMO),

and it significantly improves the identification of location relevant to tweets as measured by the classification accuracy score.

- #4: This thesis work presents facets, including topics categories, emotions and sentiments, to get the *Pulse of a City*.
- #5: The right location and tweet pairing using *DigiCities* and the use of appropriate visualizations help in developing a relatively accurate *Pulse of a City* as expressed by users of a geographical location in their tweets.

1.6 Contributions

As noted above, this research aims to identify location, relevant to a tweet, by harnessing content included in tweets, and thereby enhancing classification accuracy of tweets into relevant location categories. Once the relevant location is identified, this study further aims to visualize users' emotions and sentiments with the associated keywords and topics discussed by them. This research makes multiple contributions:

- It develops and presents a novel approach, *DigiCities*, which uses elements of the POP Framework to map representation of real world locations in the digital world.
- The proposed approach helps in mitigating some of the challenges, such as the data sparsity problem associated particularly with explicit naming of location, in tweet content. This approach helps in feature convergence, all the entities represented in the POP framework will converge to one semantic concept i.e., a city name, and thereby reducing the data sparsity problem.
- This research proposed two types of strategies for feature convergence and these include append strategy and replace strategy. The append strategy would add the city name after the identified POP element in the tweet and the replace strategy

would replace the POP element with the city name. The research compares results using both strategies to identify which approach would be better suited for managing the data sparsity problem.

- This research aids in increasing location-based classification accuracy even with the use of traditional classification algorithms (e.g., NB, kNN, SMO).
- The thesis work identifies different facets, including topics categories, emotions and sentiments, to get the *Pulse of a City*. The research work used eight (including 'Others') high-level topic categories for categorizing tweets, nine emotions and three sentiments which were identified based on tweet content. This research creates a number of visualizations which show temporal patterns of different facets of the *Pulse of a City* i.e., topic categories, emotions and sentiments as expressed by Twitter users in a city. The analysis of sentiments and emotions related visualizations further strengthens the rationale for using the proposed *DigiCities* approach to find a more accurate reflection of the *Pulse of a City*.
- This research also evaluates and compares the use of Google Knowledge Graph and WordNet in identifying topic categories for tweets.

1.7 Organization of Thesis

This thesis work has a number of chapters focusing on different facets of the research work; the following paragraphs provide an overivew of each chapter.

Chapter 2 – Literature Review

This chapter contains a review of the accumulated research with a focus on several topics including geolocation detection, topics and trend detection, emotion and sentiment analysis, epidemiology, topic categorization, Google Knowledge Graph and

WordNet, and visualizations. The chapter will consider the research work done by the leading researchers in these domains and will present salient points including a discussion on the use of techniques and methodologies, as well as a more thorough discussion on some of the key findings from their papers.

Chapter 3 – Research Questions and Methodology

This chapter is divided into multiple sections. The first section provides an overview of the chapter while the second section focuses on research objectives. The third section provides a detailed discussion on the proposed approach, *DigiCities*, the use and the operationalization of the different facets of the POP framework. The fourth section in the chapter provides step-by-step insight into the creation of digital profiles of cities. The fifth section discusses feature management and the two strategies used in feature management. The sixth section provides information related to the overall experimental design setup to deduce the location from relevant tweets. This section presents details related to Twitter data used in the research work, the pre-processing done on the dataset, software applications used to run the classification experiments, and methods used to analyze classification results. Finally, the last section discusses details related to the visualizations, including the algorithms used to generate the *Pulse of a City* (i.e., topics, emotions and sentiments) and the identification of relevant topics to the cities, as well as the applications used to create visualizations.

Chapter 4 – DigiCities: Implementation of the POP Framework

This chapter has three key sections. The first section provides an overview of the city profiles developed as represented by the POP elements on Twitter. The second section provides insight into different datasets generated and the number of times city names were appended/replaced in the tweets related to different cities. The third section

presents statistics emerging after the implementation of the digital profile on the tweet-dataset. The statistics presented in this section include the total number of times city names were appended or replaced in the tweet dataset of different cities, the number of unique and most commonly occurring terms from the digital profiles of different cities.

Chapter 5 – Experimentation Results and Discussion

This chapter discusses findings emerging from the classification experimentations related to the location identification, and includes the impact of the approach on the tweet classification followed by a discussion of results emerging from the implementation of append and replace strategies, as well as the impact of having or not having stopwords, and/or implementing and not implementing stemming on the dataset.

Chapter 6 – Emotions and Sentiments Visualizations

This chapter provides results and discussion related to the identification of emotions and sentiments, and topics from the dataset. This chapter provides screenshots from the visualizations that were developed through applying the *DigiCities* approach; the included screenshots of visualizations are based on a few scenarios such as conducting temporal reviews of different emotions as well as sentiments in a city, and; comparing sentiments as well as emotions between two cities over different periods of time.

Chapter 7 – Topic Categorization and Visualizations

This chapter provides results and discussion associated with the categorization of sample tweets into topic categories using a manual approach and automated approaches using Google Knowledge Graph and WordNet. This chapter also provides sample screenshots from the visualizations created to present visuals of topic categories.

Chapter 8 – Conclusion and Future Work

This final chapter of the thesis focuses on summarizing the proposed approach, reflecting on key findings and the key contributions of the study. This chapter also discusses key limitations and challenges associated with this work and ends with identifying potential future work directions.

Appendices

This section contains a number of result tables including confusion matrices obtained from the classification experimentations.

Chapter 2

Literature Review

2.1 Overview

Social media platforms provide opportunities to consumers to connect to digital environments with other users, friends or community members. There are a number of social media tools available such as Facebook, Instagram, LinkedIn, Snapchat and Twitter that are used for multiple purposes. Twitter is one of the popular social media platforms with a user base of over 300 million consumers [114]. It is a micro-blogging application launched in 2006 [74] and allows users to post messages (known as tweets) of up to 140 characters [49]. Recently, Twitter has changed the limit on the number of characters from 140 to 280 that can be used to post tweets [74].

Twitter is popular among all user groups including individual users, organizations, both in the private and public sector (e.g., government), and among researchers. Users use this platform for a variety of reasons such as identifying information in real-time [64], sharing diverse information (e.g., opinions and events happening in a city), and learning about current local traffic and weather conditions [113]. Organizations specifically use this platform to create outreach with customers to promote organizational products and services, and to manage customer experiences [103].

Researchers from different disciplines are interested in exploring the Twitter domain from various perspectives. For example, researchers are using Twitter data to conduct research in areas such as geolocation detection (e.g., [24][34]), trending topic

identification (e.g., [51]), sentiment analysis and opinion mining (e.g., [26], [39] and [108]), epidemiology (e.g., [2], [106] and [112]), topic categorization ([101][139]), and visualizations ([38][125]). These areas are relevant in the context of this research project and the following sub-sections will highlight the research work that has been done, which includes approaches used in the above noted areas.

2.2 Geolocation Detection

Researchers have highlighted the issue of location sparsity in social media data, and in particular, Twitter data [24][63]. Researchers such as Cheng et al. [24] and Lee et al. [63] further noted that geolocation detection is challenging to solve in the context of Twitter. Further, there is limited geolocation information associated with a tweet in its metadata, and only a limited number of tweets would have correct geolocation information included in a tweet's metadata records. For example, Graham et al. [34] collected over 19 million tweets over a period of nineteen days in 2011 and found that only fraction of tweets (~0.7%) had geolocation information which is primarily either from users' devices or through users' Internet Protocol addresses. Further, such geolocation information captured in the metadata may not be relevant to topics discussed in the posted tweets. Similarly, Lee et al. [63] noted that 0.58% tweets out of 37 million tweets posted each day are geo-tagged. This problem is further compounded as extracting any information about location from tweets is complicated by a number of reasons such as noise in the dataset, limited numbers of terms in the data due to character limitations imposed by the platform and the use of Out Of Vocabulary (OOV) terms [24]. Researchers are investigating geolocation related aspects of Twitter context. A number of papers have been published in this area, and some of these are discussed in the following paragraphs. The discussion will also reflect on the approach used by

authors in detecting geolocations, and that would help in establishing the novelty of the proposed approach.

Paradesi [89] used a multi-step framework to detect location and to disambiguate geo and non-geo locations relevant to tweets. This research work implemented Part of Speech (POS) tagging with a focus on noun phrases. The author identified noun phrases which were tagged with location names by drawing upon the location-name data from the United States Board on Geographic Names maintained by the U.S. Geological Survey (USGS) (Step1). The tagged labels were then distinguished (i.e., geo location vs. non-geolocation) if the noun phrase is referencing to a location or to a non-location by checking for spatial indicators prior to noun phrases (Step2). The outcome of Step2 helped in handling disambiguation between two geo locations (i.e., geo/geo disambiguation) (Step3). Their implementation of Step2 and Step3 improved the result over the implementation of Step1 only.

Researchers (such as Davis et al. [27], McGee et al. [77] and Li et al. [71]) investigated ways to harness the strength of social network relationships of users in Twitter to detect locations of users. Davis et al. [27] used the phenomenon of reciprocal relationship features created due to the follower-following model implemented in Twitter to detect users' locations. Their approach is built on the premise that users have some reciprocal relationships as followers and followee on Twitter, and such relationships exist for various reasons including for networking and information sharing purposes. They used the information created through the reciprocal relationship model to deduce a user's location based on the locations of other users in his/her network. Their work on identifying geolocations was relatively more suited to produce location information in the metadata record associated with a user's profile, and relatively less relevant to the

identification of the location discussed in the tweet's content. Similarly, McGee et al. [77] focused on harnessing the power of social network. They used information embedded in users' social network to predict the location of users based on location information of other users who are in their network. Their approach was a multi-step approach starting with the identification of factors that could play important roles in location detection (e.g., number of followers, level of interaction among users). This was followed by the implementation of a decision tree to identify pairs of users closer to, and pairs of users farther from one another, and this step was then used to predict users' locations using a maximum likelihood estimator. Also, Li et al. [71] harnessed the power of the social media network which they labelled as the 'following network' along with the content of tweets to detect location of users. This research has dual focus, identification of multiple locations which are also long term locations of users, (and not the temporal locations i.e., places the user is traveling for short durations), and identification of users' different relationships in connection to different locations.

Authors such as Chang et al. [22], Cheng et al. [24] and Hong et al. [42] focused on exploiting the variations in languages and terms used in tweets by users in different geographical areas. Cheng et al. [24] analyzed tweet content terms to detect location relevant to tweets. Their research primarily focused on the context and paid little attention to geolocation included in a user's profile or the metadata record associated with tweets. The foundation of their research work was on the idea that certain terms will be more 'local' as compared to other terms i.e., some terms would associate with a geographical location more than others would. For example, "howdy" which is a typical greeting word in Texas, "may give the estimator a hint that the user is in or near Texas" (p.763). They used location estimation algorithm to identify 'local' words, and they

implemented additional refinements (e.g., Laplace, Lattice-based neighborhood smoothing) to further improve 'local' words for different geographical locations. They produced multiple possible locations with varying confidence levels, and were able to approximate location of "51% of [5,190 users in their test data] Twitter users within 100 miles of their actual location" [24, p.767]. The authors shortlisted users who were active, have posted over 1000 tweet and have included their location using latitude and longitude coordinates. They filtered users such as spammers and promoters from their user consideration set leading to a shortlist of 5,190 users and around 5 million tweets posted by them.

Similarly, Hong et al. [42] focused on harnessing term diversity due to variability in topics discussed in different geographical locations. The authors noted that users in different regions of the world might be interested in different subject content (e.g., Holi, the festival of colours in India vs. Halloween in North America), and thus, are likely to have variations in language used while discussing topics on Twitter. Such variations in language and terms used in discussing topics can be harnessed to identify geographical location. The authors proposed "a novel sparse generative model, which utilizes both statistical topic models and sparse coding techniques to provide a principled method for uncovering different language patterns and common interests shared across the world" [42, p.777].

Chang et al. [22] also used the language diversity to identify geolocation relevance in tweets. The premise of their approach was, if the same words are tweeted multiple times from a given latitude and longitude, then it is likely that such words belong to that location. They labelled such terms as 'local words'. They argued that the use of an unsupervised approach over the supervised approach to find 'local' words was better for

a number of reasons such as the difficulty in creating a ground truth dataset for training/evaluation purposes. They proposed two unsupervised-based approaches labelled as 'non-localness' (NL) and 'geometric localness' (GL) to identify local words. In their 'non-localness' approach the authors harnessed the connection of terms with stopwords. The authors "propose to use the stop words as *counter examples*. That is, local words tend to have the farthest distance in spatial word usage pattern to stop words" while their 'geometric localness' approach relied on the assumption that "a local word should have a high probability density clustered within a small area" [22, p.118]. They implemented Maximum Likelihood Estimation (MLE) and Gaussian Mixture Model (GMM) based algorithms to predict location. Like the above mentioned research, Rakesh et al. [98] also focused on using terms included in tweets to help in location detection. They identified tweets relevant to a particular location and also summarized location-based topics. The authors proposed a framework known as the Location Centric Word Co-occurrence (LCWC). This framework analyses users' network information and tweet's content to identify location relevant tweets [98]. The LCWC approach focuses on capturing features from tweets relevant to a geographical location in the form of bi-grams. The authors used bi-gram sequences of terms and established a weighting scheme by calculating the point-wise mutual information (PMI), term frequency (TF), inverse document frequency and the network score of each tweet [98].

Research conducted in the geolocation domain primarily focused on identifying a Twitter user's location at city level. Mahmud et al. [75] also focused at identifying users' locations at multiple levels such as geographical region, city as well as at time zone level. Mahmud et al. [75] used "a dynamically weighted ensemble method to create an ensemble of the statistical and heuristic classifiers" (p.47:12) and their work involved

the use of a geographic gazetteer for location name identification. They did location classification through a series of hierarchical classification steps starting from a high level classification at time zone level followed by province (or state), regions and city levels.

One of the interesting research problems in geolocation detection is the handling of location ambiguities. These are primarily of two types: geo/geo ambiguity and geo/non-geo ambiguity [44]. An example of geo/geo ambiguity–'Memphis' as a location name in Egypt and the US, and 'Waterloo' in Canada and Belgium. Examples of geo/non-geo ambiguity are 'Berlin' as the name of a person and also a location name in Germany, and 'Adelaide' as the name of a person and a location name in Australia. As noted above, Paradesi [89] also focused on these issues, and investigated how to resolve both geo/non-geo and geo/geo ambiguities. Further, Inkpen et al. [44] also highlighted the challenges associated with both types of ambiguities in location detection. They proposed a two-step approach to detect location and to handle location ambiguities. In the first step, they used a Conditional Random Fields (CRF) classifier using different features (e.g., bag of words, parts of speech, adjacent token) to detect location names from tweets, and in the second step, they developed heuristics involving a five-step disambiguation process to handle location ambiguities [44].

Xia et al. [136] highlighted the importance of user location detection in real-time (or nearly real-time). They used an interesting approach by using data from external sources (i.e. Instagram data) to identify user location from Twitter content. The authors used data from Instagram along with data from Twitter and proposed a three-step framework to detect events happening in a real-time in a city from heterogeneous data extracted from Twitter and Instagram post-streams in real-time. They proposed a multi-step

framework which included event signal discovery, event signal classification and event summarization to detect events location. Their results showed that using and combining data from more than one social media application (Twitter and Instagram in this case) can improve the location detection.

Duong-Trung et al. [29] also developed a tweet content-driven regression model to solve real-time location prediction problem. The authors availed three different tweet datasets and each dataaset had relatively different geographical focus i.e., US, North American and global focus. The authors applied their proposed approach using regression model on these datasets and compared the results with results obtained using Linear regression model, SVM and Factorization Machines. They conlcuded that their proposed approach did relatively better than all the three other approaches.

2.3 Twitter Trends and Classification

Twitter trends are not just a way of saying what is trending at present in a given geographical location but they are also a way to understand what people (or a group of users) are thinking in a given geographical location during a particular time period. A number of researchers have explored the trending of topics on Twitter and the following paragraphs will discuss a number of them with some reflection on the different methodologies used in that research.

Petrovic et al. [94] proposed the use of a modified locality sensitive hashing (LSH) approach for first story detection (FSD) on Twitter. The premise for this research was that if something noteworthy is happening in a given area, people (in this case Twitter users) will discuss that event a bit more. Among other findings, they suggested that

news related to celebrity deaths are the fastest to proliferate on Twitter [94]. The FSD approach can help epidemiologists to identify or pin point a possible outbreak of disease.

Cataldi et al. [21] argued that a large number of conventional clustering and classification strategies may not be useful in identifying emerging topics on Twitter because they do not factor in the temporal relationships among tweets. Thus, they used novel aging theory to create keyword life cycle. This was a unique approach and it helped them to identify "emerging terms by ranking the keywords depending on their life status", and these terms were then used to identify co-occurring terms to get emerging topics [21, p.2].

Kwak et al.'s [59] research work on Twitter data was a multi-faceted research and one of the areas in it was on topic trends on Twitter with a focus on the "topological" features of Twitter. Within the context of topological features, they analyzed tweets that were related to trending topics to get insight into the temporal aspect of trending topics and user participaton in them. The findings include: a large number of (over eight million) users participated in trending topics; nearly 15 percent were very active and participated in over ten topics during a four-month window, and found out that trending topics were aligned with current news headlines.

Researchers such as Rosa et al. [101] and Ozdikis et al. [86] harnessed the power of hashtags to detect events from Twitter data. The use of hashtags was on the premise that they are a good indicator of topics [101]. Rosa et al. [101] reviewed and compared unsupervised and supervised approaches; they found that they got better results with the supervised approach. They also identified that coarse topics (high level themed topics) are easier to detect then fine grained topics (e.g., Sports as high level theme vs. NFL and NBA as fine grained themes). While Ozdikis et al. [86] used agglomerative

clustering algorithm to cluster tweet topics or themes based on semantic similarity of hashtags, the clustering of tweets was used not only to identify topic clusters but also to determine the opinion or sentiment of public on clustered topics. The authors [86] concluded that the use of hashtags to cluster tweets into themes improved the event detection accuracy as compared to the clustering of tweets using whole tweet content (this research was done in prior work by the same authors).

Abdelhaq et al. [1] introduced a novel framework to identify events in real-time, in local context, and proposed a comprehensive framework labelled as EvenTweet for such detection. Their proposed approach was based on "*spatial signature* for each keyword" i.e., a "spatial signature is the spatial density distribution over the usage ratio of a keyword at a particular location" in a particular temporal period (p.1327) . According to Abdelhaq et al. [1], "[k]eywords related to the same localized events tend to show some spatial proximity, meaning that they have similar spatial signatures" (p.1327). They used a "single-pass" clustering approach to cluster keywords relevant to events.

Aiello et al. [4] used multiple methods to identify trending topics and events on Twitter using three different dataset having different characteristcs (e.g., time scale). They used two fundamental approaches and that include document based clustering and keyword based clustering that they labelled as Document-Pivot Method and Feature-Pivot Method. They also evaluated the impact of pre-processing steps including tokenization, stemming and aggregation on the topic detection.

Lu and Yang [73] used a modified version of MACD (Moving Average Convergence-Divergence) indicator to analyze trending news topics. MACD is widely used in stock analysis and is known to compute both trends and momentum. The authors wanted to identify what types of topics do trend. They focused on identifying reasons for emerging

change in the topic trends and noted that a topic may start trending up if a topic gets connected with some other new event or is picked by influential users, and may start trending down if other topics are picking up and attracting the attention of a large number of users.

Naaman et al. [83] created a typology of tweets using multi-level dimensions such as exogenous trends (e.g., broadcast-media events, local participatory and physical events) and endogenous trends (e.g., memes, retweets). The authors identified a number of features which they used to identify and create characteristics for typology. The features were divided into five broad categories and they were: content features (e.g., average number of words/characters, hashtags, etc.), interaction features (e.g., retweets and replies related information), time-based features (e.g., exponential fit), participation features (e.g., no. of messages by each author), and social network features (e.g. level of reciprocity). Li et al. [70] also used features associated with Twitter to identify tweets for relevance to crime and disaster related events (CDE) and to predict location in context of these CDE tweets. The features they focused on include content features with a focus on having URL and specific terms (e.g., death), user features which focused on data points such as account age and the number of tweets associated with a user, and usage features with a focus on particular data such as specific hashtags in different tweets.

In predicting trends on Twitter, time series analysis is done on the various facets of past events to compute estimated future trends of a topic, but researchers like Gupta et al. [35] proposed to use various versions of regression models, such as linear regression, auto regression (AR), auto regressive moving average (ARMA) and vector auto regression (VAR) to "capture time dependencies (periodicity and trend)". The regression

model however can "capture time dependencies [...] but cannot learn across instances" (p.7) which can be done by classification models, and these models have their own problems. Thus, the authors [35] proposed a hybrid approach which uses both a regression model and a classification model to predict popularity trends.

Twitter trends suggests what is going on now and what users are talking about. However, spammers are harnessing the benefit of this social feature and spam tweets, which makes it difficult for users to separate topics that are organically trending and topics that are trending due to spamming. To mitigate this issue of trend-stuffing, Irani et al. [46] noted that a number of tweets have (a) link(s) to external webpage(s) and they used these external website links to create training models for algorithms. The authors problematized it as text classification problem and used multiple classifiers such as Naïve Bayes, C4.5 Decision Tree and Decision Stump implemented in Weka to conduct classification experiments. They developed three different training models using text from tweets and linked websites to identify organically trending topic tweets and trend-stuffing tweets. The first training model was developed using text from tweets alone, the second model was developed using text extracted from web page(s) whose link(s) were inlcuded in tweets, and the third model was developed by combining both text from tweets and text extracted from web page(s).

Other researchers such as [30] and [62] have used some other approaches to detect trends on Twitter. Dykov and Vorobkalov [30] used grammatical relations in tweet text to detect trends on Twitter. Lau et al. [62] used a topic modelling based approach to identify trending topics. Lau et al. [62] believed that approaches like using simple keywords and hashtags helped in identifying the trending topic but could not provide users with details on the trending topic. There were other challenges associated with

trending events or topic detection. For example, the same event might have multiple names as well as the different users are reporting an event from different perspectives. Such variations in names and perspectives create ambiguity around the event which might create problems in proper event detection. Zhou and Chen's [144] work focused on handling such ambiguities in order to further enhance event detection. They "proposed a novel graphical model called location-time constrained topic (LTT) to capture the social media data information over content, time, and location, and describe each message as a probability distribution over a number of topics" (p.382).

2.4 Tweets Sentiment Analysis and Emotion Mining

Sentiment analysis is "used to extract opinions, sentiments, and subjectivity in unstructured text" leading to mood identification of being favourable (positive sentiment) or unfavourable (negative sentiment) towards a topic or subject [7, p.2524]. Emotion mining focuses on extracting emotions from the data and the emotions could be expressed in these following states i.e. "anger, fear, sadness, enjoyment, disgust and surprise" [31, p.170]. A number of studies have been conducted in the area of sentiment analysis and emotion mining such as to identify the writers' moods or opinions towards a topic [26] or public sentiments in context of stock market [18]. Both sentiment analysis and emotion mining from tweets can be a challenging task, for example, due to presence of more than one emotions or both negative and positive sentiments in a tweet [14]. Researchers (such as [3], [8], [10], [11], [14], [26], [49], [87] and [122]) have used different approaches to enhance emotion and sentiment detection from Twitter data and the following paragraphs will reflect on research diversity in these areas.

Pak and Paroubek [87] used three sentiments, 'positive', 'negative', and 'neutral' instead of two categories. The focus of authors [87] was to evaluate the impact of various

n-grams models (e.g., unigrams, bigrams and trigrams) on sentiment analysis and they used Multinomial Bayes, SVM and CRF (Conditional Random Field) based classifiers in their research work. Pak and Paroubek [87] claimed that they got the best sentiment identificaton results by the combined use of bigrams and the Multinomial Bayes classifier.

Davidov et al. [26] proposed a supervised sentiment classification framework which contain four features including "single word features, n-gram features, pattern features and punctuation features" (p.243) and used the similar approach like kNN algorithm. The authors created sentiment labels based on "50 twitter tags and 15 smileys as sentiment labels" drawn from the Twitter data (p.241), and used these to train their algorithm. The results shows that a high number of co-occuring tags contain contrasting sentiments and all the features used such as n-grams, punctuations and words adds to the classification accuracy.

Jiang et al. [49] worked on Twitter sentiment classification and they identified sentiments as positive, negative or neutral towards its target in tweets. They used "syntactic features to distinguish texts used for expressing sentiments towards different targets in a tweet" and also incorporated the context of the tweet (e.g., what were the sentiments in tweets prior to and after a given tweet) [49, p.159]. Agarwal et al. [3] also conducted research in the area of sentiment analysis on Twitter data. The authors used three different models i.e., a unigram model which was the baseline model, a feature based model which used features drawn both from earlier research work and 100 new features (not included in earlier studies), and a tree kernel based model which they specifically designed for this work. Their findings suggested that both the feature based model and the tree kernel model fared better than the unigram baseline model.

Barbosa and Feng [10] developed an automatic sentiment detection approach that would use syntactic features (e.g., retweets, hashtags, and emoticons) and meta-features (e.g., part of speech) of words included in a tweet. The proposed approach was a two-step classification process which includes classification of tweets into subjective and non-subjective categories in the first step, followed by classifying tweets that were identified as subjective tweets into positive or negative sentiments. They used various Weka's learning algorithms and found out that SVM-based algorithm worked better for them (though the names of other algorithms were not explicitly listed in the paper and they shared results from the use of this algorithm only) and also claimed that their approach works on dataset having noise and bias [10]. Kouloumpis et al. [57] conducted an interesting experimentation to detect sentiment from tweets. They problematized it as classification problem. First they created traninig datasets, labelled as 'HASH' dataset and 'EMOT' dataset to train their classification algorithm on sentiment categories (e.g., postive, negative). They identified a number of top hashtags from the Edinburgh corpus and assigned each hashtag to a sentiment class i.e., positive, negative and neutral. They then used these polarity-labelled hashtags to identify sentiment of an individual tweet i.e., hashtag in tweet were used to determine tweet's polarity. The other training dataset, labelled as 'EMOT' dataset was based on emoticons developed for another project at Stanford Univeristy. They evaluated the impact of different features such as n-gram (e.g., uni-gram and bi-gram), lexicon feature, Part-of-Speech (POS) and other features which they labelled as micro-blogging features (e.g., emoticons and abbreviations) on sentiment identification from tweets. The authors [57] noted that the best results were achieved with the use of n-grams along with the use of both the lexicon and microblogging features. They did not recommend the POS features for sentiment

analysis but suggested that more research is needed to further evaluate POS feature in sentiment analysis.

Bae and Lee [8] used sentiment analysis approach to measure the popularity of key or influential people with Twitter accounts by analyzing tweets made by such influential people and their followers. The authors used a lexicon-based approach which uses a word-based matching approach derived from the list of words which have been pre-assigned polarity (i.e., each term in the list is associated with positive or negative sentiment). They used the Linguistic Inquiry and Word Count (LIWC) pre-coded dictionary and three correlation methods including Pearson correlation analysis, Spearman rank Correlation analysis and Granger Causality analysis.

Continuous streaming of tweets indicates many emotions every second. Measuring people's emotions could be used in assessing their overall well-being. Identifying the emotional state of one's mind could be helpful for users working in different professions such as healthcare professionals and counselling agencies [39]. A number of researchers (e.g., [9], [17], [57], [61], [100], [116], and [119]) worked on mining emotion patterns from Twitter data, and a few are discussed in the following paragraphs.

Balabantaray et al. [9] did linguistic analysis of Twitter data for opinion and emotion analysis using SVM-based classifier. The authors [9] aimed to identify basic emotions (e.g., happiness, anger, surpise and fear), embedded in tweets, which are related to facial expressions as noted by Ekman [31]. They [9] used a large number of features such as unigrams, bigrams, POS, personal-pronouns and adjectives in their emotion identification and classification experimentation. They were able to achieve an accuracy of over 70% in emotion identification and claimed that this accuracy score is relatively higher than scores presented in some of the other earlier works.

Bollen et al. [17] citing McNair, Lorr, and Droppleman (1971) suggested six mood states i.e. tension, depression, anger, vigor, fatigue, confusion. They measure the sentiments using Profile of Mood States (POMS) and suggested that sentiment analysis of micro texts, such as tweets, does not require machine learning approaches and could be obtained by a syntactic or term based approach. Machine learning approaches are good when a large data set is available. The authors concluded that public mood changes, with happenings in their surroundings, whether they are social, political or cultural [17].

Hasan et al. [39] proposed an automated approach 'Emotex' to label tweets based on the emotions they contain, and claimed that they achieved high accuracy in labelling their emotions. The authors [39] also take the high dimensionality problem of twitter data into account by considering only emotional words from lexicon LIWC (Linguistic Inquiry & Word Count). Emoticons could be replaced by emotional words or emotions expressed by the writer. Soranaka and Matsushita [116] worked to identify the relationship between emotional words and emoticons in tweets. They worked towards understanding whether the emotion used by the sender is consistent in the understanding of the receiver and found that often there was mismatch between sender's intentions and receipient's interpretations. In their findings, they observed that there was mismatch between emotions emerging from analysis of terms in tweets and emotions emerging from analysis of emoticons.

Larsen et al. [61] developed a system "We Feel" which analyses emotions expressed through Twitter. This tool accesses data from Twitter API and location is considered based on the time zone stated in a user profile. The authors followed three steps, first, they displayed data based on day-to-day variations, and then they link data to see if

patterns are stable. The third step is to determine if there is any substantial change in the mood with the changing subject on Twitter. To identify this association, the authors used PCA (Principal Component Analysis) which is a data driven approach. The authors also provide insight into how these emotional tweets could help in understanding the anxiety or mental health of the community as a whole.

Mohammad and Kiritchenko [81] used hashtags and showed that dataset made up of hashtags will help in automatic detection of emotions in tweets, as well as in personality detection. The authors [81] used word – emotion associated lexicon and gave a score to each word associated with the emotion. The higher the score, the higher the association is. Wang et al. [131] harnessed the power of hashtags found in Twitter data. They highlighted that most of the studies lack accuracy because of relatively smaller datasets. However, to overcome the problem of smaller datasets, the authors automatically created a large dataset using hashtags related to emotions from Twitter data. They used two machine learning algorithms LIBLINEAR and Multinomial Naive Bayes and found out that the best emotion and sentiment results can be achieved by combining different features such as unigrams, bigrams, POS (part of speech), and emotions and sentiments relevant terms.

Wang et al. [132] conducted an interesting study to identify trends in work-related emotion and stress, and how people recover from such stress. The authors used context analysis and the LIWC (Linguistic Inquiry and Word Count) technique in this research. The weekly analysis of emotion and stress as posted on Twitter showed interesting but predictable patterns. They found that Mondays, the beginning of the week, had the highest negative emotions and stress, and they decreased as the week progressed with a dip on Fridays. Wang et al. [132] also reflected on positive emotions and found that they

were relatively low on Tuesdays, Wednesdays, and Thursdays but increased from Fridays to Sundays. Further, they noted that there was not much difference in trends related to stress, negative and positive emotions in context of work and non-work.

2.5 Epidemiology

Health, among other topics such as politics, economy, sports, travel and physical activity (e.g., Yoon et al. [138]), is widely discussed among users on Twitter. Health is discussed in numerous contexts such as physical fitness, seasonal illnesses and pandemic outbreaks of disease(s) in a given geographical location (e.g., Ebola outbreak). Disease outbreaks can affect the growth of society, but timely intervention by government and health organizations can stem the spread of diseases within or outside of the affected community. They can deploy significant measures to contain disease outbreaks if initial signs of outbreak are caught early enough. It is believed that users' conversations on Twitter can help in detection of disease outbreak. For example: Researchers such as [2], [60], [55], [104] and [112] have focused on predicting or assessing pandemic situations by analyzing users' conversations on Twitter.

Achrekar et al. [2] developed a system called Social Network Enabled Flu Trends to predict the spread of flu in the US using Twitter. The authors compared their Twitter data analysis with the Centre for Disease Control and Prevention (CDC) analysis, and had their assessment of the outbreak confirmed by CDC. Thus, highlighting the importance of Twitter data in predicting flu outbreaks in real-time.

Similarly, Sadilek et al. [104] and Signorini et al. [112] also worked to identify the spreading of flu from tweets. Sadilek et al. [104] proposed that the network feature of Twitter can be used to predict illness with some success. According to the authors, the

probability of a user, John (pseudonym for a user) falling sick will be high if it is observed that the more friends are falling ill in John's social network. They "estimate the physical interactions between healthy and sick people via their online activities, and model the impact of these interactions on public health" (p.323). Signorini et al. [112] tracked information on the specific type of flu i.e., swine flu (or H1N1) on Twitter and predicted the outcomes; they argued that discussion on Twitter can help in identifying community members' concerns around health-related issues. The authors noted that results obtained from analysis of Twitter data related to Influenza can help in tracking disease in real-time which can be faster than tracking of disease done using existing approaches. In addition to this, researchers like Lamb et al. [60], suggests that if we are able to create more distinction between tweets suggesting awareness about a disease, and tweets noting infection or sickness with a disease, then we can better predict the size of outbreaks and estimate when they will occur.

The estimation of pandemic conditions and the use of such analysis for public benefit can be more effective only if the location where such outbreaks are likely to happen is accurately known i.e., if we are accurately able to pinpoint tweets to the right geographical location. Thus, identification of location is equally crucial in outbreak situations for taking the right steps to impede the disease from spreading further.

2.6 Topic Categorization

Users post a large amount of information online on different platforms including social media platforms such as Twitter and Facebook. For example, Twitter users post a combined total of over 500 million tweets every day worldwide (e.g., [7], [127]) leading to approximately 6,000 tweets per second. This is a large number of tweets, even a fraction of these tweets can be overwhelming for users and cause information overload

([101][139]). Users post tweets on a wide range of topics but a large number of users are interested in limited set of topics. Zhou et al. [143] argued that it is challenging to find important and exciting events from Twitter, and thus categorization of tweets into topic categories is important as it will help users to find information on topics that are of interest to them.

Manual categorization is challenging and next to impossible considering the number of tweets posted each day. Researchers are exploring new ways to improve automated categorization of tweets into topics or themes for user benefit. Some researchers focused on categorising tweets in general (e.g., Lee et al. [64] and Rosa et al. [101]) while other focused on domain specific tweets (such as Sutton et al. [120] focused on disaster-related tweets and Hewis [41] focused on MRI Patients tweets). The following paragraphs discusses some work in the area of categorization.

Rosa et al. [101] conducted research in the area of categorization using Twitter data. The authors used pre-determined list of topics such as News, Sports, Entertainment, Science, Money, and "Just for Fun" to categorize tweets. They used both unsupervised approaches (e.g., k-means and LDA) and supervised approach such as Rocchio classifier to classify tweets into categories, and inferred that supervised approach led to a good outcome. Lee et al. [64] focused on categorizing tweets into 18 broad pre-determined categories (e.g. sports, politics, etc.). They used two different supervised classification approaches, the text based approach (which uses Bag-of-Words approach) and the network-based classification approach (which uses social network information such as data on friend-follower network) to classify tweets into pre-determined categories. The authors [64] noted that the network-based approach had slight edge over the text based

approach in categorizing tweets into topic labels. They used manually labelled data to evaluate the outcome from their two automated approaches.

Zubiaga et al. [145] aimed at identifying social trends at the earliest stage. Based on personal experience and observation of data, they created a typology of trending topics and argued that all trending topics could be categorized into four broad categories i.e., news, ongoing events, memes, and commemoratives. Researchers hypothesized that user behaviour in the dissemination of information will be different for different typologies. The authors [145] identified 15 social features (e.g., hashtags, tweet length, links in tweets, retweet related information, etc.) to understand such behavioural variations and to better predict trending topics. They used an SVM based classifier in their experimentation, and concluded that the features they proposed gave more accurate classification results than using the content of tweets.

Sutton et al. [120] focused on exploring the phenomenon of serial transmission on Twitter, and one part of their work focused on thematic mapping of tweets, and these tweets were domain specific. They collected tweets relevant to disaster particularly in the context of the Waldo Canyon fire event. They categorized the disaster-focused tweets thematically into nine primary disaster-relevant themed categories (e.g., closures, advisories). They also included two additional categories which captured tweets that did not fit into any of the primary categories. Unlike researchers like Lee et al. [64] and Rosa et al. [101], these authors did manual coding of tweets for thematic analysis, and to identify disaster-related categories of tweets, which would help other researchers who would be focusing on conducting research in such focused area.

Zhou et al. [143] proposed a framework which used unsupervised approaches to filter noisy tweets out of relevant tweets, followed by event extraction of tweets from filtered

non-noisy tweets, and then categorizing them using a Bayesian model to detect events and further, discover topic categories of events.

Andrienko et al. [5] worked to thematically organize geographically-focused tweets into topic categories. They were interested in understanding users' interests in learning about topics that are confined to a specific geographical city. The authors collected tweets specific to the city of Seattle. It was a multi-faceted research and one of the facets focused on classifying tweets into topic categories (e.g., food, love, family). They operationalized each topic category by identifying a number of keywords that users might use in relation to a topic category (e.g., users may use 'father' and 'mother' for 'family' category), and used this information to categorized tweets into topic categories. Conceptually, our thesis work also uses a similar approach i.e., a number of keywords for each topic category were extracted by querying Google Knowledge Graph (GKG) and was used to identify topic categories of tweets by matching keywords drawn from GKG with terms in each tweets. The following section will discuss work done using GKG and WordNet.

2.7 Knowledge Graph and WordNet

Knowledge Graph term was coined by Google in 2012 [90] and it "store[s] factual information in form of relationships between entities" [85, p.11]. Nodes and edges in a knowledge graph represent topics and relations between topics respectively [51]. A number of implementations of knowledge graphs have emerged over a period of time including Google Knowledge Graph, DBpedia, YAGO, Freebase and Probase ([51], [85] and [90]). A number of studies (e.g., Karidi [51] and Nickel et al. [85]) have used Knowledge Graph for various reasons.

Karidi [51] focused on finding interest based similarity among users. The author used knowledge graph like Google Knowledge Graph, YAGO and DBPedia to construct a Topic Graph which was used to enhance Twitter user's profile using the Steiner Tree and the InterSim algorithm, and to identify topics that are of common interests among users leading to the identification of interest similarly among users.

Zhao et al. [142] conducted research in the Google+ domain. They argued that users have interests in different topics and they show different behavior for different topics. They suggested that if users' multiple profiles can be created for each behavior type, that they show for each topic type, would help in recommending customized content that are of interest to them. They identified four different behaviour types by analysing "Create Post, Reshare, Comment and +1" data (p.1408) (+1 is a Google+ feature of recommending someone). They used Google Knowledge Graph to "extract higher-level semantic concept from the [Google+] post in the form of entities using Google's Knowledge Graph" and used them as topics from a Google+ post (p.1408). Other researchers like Paulheim [90] presented a survey of knowledge graph refinement approaches used to further improve knowledge graphs as they are either not complete or may have errors.

WordNet is a registered tradename of Princeton University and it is a "lexical database for English" [79, p.39] in which words are grouped based on meanings [80]. WordNet has been used by researchers for multiple purpose. For example, the authors such as [12], [32], [68] and [102] used WordNet to categorize textual documents (i.e., non-social media data) into topic categories.

Specia and Motta [117] in their research work used WordNet in the preprocessing stage of their research work which focused on integrating folksonomies and the sematic

web. One of the steps in the preprocessing stage required categorizing "morphologically very similar tags" (which were assigned by users on Flickr and del.icio.us websites) into groups (p.630). Each group representing similar tags was assigned a representative tag. The main criteria of picking a tag (as the group representative) was its presence in the WordNet corpus.

Other Researchers like Hamdan et al. [36] and Chen et al. [23] conducted research in social media domain and used WordNet. Hamdan et al. [36] research focused on identifying sentiments from tweets. They used WordNet to extract "the synonyms of nouns, verbs and adjectives, the verb groups" (p.457) to reduce ambiguity in and diversity of terms, and thereby improving dataset features which would help in improving sentiment identification accuracy. Chen et al. [23] conducted research to compute similarity between tweets so that similar tweets can be recommended to users. They compared LDA and WordNet to identify similarities between tweets for recommendation purposes. The authors claimed that their WordNet-based approach gave better results in finding similarities between tweets.

2.8 Visualization

Social media applications generate a large amount of data that has a lot of informational value. It can be very challenging to process data and analyze findings from a large amount of data such as from social media applications. Visualization can play a critical role in deciphering information and analyzing results from social media data on different perspectives such as emotions and sentiments, and trending topics or trends on specific topics such as health and sustainability.

Card et al. [20] defined visualization as "the use of computer-supported, interactive, visual representations of data to amplify cognition" (p.6). A number of researchers such as Peterson et al. [93], Khan and Khan [54], Shiroi et al. [110] and Valkanova et al. [129] discussed the benefits offered by visualization and noted that visualization has the potential to help in converting large amounts of data into comprehensible forms which enhances aesthetics and human interaction with information, supports easy analysis and increases ability to discern patterns effortlessly. Shneiderman [111] paraphrased his Professor Richard Hamming's quote and stated that, "the purpose of visualization is *insight*, [emphasis added] not pictures" (p.3). The use of the term "insight" suggests that visualization is important for analyzing data, drawing meaning, and gaining critical perspective leading to discovery of patterns and phenomena [110]. A number of researchers underlined the importance of using visualization for social media data and some of them are discussed in the following paragraphs.

Hao et al.'s [38] work focused on analyzing tweets to extract customers' opinions (positive or negative) and visualized their results using calendar view and geo-map. The calendar view visualization included the temporal periods shown in columns and topics in rows, and each cell then showed sentiments which were colour coded using red for negative, gray for neutral, and green for positive sentiment, and the geo-map showed the areas from where users are posting tweets (Screenshot in Figure 2.1).

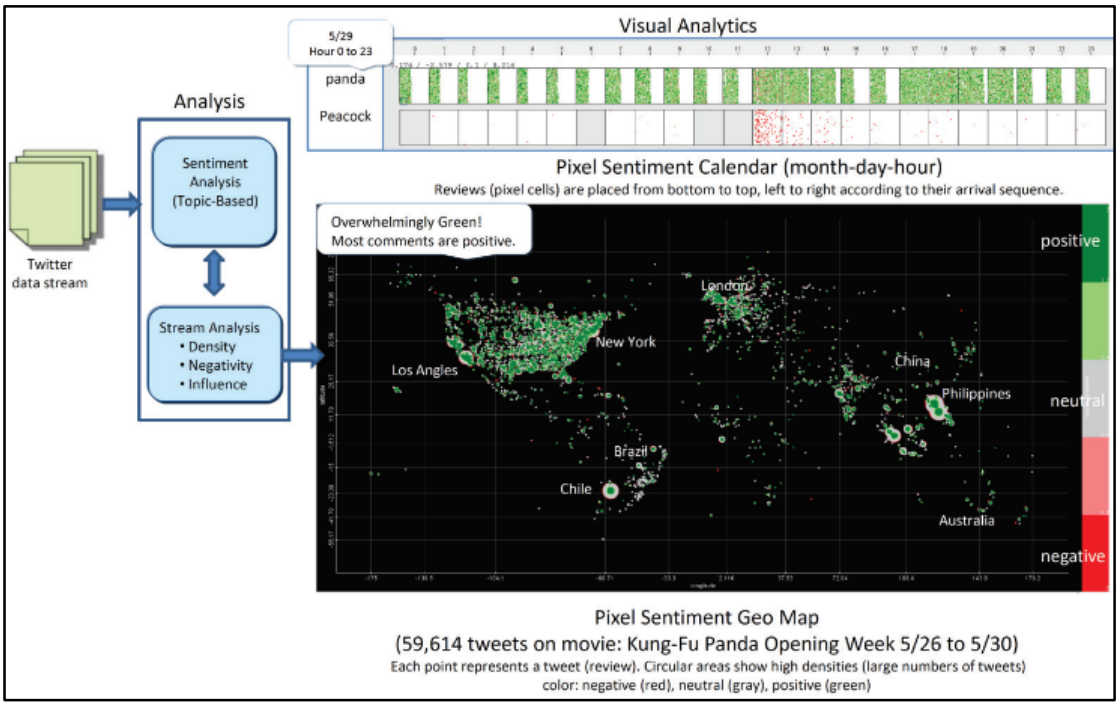


Figure 2.1: Visualization Screenshot from Hao et al. [38]

Similarly, Torkildson et al. [125] also conducted research on Twitter to identify emotions and sentiments, and used visualization for better analysis and to make sense of their dataset. They used stacked area charts to visualize eight different emotions. Each emotion was plotted on a separate stacked area chart and all the charts were then rendered in the same pane so as to have a comparative evaluation. In order to get enhanced clarity of their results, they used different colours for value bands (Screenshot in Figure 2.2). Research in the emotion and sentiment domain would generally focus on aggregating emotions and sentiments on an event or a topic, or a product over a specified period of time, by drawing upon multiple users' tweets.

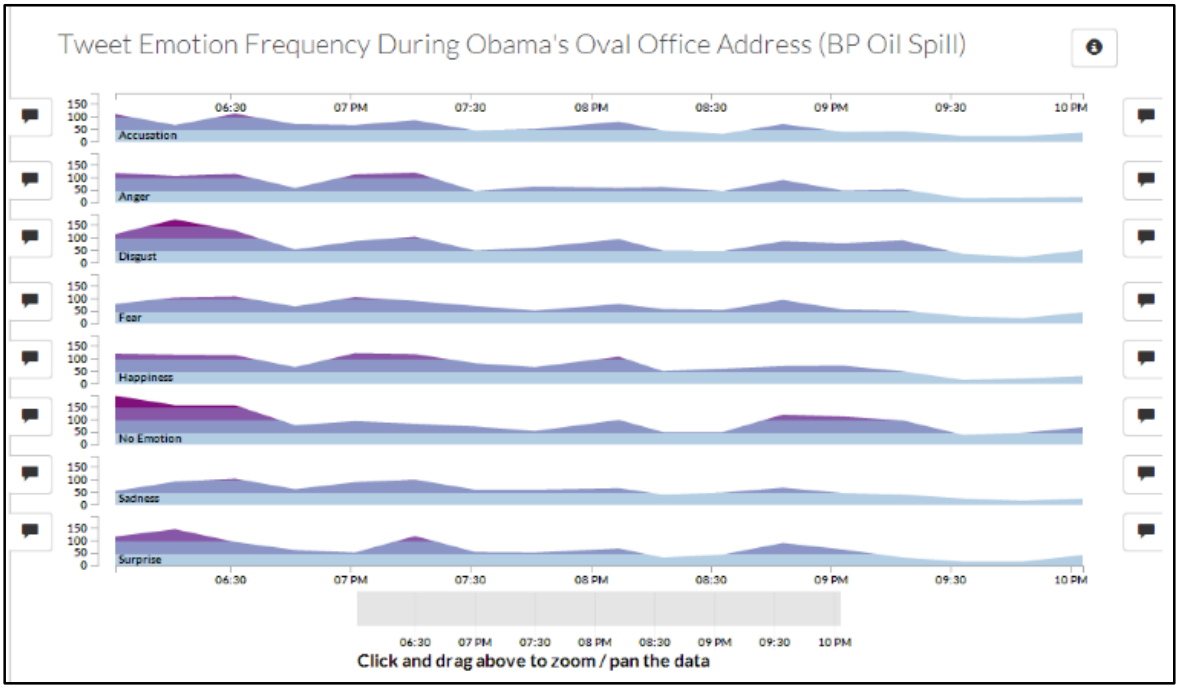


Figure 2.2: Visualization Screenshot from Torkildson et al. [125]

However, emotions and sentiments of users who are posting tweets are rarely analyzed at an individual user level. Zhao et al. [141] identified this gap and primarily focused on extracting and visualizing individual user's emotions over a period of time. They developed a visualization tool (labelled 'PEARL') to show emotions of an individual, based on their tweet postings. The visualization tool was developed using the D3 JavaScript toolkit. Like Torkildson et al. [125], they also used stacked area charts for showing one individual user's emotions over a period of time. The tool was created to visualize an individual user's data from multiple perspectives such as an overview of an individual user's emotional profile, insight into shifts in emotion and mood over a period of time and also to get access to specific tweets posted by an individual user (i.e., the original tweet). The authors argued that they were able to present a huge amount

of information through the use of visualization in a format that is easily readable and understandable (Screenshot in Figure 2.3).

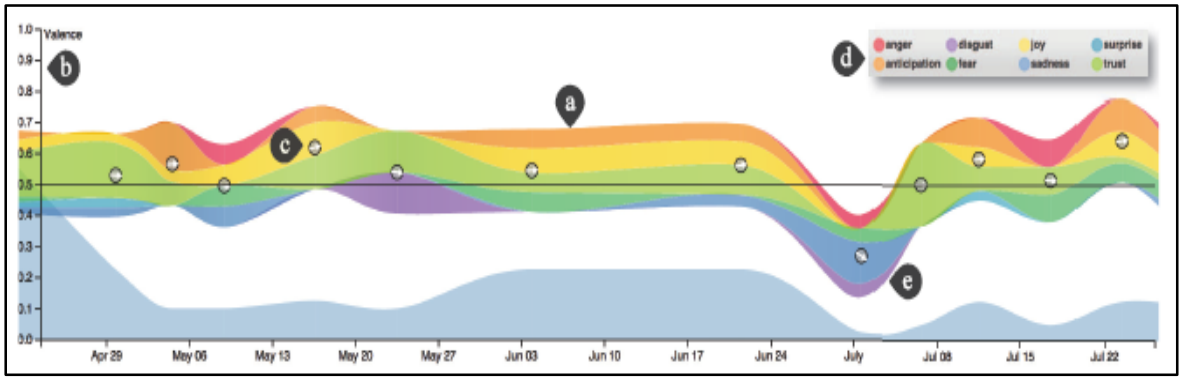


Figure 2.3: Visualization Screenshot from Zhao et al. [141]

Morstatter et al. [82] developed a visualization tool called 'TweetXplorer' which analyzed Twitter data to present information on events that are of interest to a user. The tool was developed using the D3 JavaScript toolkit. The tool allowed users to search for information about an event using multiple keywords (Screenshot in Figure 2.4). The information extracted about a specific event was then used to plot different types of visuals. For example, the tool used a heatmap visualization to show the number of tweets posted that was related to the event in different geographical regions. The heatmap visualization was meshed with the geographical map of a country (e.g., Map of USA) to show the distributions of tweets. The interface allowed users to view tweets that are of significance to an event and to provide a social network of important users i.e., the tool plotted the network graph to identify both important users and their tweets related to the inquired events. By using an on-click feature, additional information such as prominent users' ids, their tweets and hashtags, days and times when they would

normally post and location from where tweets were coming from if they were geotagged. Further, a tweet could also be viewed on a geo-map by using the zooming out feature. The authors used Hurricane Sandy as an example event to evaluate their visualization tool, and noted that their proposed visualization helped in analyzing huge datasets and facilitated in identifying important information relevant to Hurricane Sandy with relative ease.

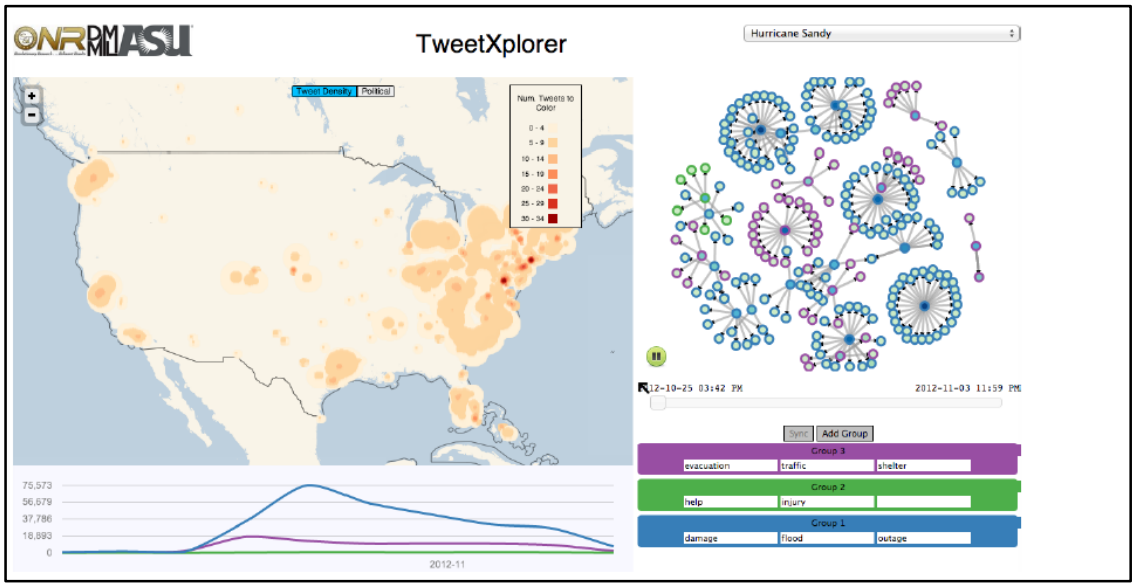


Figure 2.4: Visualization Screenshot from Morstatter et al. [82]

Meyer et al. [78] work also aspired to visualize Twitter data with the focus on identifying important news in real-time. The researchers identified important topics that could make breaking news, particularly on disaster topics (e.g., earthquake, tornadoes, terrorism). The authors used Google map visualization APIs to develop their visualization (Screenshot in Figure 2.5). The application was developed with a number of interesting features such as allowing users to pick a topic from a pre-defined list of disaster topics from a drop-down menu, setting the temporal period with a start and

end date, and showing topics that had tweets above a minimum threshold, as set by the user. The results were displayed on a geo-map in the main window and the side panel displayed the topic and time. The authors, also colour-coded different disaster types (e.g., red for man-made disaster) and depending upon the number of tweets on a topic, the colour intensity will vary. Thus, helping users to process and assimilate information. The visualization also had other features such as a zoom-out feature to further drill down on specific topics or geographical regions (Screenshot in Figure 2.5).

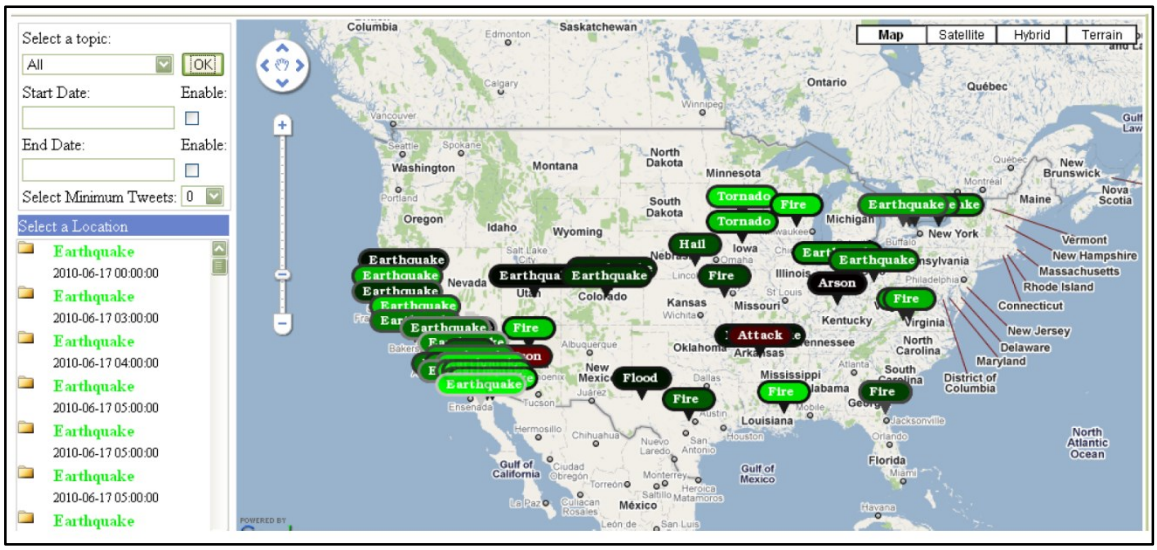


Figure 2.5: Visualization Screenshot from Meyer et al. [78]

Mathioudakis and Koudas [76] developed an interactive visualization tool called TwitterMonitor to detect topic trends from Twitter data. They first identified frequently occuring keywords; and their criteria to identify such keywords were based on the number of tweets posted in a minute on a given topic. The front end of the application was a webpage, which presented the results on emerging trends in real-time. Their visualization had a number of features such as: a list of trending topics with date and

time associated with them; topic details (using the 'more' feature) which included keywords, key locations and sources of information; a graph displaying "the evolution of a trend's popularity" with time stamp on x-axis; a mouse hover feature which showed specific tweets related to the trending topic; and some user engagement by allowing them to "submit their own description" for an event (p.1157). The visualization also allowed users to view not only the recent trends but also the daily trends on topics "that have emerged within the last day, ranked by an aggregate volume of tweets" (p.1157) (Screenshot in Figure 2.6).

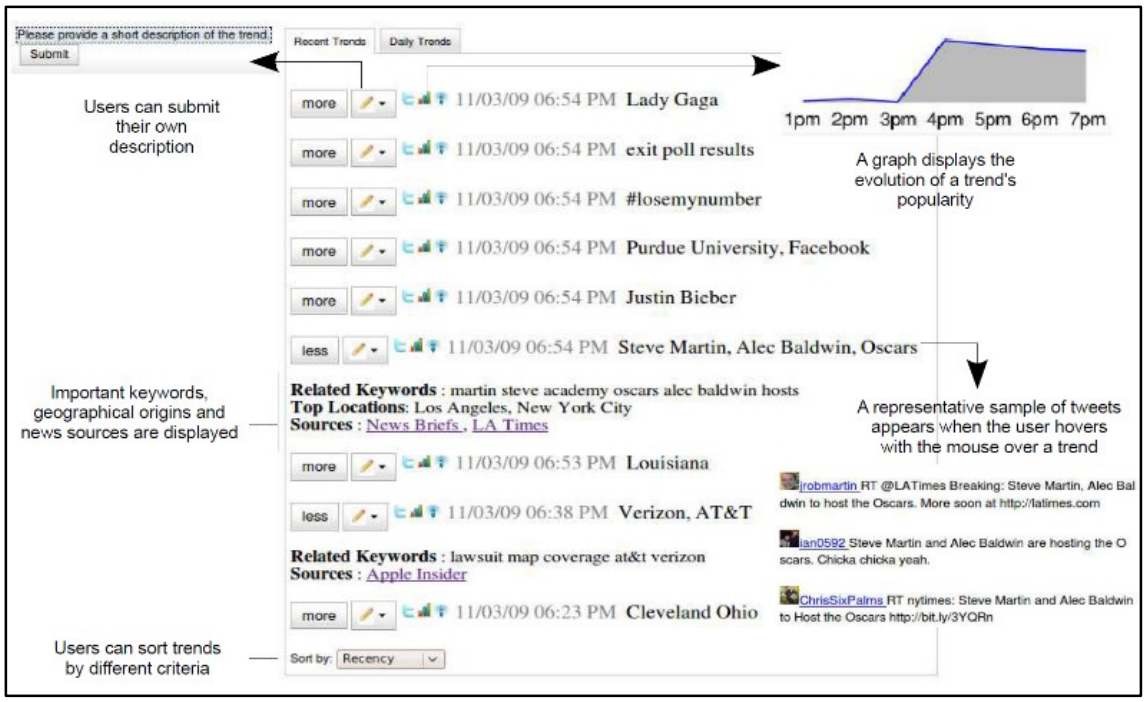


Figure 2.6: Visualization Screenshot from Mathioudakis and Koudas [76]

Wang and Cosley [133] proposed an interesting use of visualization on the topic of 'sustainability'. Sustainability is important, but is not discussed regularly and is not thematically as strong as other topics on Twitter. For example, they noted that though

hundreds of tweets were tweeted every week on sustainability issues but finding and reading such tweets among millions of tweets can be a challenging task. Thus, the author proposed a tool (labelled "TweetDrops') "to draw people's attention to the issue of sustainability" (p.33). The authors argued that the use of appropriate and aesthetically powerful visualization can help people to become more knowledgeable on such an important topic (Screenshot in Figure 2.7).

Figure 2.7: Visualization Screenshot from Wang and Cosley [133]

Chapter 3

Research Methodology and Design

3.1 Overview

There are number of sub-sections in this chapter and they discuss the following:

a) The first section (titled: "*Overall Research Objectives*") presents overall research objectives;

b) The second section (titled: "*DigiCities* and the POP Framework") provides detailed discussion on *DigiCities* and the digital representation framework;

c) The third section (titled: "*Creating DigiCities: Digital Profile of Cities*") discusses the process of developing city profiles from Twitter;

d) The fourth section (titled as "*Feature / Dimensionality Management*") provides details of feature management in tweets;

e) The fifth section (titled: "*Experiment Design – Data, Decision Making and Data Pre-processing, Software Used, and Data Analysis*") provides overall experiment details including an overview into data, data pre-processing, use of algorithms, and data analysis approaches used in this study, and;

f) The sixth section (titled: "*Visualizations - Emotions, Sentiments and Topics*") provides details related to identification of emotions and sentiments, and topic categories from tweets, and examples of visualizations developed for showing temporal patterns of emotions, sentiments and topic categorization results.

3.2 Overall Research Objectives:

There are multiple research objectives for this research work, and they are:

a) To develop digital profiles of cities i.e., *DigiCities* in the Province of Alberta in Canada;

b) To implement the proposed novel POP (people, organization, places) framework of location digital representation and to analyze the impact of such framework on tweet classifications;

c) To compare and contrast the impact of the novel POP framework on the accuracy of different classifiers through the implementation of '*replace*' and '*append*' feature convergence strategies;

d) To identify and visualize the *Pulse of a City*, that includes sentiments, emotions, and topics categories as expressed by Twitter users and reflected in location-relevant tweets.

3.3 DigiCities and the POP Framework

Cheng et al. [24] highlighted that the use of various levels of granularity related to geospatial information in users' profiles (e.g., city names such as Calgary or just a province name such as Alberta). Thus, location identification, particularly in users' profiles, could be done on a scale of geographical size (e.g., country, state (or province) or city level) or also on other parameters such as time zones [75]. For this research, the concept of location is used in the context of geographical boundaries and not on the other parameters such as time zones. The term location in this research primarily represents a geographical boundary as associated with the municipally defined

boundaries for a city or town. Though geographical locations and cities can be defined as significantly different concepts, for this research they are used interchangeably.

DigiCity is the digital identity or profile representing the real world geographical location on the Web, and in context of this research, it is the digital presence of key actors associated with a city on Twitter. The *DigiCities* are developed using the POP Framework which is discussed in the following sub-sections.

3.3.1 DigiCities Framework: Represented by POP

This research proposes a novel framework known as the POP Framework which helps in creating digital identities of cities i.e., *DigiCities*. The POP acronym stands for People, Organizations and Places. The proposed framework draws inspiration from the work of Kindberg et al. [56] and Warf and Sui [134].

As noted in Section 1.4 in Chapter 1, Kindberg et al. [56] divided physical entities into three key categories: people, places, and things. They used this categorization to bridge the gap between physical and virtual worlds by mapping them with their web presence; they presented the idea of linking the physical world with the virtual world as shown by the graphical representation in Figure 3.1.

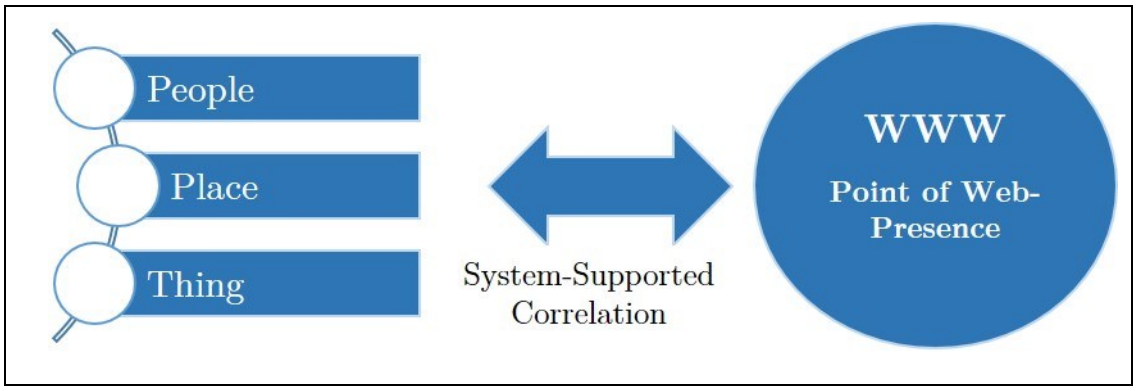


Figure 3.1: Web Presence (Source: Kindberg et al. [56])

In drawing inspiration from the work of Kindberg et al. [56] and Warf and Sui [134], this research focuses on having a digital representation i.e., DigiCity of a physical geographical location, particularly on a social media platform such as Twitter, by using the three key facets of People, Organizations, and Places (or POP Framework which will be used in the future discussion in this thesis).

- ***P****eople* – This facet represents public figures and the prominent members of a community and thus, are the face of a city. For example: City Mayor, Members of Parliament (MPs) and Members of Legislative Assembly (MLAs) representing various ridings in a given city.
- ***O****rganizations* – This facet represents key organizations and institutions in a city. Examples of such units include local radio channels, private organizations having a presence in a city (e.g., Fairmont Banff Hotel in Banff), museums, public libraries (e.g., Edmonton Public Library in Edmonton), and educational institutes (e.g., NAIT in Edmonton). This facet may also capture sub-units of a larger unit (e.g., Faculties or Departments in the University of Alberta).
- ***P****laces* – This facet represents a city by its name or airport code or through the prominent spaces and landmarks. Examples of such units include legislative buildings, sports arenas (Rogers Place in Edmonton), recreation centres, local parks and entertainment spots (e.g., Edmonton zoo in Edmonton or Calgary zoo in Calgary).

This research uses Twitter data, and within Twitter context, the facets included in the POP framework are represented by: a) handles, also known as 'user-ids' and they are denoted by '@' at the start of a term which may have one (e.g., @yeg) or more words (@calgarystampede) combined into a unigram, and; b) hashtags which are keyword(s) and they start with '#' sign (e.g., #calgarystampede).

3.3.2 *DigiCities – Implementation of the POP Framework (A City of Edmonton Example)*

The use of the three key facets, and the presence of these facets through handles and hashtags on Twitter are further explained by examples in the following paragraphs. Figure 3.2 provides a visual representation of the POP framework as applied to the city of Edmonton. Figure 3.2 shows examples of real Twitter handles or hashtags associated with the city of Edmonton and demonstrates the different facets of the POP Framework as represented through handles (e.g., @UAlberta) and hashtags (e.g., #yeg, #yegfreeride) leading to a city's presence on Twitter. Each element of the POP framework is further described through examples.

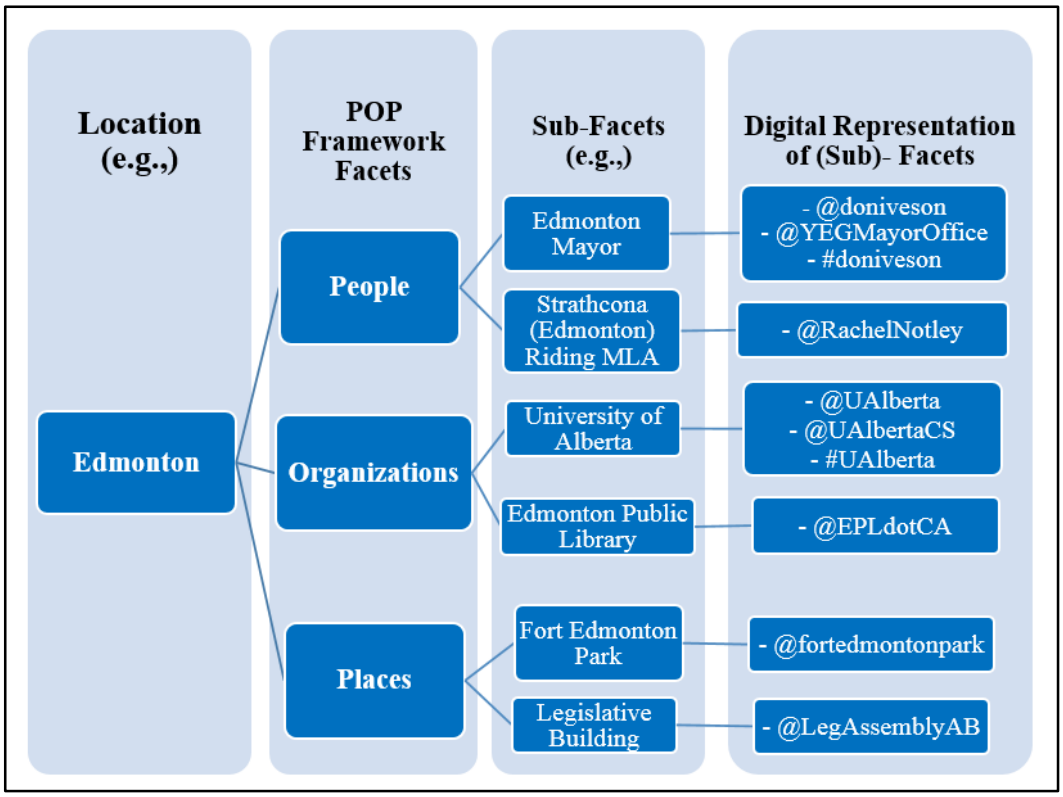


Figure 3.2: An Example of the POP Framework and the Digital Representation of Edmonton on Twitter

- ***P**eople (A City of Edmonton example):* One of the key people in the city of Edmonton would be the Mayor whose digital presence on Twitter would build the city's presence on Twitter. Both the Mayor and the Mayor's office have Twitter presence through multiple handles and/or hashtags such as: '*@doniveson*' is the (personal) handle (or user id) of Mr. Don Iveson, the current mayor of Edmonton, and '@YEGMayorOffice' is the mayor's office handle on Twitter (Figure 3.2 and Figure 3.3). Both these handles, associated with the prominent public figure from the city of Edmonton, are establishing the digital presence of the city on Twitter.

- ***O**rganizations (A City of Edmonton example)*: The University of Alberta (UofA) has presence on Twitter through handles such as '*@UAlberta'* and @UofAResearch. The University of Alberta is one of the major educational institutes in Edmonton, and thus, represents the city on Twitter under the *Organizations* category of the ***POP Framework*** (Figure 3.3). Large organizations such as the UofA may have multiple sub-units (e.g., faculties and/or departments), and each sub-unit may have their own handle(s) and/or hashtag(s). These sub-units are representing the UofA on Twitter through their individual handles and/or hashtags. For example: The Department of Computing Science is part of the Faculty of Science at the University of Alberta. The department has '*@UAlbertaCS'* as its Twitter handle while the department's parent faculty i.e., the Faculty of Science has '*@ualbertaScience*' as its Twitter handle. All such handles and/or hashtags represent the UofA on Twitter, which in turn establishes Edmonton's presence on Twitter. Another example of the organization in Figure 3.3 is the Edmonton Public Library whose handle is '@EPLdotCA', and thus EPL is also establishing the city's presence on Twitter.

- *Places (A City of Edmonton example)*: The legislative building is located in the city of Edmonton and is represented on Twitter by '*@LegAssemblyofAB*' handle. This landmark building in the city is representing the **Places** category of the ***POP Framework***.

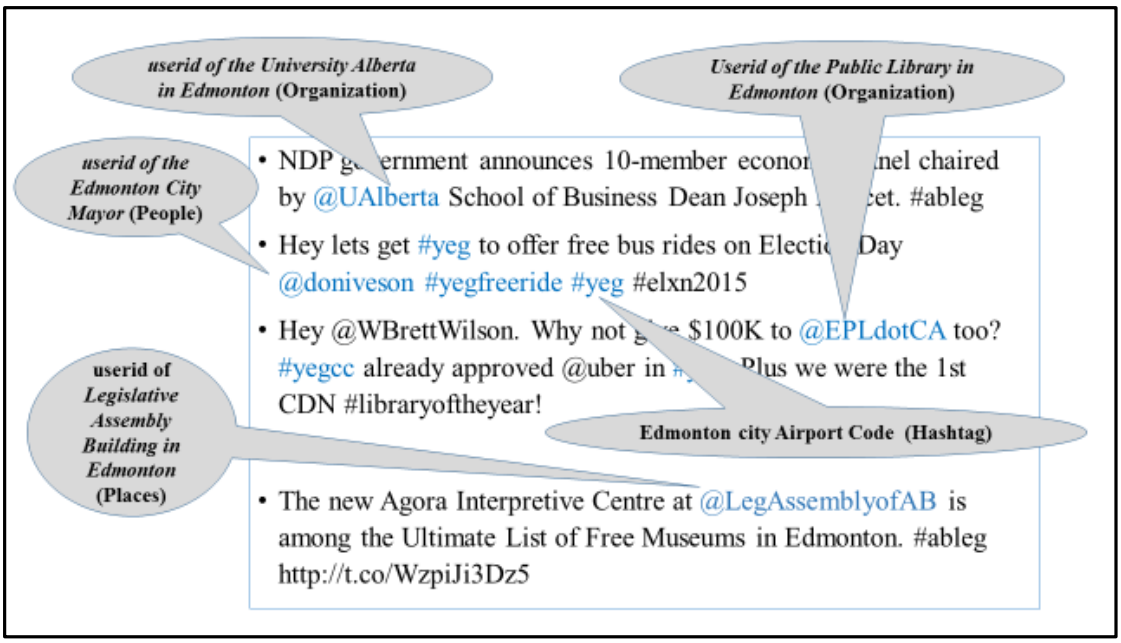


Figure 3.3: Examples of Tweets Related to Edmonton Reflecting the POP Framework

In addition, cities such as Edmonton are also represented on Twitter by their name (or short form of their names) and airport code, if there is an airport in the city. For example, Edmonton has digital presence on Twitter by hashtag (e.g., #edmonton) and handle (@edmonton), and also with the airport code (e.g., #yeg is the airport code for Edmonton) on Twitter.

3.4 Creating DigiCities: Digital Profile of Cities

There is not much in the literature to use as the guiding principle on the ways to develop the digital profile of cities from Twitter. *Digicities*, the digital profile for each city, is developed primarily using snowball sampling technique and additional profile terms involving the city name and the city airport code, are captured by using regular expressions.

3.4.1 Handles and Hashtags Identified Using Snowball Sampling Technique

Handles and hashtags are manually captured using a self-created "iterative multi-step" approach which used the concept of a snowball sampling technique (described in the sub-section "Snowball Sampling"). The details of the iterative steps are in the following paragraphs. Figure 3.4 provides the graphical representation of the iterative multi-step process.

- ***Step 1***: In the first step, the Google search engine was used to identify handles and hashtags. This started with the use of a keywords query in Google including words such as "cityname Twitter" (e.g., "Edmonton twitter"). Google returned the set of results, which are further used as the initial seed to identify city relevant handles, and to start the process of developing the digital profile of a city. These results led to the Twitter website to review the information related to handles and collect handles/hashtags relevant to a given city.
- ***Step 2***: After the initial seeding from the Google search results, the next set of handles were selected using the snowball sampling technique (see note on this in sub-section "Snowball Technique") i.e., Twitter recommended other handles under the

'You may also like' section. The information associated with each recommended handle is reviewed, and if relevant to a city, such handles are then captured to build a city's Twitter-based digital profile.

- ***Step 3***: Once a new set of handles is collected using the '*You may also like'* section on the new handle's page, Twitter has a '*Refresh*' option that is used to generate a fresh set of handles which are then reviewed for relevancy to the city, and then added to the list, if they are relevant to the city. The process initiated through '*Refresh'* and *'You may also like'* continued until the recommended handles either started repeating themselves or are no longer relevant to the city. The assessment, whether handles are relevant, was done by us.

In summary, the city profile was created with Step 1 which involved querying the search engine (i.e., Google) and followed by iterations of Step 2 and Step 3. The implementation of the process is further explained with the City of Edmonton example in Figure 3.4.

3.4.2 Variations of Handles and Hashtags Based on Key City Related Terms:

During a city profile development and the review of tweets, it is noted that there are a number of handles and hashtags which has the city name and/or airport code in them (e.g., calgary in @calgarytoday and @cowboyscalgary and city airport code 'yyc' in #yyctraffic, #yyczooteambuilder, #yegbellydance, and @newscoopyyc). Many such variations are not captured through the above mentioned process of creating city profiles. However, such handles and hashtags are also representative of the city's digital profile on Twitter and thus needed to be identified. The variants are identified by using

regular expressions with the name of the city and airport code except in the case of cities such as Banff and St. Albert which have no airports.

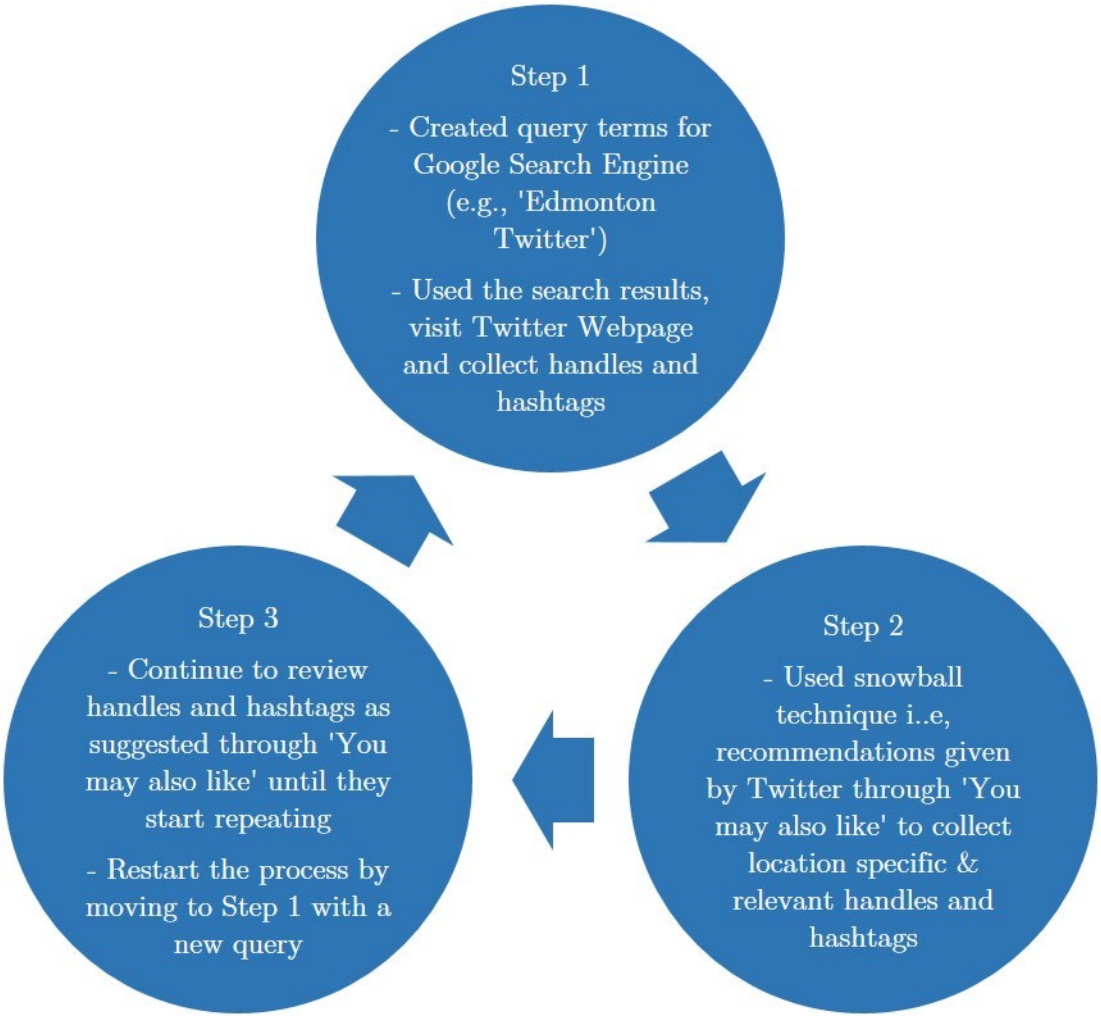


Figure 3.4: Steps to Create Digital Profile of a City

3.5 Implementation of Iterative Steps – A City of Edmonton Example

This section provides the implementation of the steps to create the digital profiles of cities as described in the above sub-section. The implementation is demonstrated

through an example related to the city of Edmonton. The screen shots are from Google and Twitter.

Step 1

- *Search Engine*: Google;
- *Search Term (e.g.)*: Edmonton Twitter
- Google returned the following results, and these results served as the seeding handles/hashtags.
- All the relevant results fitting in the POP Framework are clicked and the Twitter page for each was visited (e.g., City of Edmonton page on Twitter)

Step 2

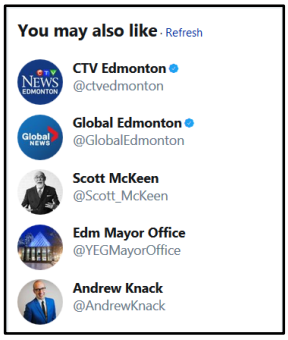
- Started with '@CityofEdmonton' handle from the seed page.
- Then, by using the snowball technique i.e., recommendations given by Twitter through '*You may also like*' are reviewed for location relevant handles.
- Relevant handles are captured and added to the city profile list.
- Examples of handles and/or hashtags are reviewed and collected based on the recommendation through '*You may also like*' include:

  ✓ @doniveson
  ✓ @GlobalEdmonton
  ✓ @edmontonjournal
  ✓ @CBCEdmonton

Step 3

- '*Don Iveson*', for example is clicked based on the recommendation in the above step and the webpage is reviewed including recommended handles.
- The examples of hashtags such as '#Edmonton', '#YEGCC', and "YEGmetro' are captured.

- Don Iveson's page also recommended a number of other handles through '*You may also like*' (as shown below)
- The '*Refresh*' option in the '*You may also like*' is also used to generate additional new handles



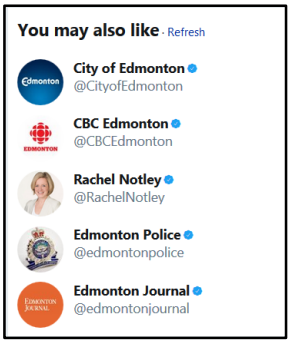


Figure 3.5: Digital Profile Development Example

3.6 Snowball Sampling

Snowball sampling, also known as chain referral sampling and is a popular method used in conducting research in the social sphere [13]. According to Biernacki and Waldorf [13], this approach helps in recruiting participants for study "through referral made among people who share or know of others who possess some characteristics that are of research interest" (p.141). This approach can play an important role in identifying the

participants when potential study participants are not known. The sample recruitment starts with "a convenience sample of initial" participants (P1) and these initial (conveniently sampled) participants (P1) then "serve as seeds," through which the next set of participants (P2) are drafted, and then these new participants (P2) help in recruiting the next set of participants (P3), and this process continues until it is stopped; thus, the sample size grows "like a snowball growing in size as it rolls down a hill" [40, p.356]. This research followed the snowball sampling technique. A 'convenience sample of initial' participants was recruited (P1) [40] i.e., and then seed handle(s) were identified by querying the Google search engine. As noted above, it started with a query 'cityname Twitter" (e.g., 'Edmonton Twitter') and the returned results created the initial sample (P1), that led to seeding of the next set of handles (P2) and so on (as demonstrated in the Figure 3.5 using the city of Edmonton example).

3.7 Feature / Dimensionality Management

Saif et al. [105] noted that there are two key directions of research in the concept of sentiment analysis on Twitter type data – a) identifying new approaches to conduct analysis (e.g., "performing sentiment label propagation") and; b) identifying novel features to strengthen the models (e.g., hashtags) (p.508). The authors, in their work on sentiment analysis, used the concept of semantic features as well as a new feature set "derived from the semantic conceptual representation of the entities that appear in tweets" (p.509). According to them "[t]he semantic features consist of the semantic concepts (e.g. "person", "company", "city") that represent the entities (e.g. "Steve Jobs", "Vodafone", "London") extracted from tweets" (p.509). The facets in the POP Framework i.e., people, organizations and places (POP) are also digitally reflected in tweets by handles (or user-ids, starting with '@') and hashtags (starting with '#'), and

drawing from Saif et al.'s [105] concept, these facets are semantically representing an entity i.e., a geographical location (e.g., Edmonton, Red Deer, and Calgary). Such representation helps in feature convergence and/or feature strengthening, for example, handles and hashtags associated with an individual entity in the POP Framework are referring to (a) a geographical location(s) and thereby, converging to one semantic concept i.e., a location or a city (e.g., Edmonton as shown in Figure 3.6).

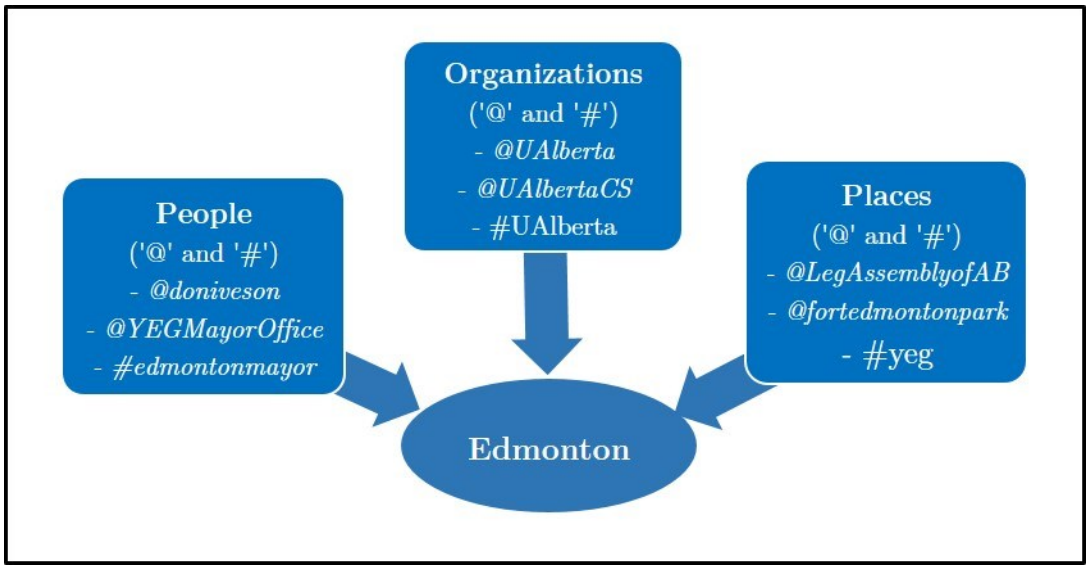


Figure 3.6: Feature Conversion to the City Name

As noted in the Chapter 1, data sparsity is one of the major challenges in Twitter data ([44][63]) and thus, has implications in location detection. The feature convergence approach will help in overcoming the data sparsity issue. Two approaches i.e., 'Replace Strategy' and 'Append Strategy' are applied to implement the feature convergence and feature strengthening in tweets where a location is represented semantically through various facets of the POP Framework. Both approaches are described in the following sub-sections.

3.8 Append Strategy and Replace Strategy

As noted in earlier discussion, city profiles are created that have a list of handles and hashtags relevant to a city. Handles and hashtags in the city profiles are then matched with handles and hashtags in tweets. Whenever such matches occurred, the city name (e.g., Edmonton, Calgary) is then appended after the matching term in tweets. For example, in Edmonton profile, there are a number of handles and hashtags including @UAlberta and #ableg respectively. Table 3.1 provides an example of a tweet relevant to Edmonton and is noted in the Original Tweet row. Both the terms (i.e., @UAlberta and #ableg) in the tweet matched with the terms in the city of Edmonton profile.

In the append strategy, city name, 'edmonton', is appended after the handle, '@UAlberta' and the hashtag, '#ableg' (Row Append Strategy in the Table 3.1). In the replace strategy, both the handles '@UAlberta' and the hashtag, '#ableg' are replaced by the city name, i.e., 'edmonton' (Row Replace Strategy in the Table 3.1).

Table 3.1: Example of Implementation of Replace Strategy and Append Strategy

Original Tweet	NDP government announces 10-member economic panel chaired by ***@UAlberta*** School of Business Dean Joseph Doucet. ***#ableg***
Append Strategy	NDP government announces 10-member economic panel chaired by ***@UAlberta*** **edmonton** School of Business Dean Joseph Doucet. ***#ableg*** Edmonton
Replace Strategy	NDP government announces 10-member economic panel chaired by **edmonton** School of Business Dean Joseph Doucet. **edmonton**

3.9 Experiment Design – Data, Decision Making and Data Preprocessing

This section will provide the experimentation details that include an overview of the tweets' dataset and data preprocessing, software, and algorithms used in this research.

3.9.1 Tweet Dataset

Twitter data was collected intermittently for approximately 12 months, January 12, 2017 to December 30, 2017, using API developed by Samuel Hamman while working with the Alberta Innovates Centre for Machine Learning (AICML) (now known as the Alberta Machine Intelligence Institute (AMII)). Tweets related to the Province of Alberta were shortlisted and the initial selection criteria of tweets (i.e., relevant to the Alberta) was based on the geolocation information included in the tweets metadata associated with the user's profile. The initial corpus includes over 700,000 tweets.

A total of eight urban centres in the province of Alberta were shortlisted for this study. These urban centres (or cities) include Calgary, Edmonton, Red Deer, Lethbridge, St. Albert, Medicine Hat, Banff, and Fort McMurray. These geographical locations are a mix of different sized urban population centres (Table 3.2 lists population in each city), including the provincial capital (Edmonton), the largest city in Alberta (Calgary), a popular tourist destination (Banff), the twin-city of a larger population centre (St. Albert), an industrial centre (Fort McMurray), and other key cities in the Province of Alberta (Red Deer, Lethbridge, Medicine Hat).

Table 3.2: Population of Shortlisted Cities

City Name	**Population (as per 2016 Census, Statistics Canada)**
Banff	7,851
Calgary	1,237,656
Edmonton	1,062,643
Fort McMurray	66,573
Lethbridge	87,572
Medicine Hat	62,935
Red Deer	99,718
St. Albert	65,589

There were varying numbers of tweets for each of the eight cities (see Table 3.3) and subsequently 500 tweets were selected for each city using purpose or the criterion sampling approach [123]. The criteria for selecting a tweet for a city incorporated that tweets should be in the English language, and they should contain the city relevant content such that a human adjudicator should be able to detect the relevancy of the tweet to a given city. The rationale for selection of tweets was the relevancy the tweet had to a given city and to test the efficacy of the proposed POP Framework on the classification. All the tweets for each city were manually reviewed and selected by us until the count of 500 was reached. There were primarily two reasons for selecting 500 tweets for each city (i.e., class). First, to balance each class - there were varying number of tweets in the collected dataset for selected cities i.e., smaller cities (e.g., St. Albert) had relatively a fewer number of tweets as compared to larger cities (e.g., Edmonton) which had a very high number of tweets. Thus, 500 tweets for each city helped in balancing classes. Second, to manage the scope of the work - the tweets were selected

manually for each city from a large pool of tweets as noted above and it was labour-intesive work to select a total of 4,000 tweets (= 8 cities x 500 tweets/city) in addition to 500 tweets for the "Others" category for this research.

Table 3.3: Dataset and Data Sampling

City Name	**Total Tweets in Dataset**	**No. of Tweets Selected**
Calgary	188,342	500
Edmonton	163,121	500
Red Deer	10,992	500
Lethbridge	11,006	500
Banff	9,341	500
St. Albert	8,163	500
Medicine Hat	5,434	500
Fort McMurray	4,798	500
Others	398,516	500
Total	***799,713***	***4,500***

3.9.2 Decision Taken and Data Pre-processing

A few key decisions were made in the context of data preprocessing and city name handling in append and replace strategies implementation, and these include:

a) More than one term (i.e. city name) are concatenated to create a single term city name. This was used in the implementation of append and replace strategies. For example: cities such as 'Red Deer', 'Medicine Hat' and 'Fort McMurray' have two

terms, and in this research all the terms were concatenated to 'RedDeer', 'MedicineHat' and 'FortMcMurray' respectively.

b) Tweets have a various number of special characters and/or symbols, such as new line character "\n", question mark "?", brackets like "(" ")", exclamation mark "!", quote "", etc., which are removed from the shortlisted tweets.

c) The analysis of tweets revealed that there are multiple instances of whitespace(s) between "@"and "#", and the terms following them. A decision was made to remove such whitespace(s) and concatenate '@' and '#' with the terms following these symbols. This decision is based on the assumption that whitespace is created during tweet processing (e.g., downloading and subsequent handling) or that it could be due to the user's typographical error.

d) Removed characters such as ',' and '.' that were prefixed '@' and '#' (e.g., '.@') and '.#').

e) Stopwords are not removed in the data preparation stage i.e., pre-processing stage but are removed in the experimentation stage to understand the impact of the presence or absence of stopwords in tweets on classification accuracy.

f) Similarly, stemming is not implemented in the pre-processing stage but is applied during the experimentation stage for the same reasons as discussed in the presence or removal of stopwords.

3.9.3 Application – Algorithms, Stemming and Stopwords List Used

Weka3.6 (Waikato Environment for Knowledge Analysis) machine learning software application was used for classification experimentations.

The software was downloaded from the 'Weka 3: machine Learning Software in Java' webpage[1]. This research used kNN, NB and SMO classificaiton alogirithms, Lovins stemming algorithm, and (Rainbow) stopword list as implemented in the Weka Application.

The Lovins algorithm [72] is a "two-step stemming algorithm" (Lovins [72] as referenced in Schofield and Mimno [107]) [107, p. 288] while the Porter algorithm [97], as an example, is a five-step stemming algorithm [107][47]. According to [107], the Lovins algorithm uses a "long lists of rules", and it is fast and easy to execute [p. 288]. Both Lennon et al. [65] and Hull [43] evaluated various stemming algorithms including Lovins and Porter stemming algorithms. Goldsmith et al. [33] noted that both [65] and [43] "found no overall consistent differences between stemming algorithms of various types, though on a particular query one algorithm might outperform other, but never consistently" [p. 275]. They further stated that "[m]ost studies note that stemming performance varies on different collections" [33 p. 275-276].

The choice of using both the stemming algorithm and the stopword list was inspired by research conducted in the Twitter domain. Armentano et al. [6] and Ji et al. [48] used Twitter-based data and they both used Lovins stemming algorithm in their research work. The authors [6] used opinion mining approach to recommend movies by harnessing information from tweets while Ji et al. [48] focused on using Twitter as health surveillance tool, and they used sentiment classification approach to measure public health concerns. In terms of stopwords, authors such as Iosifidis and Ntoutsi [45] and

[1] http://www.cs.waikato.ac.nz/ml/weka/

Krouska et al. [58] used (Rainbow) stopword list as implemented in the Weka application and their work was also in the Twitter domain.

Weka3.6 has implemented a number of classification algorithms such as kNN, SMO and NB, and these three algorithms were used in this research to conduct classification experimentations. Naïve Bayes (NB) classification algorithm has been one of the popular classifiers [67] and it uses examples to learn in order to classify new data [140]. It is implemented in Weka3.6 applications as well. NB algorithm, is based on Bayes' theorem which uses the concept of conditional independence [121] i.e., it assumes that all features in a dataset are "independent given the value of the class variable" [140, p.591]. Though the conditional independence assumption used in the implementation of NB algorithm is not reflective of real world situation (i.e., attributes may be dependent) yet this assumption works well and helps in creating a model with the reasonable set of parameters [67][140].

kNN (k-nearest neighbor) classification algorithm is a lazy learner algorithm [121] as it stores training examples and does not build model unless classification of test example is required. The algorithm requires values for 'k' (e.g., k=1, 2, 3, etc.). When test sample is given to the algorithm to classify, it searches for 'k' number training examples that are closest to the given test sample. Based on the assigned value of 'k', the algorithm assigns the test example to "the majority class of its nearest neighbours" if the value of k is more than 1 [121]. If there is tie among various classes, other rules may come into play such as random assignment to one of the majority class [121]. The algorithm calculates the closeness of the test example with training examples by using distance metrics such as Manhattan, Supermum and Euclidean [37][53][121]. The kNN

implementation in Weka3.6 uses various distance metrics such as Euclidean, Manhattan etc. For this research work, we used Euclidean as a distance metric for kNN.

Platt [96] proposed the Sequential Minimal Optimization (SMO). SMO is an iterative algorithm which simplifies the quadratic programming (QP), used in SVM (Support Vector machines), by breaking large QP into a number of smaller QP problems by using Osuna's theorem. SMO, to optimize the smaller quadratic problems, uses two Lagrange's multipliers such that they both satisfy the inequality constraint and linear equality constraint. SMO is relatively fast and computationally efficient [96][52].

3.9.4 Data Analysis

The evaluation of classification results was performed using the overall accuracy of correct and incorrect classification of tweets. The results were analyzed and reported using descriptive statistics such as percentage, and inferential statistics using t-tests to evaluate the statistical difference in the accuracy score under different experimental conditions.

3.10 Visualizations of Emotions and Sentiments with their Topics

In this research, both emotions and sentiments, and topics associated with them were identified from the tweets for all eight cities used in the study (Calgary, Edmonton, St. Albert, Red Deer, Fort McMurray, Lethbridge, Medicine Hat and Banff), and the results were visualized using the JavaScript-based application.

3.10.1 Emotions

Shahraki and Zaiane [108] suggested nine categories of emotions which include joy, fear, anger, disgust, sadness, surprise, love, thankfulness and guilt. This research work

used these nine categories of emotions to identify location-relevant emotions from the tweets for eight cities (Edmonton, Calgary, Banff, Red Deer, Fort McMurray, Lethbridge, and St. Albert). The emotions were identified by using algorithm initially developed by [108]. The algorithm evaluates one or more emotions for each tweet and each emotion is assigned a value in percentage. Table 3.4 provides an example of emotions-related results using five tweets (T1, T2, T3, T4, and T5) that are related to the city of Banff. The example shows that three tweets (T1, T2, and T3) were posted on 2017/10/01 (column heading: "Date") and two tweets (T4 and T5) posted on 2017/11/15 for Banff (and the Example tweets (T1, T2, T3, T4, T5) are included in Appendix D). The algorithm identified multiple emotions relevant to each tweet, and the associated percentage value for each emotion. Both emotions and the percentage value (in brackets) for each emotion are noted in the third column (column heading: "Emotions (values in percentage)") of Table 3.4.

Table 3.4: Example of Emotions as Reflected in Tweets

Date	**Tweets for Banff**	**Emotions (values in percentage)**
2017/10/01	T1	Surprise(100), anger(90), sadness(82), love(68), disgust(58), guilt(51)
2017/10/01	T2	Love(100), Joy(72), Sadness(48), Disgust(45), Surprise(41)
2017/10/01	T3	Thankfulness(100), Fear(75), Sadness(58), Love(56), Surprise(52), Disgust(51), Anger(43)
2017/11/15	T4	Sadness(100), Anger(98), Love(88), Fear(87), Thankfulness(61), Guilt(55)
2017/11/15	T5	Love(100), Joy(70), Sadness(66), Disgust(65), Fear(51)

City related emotions were evaluated using the following criteria:

a) Emotion(s) showing strength of 50% or more were considered. Emotion(s) having percentage values less than 50% were ignored as it was considered that, that particular emotion was not strong emotion for that specific tweet.

b) In this research, emotions relevant to a city were cumulatively evaluated for each day. The rationale was to get a full day's picture of overall emotions emerging in a city. The full day's emotions were calculated in the following way.

- Identified number of times a particular emotion (=a) emerged (and having the percentage value of more than 50%) on a given day (= count of 'a' on day x)
- Identified total number of tweets posted on that day (= total tweets on day x)
- Emotion A occurring in a city on a day is equal to: (count of 'a' on day x)/ total tweets on day x.
- Using the values from Table 3.4, Table 3.5 provides an example of how emotions for each day for a city were calculated. For example, on 2017/10/01, the total number of tweets posted was three, and the count of emotions on tweets for day D1 is Surprise: 2, Anger: 1, Sadness: 1, Love: 3, Disgust: 2, Guilt:1, Joy:1, Thankfulness: 1, Fear:1. Thus, the Love emotion was 100% (=3/3), Surprise and Disgust emotion was 66.67% (=2/3), and the Guilt, Anger, Sadness, Fear, Joy and Thankfulness emotion was 33.33% (1/3) for Banff on 2017/10/01.

Table 3.5: Example of Emotions in a City on Different Days

Date	**Total Tweets / day**	**Emotions with 50% or more value**	**Emotions less than 50% value**	**Emotions of the Day for the City Banff**
2017/10/01	3 (T1, T2, T3)	T1: Surprise (100), Anger (90), Sadness (82), Love (68), Disgust (58), Guilt (51)	T1: NONE	Count for: Love = 3 Love Emotion is 100% **(=3/3)**
		T2: Love (100), Joy (72)	T2: Sadness (48), Disgust (45), Surprise (41)	Count for: Surprise = Disgust = Sadness = 2 **=> is 100% (=2/3)**
		T3: Thankfulness (100), Fear (75), Sadness (58), Love (56), Surprise (52), Disgust (51),	T3: Anger (43)	Count for: Anger = Guilt = Joy = Thankfulness = Fear = 1 **=> is 33.33% (=1/3)**
2017/11/15	2 (T4, T5)	T4: Sadness (100), Anger (98), Love (88), Fear (87), Thankfulness (61), Guilt (55)	T4: NONE	Count for: Sadness = Love = Fear = 2 **=> 100% (=2/2)**
		T5: Love (100), Joy (70), Sadness (66), Disgust (65), Fear (51)	T5: NONE	Count for: Thankfulness = Guilt = Joy = Disgust = Anger = 1 **=> 50% (=1/2)**

3.10.2 Sentiments

This work uses sentiments identified as positive, negative and neutral as used by Pak and Paroubek [87] and Thelwall et al. [124]. This research uses '*SentiStrength*' algorithm as proposed by Thelwall et al. [124] to compute the sentiments for the location-relevant tweet data. The value ranges from -5 to 5. According to the authors [124], -1 is not so negative and 1 is not so positive, while -2 to -5 is negative and 2 to 5 is positive. As the number value changed for each sentiment the relative strength changed. For example, the tweet with value of '-5' is more negative than the tweet with value of '-3' and similarly, the tweet with value of '4' is more positive than the tweet with value of '2'.

The '*SentiStrength*' returned two polarity-related values for each tweet reflecting both positive and negative sentiment in a tweet. For this research purpose, both these values were combined i.e., were summed and the summed value was used to identify the overall relative sentiment associated with tweets. For example: Tweet 'T1' received sentiment value as '1' and '-1'. This tweet is reflecting both positive and negative sentiments at equal level. The combined value is '0' [= (1) + (-1)] and thus, such a tweet was considered as the tweet with 'neutral' sentiment. Similarly, another tweet 'T2' received sentiment values at '5' and '-2'. This tweet is reflecting very strong positive sentiment and is stronger than the negative sentiment ('5' vs '-2'). The combined value is '3' and thus, such a tweet was considered as the tweet with 'positive' sentiment. Overall, each tweet had only one sentiment, unlike in emotions where more than one emotion could be associated with a tweet.

Like emotions, sentiments in a city were also calculated for the full day using the same approach as discussed in the 'Emotion' sub-section above. For example, if there

were five tweets in a day and out of these five tweets, two tweets had positive sentiments, three tweets had neutral sentiment, and zero tweet had negative sentiment, then sentiment strength for that day will be: Positive 40% (=2/5), Neutral 60% (=3/5) and Negative 0% (=0/5).

3.10.3 Topics for Emotions and Sentiments

Topic modelling helps in identifying themes from a set of documents. The basic assumption in topic modelling is that a document might have a number of topics and each topic has one or more terms associated with it [137]. A number of specific approaches such as LDA [16] and LSA [28] related to topic modelling from documents have been developed. However, the data is very small and is informal text, this study uses the term frequency count approach i.e., topics relevant for emotion and sentiment types were identified by using simple term frequency count.

The tweets belonging to an emotion (or a sentiment) on a given day were combined and then the frequency count for each term in the combined set of tweets was calculated and the terms with the highest counts were extracted and included in the dataset prepared for the visualization. The topics were identified only for those emotions which were selected to define emotions for the day (i.e., having value of more than 50%). Using the example in Table 3.5, the topics for emotion (and similar approaches were used to calculate topics for sentiments as well) was calculated in the following way:

- For the date 2017/10/01, the 'Love' emotion was reflected in tweet T1, T2, T3. Hence, tweets T1, T2, T3 were combined to do a frequency count for all the terms occurring in these tweets. The top five terms based on the frequency counts were

identified and thus, were included as the topics for the 'Love' emotion for 2017/10/01 date.

- Similarly, for the same date tweets for the 'Surprise', 'Sadness' and 'Disgust' emotions, two tweets 'T1' and 'T3' were associated. Hence, tweets 'T1' and 'T3' were combined to calculate frequency counts for all the terms occurring in these tweets. The top five terms based on the frequency counts were identified and thus, were included as the topics for the 'Surprise', 'Sadness' and 'Disgust' emotions for the date 2017/10/01.
- Following the above process, the count for other emotions such as 'Anger' and 'Guilt' from tweet T1, 'Joy' from tweet 'T2', 'Thankfulness' and 'Fear' from tweet 'T3' were also computed.
- City name from each city-specific tweets were removed while doing term frequency count to identify topics for both emotions and sentiments for a city in a given day. It was likely that in the city-relevant tweets, the city name and/or its variants (e.g., hashtags like #yeg for Edmonton) were likely to occur heavily and would suppress the other terms. Thus, to avoid such situation, decision was taken to remove city names from the city-relevant tweets.

3.10.4 Dataset Used to Calculate Emotions and Sentiments

Emotions and sentiments were calculated for all the eight cities used in this research, (Calgary, Edmonton, St. Albert, Red Deer, Fort McMurray, Lethbridge, Medicine Hat and Banff). Each of these cities had 500 tweets which were manually identified by the researcher as the location-relevant tweets. This dataset is considered as the 'gold

standard' data of tweets. Thus, emotions and sentiments were calculated on this (gold-standard) dataset as these tweets would reflect emotions and sentiments of each city.

In order to evaluate the efficacy of the *DigiCities* approach, the visualizations of emotions and sentiments generated from the gold standard data were compared with visualizations of emotions and sentiments generated from the post classification dataset. This helped in comparing the emotions and sentiments of a city pre-and-post classification. The city of Calgary was chosen from the eight cities to compare the emotions and sentiments emerging from the gold standard data and post classification datasets. Calgary was selected as the sample city as it is the largest population centre among the cities included in this research. The following dataset of tweets for Calgary were prepared to compare emotions pre-and-post classification:

- ***Gold Standard***: The gold standard dataset as created by us (as discussed above).
- ***NB_B***: This dataset contain tweets as classified in Calgary bin by the NB algorithm. In this research this experiment is labelled as NB_B. On this dataset - no stopwords were removed, no stemming algorithm was applied and no implementation of *DigiCities*.
- ***NB_A_SA_WS***: This dataset contained tweets as classified in Calgary bin by the NB algorithm. On this dataset – stopwords were removed, stemming algorithm was applied and the *DigiCities* approach using append strategy was implemented.

3.10.5 Visualizations Development

The computed emotions and sentiments results were visualized using the column chart and the line chart but only the column chart visuals are included in this thesis. However, additional visualizations using the Heat Map was done to plot sentiments. The

topics associated with each emotions and/or sentiments are accessible 'on click'. The emotion and sentiment visualizations were implemented using 'Highcharts' libraries[2].

The visualization of results helped in conducting the following comparisons and the findings emerging from such comparisons are discussed in Chapter 6: Visualizations.

Comparison of different emotions emerging in each city during different temporal periods (e.g., weekly, monthly).

- Comparison of emotions between two cities (e.g., Edmonton and Calgary or Calgary and Banff, etc.) during the same temporal periods.
- Similar visualizations and comparisons are developed for sentiments as well.
- Comparison of emotions for the city of Calgary on pre-and-post classification dataset of tweets. The pre-classification dataset was the Gold Standard dataset, and the post-classification datasets were NB_B and NB_A_WS_SA (as discussed above).

3.11 Topic Categorization of Tweets

Categorization of location-relevant tweets into topics will assist in identifying range of topics that users are discussing, facilitating in learning about topics that are trending during different time periods, and help in better understanding of users' reaction and opinion on different topics. Also, categorization can help in removing non-relevant and noisy tweets which are important to "extract higher-level, meaningful information from them [tweets], such as general topics or sentiments" [5, p.72].

[2] https://www.highcharts.com/

As noted in the Motivation sub-section (Section 1.2 in Chapter 1), the majority of users are interested in tracking or following a small number of topics (e.g., sports, health, weather) on social media including Twitter [69] while the volume of tweets posting is enormous and creates information overload for users [101]. Thus, it is important to identify and categorize tweets into topics (or themes) so that users are able to track tweets relevant to their interests and allow them to keep abreast on topics of their interests. The categorization of tweets into topics along with the visualizations of sentiments and emotions will help users to have a more accurate insight into the *Pulse of a City*.

It is critical that an automated tweet categorization approach is implemented because it is known that a large number of tweets are posted each day and manual categorization of a high volume of tweets is not practical. Thus, this research aims to use Google Knowledge Graph[3] (GKG) and WordNet[4] to help in automated categorization of tweets into topic categories. The categorization results obtained by the use of WordNet and GKG are evaluated by comparing with manual categorization of tweets done by us. The following paragraphs provide experimentation details related to the categorization of tweets using manual, GKG and WordNet approaches. The categorization of tweets drew inspiration from the work of authors such as Piao and Breslin [95], Ray and Singh [99] and Sutton et al. [120] in the field of topic categorization.

[3] https://developers.google.com/knowledge-graph/

[4] https://wordnet.princeton.edu/

3.11.1 Manual Topic Categorization of Tweets

The topic categorization was done on the dataset comprising of 400 tweets which was created by drawing 100 tweets each from the four cities (i.e., Banff, Calgary, Edmonton and Red Deer) out of the seven cities used in this research work. A sample of 100 tweets each from four cities were selected for topic categorization for the two keys reasons: a) to manage the volume of manual coding of tweets with topic categories, and; b) to evaluate if Google Knowledge Graph or WordNet can be used to automatically assign topic label(s) to tweets. Also, these four cities added to the data heterogeneity because the cities, Edmonton and Calgary, were the two largest cities in Alberta, Red Deer represents relatively small city (as compared to Edmonton and Calgary) and Banff represents the city with a transient population.

100 tweets from these four cities were manually categorized into eight categories and these include weather, sports, jobs, entertainment, health, news, traffic and urgent events, and others. Each of these categories had no major sub-category but for the 'entertainment' category which had sub-categories such as music, movies, food, festival, and tourism. However, only 'entertainment' as a category, which subsumed these sub-categories, is used in analysis and reporting purposes. The labelling into topics category was performed by using thematic analysis approach which allows to encode "qualitative information" [19, p.4]. These categories and sub-categories were identified based on the author's assessment of sample tweet dataset, and in consultation with thesis supervisor, Dr. Osmar R. Zaiane and literature review (e.g., [5], [101], [99]).

Table 3.6 provides examples of manual categorization of tweets into topic categories. Tweet#1 explicitly uses terms including hashtags such as 'hiring', 'Job', and other terms such as 'apply', and thus leads to categorization of the tweet in the 'Job' category while

Tweet#2 implicitly reflects about tourism and travel to the city of Banff, and thus, categorized in the 'Entertainment' category. There were other tweets such as Tweet#3 which could not be categorized into any of the topic category and thus, were assigned to 'Others' category.

Table 3.6: Example of Tweets Categorizations into Topics

No.	**Tweet**	**Topic Category**
Tweet#1	*We're #hiring! Click to apply: Registered Nurse (2022.75) - https://t.co/7NdCT8D0EU #Job #Nursing #Edmonton AB*	Jobs
Tweet#2	*Walking around Banff is a dream no matter the season of the year I think I'm gonna print*	Entertainment
Tweet#3	*Dumpster Bin Rentals Calgary Garbage Bin Rentals in Calgary Construction Bin Rental https://t.co/2EwbuB-iNgT*	Other

3.11.2 Topic Categorization of Tweets Using Google's Knowledge Graph

Google Knowledge Graph (GKG) was used to identify keywords associated with the seven topic categories (i.e., Weather, Sports, Jobs, Entertainment, Health, News, Traffic and Urgent Events) plus each of five sub-categories (i.e., Food, Movie, Tourism, Music, and Festival) in the 'entertainment' category. Each category and sub-category were operationalized by populating them with 50 keywords except for 'Music' sub-category which had less than 50 keywords. The sub-categories keywords were then combined with and added to the 'entertainment' category. All the keywords were extracted from GKG API. The keywords had varying number of terms including one term (e.g., 'jobs' in the 'Jobs' category), two terms (e.g., 'jobs corner' in the 'Jobs' category) and three or more

terms (e.g., 'conference assistant jobs' in the 'Jobs' category). It was decided to seek matching of any term from a multi-term keyword (e.g., 'conference assistant job') with terms in tweets. For example, if any one term (i.e., either 'conference' or 'assistant' or 'job') out of the three-terms keyword, 'conference assistant job' matched with a term in a tweet, then such tweet would be assigned the 'Jobs' category.

A topic category label was assigned to a tweet if one or more terms from each topic category matched with the term(s) in the tweets, and if there was no matching of keywords with terms in tweets, then the tweets were assigned to 'Others' category. For example, the following tweet due to the keywords 'job' and 'healthcare' got assigned the 'Health' and the 'Jobs' categories as these keywords are included in these categories.

> *"Join the AHS team! See our latest #job opening here: https://t.co/PcVnks4nvW #Healthcare #Calgary AB #Hiring https://t.co/dk9HQuNHtQ"*

Tweets have hashtags (starting with '#') and user-ids (starting with '@') but GKG does not return keywords staring with '#' and '@'. Thus, the decision was made to remove these characters but the terms associated with these special characters were retained in the dataset (e.g., the term "*#hiring*" in Tweet#1 above was pre-processed to "*hiring*"). In addition, there were other noise elements in the dataset which were removed during the pre-processing stage including removal of special characters, HTML elements like '#amp', and stopwords.

3.11.3 Topic Categorization of Tweets Using WordNet

The second categorization approach involved the use of WordNet[5] which "is an online lexical database" developed at the University of Princeton [79, p.39]. Pedersen et al. [92] developed a 'WordNet::Similarity' module which draws upon "the structure and content of WordNet". The implementation helps in finding the "semantic similarity and relatedness between a pair of concepts (or synsets)" and provides options to calculate similarity and relatedness based on multiple measures (p.1). A web-based implementation developed by Pedersen and Michelizzi [91] (available online[6]) is used in this research to compute relatedness between two terms.

A list of all the unique terms (say, 'n' terms) from 400 tweets (after pre-processing the tweets as noted in the section 3.11.2) was created. Each term in this list was paired with each topic category/sub-category (say, 'm' categories) and a score for such pairing was calculated using the 'WordNet::Similarity' module as implemented by Pedersen and Michelizzi [91]. The 'Adapted Lesk' (also known as 'Lesk') measure was used to calculate the relatedness between each term as drawn from the list and topic category/sub-category label, thus creating n x m matrix of scores. The score ranged from 'no score' returned to zero (0) and above. The 'no score' results were assigned 'NA' (Not Applicable) meaning such combination of keywords and topic labels have no relationships.

A number of descriptive statistics related values (e.g., median, mode) were generated for the returned scores (i.e., value >=0) and summary is included in Table 3.7. A total

---

[5] https://wordnet.princeton.edu/

[6] http://maraca.d.umn.edu/cgi-bin/similarity/similarity.cgi

of 7,929 category-pair received a score of greater than '0'. The category- keyword pair for which the score returned with 'no value', such pairs were assigned 'NA' label and there were 18,111 such category-keyword pair which were removed from the descriptive statistics calculations.

Table 3.7: Descriptive Statistics of 'Topic Categories and Keyword' Pair for WordNet

Total Number of Categories (m)-Keywords (n) and their Pairs	
Number of categories (including "Entertainment' sub-categories) (m)	12
Number of unique terms from tweets (n)	2,170
Total Category-Keyword Pairs Generated (n x m)	*26,040*
Category – Keyword Pairs - Score vs No Score Received	
Category-Keyword Pairs Received ***Score*** *(>=0)*	7,929
Category-Keyword Pairs Received ***No Score*** (i.e., labelled "NA")	18,111
Descriptive Statistics Results of 7,929 *Category-Keyword Pairs*	
Minimum	0
Maximum	1489
Mean	7.58
Median	4
Mode	2

A four-step process was used to assign category or categories to tweets, and the implementation of the steps is explained using an example tweet as noted in the Section 3.11.2. The keywords in the example tweet are: join, team, job, opening, healthcare, hiring, AHS (as listed in the columns while categories are in the rows in Table 3.8). The category-keyword pairs (Lesk) scores were calculated using the 'WordNet::Similarity' module, and included in Table 3.8.

- *Step 1 - Median Score Cutoff*: The median was equal to '4' (Table 3.7) and was selected as the minimum score value to filter category-keyword pairs.
    - From Example: The circled scores in the cells in Table 3.8 are more than the median score of '4', and are thus the filtered category-keyword pair candidates for assignment of keywords to categories.
- *Step 2 - Assigning a Keyword to (a) Category(ies)*: The keyword was assigned to a category that had the highest category-keyword pair score. For a keyword, if multiple categories got the same highest score, then such keyword was assigned to all those categories which had same highest score.
    - From Example: The scores in bold and with asterisk (superscript) (among the circled score) are the highest scores for each category-keyword pair after filtering them based on Step 1, and thus, such keyword is assigned to the pairing category. Therefore, the keywords with highest scores are assigned to their respective categories (e.g., 'join' is assigned to the category 'food' with highest score of 9).
- The keyword 'job' assigned to the category 'jobs' with highest score of 440.
- The keyword 'healthcare' is assigned to the two categories – 'health' and 'tourism'.
- *Step 3 - Combining Categories*: All the keywords generated under the sub-categories - festival, food, movie, tourism and music were then combined and added under the 'entertainment' category.
    - From Example: 'food' and 'tourism' are the sub-categories of the 'entertainment' category (as demonstrated by the indentation in Table 3.8). Thus, the terms 'join', 'team', 'healthcare' and 'hiring' were assigned to the 'entertainment' category.

- *Step 4 - Assigning Tweets to Categories*: A tweet was assigned to a category if any term in the tweet matched with a keyword associated with a category. A tweets was assigned multiple categories if terms from tweets were found matching with keywords from different categories.
    - From Example: The example tweet was assigned three categories – 'entertainment', 'health', and 'jobs' due to multiple terms in the tweets matched with keywords in these multiple categories.

Table 3.8: Example of Topic Category & Keyword Pair with their Relatedness Score

Categories (↓)	**Keywords from the Example Tweet**						
	Join	*Team*	*Job*	*Opening*	*Healthcare*	*Hiring*	*AHS*
weather	3	2	7	5	1	1	NA
sports	NA	NA	73	5	NA	NA	NA
jobs	4	9	**440***	8	NA	NA	NA
health	1	1	7	2	**4***	1	NA
news	NA	NA	NA	NA	NA	NA	NA
traffic	NA	NA	4	5	1	4	NA
entertainment	NA	6	4	6	NA	NA	NA
festival	NA	NA	3	4	NA	1	NA
food	**9***	**12***	9	**32***	2	NA	NA
movie	6	11	27	18	NA	NA	NA
tourism	1	NA	12	4	**4***	**5***	NA
music	NA	NA	NA	14	NA	NA	NA

Chapter 4

DigiCities: Implementation of POP Framework

This chapter will present findings from the implementation of the POP framework on the Twitter data and provide insight into the city profiles that are created for this research study. In addition, we will also discuss the outcome of the implementation of the city profiles on the Twitter data, in particular when the terms in the tweets were matched with the terms in the city profiles.

4.1 City Profile Overview and Analysis

Eight cities from the Province of Alberta were selected to test the efficacy of this approach. The terms used in a city profile includes the handles and hashtags generated using the snowball sampling technique in addition to variations of those handles and hashtags. During the city profile development and data collection, it was observed that many handles also have equivalent hashtags. For example: the city of Banff handle '@banff' also has an equivalent hashtag '#banff', similarly, for the city of Calgary, a handle such as @calgarystampede also has an equivalent hashtag i.e. '#calgarystampede'. In order to capture such occurrences, all the handles were also converted into equivalent hashtags. The number of terms used, in this research, for the city profiles are summarized in Table 4.1:

Table 4.1: Number of Terms in Each City Profile

City Name (A)	**Total Number of Handles and Hashtags in a City Profile (B)**	**City Specific Keywords (C)**
Banff	114	'banff'
Calgary	214	'calgary' and 'yyc'
Edmonton	198	'edmonton' and 'yeg'
Fort McMurray	100	'fortmcmurray' and 'ymm'
Lethbridge	98	'lethbridge' and 'yql'
Medicine Hat	46	'medicinehat', 'mhat', 'medhat' and 'yxh'
Red Deer	112	'reddeer' and 'yqf'
St. Albert	72	'stalbert'

Terms in city profiles include handles and hashtags identified using snowball sampling plus variations of such handles and hashtags created by the use of term(s) noted in column (C) (labelled as 'City Specific Keywords') in Table 4.1.

In general, most of the cities have one name on Twitter (e.g., Edmonton, Calgary, Banff) but Medicine Hat is a unique example in the dataset as it has multiple variations of the city name such as 'medhat' and 'mhat'. Such variations were also taken into account and different variants were matched using regular expressions while implementing append and replace strategies. Also, two cities, Banff and St. Albert, have

no airport and thus, variations of handles and hashtags having city airport codes were not identified for such cities.

One of the interesting findings, noted in the evaluation of the city profile terms, is that the number of terms in the profile for each city was relatively proportional to the population size of each city. The population for these cities was collected from the Statistics Canada (para.1) [118] and noted in Chapter 3. This analysis was more anecdotal and no statistical test was performed for the correlation between population size of a city and the number of elements in that city's profile.

Cities having a relatively higher number of people have higher number of digital presence, leading to more profile terms associated with such cities. For example, Calgary has the highest population (=1,237,656) among the selected cities and also had the highest number of profile terms in the POP Framework. Similarly, Edmonton with the second highest population (=1,062,643) among the chosen cities also had a relatively higher number of profile terms. Two cities out of eight cities, Red Deer (=99,718) and Lethbridge (=87,572) did not have a significant difference in population sizes and neither did the number of terms in their profiles; the number of terms were relatively similar in number for both the cities. Medicine Hat has one of the lowest population sizes, and coincidently, this city also has the lowest number of profile terms. The other three cities were also interesting case examples, Fort McMurray (=66,573) and St. Albert (=65,589) which have relatively similar population levels and Banff with a population of 7,851. Banff had a relatively higher number of profile terms even though, according to the census, it has the lowest number of residents. The most plausible explanation of this anomaly could be due to the fact that Banff is a tourist destination. Banff, as a tourist destination has a large floating population (i.e., visitors visiting Banff) with a good

number of organizations and places, as well as the potential for more events and occasions happening there, thus leading to a relatively higher digital presence in contrast to the population size. Fort McMurray has a population in the same range as St. Albert has, but had a relatively higher number of terms in profile than St. Albert did. The possible reason for such a difference could be attributed to Fort McMurray being an industrial town and not in close proximity to any larger population and economic centre (e.g., near Calgary or Edmonton) while St. Albert is in close proximity to Edmonton which is a larger population centre, and has more tourist attractions and events taking place (e.g., food festival, Heritage festival, etc.). There are also more organizations operating from Edmonton, and serving both Edmonton and St. Albert.

This analysis provided opportunity for further investigation to explore the correlation between the population sizes and the number of terms in the city's digital profile.

4.2 Implementation of the POP Framework - Tweet Data Types

The Tweet dataset was pre-processed as discussed in the Chapter 3 and three types of dataset were created using the preprocessed data. These datasets were labelled as *Baseline Dataset*, *Appended Dataset*, and *Replace Dataset*.

Baseline Dataset – this dataset includes tweets on which the elements from the POP Framework created for different cities were neither appended nor replaced with the city names.

Appended Dataset – this dataset is a processed dataset of tweets on which the POP Framework was implemented by appending the city name in the tweets next to the term that matched with the term included in the city's profile.

Replaced Dataset – this dataset is a processed dataset of tweets on which the POP Framework was implemented by replacing the term in the tweets with the city name when there was a match of the term in the tweet with the term in the city's profile.

The examples of baseline data, appended data and replaced data are shown in Table 4.2 and Table 4.3 which provides sample tweets from the city of Edmonton and the city of Calgary. The 'Baseline Data' row in both the tables provide examples of baseline data having original tweets, but after pre-processing as done in initial data preparation, and without implementation of append and/or replace strategies.

The 'Appended Data' row in the tables provides examples of the city names as appended. For example, in Table 4.2, city name 'edmonton' was appended after '@edmontonjournal' and '@yegcc' as these terms matched with the terms in the profile of Edmonton.

The 'Replaced Data' row in the tables provides examples of the city names replaced. Using the example in Table 4.2, the city name 'edmonton' replaced both the terms '@edmontonjournal' and '@yegcc' in this specific tweet as these terms matched with the terms in the city of Edmonton profile.

Table 4.2: Example from the city of Edmonton after applying Append and Replace strategies

Data Type	Tweet Example
Baseline Data	@edmontonjournal the current mental capacity of @yegcc is that if an ashtray - so no they should not be making any decisions whatsoever
Appended Data	@edmontonjournal **edmonton** the current mental capacity of @yegcc **edmonton** is that if an ashtray - so no they should not be making any decisions whatsoever
Replaced Data	**edmonton** the current mental capacity of **edmonton** is that if an ashtray - so no they should not be making any decisions whatsoever

Table 4.3: Example from the city of Calgary after applying Append and Replace strategies

Data Type	Tweet Example
Baseline Data	Thanks to @OpenStreetsYYC for making this moment possible #YYC #yycbike #itsastampedething https://t.co/i3UDP6rbal
Appended Data	thanks to @openstreetsyyc **calgary** for making this moment possible #yyc **calgary** #yycbike **calgary** #itsastampedething https://t.co/i3udp6rbal
Replaced Data	thanks to **calgary** for making this moment possible **calgary** #itsastampedething https://t.co/i3udp6rbal

A handle such as '@yegcc' is also an example of how Edmonton airport code is incorporated in user-ids. Similarly, Table 4.3 also includes examples where Calgary airport code is included in user-id (e.g. @OpenStreetsYYC) and hashtags (#yycbike).

Such handles and hashtag terms were detected by the inclusion of such terms in the city profiles and supported by the use of regular expressions and findings.

4.3 Append and Replace Strategy – Implementation and Analysis

The append strategy implementation led to the inclusion of the city names in tweets when the terms of tweets matched with the terms in the city profiles. The replace strategy implementation led to the replacement of terms in tweets which matched with the terms in the city profiles by the relevant city name (Table 4.2 and Table 4.3).

Table 4.4 provides a summary of the number of times the city names were appended after the profile matching term (or number of terms replaced with the city names in tweets). These numbers include both handles and hashtags identified using snowball sampling technique, and variations of handles and hashtags based on the key city-related terms. Further, given that both the strategies should identify the same number of terms in the dataset that would match with terms in the city profiles, there would be an equal number of city names getting appended or city names getting replaced in the dataset. In the remaining discussion of this chapter and chapters thereafter, the term append will be used more often than replace.

A profile of each city was used to append the relevant city name in every city's tweet dataset. A rule was applied in the implementation of city profiles on tweets was that the first city profile that would run on that particular city's tweets would be the same city's profile and after that there was no particular order. For example, in appending 'Banff' as the city name in tweets that were categorized as Banff tweets, Banff profile was first implemented on such tweets and this was followed by the implementation of the other seven cities' profiles.

A total of eight cities were used in this research work and each city has a total of 500 tweets. There was a ninth category labelled as 'Others' that also has 500 tweets. There was no term in the 'Others' category tweets that matched with the terms in any of the eight city profiles i.e., zero time any city name was appended (or replaced). Note that this category is not included in further analysis or discussion in this Chapter.

4.3.1 Overall Number of Times City Names Appended in Tweets

In 4,000 tweets (500 tweets per city * 8 cities), a total of 3,780 terms matched with the terms in eight city profiles. Three predominant categories of cities emerged, based on the total number of times city names were appended (or city names were replaced), in each city's tweet dataset.

The number of times city names were appended in Banff and Red Deer tweets were almost similar, 340 and 341 respectively. Similarly, for Medicine Hat and St. Albert, the number of times city names were appended were 469 and 467 respectively. In all the other cities (Edmonton, Calgary, Fort McMurray and Lethbridge), the number of times city names were appended was more than 500 times. These numbers include both handles and hashtags identified using the snowball sampling technique, and variations of handles and hashtags based on the key city-related terms.

Another interesting phenomenon observed during this stage of data processing was that there were terms from other city profiles that were included in other cities' tweets as well. For example: in tweets from Banff, Banff and Calgary names were appended and a total of 340 city names were appended in Banff tweets and out of these, the name of the city of 'Banff' was appended 332 times and the 'Calgary' was appended eight times.

In general, the overall percentage of the presence of terms from the profile (i.e., user-ids and hashtags) of one city in another city's tweets was not so high except for one city and that is St. Albert. In St. Albert's dataset, other than St. Albert's name, Edmonton and Calgary names were also appended a total of 467 times. Out of 467, 'St Albert' was appended 386 times, 'Edmonton' was appended 80 times and 'Calgary' was appended once. The inclusion of 'Edmonton' 80 times in St. Albert tweets is quite a high number compared to other cities, including Fort McMurray which had Edmonton and Calgary related user-ids and hashtags 24 and eight times respectively. St. Albert could be a unique city as it is very close to Edmonton; both cities could be seen as twin cities. Edmonton is a larger economic centre and hosts more events and activities that are attended by residents of both cities, and thus, Edmonton finds its way into St. Albert related tweets.

Table 4.4 provides an overview of the number of times a city name was appended (or terms were replaced by city names) in different city tweets. The row is the number of tweets for a given city (e.g., Lethbridge) while the column defines the number of times city names were appended (e.g., in Lethbridge, Calgary was appended once, Fort McMurray was appended twice, Lethbridge was appended 549 times and Medicine Hat was appended once) by identifying both handles and hashtags and their variations.

Table 4.4: Number of Terms Appended and Replaced in Tweets for Each City

City Tweets (↓)	***City Names Appended / Replaced in Each City Tweet Dataset***									
	Banff	*Calgary*	*Edmonton*	*Fort McMurray*	*Lethbridge*	*Medicine Hat*	*Red Deer*	*St. Albert*	*Others*	*Total Terms Appended / Replaced*
Banff	**332**	8	0	0	0	0	0	0	0	**340**
Calgary	0	**498**	9	0	0	0	0	0	0	**507**
Edmonton	1	5	**542**	2	0	0	0	4	0	**554**
Fort McMurray	0	8	24	**517**	0	0	0	0	0	**549**
Lethbridge	0	1	0	2	**549**	1	0	0	0	**553**
Medicine Hat	0	6	0	0	6	**457**	0	0	0	**469**
Red Deer	0	2	1	0	0	2	**336**	0	0	**341**
St. Albert	0	1	80	0	0	0	0	**386**	0	**467**

4.3.2 Unique City Profile Terms in Tweets

A list of unique terms count was calculated and these unique terms count (see Table 4.5) include both handles and hashtags created using snowball sampling technique and variants of city name and airport code (if city has airport code). The break-up of the unique terms in each of these categories are included in Table 4.6.

The count of the unique terms showed that the smaller population centres (e.g., Medicine Hat, Red Deer, St. Albert, Lethbridge, Fort McMurray and Banff) had relatively fewer numbers of handles and hashtags that were used more often by users as compared to other handles and hashtags. For example: In Banff, there were a total of 48 unique user-ids and hashtags (42 from Banff profiles and 6 from Calgary profiles) counted for 340 terms found in Banff related tweets that matched with terms in different

city profiles. Thus, for smaller population centres, approximately 10% to 14% of total terms that were replaced were unique user-ids and hashtags (e.g., Banff - 48/340 = 14%; Red Deer – 43/341 = 12.6%; Fort McMurray – 55/549 = 10%).

The larger population centres (Calgary and Edmonton) had tweets of relatively higher numbers of unique user-ids and hashtags in their dataset, values ranging from 20% to 22%, when compared to 10% to 14% for smaller population centres. For example: 101 user-ids and hashtags (97 from Calgary and 4 from Edmonton) out of 507 total user-ids and hashtags identified and appended in Calgary dataset were unique. The unique user-ids and hashtags were 20% (=101/507) of the total user-ids and hashtags that were identified and appended in the dataset for Calgary. Similarly, for Edmonton, 122 user-ids and hashtags (113 from Edmonton, one from Banff, two from Calgary, two from Fort McMurray and four from St. Albert) out of 554 total user-ids and hashtags were identified and appended in Edmonton dataset were unique. The unique user-ids and hashtags were 22% (=122/554) of the total user-ids and hashtags that were identified and appended in the dataset for the city of Edmonton.

Out of the eight cities, St. Albert was an interesting case example. A total of 63 unique handles and hashtags were identified in tweets associated with St. Albert and out of these 63, 36 terms matched with terms in St. Albert's profile, one term matched with the profile of Calgary, and an eye-catching number of 26 terms matched with the profile of Edmonton. This example demonstrates a potential significant influence that a large population and economic centre can have on nearby smaller population centres. The number of unique terms for each city profile collected through snowball sampling is noted in Table 4.5. The counts for the unique terms and unique variants derived from the city names and airport codes, and their breakdown are included in Table 4.6.

Table 4.5: Unique POP Terms Appended/ Replaced in Tweets

City Tweets (↓)	***Unique Terms Appended / Replaced in Each City Tweet Dataset***									
	Banff	*Calgary*	*Edmonton*	*Fort McMur-*	*Lethbridge*	*Medicine Hat*	*Red Deer*	*St. Albert*	*Others*	*Unique Terms Appended / Replaced*
Banff	**42**	6	0	0	0	0	0	0	0	**48**
Calgary	0	**97**	4	0	0	0	0	0	0	**101**
Edmonton	1	2	**113**	2	0	0	0	4	0	**122**
Fort McMurray	0	3	6	**46**	0	0	0	0	0	**55**
Lethbridge	0	1	0	1	**76**	1	0	0	0	**79**
Medicine Hat	0	4	0	0	4	**56**	0	0	0	**64**
Red Deer	0	2	1	0	0	2	**38**	0	0	**43**
St. Albert	0	1	26	0	0	0	0	**36**	0	**63**

Data analysis of the unique terms count in Table 4.6 further shows that the users used a number of handles and hashtags which were variants of city names and airports. Such handles and hashtags primarily had a city name or an airport code, either at the end or beginning of the term (e.g., #yycbike or @ OpenStreetsYYC). All of them might be difficult to capture with the snowball sampling approach but they are important to capture to create effective city profiles as they help in identifying significant numbers of terms in the city-related tweets. For example, approximately, one-fifth (~21%) of terms in tweets for cities like Banff and Calgary were identified as variants of the city name and/or airport code. In cities, such as Lethbridge (16.4%), Red Deer (11.3%), Fort McMurray (10.4%) and St. Albert (8.8%) there were relatively less numbers of variants used in tweets but significant enough that they should be captured. Surprising data was

related to Edmonton and Medicine Hat as in these cities a significantly higher number of handles and hashtags (30.6% for Edmonton and 60% for Medicine Hat) had city name and/or airport code in them. This could be further investigated if a greater number of handles and hashtags, particularly for cities like Medicine Hat, have city name/airport code in them due to additional versions of the city name (e.g., Medicine Hat has multiple versions such as @/#medicinehat, @/#mhat, @/#medhat, @/#YXH) as compared to Edmonton (e.g., @/#edmonton, @/#yeg) and other cities in this list.

Table 4.6: Variation of Handles and Hashtags based on City Name and Airport Code used on Twitter

City Name	**Variants of City Name and Airport Code – Unique Count and Total Count Appended/Replaced**		**Profile Terms from the Snowball Sampling – Unique Count and Total Count Appended/Replaced**	
	Unique Count of Variants	*Total Count of Variants Appended / Replaced*	*Unique Count*	*Total Count Appended / Replaced*
Banff	23	71	19	261
Calgary	70	106	27	392
Edmonton	79	166	34	376
Fort McMurray	20	54	26	463
Lethbridge	51	90	25	459
Medicine Hat	32	274	24	183
Red Deer	23	38	15	298
St. Albert	24	34	12	352

Note related to Table 4.6: The numbers noted in this table are based on the terms identified in the specific city-labelled tweets, for example, 23, 71, 19 and 261 in the 'Banff' row from the tweets shortlisted as the city of Banff tweets. If any variant of

the profile term was found, in say other city's tweets (e.g., Calgary), it is not included in these counts.

4.3.3 Top Ten Handles/Hashtags

The top ten most popular user-ids and hashtags were also extracted for each city, and the results are included in Table 4.7 and Table 4.8.

The two most popular handles and/or hashtags for each city were city names and their airport codes for cities such as Calgary, Edmonton, Fort McMurray and Lethbridge, as in: #calgary (146) and #yyc (92) (Calgary airport code) for Calgary; #edmonton (158) and #yeg (127) (Edmonton airport code) for Edmonton; #ymm (211) and #fortmcmurray (123) for Fort McMurray, and; #yql (209) and lethbridge (96) for Lethbridge. This is in contrast to the four other cities where the two most popular handles and hashtags were: #banff (111) and @banff (81) for Banff; #reddeer (179) and @reddeer (99) for Red Deer; #stalbert (216) and @saintalbert (104) for St. Albert, and; #medhat (209) and #medicinehat (88) for Medicine Hat.

Analysis of the top ten terms provide an interesting insight that there are a number of variations of city names that are used in Twitter. For example: the variations for St. Albert include #stalbert, @stalbert, @cityofstalbert and @stalbert, and for Medicine Hat, the variations based on the city name include #medhat, #medicinehat, @medicinehat and @medicinehatcity. There are similar variations for other cities as well.

Among the top ten terms, there are handles and hashtags based on the city name and/or city airport code name used on Twitter that can reflect events happening in those cities. For example: the terms @calgarystampede (37) and #calgarystampede (27)

reflect the popular annual stampede event in the city of Calgary, and #yegfood (12) provides a context for food related events in Edmonton.

Table 4.7: Top 10 Terms Appended/Replaced in tweets from Banff, Calgary, Edmonton, and Fort McMurray

S. No.	City Name: Banff		City Name: Calgary	
	Terms	**Count**	**Terms**	**Count**
1	#banff	111	#calgary	146
2	@banff	81	#yyc	92
3	#banffnationalpark	30	@calgarystampede	37
4	@fairmontbanff	25	@calgary	27
5	@skilouise	11	#calgarystampede	27
6	#mybanff	11	#yyctraffic	16
7	@banff_squirrel	10	#stampede	15
8	@fairmontcll	5	@calgarytransit	8
9	@sunshinevillage	4	@calgarypolice	8
10	@banffavebrewing	3	@nenshi	6
S. No.	**City Name: Edmonton**		**City Name: Fort McMurray**	
	Terms	**Count**	**Terms**	**Count**
1	#edmonton	158	#ymm	211
2	#yeg	127	#fortmcmurray	123
3	@edmonton	17	@fmpsd	61
4	#yegwx	15	#yeg	18
5	@cityofedmonton	13	#ymmnews	15
6	#yegfood	12	#ymmarts	14
7	#yegbike	11	@fortmcmurray	9
8	#yegtraffic	10	@fortmactoday	9
9	@edmontonesks	9	@rmwoodbuffalo	8
10	#yegevents	9	#yyc	6

Table 4.8: Top 10 Terms Appended/Replaced in tweets from Lethbridge, Medicine Hat, Red Deer, St. Albert

S. No.	City Name: Lethbridge		City Name: Medicine Hat	
	Terms	**Count**	**Terms**	**Count**
1	#yql	209	#medhat	209
2	#lethbridge	96	#medicinehat	88
3	#lethbridgeab	55	@medicinehat	38
4	@lethbridge	30	@medhatmovember	16
5	@lethbridgecity	16	@medicinehatcity	8
6	#yqltraffic	10	@medhatfoodbank	8
7	@wots_lethbridge	9	@medicinehatspca	6
8	@ulethbridge	8	#yxh	6
9	@downtownleth	7	@medicinehatymca	6
10	@globalleth	6	@cjcyfm	5
S. No.	**City Name: Red Deer**		**City Name: St. Albert**	
	Terms	**Count**	**Terms**	**Count**
1	#reddeer	179	#stalbert	216
2	@reddeer	99	@saintalbert	104
3	#infinityphotog-raphyreddeer	10	#yeg	40
4	@cityofreddeer	6	@cityofstalbert	11
5	@smbreddeer	3	@stalbert	7
6	@rdnewsnow	2	#edmonton	5
7	@reddeercollege	2	#yegfood	5
8	@9roundreddeer	2	@stalbertpublic	4
9	@reddeerviews	2	@lifeinstalbert	3
10	@reddeervipers	2	@stalberthomes	3

Chapter 5

DigiCities Experimentation Results and Discussion

5.1 Overview

Weka application was used to conduct classification experiments, and the three algorithms NB, kNN and SMO (Sequential Minimal Optimization) were used. Previous research work in the classification area suggests that all three algorithms (i.e., NB, kNN & SMO) can achieve good results in text classification (e.g., [15], [50]). These algorithms were used to see if they would accomplish excellent results on tweets as well.

In the literature, it is noted that removal of stopwords and stemming of terms can help in improving classification accuracy of text data (e.g., [128]). This research, using the three different algorithms, also aimed at evaluating the impact of the removal of stopwords and stemming of terms on classification accuracy. During the pre-processing stage, stemming was not implemented and stopwords were not removed. This was done to understand the implications of both the stemming and stopwords removal in the context of append and replace strategies on the classification accuracy. Twelve different datasets were created based on: no implementation of stemming and removal of stopwords; a combination of stemming implementation and stopwords removal; and the implementation of append and replace strategies.

Further, these datasets were used with the three algorithms (NB, kNN, SMO) and classification experiments were conducted. Similar datasets were created on which

append and replace strategies were implemented. The classification results from the baseline dataset created the initial accuracy score benchmark, and these results were compared with the datasets with the implementation of append and replace strategies. Overall 36 experiments were conducted using the different combinations of algorithms and variants of datasets. A number of t-tests were conducted on different result combinations to check for statistical difference in the accuracy score achieved through various combinations of algorithm dataset and implementation of pre-processing techniques. The findings are organized into two broad areas – a) Impact of the POP Framework and its implementation through append and replace strategies; b) Impact of pre-processing (includes removal of stopwords and applying stemming algorithm) in context of the append and replace strategy implementation. Table 5.1 and Table 5.2 provide experiment nomenclature which will be used in the discussion in this chapter.

Table 5.1: Example of Experiment Names having Algorithm, Data, and Preprocessing Information

Experiment Name	**Elements**	**Description**
NB_B	NB	Algorithm used is Naïve Bayes
	B	Dataset used in the experimentation is baseline data. **Note**: Stopwords were not removed and stemming algorithm was not implemented.
SMO_A_SA	SMO	Algorithm used is SMO
	A	Dataset used in the experimentation is the one in which city names were **appended (A)** (by city names) when terms in the tweets matched the city profile POP Framework elements
	SA	Stemming algorithm implemented on the dataset in Weka
kNN_R_WS_SA	kNN	Algorithm used is kNN
	R	Dataset used in the experimentation is the one in which city names were **replaced (R)** (by city names) when terms in the tweets matched the city profile POP Framework elements
	WS	Without Stopwords (WS) i.e. Stopwords were removed from the dataset
	SA	Stemming Algorithm implemented on the dataset i.e., terms in this dataset were stemmed.

Table 5.2: Combination of Datatypes and Algorithms

Experiment Name	**Algorithms**			**Pre-Processing Strategies**			**POP Implementation Strategies**		
	NB	*kNN*	*SMO*	*WS*	*SA*	*WS and SA*	*None (Baseline)*	*A*	*R*
NB_B_SA	•				•		•		
NB_B_WS	•			•			•		
NB_B_WS_SA	•					•	•		
NB_A_SA	•				•			•	
NB_A_WS	•			•				•	
NB_A_WS_SA	•					•		•	
NB_R_SA	•				•				•
NB_R_WS	•			•					•
NB_R_WS_SA	•					•			•
kNN_B_SA		•			•		•		
kNN_B_WS		•		•			•		
kNN_B_WS_SA		•				•	•		
kNN_A_SA		•			•			•	
kNN_A_WS		•		•				•	
kNN_A_WS_SA		•				•		•	
kNN_R_SA		•			•				•
kNN_R_WS		•		•					•
kNN_R_WS_SA		•				•			•
SMO_B_SA			•		•		•		
SMO_B_WS			•	•			•		
SMO_B_WS_SA			•			•	•		
SMO_A_SA			•		•			•	
SMO_A_WS			•	•				•	
SMO_A_WS_SA			•			•		•	
SMO_R_SA			•		•				•
SMO_R_WS			•	•					•
SMO_R_WS_SA			•			•			•

5.2 Summary of Key Findings

The first key finding is that the appropriate selection of algorithm can help in achieving relatively better classification accuracy even with limited data pre-processing. In case of tweet dataset used in this research, SMO proved to be the better choice from the three algorithms chosen for this study. A total of nine classes were included in the dataset. These include eight cities – Banff, Calgary, Edmonton, Fort McMurray, Lethbridge, Medicine Hat, Red Deer and St. Albert, and the ninth was 'Others' category. All of the 36 experiment accuracy scores are summarized and presented in Figure 5.1. These experiments, with all three classification algorithms, and after applying both the replace and append strategies, show varying results (i.e. 47.6% to 94.2%). SMO had limited impact on strategies or data preprocessing techniques. However, kNN and NB had shown substantial impact of both the strategies. Furthermore, applying stemming had no impact on any algorithm, however, removal of stopwords had impact on the classification accuracy of all the algorithms.

NB algorithm, with baseline data, performs really well i.e., 69.9% accuracy was achieved without applying any data preprocessing (stemming and removing stopwords). Removing stopwords enhances the results quite a bit (from 69.9% to 77.3%) and t-test suggests that the difference between these accuracy scores is significantly different (see p value for NB_B, NB_B_WS in Appendix B). However, when applying stemming, rather than increasing the accuracy, it degrades result marginally (69.9% to 69.6%) and t-test suggests that the difference between these accuracy scores is not significantly different (see p value for NB_B, NB_B_SA in Appendix B). kNN result on baseline data is relatively poor with the accuracy of 47.6%. After applying only stemming, the accuracy is 48.3% and after removing stopwords (only), the accuracy is 58.8%. Further,

after applying stemming and removing stopwords together, the accuracy improves to 60.2%. The results from t-tests suggest similar outcome as observed with NB i.e., after applying stemming, there is no statistical difference in the accuracy score (see p value for kNN_B, kNN_B_SA in Appendix B) but after stopwords removal, there is a statistical difference between the accuracy scores (see p value for kNN_B, kNN_B_WS in Appendix B). However, SMO worked amazingly well with 87.8% accuracy from the baseline data, stemming reduced the accuracy marginally to 87.4% and stopwords removal improved the accuracy to 89.1%. Furthermore, applying stemming and removing stopwords (together), the accuracy improved to 88.5%. The t-tests performed on these experiment results imply that there is no statistical difference between results achieved before and after implementing preprocessing approaches (i.e., stemming and stopwords removal).

The accuracy results of all the algorithms run on append datatype were better among the three datatypes. For example, kNN improved to 69.6%, NB gave an accuracy of around 85% and SMO performed better at 93.9% without applying stemming and removing stopwords. Applying only stemming does not improve any results further whether it is kNN, NB, or SMO (i.e. 70%, 85.2% and 93.9 respectively); however, removing stopwords does improve accuracy significantly for all to 83%, 89.9% and 94.2% respectively. The results from t-tests suggest that after the implemenation of append strategy, the removal of stopwords made no significant difference on the accruacy score for both kNN and SMO algorithm but significant difference on the accruacy score for NB (see p values for kNN_A and kNN_A_WS; NB_A and NB_A_WS; SMO_A and SMO_A_WS in Appendix B).

Replace datatype also works well with all three algorithms. NB gave 81% accuracy and kNN gave 56.1% accuracy without applying any data preprocessing. SMO algorithm does not show ample difference with the append or replace datatype. As with the replace data, the accuracy was 93.8% (while for append it was 93.9%). Applying stemming to NB further decreases to 80.3%, whereas kNN increases to 57.8% and SMO to 94%. Apparently, removing stopwords does increase the accuracy significantly for both kNN and NB algorithms i.e. the accuracy score improved from 81% (NB_R) to 88.4% (NB_R_WS), and similarly, the accuracy score improved from 56.1% (KNN_R) to 74.6% (kNN_R_WS). However there was no statistical difference in the scores for SMO as it improved from 93.8 (SMO_R) to 94.1% (SMO_R_WS) (see p values in Appendix B).

5.3 DigiCities – Impact of the POP Framework on Tweet Classification

The implementation of the POP Framework, using both append and replace strategies, had an impact on the classification accuracy. The analysis in this section focuses on the baseline data (denoted with the symbol 'B' in the experiment name) with append data (denoted with 'A') and replace data (denoted with 'R') created through the implementation of the append strategy and replace strategy respectively. This discussion does not focus on results obtained after stopwords removal and implementation of stemming algorithm as they are discussed in a separate sub-section later in this chapter.

Overall, the best classification results were achieved by the use of SMO algorithm for any data type i.e., whether it was baseline data or data created after the implementation of append strategy and replace strategy. The accuracy score for the

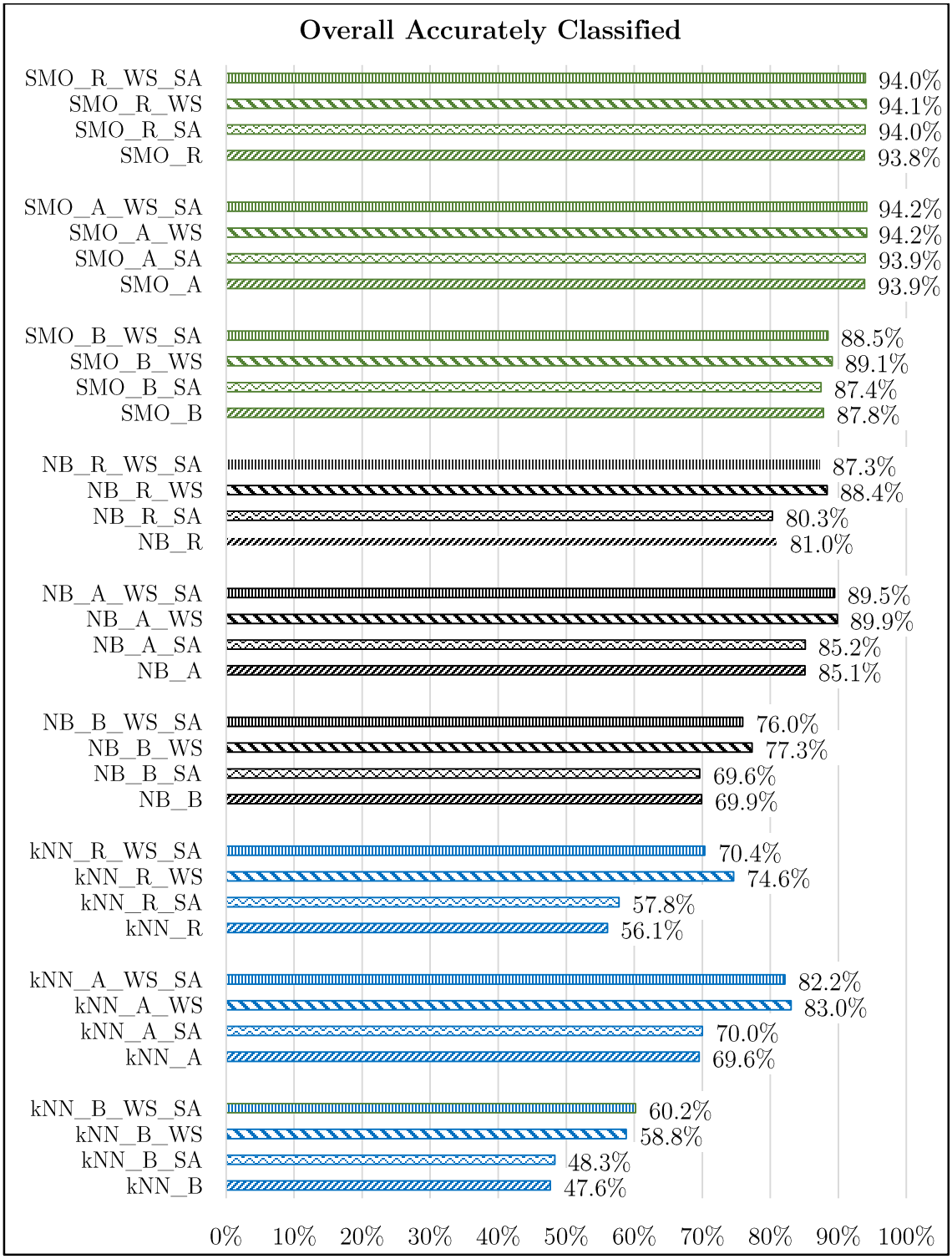


Figure 5.1: Overall Classification Score of 36 Experiments Based on Algorithm, POP Strategy and Pre-Processing Strategies Implementation

baseline data for kNN algorithm was the lowest with 47.6% (kNN_B), followed by NB algorithm with 69.9% (NB_B), and the best result was with SMO algorithm with 87.8% (SMO_B) accuracy (Figure 5.2).

The implementation of the POP Framework using both append and replace strategies improved the overall classification accuracy of tweets for all three algorithms when compared with the baseline data classification accuracy for each algorithm (Figure 5.2). The accuracy score for kNN algorithm improved significantly from 47.6% (kNN_B) to 56.1% (kNN_R) and 69.6% (kNN_A) with the implementation of replace strategy and append strategy respectively. Similarly, the accuracy score for NB algorithm improved significantly from 69.9% (NB_B) to 81.0% (NB_R) and 85.1% (NB_A) for replace strategy and append strategy respectively. For SMO, these numbers were also improved; they changed from 87.8% (SMO_B) to 93.8% (SMO_R) and 93.9% (SMO_A) (Figure 5.2).

The append strategy had the highest impact on the classification accuracy for the kNN algorithm (score improve by 21.9%) as compared to NB and SMO algorithms where scores improved by 15.2% and 6.1% respectively. The replace strategy also improved the classification accuracy for all three algorithms but had a relatively lesser impact than the append strategy.

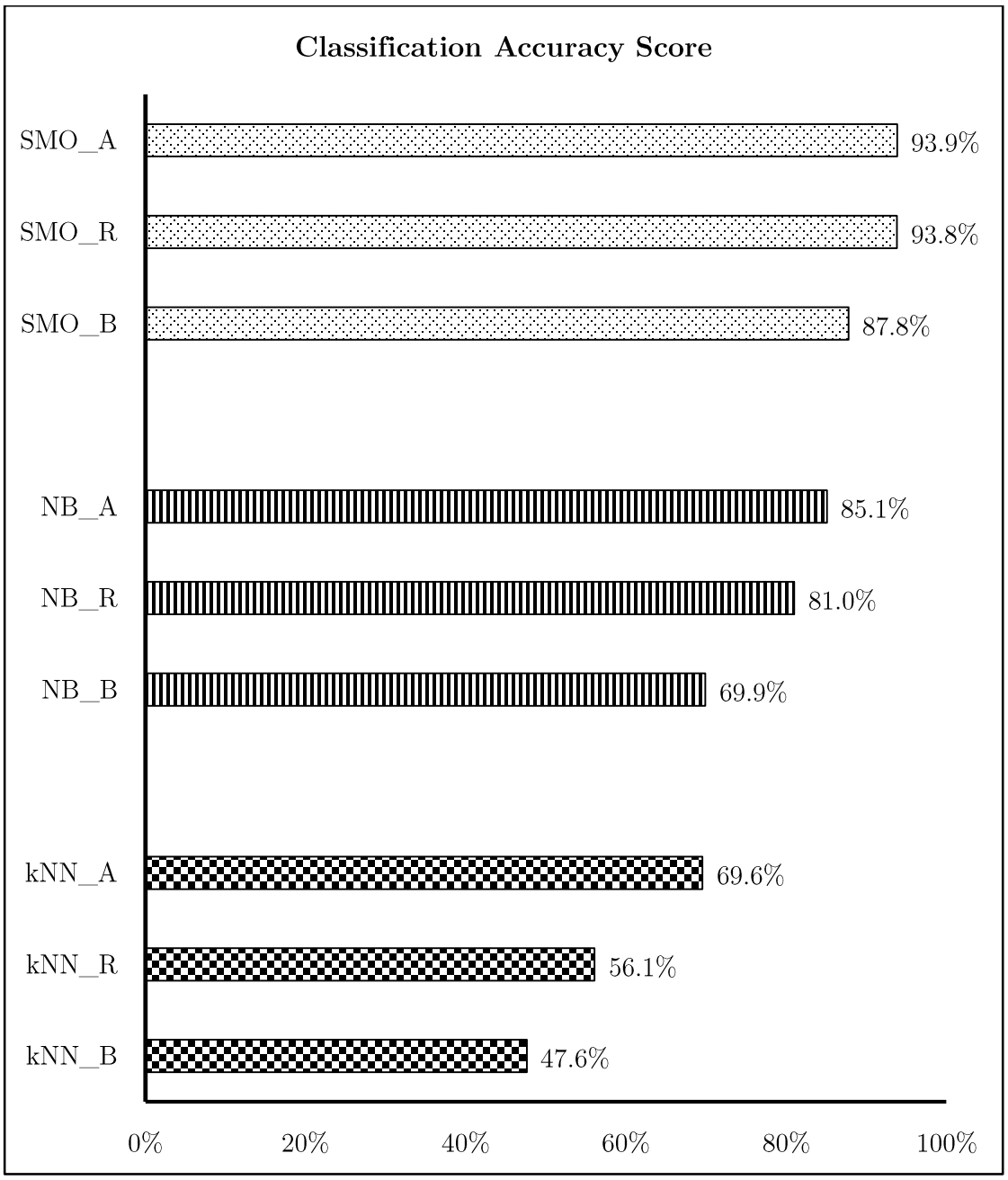


Figure 5.2: Classification Accuracy of Different Data Types and Algorithms-1

A number of t-tests were conducted to evaluate if the change in accuracy score was significant or not. Figure 5.3 provides a graphical representation of the p-scores on dual

polarity i.e., 'significantly different' or 'not significantly different'. The t-test results (p-values in the Appendix B) show that:

- There is a significant statistical difference in the accuracy scores for all the algorithms (see (datasets – algorithmname_B, algorithmname_R, algorithmname_A) between baseline data, and datasets created after the implementation of the append strategy and the replace strategy (Figure 5.3). The p-value for: kNN_B and kNN_R is 0.01134; kNN_B and kNN_A is 0.00067; NB_B and NB_R is 0.00064; NB_B and NB_A is 0.00017; SMO_B and SMO_R is 0.00071, and; SMO_B and SMO_A is 0.00063. All these values are at less than 5% significance level. Thus, demonstrating that there is a significant difference in the accuracy score achieved with the implementation of the append and replace strategies vis-a-vis accuracy results from the baseline dataset.

- There is also a significant statistical difference in the accuracy scores between the append strategy and the replace strategy implementation for the kNN algorithm (kNN_A, kNN_R – green colour bar in Figure 5.3) with p-values equal to 0.00350 (<5% significance level) and the NB algorithm (NB_A, NB_R – blue colour bar in Figure 5.3) with p-value equal to 0.00161 (<5% significance level). One of the interesting findings is that there is **no** significant statistical difference between the append strategy and the replace strategy implementation for the SMO algorithm (SMO_A, SMO_R – orange colour bar in Figure 5.3) with p-value equal to 0.824 (5% significance level).

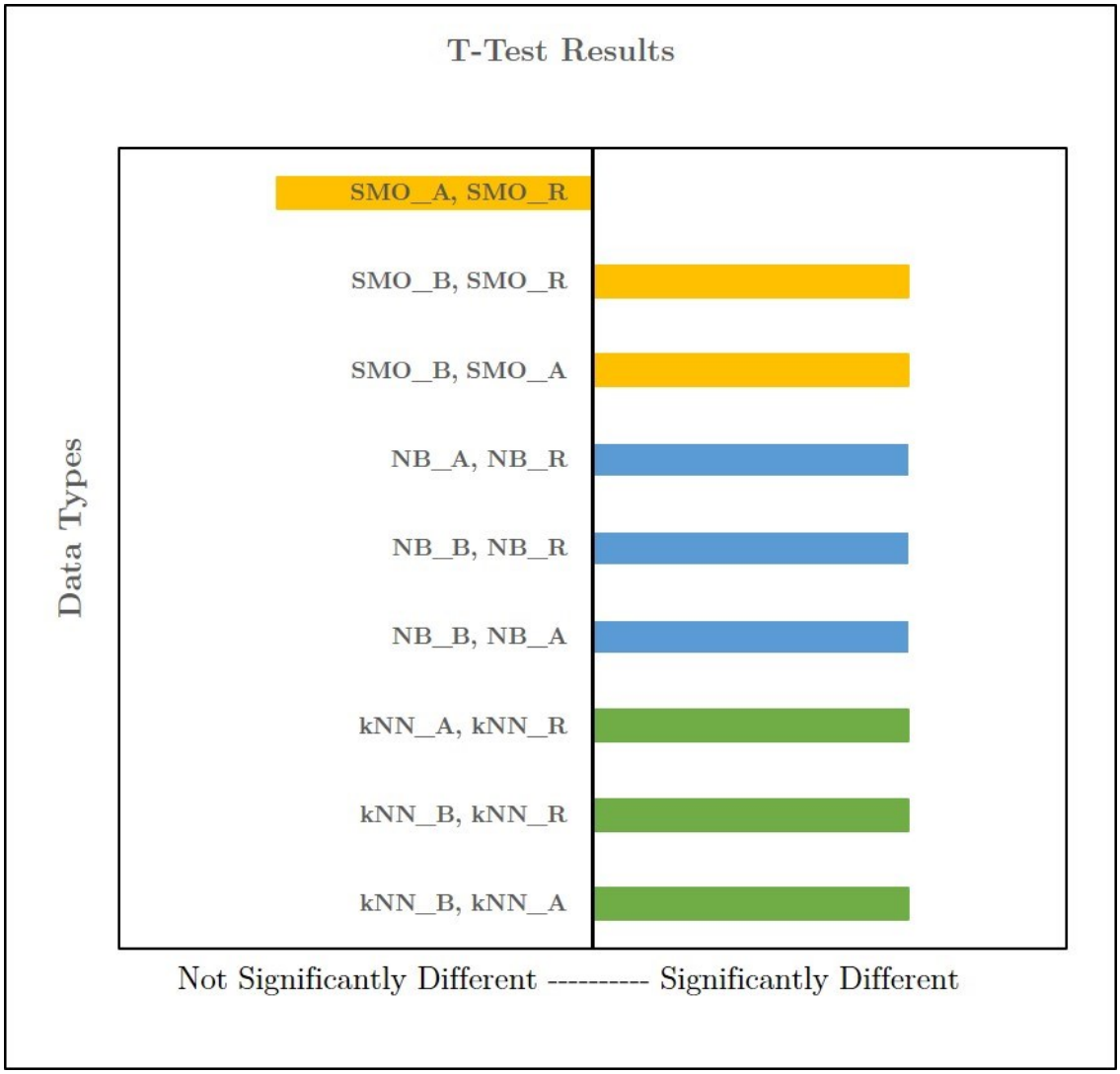


Figure 5.3: P-Values after Append and Replace Strategy Implementation

The analysis of confusion matrices (included in C.1- C.36 in Appendices) provides some interesting insight into misclassification of tweets in different classes.

- kNN algorithm had the lowest accuracy score among all the three algorithms. The kNN algorithm misclassified a large number of tweets from the 'Others' category into different cities. Out of 500 tweets in the 'Others' category, only 18, 45 and 4

tweets from the Baseline, Append and Replace datasets were classified into 'Others' respectively, the rest were classified into the selected cities. The top three cities in which 'Others' category tweets got classified were Edmonton, Fort McMurray and Medicine Hat for all the three datatypes. For example, the city of Fort McMurray had the most misclassified tweets (from Baseline – 1052 tweets, from Append Data – 481 tweets, from Replace Data – 1145 tweets) from other cities, followed closely by Edmonton (from Baseline – 915 tweets, from Append Data – 313 tweets, from Replace Data – 318 tweets).

- Classification accuracy by NB algorithm was relatively better than with kNN algorithm (as noted in this sub-section). Interestingly, a large number of tweets were misclassified in the 'Others' category by NB algorithm for all the three data types but less tweets were misclassified in this category with the implementation of the Append and Replace strategies (e.g., from Baseline – 622 tweets, from Append Data – 371 tweets, from Replace Data – 388 tweets). The other two top cities in which NB misclassified tweets were Edmonton (from Baseline – 219 tweets, from Append Data – 78 tweets, from Replace Data – 138 tweets) and Banff (from Baseline – 136 tweets, from Append Data – 77 tweets, from Replace Data – 97 tweets). For both cities, misclassification decreased for both the append and replace relative to baseline data, but the replace data had a relatively higher number of misclassified tweets as compared to the append data.

- The best classification results were by SMO algorithm. Primarily, the misclassification of the tweets from the eight cities were into the 'Others' category (e.g., from Baseline – 368 tweets, from Append Data – 213 tweets, from Replace Data – 211

tweets). There were very few tweets that were misclassified in other cities. The highest number of misclassified tweets for the baseline data was in the city of Edmonton with 41 tweets followed by Medicine Hat with 40 tweets. For both the append and the replace strategy, the highest misclassified number of tweets in other cities were 22 each.

5.4 Impact on Append and Replace Strategies in Context of Stopwords Removal and Stemming

The removal of stopwords is a widely adopted practice in the classification experiments. Stopwords are removed to enhance the text data quality by reducing the dimensionality of data [128]. A number of experiments were also performed by having or removing stopwords in the dataset. The aim was to understand the impact of the presence or absence of stopwords in the context of the proposed approach. A set of experiments were conducted, using three algorithms, to understand the impact of stopwords removal and stemming on classification when implemented alongside the append and replace strategies. The following sub-section discusses the findings from these experimentations.

5.4.1 Impact of Stopwords Removal

There was a varying degree of impact on the classification accuracy after stopwords were removed by the datasets. Figure 5.4 provides the accuracy scores on three different datasets, with use of the three algorithms, after stopwords were removed from each dataset.

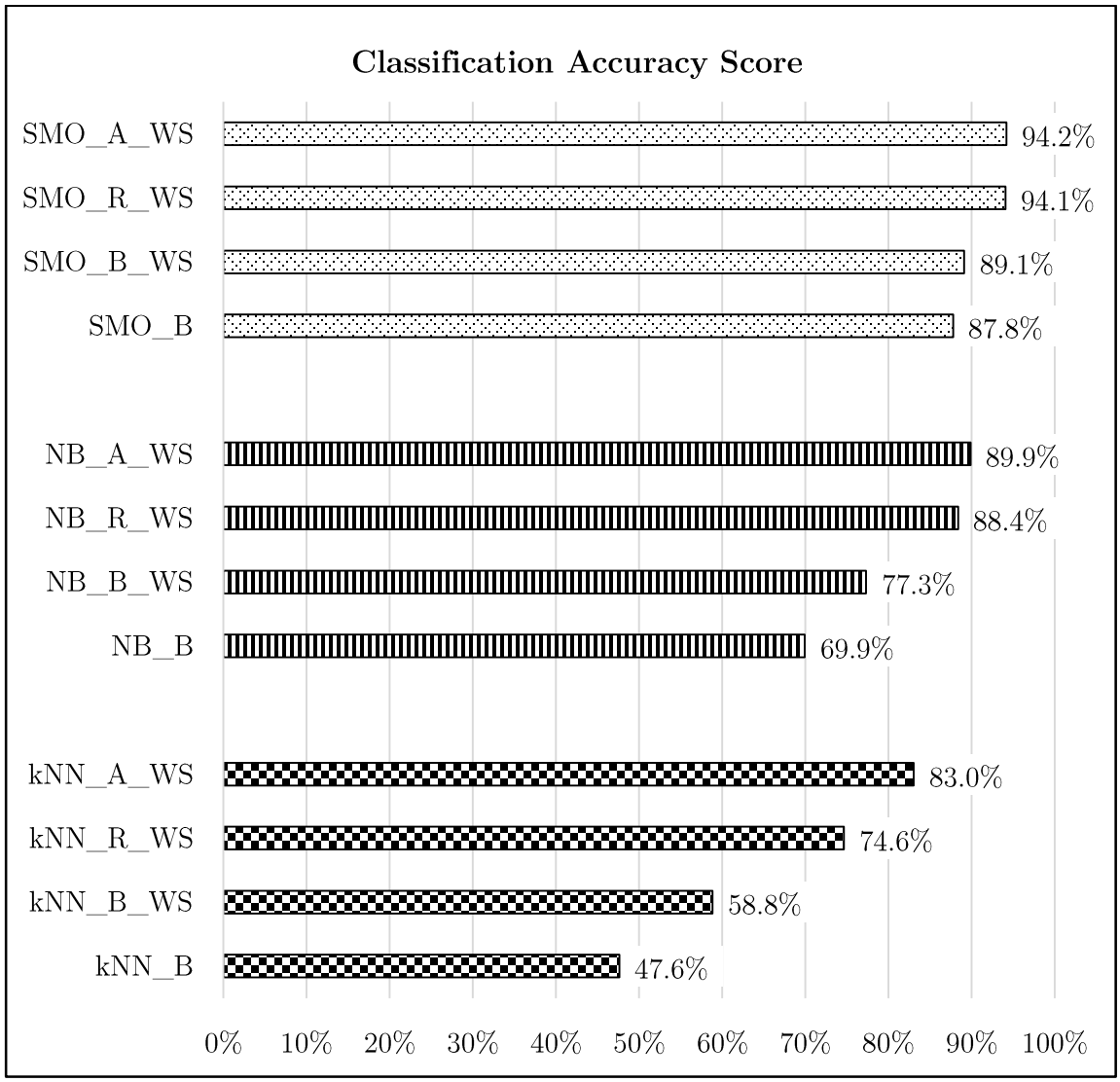


Figure 5.4: Classification Accuracy of Different Data Types by Different Algorithms-2

- There was a significant improvement in the classification accuracy for the kNN algorithm after stopwords were removed from the datasets. The accuracy score improved from 47.6% (kNN_B) to 58.8% (kNN_B_WS), 56.1% (kNN_A) to 83.0% (kNN_A_WS), and 69.6% (kNN_R) to 74.6% (kNN_R_WS) for the baseline, ap-

pend and replace dataset respectively. These improved scores were statistically significantly different and thus, these results show that the removal of stopwords had an impact on the accuracy score. With the use of append strategy, the accuracy score improved after the removal of stopwords from 69.6% (kNN_A) to 83.0% (kNN_A_WS) but surprisingly the t-test results suggest that this score is not significantly different (p-value = 0.0988 > 5% significance level). This result shows that there is no impact of removal of stopwords with the append strategy on the accuracy score for kNN algorithm. Figure 5.5 provides a graphical representation of calculated p-values if the values are significantly different or not significantly different. The p-values are included in the Appendix B.

- The removal of stopwords from the datasets also improved the classification accuracy for both NB and SMO algorithms. For the NB algorithm, the baseline data score improved after the removal of stopwords from 69.9% (NB_B) to 77.3% (NB_B_WS). Similarly, the scores for both append and replace strategies improved the accuracy for the NB algorithm after the removal of stopwords. For example, the use of append strategy improved the accuracy from 85.1% (NB_A) to 89.9% (NB_A_WS) (and see Figure 5.4 for the accuracy score for NB_R and NB_R_WS). These accuracy scores are significantly different which implies that, for the NB algorithm, the use of both append and replace strategies, and the removal of stopwords will improve the accuracy score (see Figure 5.5 and Appendix B for p-values). In case of the SMO algorithm, though there is marginal improvement in the accuracy score with the use of append and replace strategy after removing stopwords but the change in the accuracy is not statistically significantly different (i.e., SMO_A: 93.9%;

SMO_A_WS: 94.2% with p-value = 0.11 > 5% significance level; and SMO_R: 93.8%; SMO_R_WS: 94.1% with p-value = 0.71 > 5% significance level).

- Between the append and replace strategy when evaluated after the removal of stopwords, there is no significant difference in the accuracy scores for the SMO algorithm (p-value for SMO_A_WS and SMO_R_WS is 0.1388 > 5% significance level) and kNN algorithms (p value for kNN_A_WS and kNN_R_WS is 0.2614 > 5% significance level). However, for the NB algorithm, there is statistical difference in the accuracy score between the append and replace strategies when evaluated after the removal of stopwords (p value for NB_A_WS and NB_R_WS is 0.03 < 5% significance level).

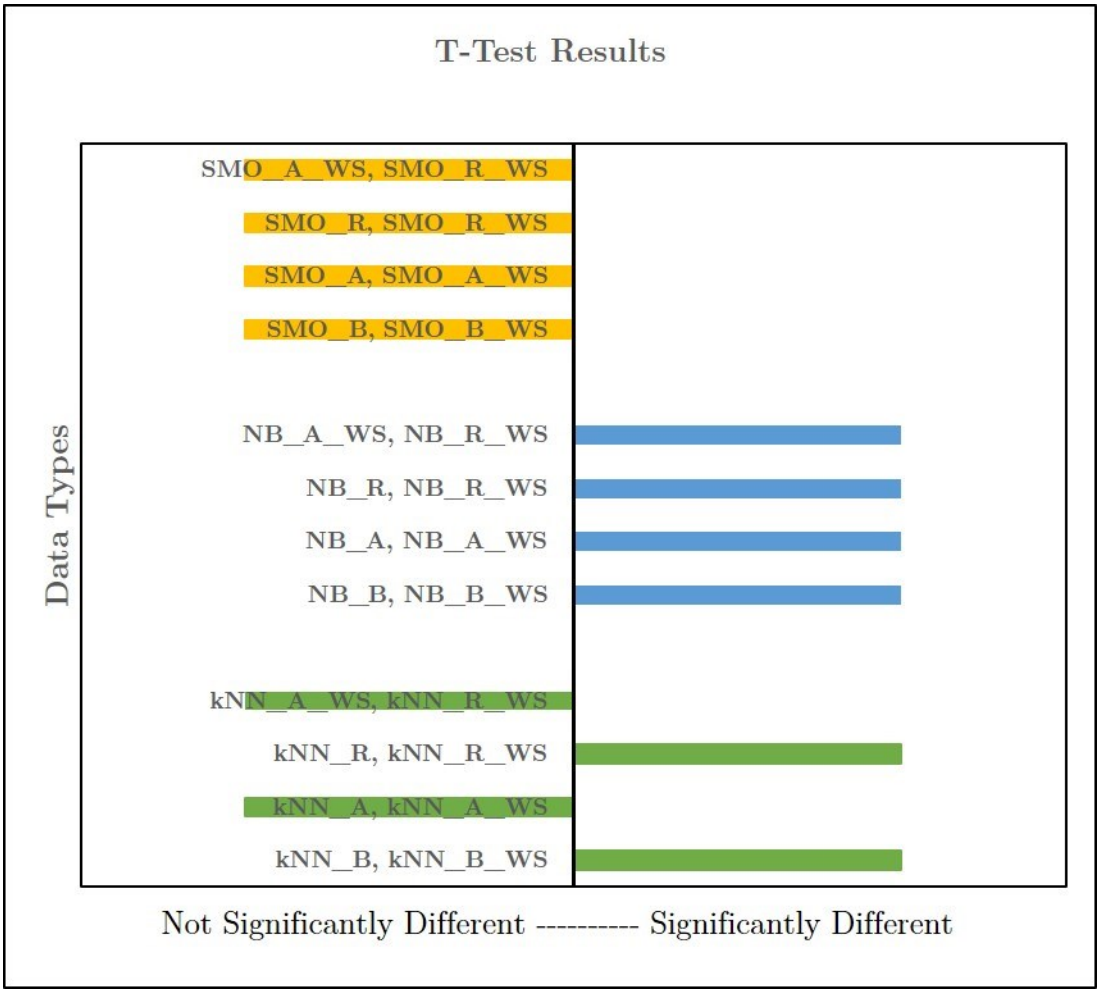


Figure 5.5: P-Values after Stopwords Implementation

The analysis of confusion matrices (included in Appendices in C.1- C.36) provides some interesting insight into (mis)classification of tweets into different classes.

- The kNN algorithm had the lowest accuracy score of all three algorithms. As in baseline data without stopwords removal, the algorithm misclassified a large number (=397) of tweets from 'Others' category into different city categories. However, there was a significant improvement in classifying 'Others' category tweets in its own category, the jump from 18 in baseline data having stopwords to 103 after removal of stopwords; but this dropped to 56 in the context of replace strategy and after

removal of stopwords. However, the append strategy enhanced the classification in the 'Others' category to 364 out of 500 after the removal of stopwords. These numbers were 18 and 45 with the baseline data and append data before the removal of stopwords. The notable numbers in the kNN algorithm classification is that some classes include: Fort McMurray acquired 1002 tweets from other classes, these were in addition to 417 correctly classified tweets from Fort McMurray; 1002 dropped to 768 and 88 in the replace and append data respectively after stopwords were removed. The second city which had the most misclassified tweets was Medicine Hat with 309 tweets for kNN_B_WS (baselines data and after removal of stopwords), and there was a significant change in number (=300 misclassified tweets) for this city with append data and after removal of stopwords but the surprising result was with the replace strategy (after removal of stopwords) with numbers dropping to 62.

- Again, classification by NB algorithm was relatively better than with kNN algorithm but not as good as SMO. There were two key categories in which the highest numbers of tweets were misclassified. First was the 'Others' category in which 541 tweets were misclassified and second was 'Edmonton' category in which 153 were misclassified. With the implementation of the append strategy and after removal of stopwords, these two categories were the biggest gainers. The numbers dropped to 277 and 57 for 'Others' and 'Edmonton' class respectively. However, with the implementation of the replace strategy (and removal of stopwords), the numbers changed to 232 and 90 respectively.

- Again, the best classification results were with the SMO algorithm. Primarily, the misclassification of tweets from the eight cities went into the 'Others' Category (e.g., from Baseline data + no stopwords: 395 tweets, from Append data + no stopwords:

207 tweets, from Replace data + no stopwords: 206 tweets). There were relatively very few tweets that got misclassified in other cities. The highest number of misclassified tweets for the baseline data (after removal of stopwords) was in Medicine Hat with 20 tweets, Calgary with 21 and 22 for append and replace strategy (after removal of stopwords) respectively.

5.4.2 Impact of Stemming

Stemming is an important technique while implementing information retrieval. Stemming also helps in reducing data dimensionality by taking any term to its root term. This study used Lovins stemmer [72], implemented in Weka. This stemmer in Weka also converted terms to lower case before stemming them. However, most of the previous research work suggests that stemming has limited effect on accuracy [43]; this has also been proven in this study. It is important to note that the findings are based on the use of one particular stemming i.e., Lovins algorithm. The accuracy scores for each algorithm (kNN, NB, SMO) were evaluated after the implementation of stemming algorithm on all the three datatypes (i.e., baseline data, append data and replace data), and the implementation of stemming algorithm did not have an impact on the classification score of all three algorithms and respective data types (See Figure 5.2 and Figure 5.6). For example:

- The accuracy score before and after the implementation of stemming algorithm did not change for the same datatype. The classification accuracy changed from 47.6% (kNN_B) to 48.3% (kNN_B_SA), from 69.9% (NB_B) to 69.6% (NB_B_SA), and 87.8% (SMO_B) to 87.4% (SMO_B_SA). Similar trends were observed on append and replace data as well (Figure 5.2 for kNN_A vs. kNN_A_SA; kNN_R vs. kNN_R_SA as well as for NB and SMO).

- The statistical tests were conducted on various combinations of results to evaluate if there were any statistically significant differences in the accuracy scores arising due to data quality emerging before and after the implementation of stemming algorithm. Figure 5.7 (a graphical representation of p-values reflecting significant difference or no significant difference) shows that the impact of stemming on accuracy score. For example, there was no significant difference in the accuracy score for SMO before and after implementation of stemming on the same data type (e.g., p values for SMO_B and SMO_B_SA; SMO_A and SMO_A_SA, and; SMO_R and SMO_R_SA are 0.3486, 0.7051 and 0.5539 respectively). However, for the NB and kNN algorithm, some deviations in the outcome was observed as compared to the SMO algorithm. For the NB algorithm, the statistical difference was observed after the implementation of append and replace strategy (i.e., the p-value for NB_A_SA and NB_R_SA is $0.0006 < 5\%$ significance level). The statistical difference in results was observed in two cases for the kNN algorithm (see Figure 5.7). In summary, there was a statistical difference in the scores due to the implementation of append and replace strategies even after the implementation of stemming algorithm for kNN and NB algorithms which was not observed in case of the SMO algorithm. The p–values for all the statistical tests (as discussed above) are included in Appendix B.

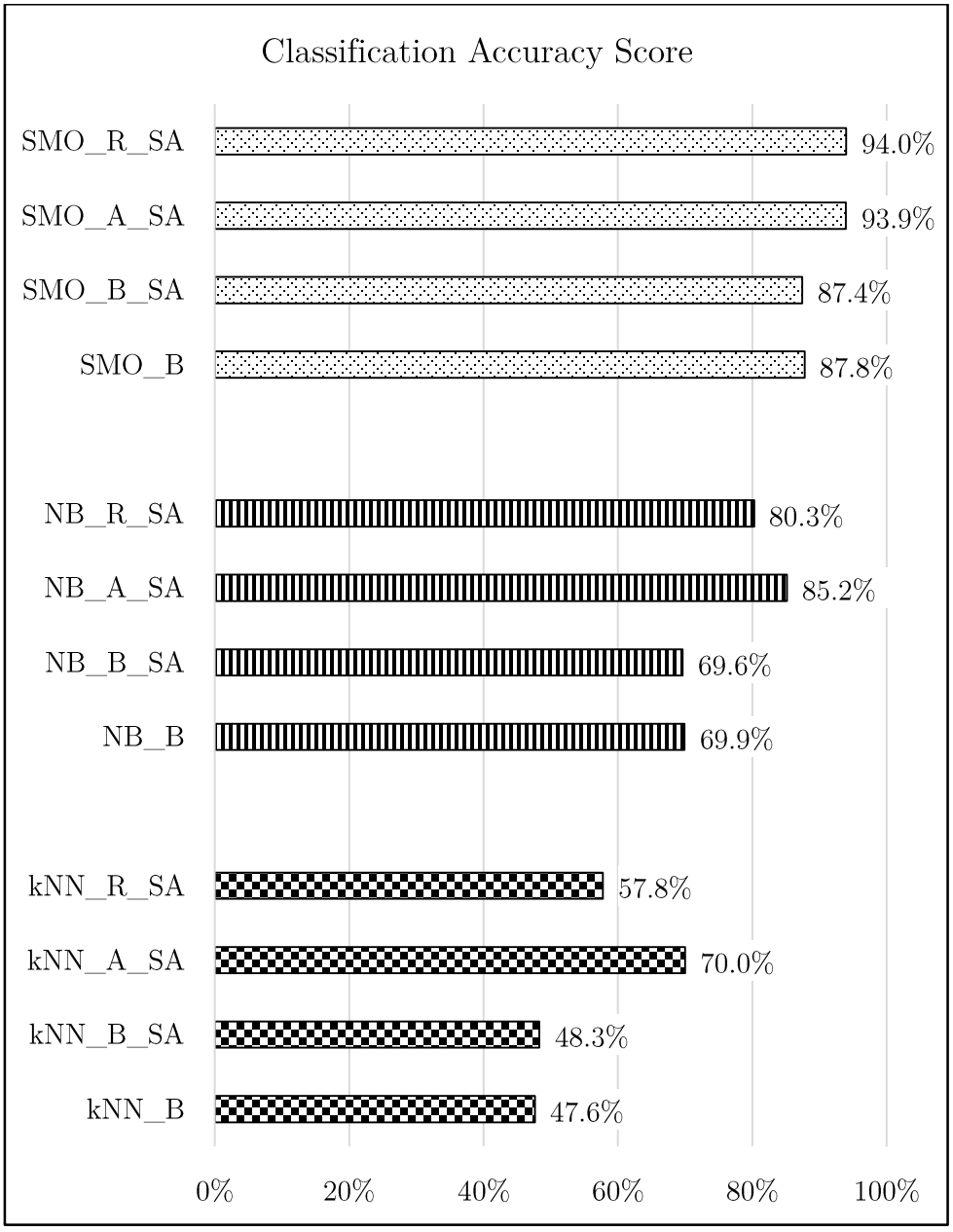


Figure 5.6: Classification Accuracy with Stemming Implementation

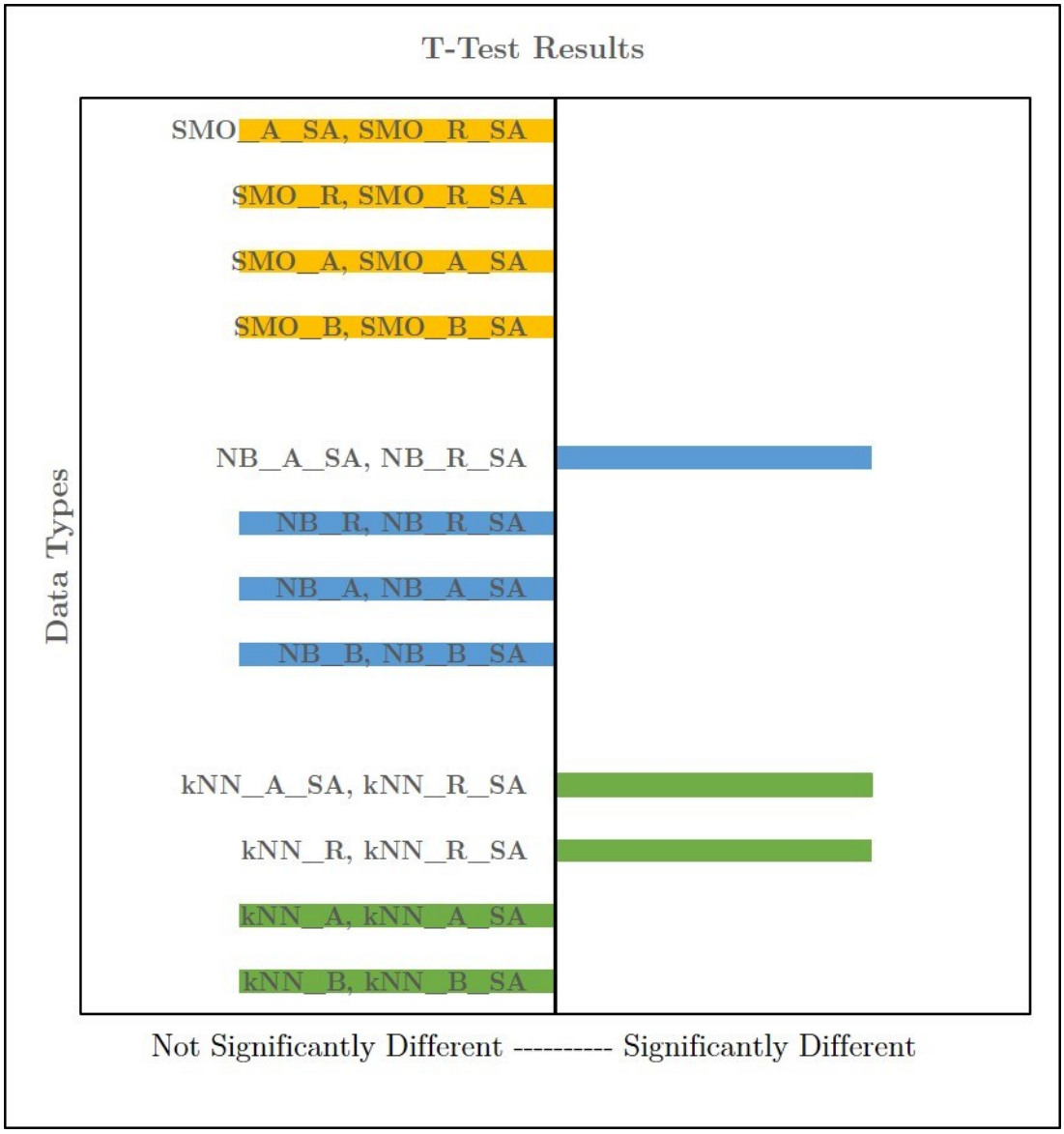


Figure 5.7: P-values after Stemming Implementation

Since, there is only a marginal change and not a significant improvement in the accuracy scores for baseline, append and replace data before and after the implementation of stemming algorithm, the pattern of tweets misclassification, of and in different cities, is similar to results obtained without the implementation of stemming

algorithm on the datasets. Examples of results from the confusion matrices in the context of kNN algorithms are discussed below.

- The kNN algorithm misclassified a large number of tweets from the 'Others' category into other cities. Out of 500 tweets in each data type only 20, 56, and 4 for baseline, append and replace dataset, after the implementation of stemming algorithm, were classified into 'Others' category respectively, the rest were classified into other cities. After implementation of the stemming algorithm, the top three cities into which 'Others' category tweets were classified were Edmonton, Fort McMurray and Medicine Hat for all three datatypes, and these cities are the same ones as before the implementation of the stemming algorithm on the datasets. Fort McMurray had the most misclassified tweets (from Baseline – 1015 tweets, from Append Data – 452 tweets, from Replace Data – 1054 tweets) from other cities including the 'Others' category followed by Edmonton (from Baseline – 894 tweets, from Append Data – 309 tweets, from Replace Data – 328 tweets).
- The pattern of misclassification for NB and SMO are similar to as discussed in Section 5.3.

5.4.3 Impact of both Stemming and Stopwords Removal

Combining application of both stemming and removing stopwords helped in reduction of feature space and data dimensions. The impact of stopwords removal and stemming was evaluated while they were implemented separately (discussed in Section 5.4). Experiments were also conducted to evaluate if both stopwords removal and stemming, when implemented together, will have any different outcome on the classification accuracy. The discussion in this sub-section will focus on the classification

accuracy achieved after both the stopwords removal and stemming were implemented together vis-à-vis classification accuracy achieved after removing stopwords only. The rationale for such focus is that the results achieved when both were implemented together were very close to the results achieved after only the stopwords removal; other results, particularly those achieved after stopwords removal, and from other data variants, have been discussed in-depth in earlier sub-sections in this chapter.

- The combined implementation of stopwords removal and stemming had mixed impact on the accuracy scores when compared with results obtained after stopwords removal alone. In general, there is a drop in the accuracy scores, for the same datatype (i.e., baseline, append and replace), after the implementation of both stopwords removal and stemming over the accuracy scores achieved after the removal of stopwords alone.

- These was no significant gain in the classification accuracy for the SMO algorithm; primarily the accuracy scores achieved after stopwords removal and stemming algorithm applied were very close (or there was a marginal drop) to the accuracy scores achieved after stopwords removal alone. For example, the classification accuracy was 89.1% for SMO_B_WS (baseline data with stopwords removed) and 88.5% for SMO_B_WS_SA (baseline data with both stopwords removed and stemming applied). The accuracy scores for the append strategy were 94.2% for SMO_A_WS and 94.2% for SMO_A_WS_SA while for the replace strategy the scores were 94.1% for SMO_R_WS and 94.0% for SMO_R_WS_SA. There was no significant difference in the accuracy scores as reflected in Figure 5.9 (orange bars), and the p-values are in Appendix B.

- The accuracy scores for the append strategy dropped marginally from 89.9% for NB_A_WS to 89.5% for NB_A_WS_SA (Figure 5.9) in case of the NB algorithm and this drop was not statistically significant as represented in Figure 5.9 (blue bars). Similarly, the drop in the accuracy score for both the baseline data and the replace-strategy based data for the NB algorithm (e.g., the score of 88.4% for NB_R_WS dropped to 87.6% for NB_R_WS_SA). In both the cases, the differences in the accuracy score were statistically significant (see Figure 5.9 with blue bars). However, the append strategy (NB_A_WS_SA: 89.5%) worked relatively better than the replace strategy (NB_R_WS_SA: 87.3%) when stopwords were removed and stemming was applied and difference was statistically significant (p value is 0.0032).

- In the case of kNN, the accuracy score dropped for both the append and replace strategy; the scores dropped from 74.6% for kNN_R_WS to 74.0% for kNN_R_WS_SA, and they dropped from 83.0% for kNN_A_WS to 82.2% for kNN_A_WS_SA (Figure 5.8). The drop in the accuracy scores for the replace strategy (former) was statistically significant (p-value = 0.0332 $<$ 5% significance level) while the drop for the append data (latter) was not significant (p-value = 0.79 $>$ 5% Significance Level) (Figure 5.9 green bars). The result for the baseline data for the kNN algorithm was interesting and the only case when the accuracy score improved with stopwords removal and stemming as compared to stopwords removal i.e., it improved from 58.8% for kNN_B_WS to 60.2% kNN_B_WS_SA. However, this increase is not statistically significant (Figure 5.9, green bar with the label kNN_A_WS_SA, kNN_R_WS_SA). Finally, there is a significant difference (p value is 0.0463) in the accuracy scores between the append and replace strategies

when used with both stemming and stopwords removal (Figure 5.9, green bar with the label kNN_A_WS_SA and kNN_R_WS_SA).

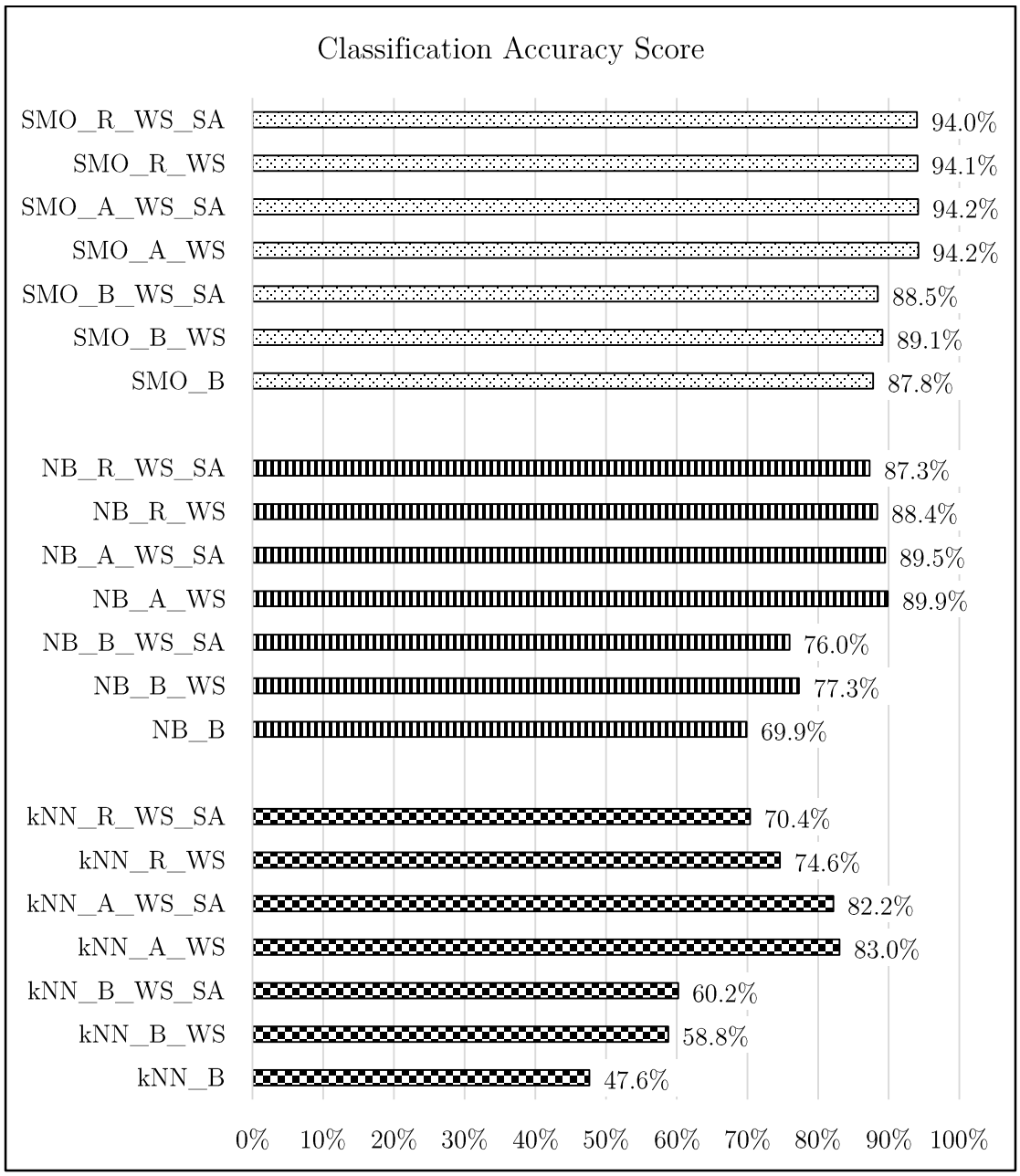


Figure 5.8: Classification score after Stemming and Stopwords Removal

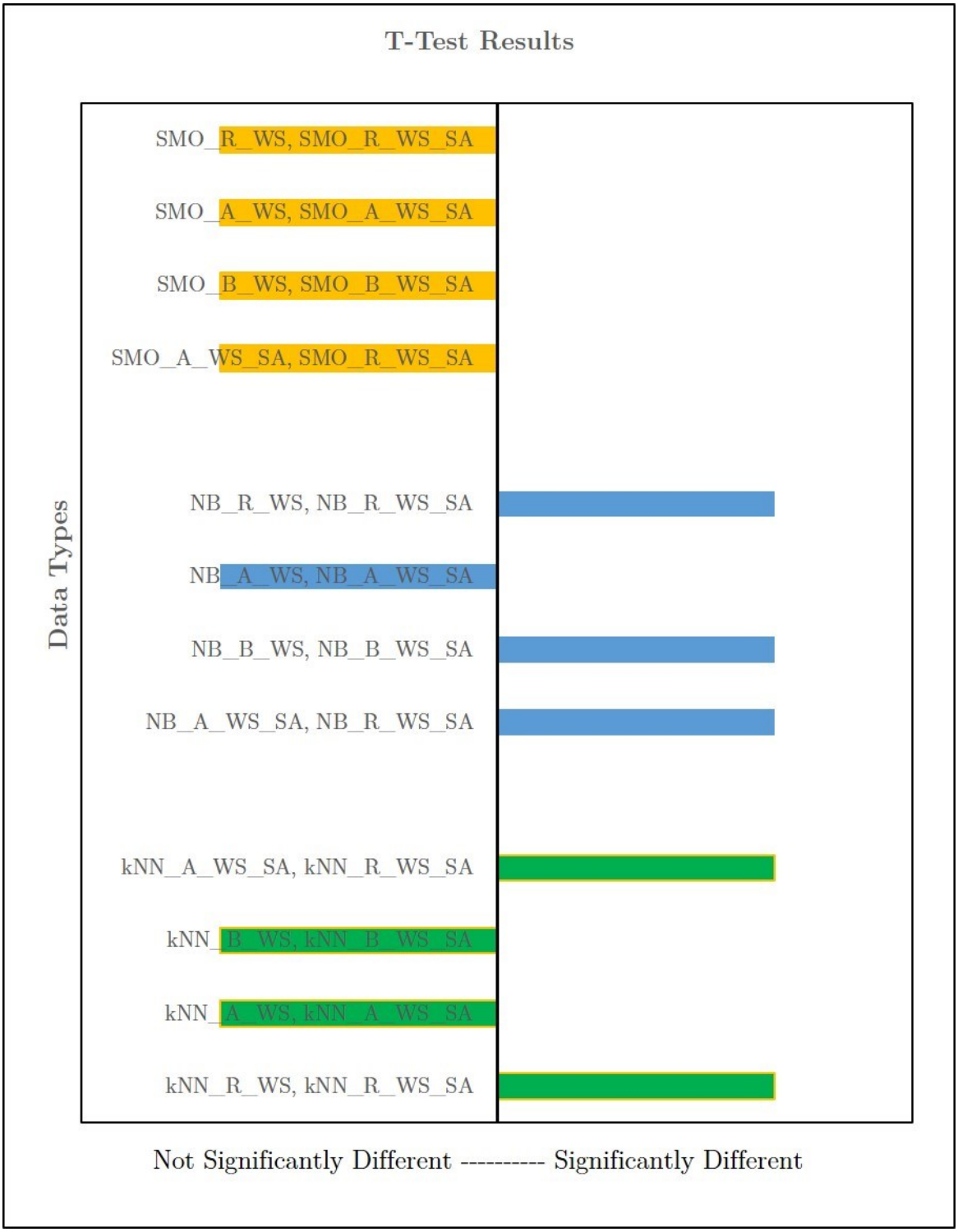


Figure 5.9: P-Value after Stemming and Stopwords Removal

5.5 Impact on Results Without 'Others' Category in the Datasets

Another set of experiments were conducted to check the impact of not having the 'Others' category in the dataset. By excluding the 'Others' category, the tweets were primarily forcibly assigned to one of the eight cities and there was no category or class to 'catch' tweets that were not highly relevant to the eight cities. No statistical tests were done on the classification scores to assess if the results are significantly different or not. However, there were some interesting sets of results obtained after removing the 'Others' category.

The classification accuracy relatively improved for all three algorithms for the same data types with no 'Others' Category vis-à-vis having the 'Others' category in the datasets (Results in Table A.1 vs Table A.2 in the Appendices). For example:

- kNN Algorithm: The accuracy score improved from 47.6% for kNN_B (note: not having 'WO' in the label indicates that this dataset had an 'Others' Category) to 54.13% for kNN_B_WO (note: WO here indicates that this dataset had no 'Others' category). Similarly, there was improvement in the accuracy score for append and replace strategy as it changed from 69.9% for kNN_A and 56.1% for kNN_R to 78.6% for kNN_A_WO and 64.08% for kNN_R_WO.
- NB Algorithm: The accuracy score for the NB algorithm also improved similar to the case of kNN algorithm. For example: 69.9% for NB_B to 76.03% for NB_B_WO. Interestingly, the accuracy score using NB algorithm crossed the 90% mark with the implementation of append strategy both before stopwords removal and stemming implementation (e.g., 90.55% for NB_A_WO), and after stopwords removal (92.18% for NB_A_WO_WS) and stemming implementation (e.g., 90.48%

for NB_A_WO_SA) while this happened only once when replace strategy was implemented (e.g., 90.43% for NB_R_WO_WS).

- SMO Algorithm: The accuracy score for SMO algorithm was higher already even with the inclusion of the 'Others' category (See Table A.2 in Appendices for SMO related results) and it improved marginally after removal of the 'Others' category from the datasets. The lowest score for SMO algorithm was more than 91% even for the baseline data (SMO_B_WO) as compared to similar type dataset of SMO_B with 87.8%. The SMO algorithm breached the 95% accuracy level with the implementation of either append or replace strategy with 95.18% being the lowest for SMO_R_WO and SMO_R_WO_SA, while the highest score was 95.85% for SMO_A_WO_WS.

Chapter 6

Visualizations - Emotions and Sentiments

6.1 Overview

Twitter has become a popular social media platform among users not only to communicate and share information with each other on different subjects but also to express emotions and share sentiments on various topics. For example, users express sentiments and emotions on their experiences related to the use of products and services [87], current events [83] and government initiatives and policies ([84][87]). These emotions and sentiments would vary by location and time. Thus, it is critical to present such information in a format that is easier to review and comprehend. Visualization plays an important role in transforming large volumes of data into forms that are easy to engage with and to identify patterns for meaningful outcomes [110]. Nazemi et al. [84] argued that the use of appropriate visualization is valuable in gaining insight from temporal data captured from social media applications (e.g., Twitter). Thus, in order to understand trends and to learn about the *Pulse of a City* i.e., emotions and sentiments along with associated topics, multiple visualizations were developed for the cities.

Work of several authors contributed in the development of visualizations included in this thesis work. For example, Torkildson et al. [125] used stacked area charts to show eight emotions in one visual. Such visualization provides opportunity to users to view multiple emotions in one view and have a quick comparison. This research work also

used similar approach by using stacked column charts to show multiple emotions together (e.g., Figure 6.1). Morstatter et al. [82] included 'on-click' feature to present additional details (e.g. user-ids) when desired by user. The 'on-click' feature in a visualization helps in minimizing visual chaos and reduce information overload by empowering user to display additional information when required. A similar 'on-click' feature was included in visualizations developed for this research work. The 'on-click' feature included in this research work's visualizations for displaying additional information. For example, the additional information in visualization of emotions includes emotions label (e.g., love), tweet date and topics associated with emotion (e.g., Figure 6.1 and 6.2). For example, Nazemi et al. [84] highlighted the importance of including temporal window feature in visualizations, particularly using social media data. They argued that such feature is important because social media data is voluminous which would necessitate filtering to gain more insight. Authors like Meyer et al. [78] incorporated temporal window feature in their visualization by allowing user to select a start date and an end date to see visual results and Shaikh et al. [109] implemented navigator (or slider) which was placed just below the x-axis. The navigator supported interactiviy and allowed users to select temporal window to narrow or widen visualization of results. Such features help users to view results for selected temporal periods as per user's needs and interests. This research work also incorporated similar functionality by including two implementations to render visuals for a temporal period. One implementation allowed user to input start and end dates and the second implementation allowed user to manipulate temporal window by using a navigator placed below the x-axis (e.g., Figure 6.1 and 6.2).

In this chapter, a number of screenshots are included to exhibit examples of different visualizations developed for this research, and more importantly, to demonstrate the practical usefulness of the proposed *DigiCities* approach, which helped in providing close to accurate reflections of emotions and sentiments. As discussed in this study, identification of location, relevant to a tweet, is critical to get the proper reflection of the *Pulse of a City*. Poor location identification would result in identification of emotions and sentiments that are not relevant to tweet(s). Such differences are demonstrated by using the city of Calgary example, where emotions and sentiments from three different datasets are compared. The visualization results also include the keywords relevant to emotions and sentiments.

6.2 Data for Emotions and Sentiments

As discussed in Chapter 3, 500 tweets relevant to each of the eight cities (Calgary, Edmonton, Banff, St. Albert, Fort McMurray, Lethbridge, Medicine Hat and Red Deer) were manually identified. For discussion (and easy reference) purposes in this chapter, this dataset is labelled as the 'Gold Standard Dataset'. Visualizations of emotions and sentiments were developed for all eight cities by using this data, as it reflect the accurate emotions and sentiments of these cities. Additional visualization examples (for both emotions and sentiments) from the city of Calgary were also developed to discuss the *Pulse of a City* emerging from the use of the *DigiCities* approach. Two example datasets of tweets were created and these datasets were from the city of Calgary. These datasets were created, based on the results obtained post classification experiments conducted on tweets, using NB algorithm (as discussed in Chapter 5).

The first dataset, labelled as *NB_B*, was created based on the outcome of the classification experiment conducted under the following experimental conditions: For

the dataset 'NB_B', it means that it had the following characteristics: a) algorithm used was Naïve Bayes; b) stopwords were not removed from the dataset; c) stemming was not applied on the dataset, and; d) neither append nor replace strategy was implemented on the dataset (see Table 5.1 and 5.2 for details of this data). Tweets classified into the city of Calgary category by means of the NB algorithm using this dataset (*NB-B*), were used to identify emotions emerging from such a classification outcome.

The second dataset, labelled as *NB_A_WS_SA*, was created based on the outcome of the classification experiment conducted under the following experimental conditions: for the dataset labelled as 'NB_A_WS_SA' it means that it had the following characteristics: a) algorithm used was Naïve Bayes; b) stopwords were removed from the dataset; c) stemming was applied on the dataset, and; d) append strategy was implemented on the dataset (please see Table 5.1 and 5.2 in Chapter 5 for details relevant to this data). Tweets classified into the city of Calgary category by means of the NB algorithm using this dataset (*NB_A_WS_SA*), were used to identify emotions emerging from this classification outcome after applying the *DigiCities* approach.

6.3 Emotions Sample Results

Nine emotions were computed by using the algorithm developed by Shahraki and Zaiane [108] and overall emotions for the day were calculated (see Section 3.10.1 for details) and keywords associated with these emotions were identified (see Section 3.10.3 for topics identification related details). The sample results for emotions identified for Calgary are shown in Table 6.1, Table 6.2, and Table 6.3.

Table 6.1 shows (sample) emotions, and topics associated with them, reflected in tweets in Calgary. These are true emotions as they were identified from the Gold

Standard Data. Table 6.2 shows emotions, and topics related to the emotions in Calgary, calculated on tweets classified by the Naïve Bayes algorithm after the implementation of the append strategy (i.e. on the dataset NB_A_WS_SA). Table 6.3 shows emotions for Calgary, and topics related to them, calculated on tweets classified by the NB algorithm on the baseline data (i.e., on dataset NB_B) with no implementation of append strategy. In these tables, emotion with the value '0' means that the respective emotion was not identified on that particular day's tweets.

The comparison of sample results in Table 6.1, 6.2, and 6.3 shows that there were differences in emotions for the two examples of data. The following changes (or changes) occurred for each date:

Date: 1/31/2017: No difference in emotions' percentages and their respective keywords in Table 6.1 (gold standard dataset) and Table 6.2 (NB_A_WS_SA dataset). However, there were a number of differences in results in Table 6.3 (dataset NB_B) when compared to results in Table 6.1. Five new emotions and their keywords emerged – 'anger', 'fear', 'guilt', 'love', and 'sadness' in Table 6.3. These emotions did not exist originally. Also, there was a drop in percentage for two emotions – 'disgust' and 'thankfulness' in Table 6.3 as compared to the values in Table 6.1, but there was no change in keywords associated with these two emotions.

Date: 3/12/2017: There was only one change between Table 6.1 and Table 6.2. The percentage value for the emotion 'disgust' increased and there was no change in associated keywords. Other emotions, and their values and keywords remained the same. However, there were more changes between Table 6.1 and Table 6.3 results. The percentage value for emotions, 'disgust', and 'thankfulness' increased. Though the

keywords for 'disgust' emotion remained the same but the keywords for 'thankfulness' changed completely. Another emotion i.e., 'love' totally disappeared in Table 6.3.

These sample results show the importance of the *DigiCities* approach as its use in applications like emotion mining can help in identifying the closer reflection to the *Pulse of a City*.

Table 6.1: Emotions in Calgary

Date	**Emotion**	**Percentage**	**Topics**
1/31/2017	anger	0	
	disgust	100	police, nurses, killing, investigate, demanded
	fear	0	
	guilt	0	
	joy	0	
	love	0	
	sadness	0	
	surprise	0	
	thankfulness	100	police, nurses, killing, investigate, demanded
3/12/2017	anger	100	update, team, searching, police
	disgust	50	update, team, searching, police
	fear	100	update, team, searching, police
	guilt	0	
	joy	0	
	love	50	@sunrickbell, @sunlorrie, @liberal_party, @duanebratt, @cpc_hq
	sadness	100	update, team, searching, police
	surprise	0	
	thankfulness	50	update, team, searching, police

Table 6.2: Emotions in Calgary based on the Classification Results Achieved with the use of Append Strategy

Date	**Emotion**	**Percentage**	**Topics**
1/31/2017	anger	0	
	disgust	100	police, nurses, killing, investigate, demanded,
	fear	0	
	guilt	0	
	joy	0	
	love	0	
	sadness	0	
	surprise	0	
	thankfulness	100	police, nurses, killing, investigate, demanded,
3/12/2017	anger	100	update, team, searching, police,
	disgust	100	update, team, searching, police,
	fear	100	update, team, searching, police,
	guilt	0	
	joy	0	
	love	50	@sunrickbell, @sunlorrie, @liberal_party, @duanebratt, @cpc_hq,
	sadness	100	update, team, searching, police,
	surprise	0	
	thankfulness	50	update, team, searching, police,

Table 6.3: Emotions in Calgary based on the Classification Results Achieved without using Append Strategy

Date	**Emotion**	**Percentage**	**Topics**
1/31/2017	anger	50	wednesday, trends, number, largest, edmonton,
	disgust	50	police, nurses, killing, investigate, demanded,
	fear	50	wednesday, trends, number, largest, edmonton,
	guilt	50	wednesday, trends, number, largest, edmonton,
	joy	0	
	love	50	wednesday, trends, number, largest, edmonton,
	sadness	50	wednesday, trends, number, largest, edmonton,
	surprise	0	
	thankfulness	50	police, nurses, killing, investigate, demanded,
3/12/2017	anger	100	update, team, searching, police,
	disgust	100	update, team, searching, police,
	fear	100	update, team, searching, police,
	guilt	0	
	joy	0	
	love	0	
	sadness	100	update, team, searching, police,
	surprise	0	
	thankfulness	100	wednesday, trending, topic, hours, #trndnl,

6.3.1 Match and Mismatch of Emotions (A City of Calgary Example)

In this section, more detailed analyses reflect changes in emotions as they emerged in Calgary from the use of different datatypes (gold standard, NB_B and NB_A_WS_SA). Such a discussion demonstrates the application and importance of *DigiCities* to discover more accurate emotions for a location. The findings are:

In the Gold Standard Data, 500 tweets for the city of Calgary were distributed over 197 days. The dataset achieved after the classification experiments on the dataset NB_B showed only 173 days. While the dataset achieved after the classification experiments on the dataset NB_A_WS_SA showed 194 days. This means that 24 days were missing from the data in NB_B case versus only three (3) days missing from the data in NB_A_WS_SA. Dataset NB_B showed emotions for significantly fewer days as compared to the dataset NB_A_WS_SA that showed relatively more accurate emotions in Calgary.

A simple descriptive, statistics comparison was done by identifying the total number of matches and mismatches of emotions' values for each day between the gold standard dataset and the other two datasets (NB_B and NB_A_WS_SA). The results were:

The gold standard data contained 1773 (=197 days * 9) emotions for Calgary in 197 days. The total number of matches were 967 (54.5%) and the total number of mismatches were 806 (45.5%) between the NB_B dataset and the gold standard dataset. Similarly, the total number of matches were 1500 (84.6) and mismatches were 273 (15.4%) between the NB_A_WS_SA dataset and gold standard dataset. The results are shown in Table 6.4.

These mismatches were further investigated and found the following:

- ***Missing Emotions*:** Some emotions were originally present on a particular day (as identified through gold standard data) but were found to be missing from those days in one or both datasets.

Table 6.4: Statistics of Matches and Mismatches in Emotions in both the Datasets

Data Type	**Number of Mismatches**	**Number of Matches**	**Total number of Emotions in 197 days in Gold Standard Data Equals to 1773**
NB_B (Count)	806	967	
NB_B (in % out of 1773)	45.5%	54.5%	
NB_A_WS_SA (Count)	273	1500	
NB_A_WS_SA (in % out of 1773)	15.4%	84.6%	

- ***Added Emotions:*** Some emotions were originally not present on a particular day (as identified through gold standard data) but were founded to be present on those days in one or both datasets.
- ***Change in Emotion Values*:** The value of some emotions on a particular day as noted from the gold standard data changed i.e., the percentage value either *increased* or *decreased*. Table 6.5 provides the statistics of the various types of mismatches.

The findings from the analysis of data on the four parameters – Missing Emotions, Added Emotions, and Increase in Emotion Values and Decrease in Emotion Values are discussed in the following paragraphs. Table 6.5 provides distributions of such change. All the values discussed are in comparison with the gold standard dataset.

- A total of 105 (=5.92% of the total count of 1773) instances where emotions were missing with the use of the NB_B dataset while only 25 (=1.41% of the total count of 1773) instances where emotions were missing with the use of the NB_A_WS_SA dataset.
- A total of 155 (8.74%) instances where emotions were added with the use of the NB_B dataset while only 27 (1.52%) instances where emotions were added with the use of the NB_A_WS_SA dataset.
- There were 382 (21.55%) instances for which the percentage value of emotions increased with the use of the NB_B dataset as compared to 131 (7.39%) instances for which the percentage values of emotions increased with the use of the NB_A_WS_SA dataset.
- There were 164 (9.25%) instances for which the percentage value of emotions decreased with the use of the NB_B dataset as compared to 90 (5.08%) instances for which the percentage values of emotions decreased with the use of the NB_A_WS_SA dataset.

Table 6.5: Break-up of Emotions in both the Datasets

Out Come	**Datatype: NB_B**		**Datatype: NB_A_WS_SA**	
	Count	***% of 1773***	***Count***	***% of 1773***
Missing Emotions	105	5.92%	25	1.41%
Added Emotions	155	8.74%	27	1.52%
Increase in Emotion Value*	382	21.55%	131	7.39%
Decrease Emotion Value*	164	9.25%	90	5.08%
Overall Change	**806**	**45.5%**	**273**	**15.4%**
**these values are not inclusive of missing emotions and added emotions*				

The above results show that there was a higher number of matches of emotions from the NB_A_WS_SA dataset when compared to the number of mismatches. Also the erroneous inclusions ('added emotions') or exclusions ('missing emotions') were not very high (less than 3% combined) and thus showed that the use of the proposed approach can help in identifying emotions closer to reality.

6.4 Examples of and Discussion on Visualizations of Emotions and Sentiments

The *Pulse of a City* representing emotions, sentiments and topics discussed in the city can be presented in a visual form to users for better and quicker understanding. The research aim was to develop a number of visualizations to graphically present what is happening in a city in respect to users' emotions and sentiments, and topics discussions, which evolve over time. The visualizations also provide opportunities for users to learn about keywords associated with such emotions and sentiments.

The visualizations were developed separately for both emotions and sentiments using Highcharts library. The visualizations were developed in two popular formats, column stacked charts and line charts. In addition, sentiments were also plotted using heatmap. Some examples of screenshots of column stacked charts showing emotions and sentiments in a select few cities are included. Also, an example of a line chart is included to present emotions of a city in a different visual form. In the proposed visualization application, two kinds of implementation were developed, one with single pane setup, allowing users to review emotions or sentiments for a city and the second one with two-pane setup where a user can compare two cities, or the same city on different emotions or sentiments. The visualization (irrespective of number of panes) allows a user to select one city out of the eight cities (used in this research) from the top-left dropdown menu

to view emotions (or sentiments) for that city. Further, three pre-determined temporal periods were set up to help users. These include time intervals of one week labelled as '1w', one month labelled as '1m' and all labelled as 'All' which would capture data point for the whole dataset used. The visualizations have an additional feature allowing users to set the date interval based on their preferred time period; this option is located on the top-right corner of the pane and has labels 'From' and 'To'. The visualization also has a navigator below x-axis, which allows users to change the temporal window. Further, visualizations also give users the ability to control (or select) specific emotions they want to view for the selected city, by selecting or de-selecting emotions or sentiments through the use of radial buttons. Topics are also an integral part of emotions and sentiments as they help to better understand the context associated with different emotions and sentiments. Thus, associated topics for each emotion (or sentiment) are also available 'on-click'. Similar implementation is done for the two-pane setup as well. Examples of different visualization options and results are presented and discussed in the following section.

6.4.1 Emotions (or Sentiments) in a City

This visualization helps in learning about change in one or more emotions (e.g. anger, joy etc.) and/or sentiments (positive, negative or neutral) during a given temporal period in a city. The visualization also provides keywords associated with different emotions and sentiments 'on-click'. For example, Figure 6.1 shows the visualization of emotions for the city of Banff showing 'joy', 'love', 'sadness', 'surprise' and 'thankfulness' emotions for the temporal period of one month from November 30, 2017 to December 30, 2017. The visualization of emotions also shows keywords of a few select emotions such as the 'surprise' emotion had keywords such as 'louise', 'lake', 'frozen', '@lake' for December

16, 2017. Percentage values for each emotion, on any given day, are also available on hoover as shown in Figure 6.1. Line charts, an alternate visual form of the column stacked charts were also developed to view the same data. Figure 6.2 provides an example of line chart visualization of column stacked chart visualization presented in Figure 6.1.

6.4.2 Comparison of Emotions (or Sentiments) of Two Cities

The proposed visualization can also compare emotions (or sentiments) between two cities in a two-pane setup. For example, Figure 6.3 provides a comparison of different emotion types prevalent in Edmonton and Calgary during a two-month window from January 12, 2017 to March 12, 2017. Such visualization provides significant help in understanding and comparing the emotions of two cities in totality, as well as to learn the variations in different emotions occurring in two cities, by zooming out on a very short window of time. For example, the review of a window in Figure 6.4 (which is a 'zoom out' of a specific time period – January 22 to February 2 of Figure 6.3) shows that only two emotions, 'disgust' and 'thankfulness' were prevalent in Calgary while different emotions prevailed in Edmonton. For example, on January 29, only two emotions, 'love' and 'fear' were prevalent, while eight emotions for 'guilt' were predominant on January 31 and February 2. Thus, through this visualization it was relatively easy to learn the mood in two cities during different temporal periods.

In addition to the multi-emotion comparison between two cities, visualization can also be done to compare a single emotion between two cities during different temporal periods. For example, Figure 6.5 provides such a visualization, which is showing the 'thankfulness' emotion in the cities of Edmonton and Calgary during a temporal period of one-month (October 31, 2017 to November 30, 2017).

The visualization also shows some example topics associated with this emotion (See Figure 6.5). The analysis, for example, shows that the 'thankfulness' emotion is not prevalent in Edmonton between November 17 and November 22 (except for November 18) but is prevalent in Calgary during this time. The selection of a large temporal window helps in providing a comparative overview of emotions in two cities, while the zoom out feature, leading to a smaller temporal window, helps to provide a clearer picture of emotions on a day-to-day basis. It would be interesting to compare twin cities, such as Edmonton and St. Albert on various emotions. For example, Figure 6.6 provides a comparative overview of one specific emotion, 'joy' and how it varies over time. Visualizations and analysis can help us understand (dis)similarity in emotions between nearby cities (e.g., Edmonton and St. Albert) as well as help us to assess if events in a large city has an impact on emotions and topics in smaller neighbouring cities. For example, Figure 6.6 shows that the 'joy' emotion varied during the selected time period.

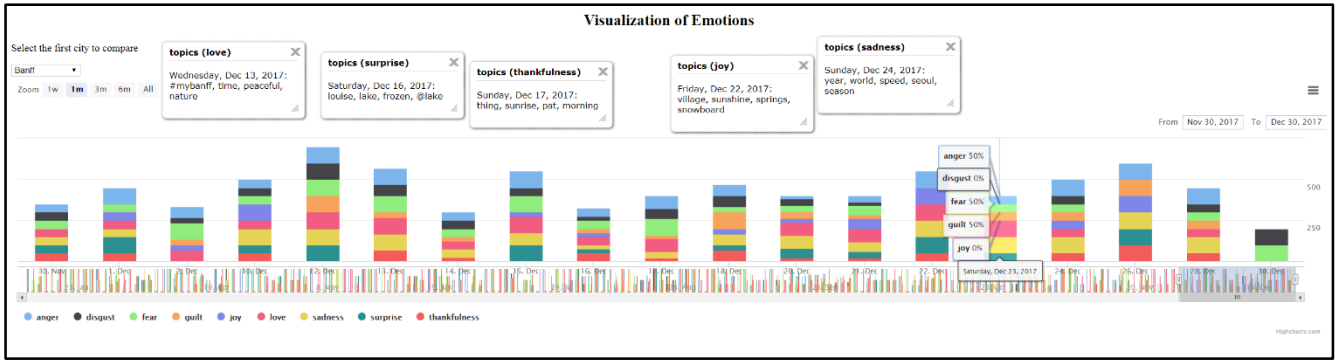


Figure 6.1: Emotions in Banff between November 30 and December 30, 2017 (↑)

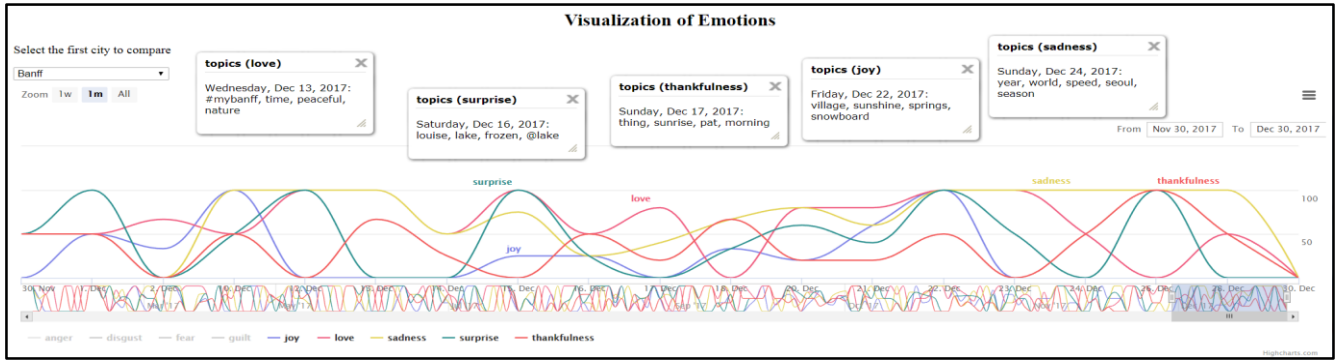


Figure 6.2: Emotions in Banff between November 30 and December 30, 2017 (↑)

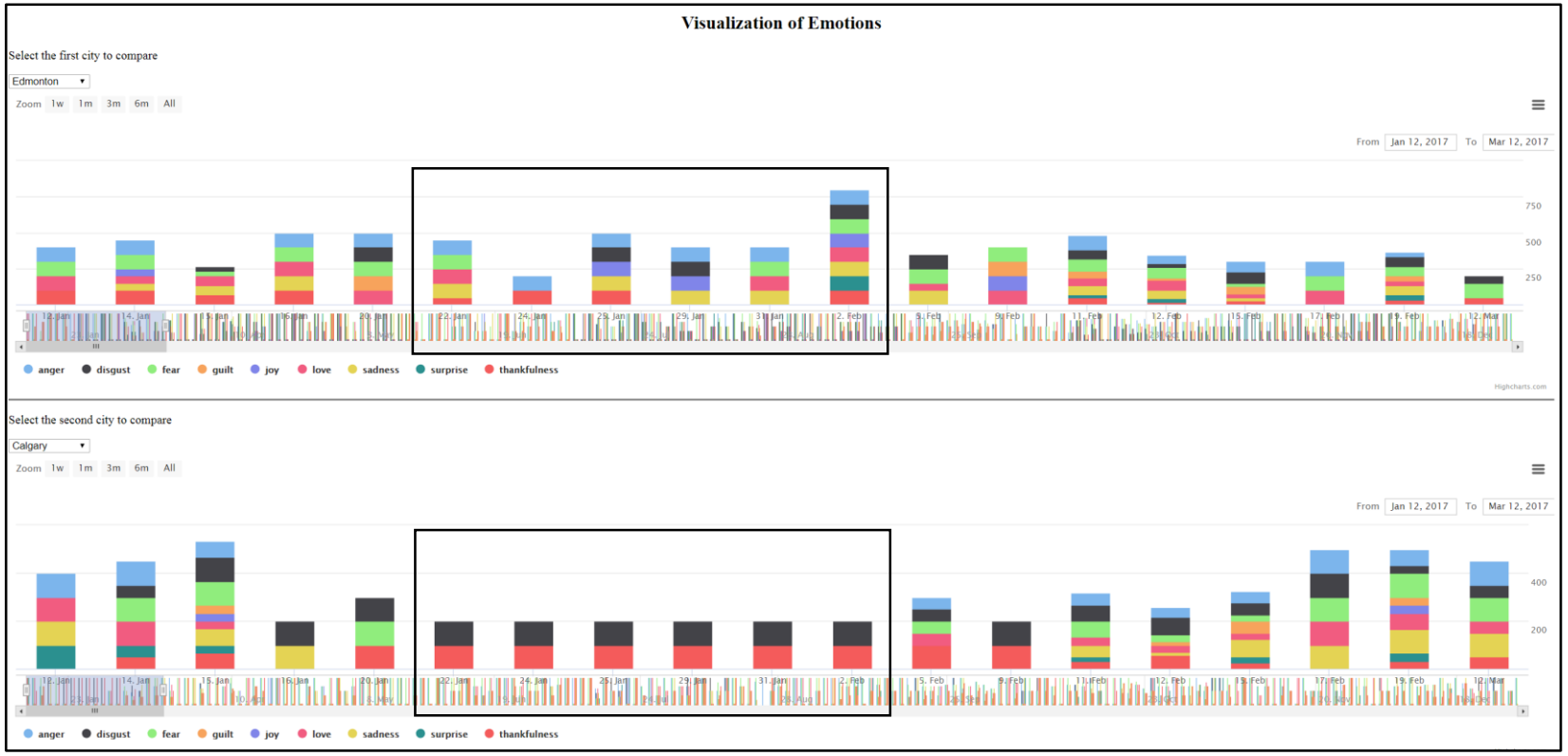


Figure 6.3: Comparison of Emotions in Edmonton and Calgary during a Two-month Window

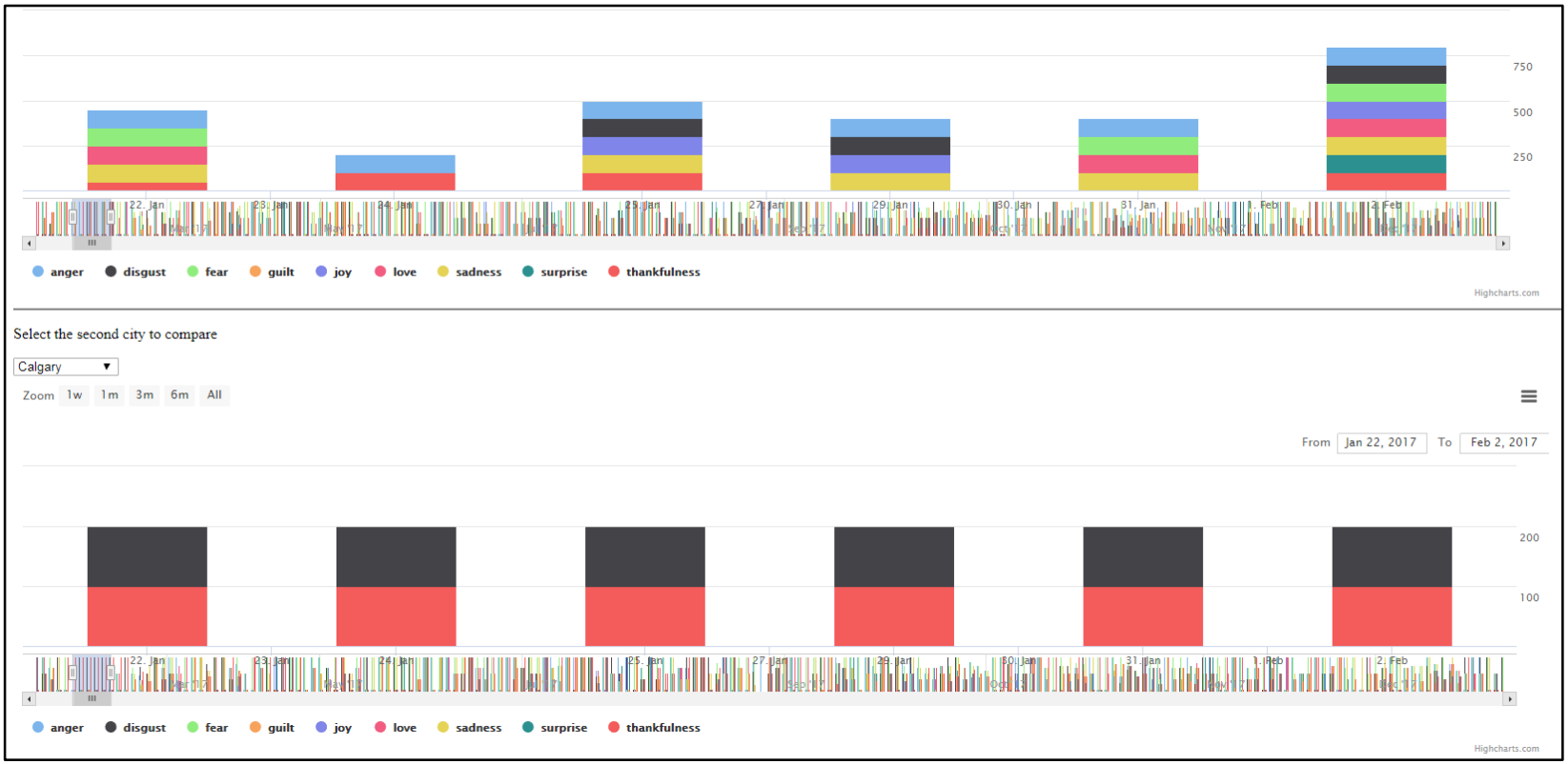


Figure 6.4: Zoom out of January 22 - February 02 window (as shown in Figure 6.3)

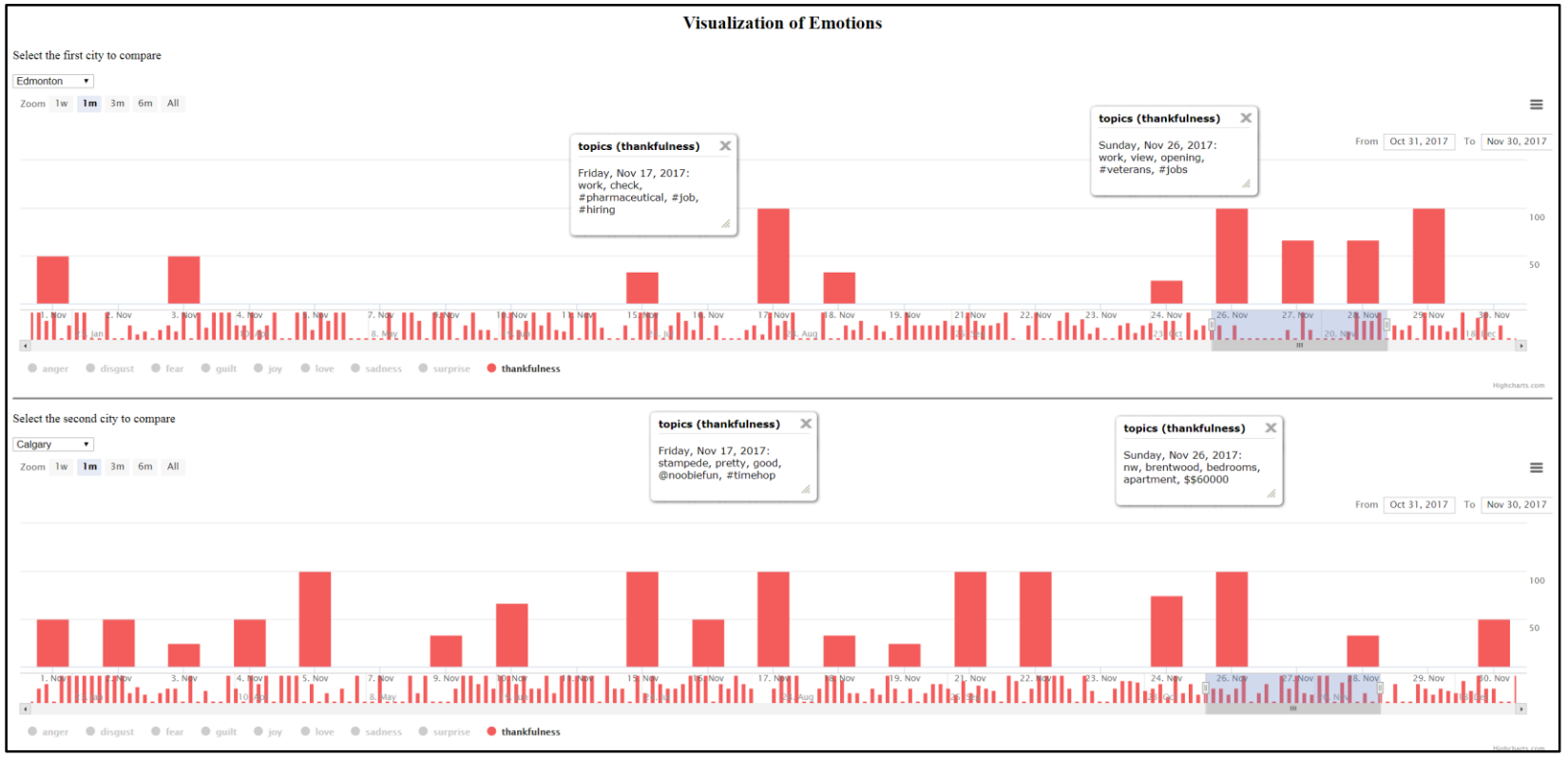


Figure 6.5. Single Emotion Comparison between Edmonton and Calgary

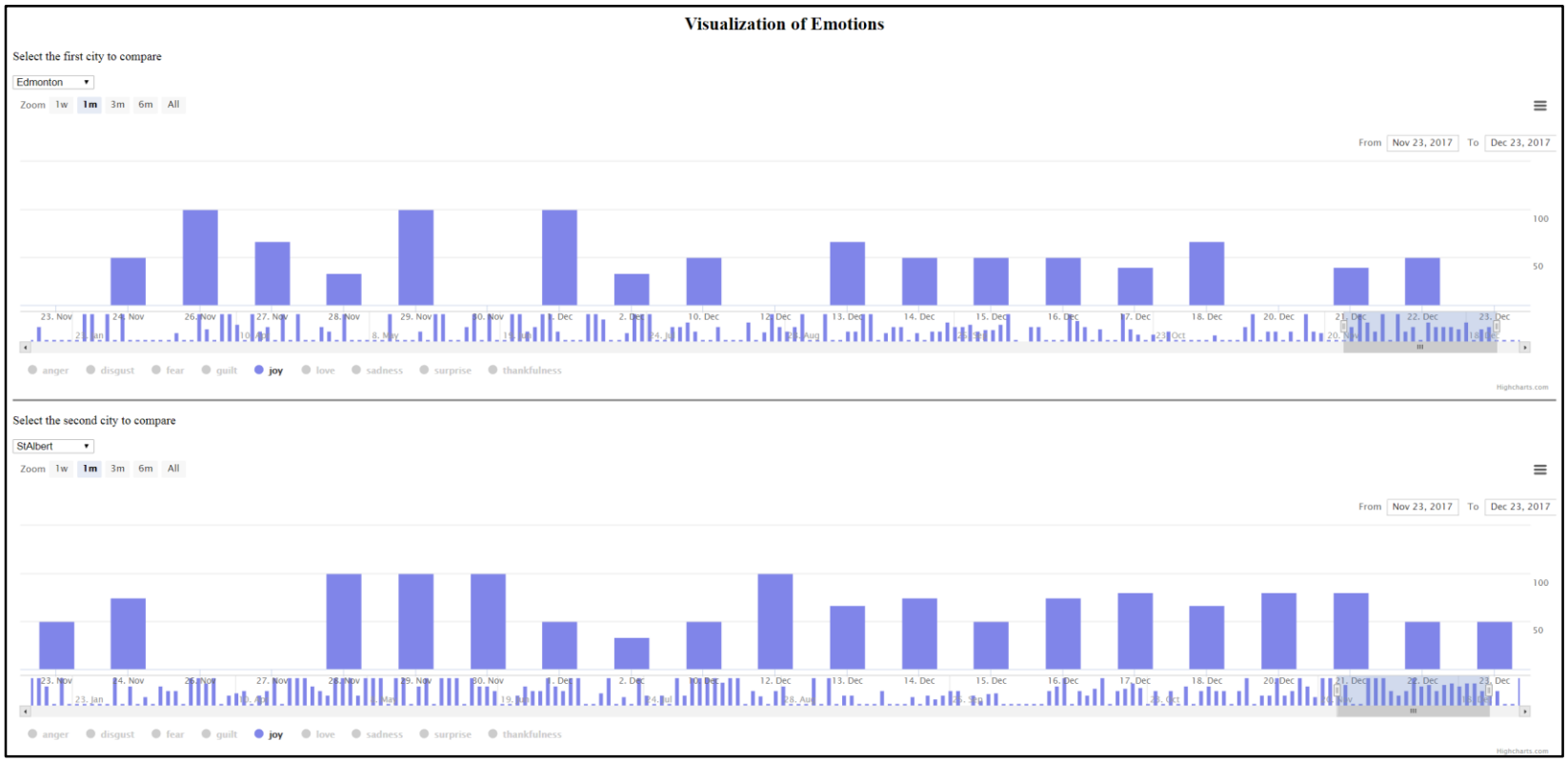


Figure 6.6: Comparison of 'Joy' Emotion between Edmonton and St. Albert

6.4.3 *Comparison of Emotions Before and After Implementation of DigiCities Approach*

Visualizations were created not only to get the *Pulse of a City* through its emotions and sentiments, but also to prove that the use of the *DigiCities* approach helps in identifying closer to true emotions of a city. In this section, visualizations comparing emotions, and how they changed after the NB algorithm classified tweet data, are shown, both before and after the implementation of the *DigiCities* approach. The *DigiCities* approach was implemented using append strategy.

The city of Calgary example is used to demonstrate the differences in emotions. Calgary is chosen because it is the largest population centre among the eight cities used in this research. The comparison is done between the emotions generated from three different data types (gold standard dataset, NB_B dataset, and NB_A_WS_SA dataset) as discussed in Sections 6.2 and 6.3. Results emerging from the comparison of these datatypes were discussed in Section 6.3.

The same set of results are discussed and reviewed through visualization. The gold standard data shows true emotions associated with Calgary. The comparison between the gold standard data and the dataset NB_B shows that emotions differ markedly. Figure 6.7 shows emotions for a two-month period, starting from January 12, 2017 to March 13, 2017. There are differences in the overall emotions as well as in the percentages of emotions; percentages varied as false positive tweets were added and false negative tweets were removed by the NB algorithm when classifying tweets using the NB_B dataset. Thus, in the process tweets were removed that truly belonged to the city of Calgary and were reassigned to a different city. For example, the 'thankfulness' emotion is the only emotion, identified from the gold standard data, between January

21 to February 08, and is shown with an arrow in the top pane of the screenshot in Figure 6.7 (Calgary). Due to incorrect classification of the tweets, different emotions also emerged such as 'joy' and 'sadness' as shown by the arrow in the bottom pane of Figure 6.7 (Calgary_NoAppend). Thus, not showing the true emotions of the city of Calgary.

Similar comparisons were done between the results from the gold standard dataset and the classification results obtained post implementation of the *DigiCities* approach by using the append strategy. The comparison shows that the use of the *DigiCities* approach helped to get more accurate emotions for Calgary. For example, from January 21 to February 08, there are relatively no significant changes in emotions for Calgary (comparison of two visualizations in two panes in Figure 6.8).

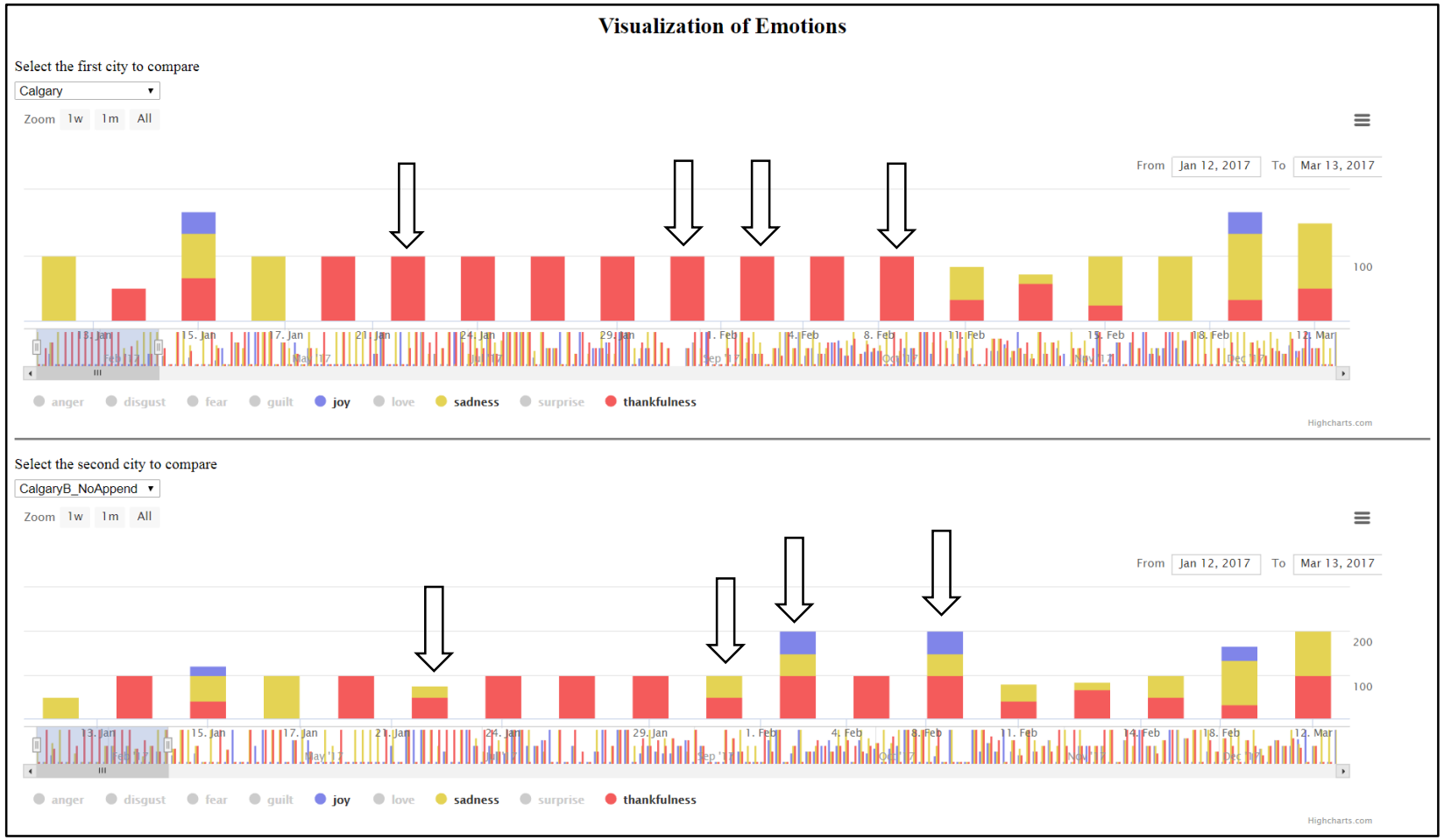


Figure 6.7: Comparison of Emotions Generated from the Gold Standard Data and the NB_B Dataset for Calgary

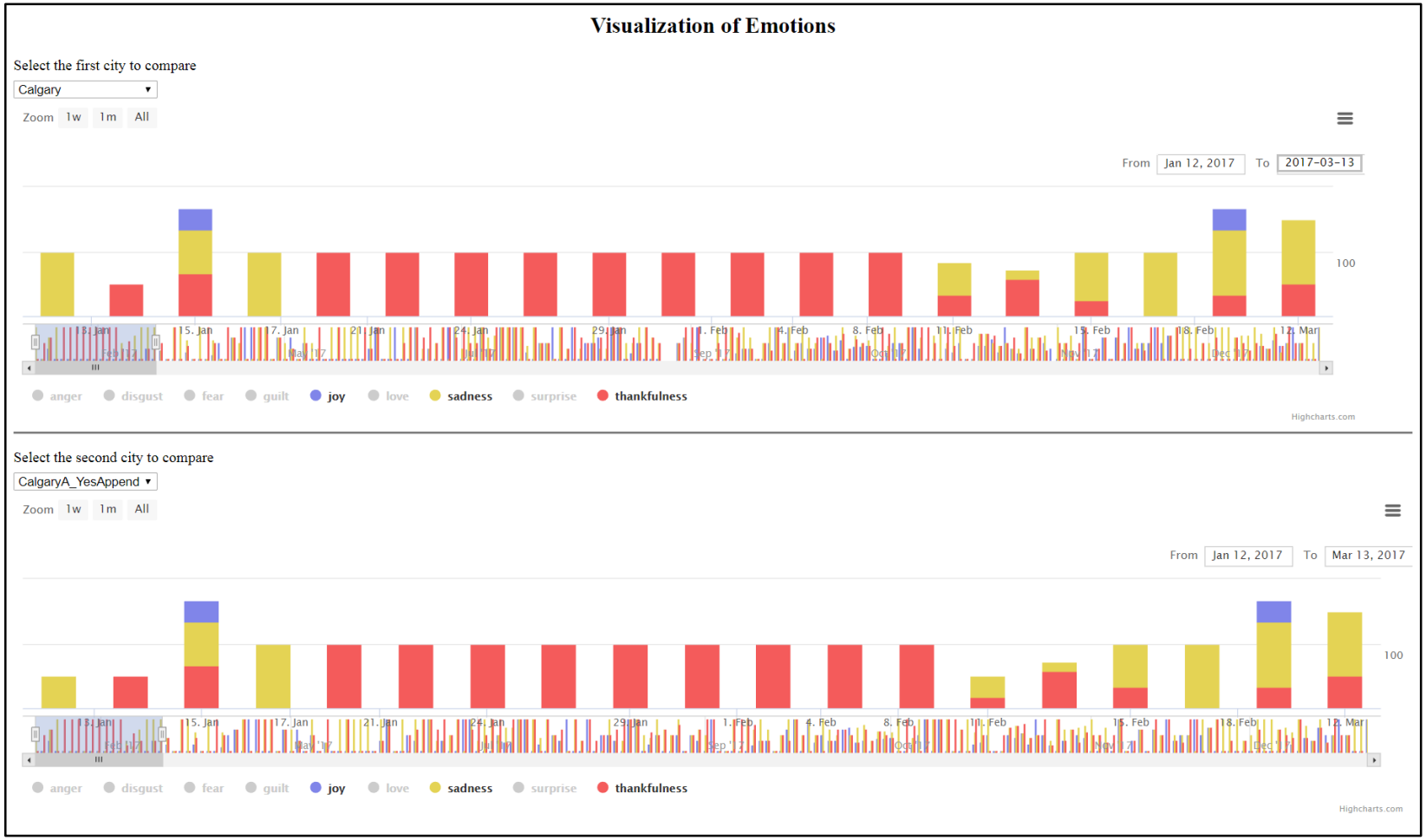


Figure 6.8: Comparison of Emotions Generated from the Gold Standard Data and the NB_A_WS_SA for Calgary)

6.5 Visualizations of Sentiments

Visualizations were also developed for reviewing sentiments in different cities included in this research, and the same chart types were used as in the emotions visualization i.e., line charts and column stacked charts. In addition, heatmap visualization was also developed to review sentiments. Heatmaps were developed for sentiments and they are represented by three features (positive, negative and neutral) Each sentiment is represented by a block and it is relatively easy to understand the pattern with three blocks as compared to nine blocks.

In Figures 6.9, 6.10, and 6.11, the green colour block represents positive sentiment, the red colour block represents negative sentiment and the grey colour block represents neutral sentiment. The heatmap visualization, like line and bar charts, also has the capability to select one or more sentiments for evaluation purposes (the default is all three sentiments). This heatmap visualization has the additional feature of showing sentiment at different hours of the day. The selection or deselection can be made by clicking on 'Neutral', 'Positive', and 'Negative' labels at the bottom of the navigator bar. Also, the topics associated with each sentiment block is available through 'on-click' (as shown in Figure 6.10). Figure 6.9 shows an example of sentiments for Edmonton for the temporal period of almost one year (January 12, 2017 to December 31, 2017) from the sample data tweets. The analysis of Figure 6.9 shows that there is more of a neutral sentiment as compared to positive or negative sentiment in Edmonton during this particular temporal period. This visualization can be further analyzed on a smaller temporal window.

This can be done three ways: a) by re-sizing the navigator window and moving it over the specific time frame; b) by using a pre-determined temporal window by clicking next 'zoom' button ('1w', '1m' and 'All' equals one week, one month and full dataset respectively), or by using the navigator at the bottom of the main chart, and; c) by specifying the start and end dates in the date section on the top-right corner. For example, Figure 6.10 shows sentiments for one-month period along with the example topics for a select few blocks. Figure 6.10 shows that neutral sentiment, followed by positive sentiment is prevalent in the city of Edmonton as compared to negative sentiment during this particular month. Figure 6.11 is a two-pane window which allows comparison of sentiments (one or more sentiments) between two cities. In this example, the comparison is between Edmonton and Calgary on negative sentiment. The analysis of the results show that there are considerably more negative sentiments in Calgary when compared with Edmonton between September 11, 2017 and October 09, 2017. This two-pane window can also be used to compare two sentiments for the same city as well (See Figure 6.11), similar to the visualization of emotions in two-pane window.

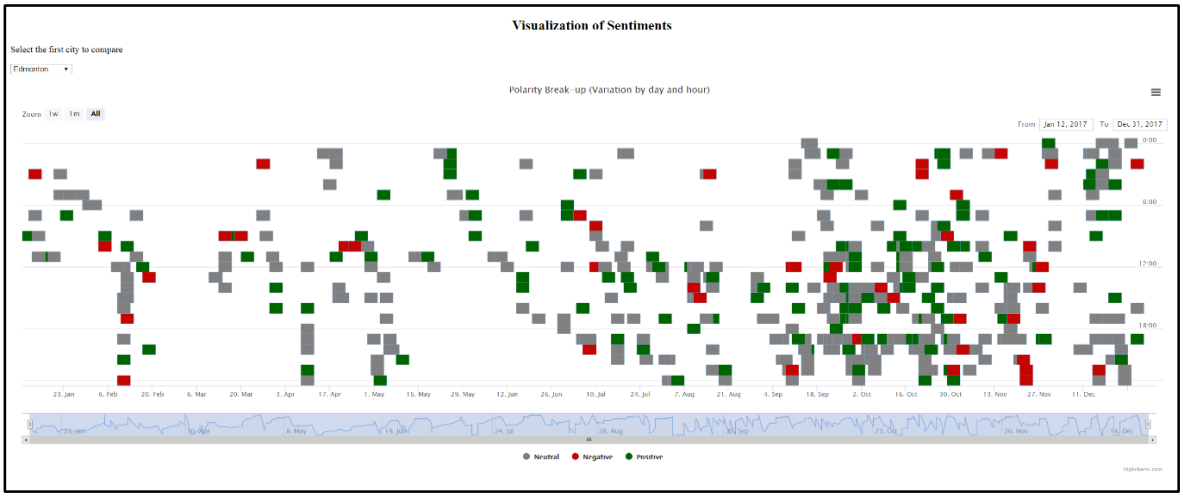


Figure 6.9. Example of Sentiments in Edmonton from Sample Tweets (Dates: January 12 - December 31, 2017)

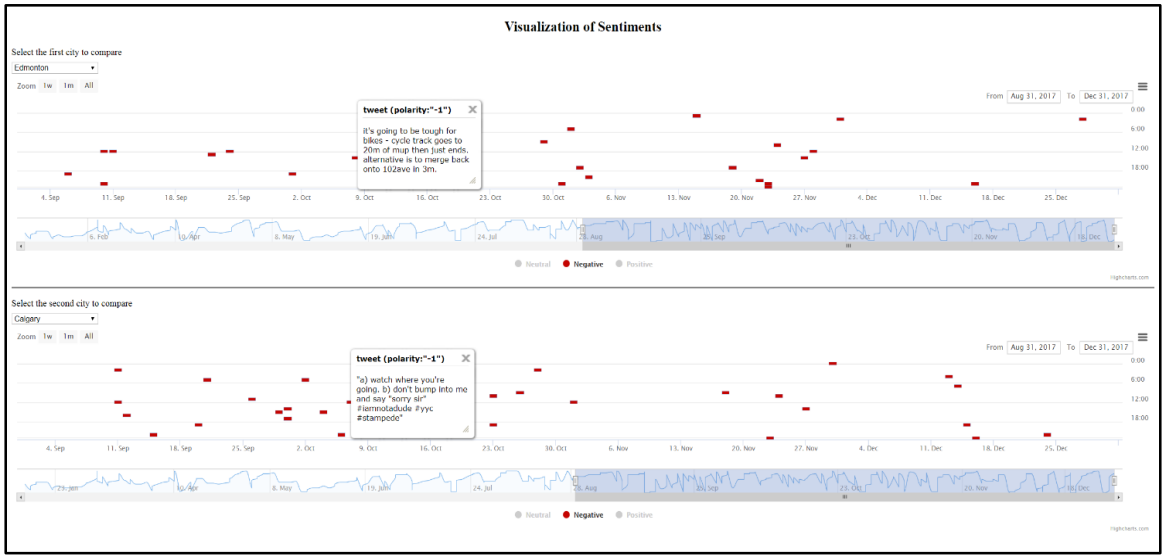


Figure 6.10. Example of Sentiment Comparison between Two Cities - Edmonton and Calgary

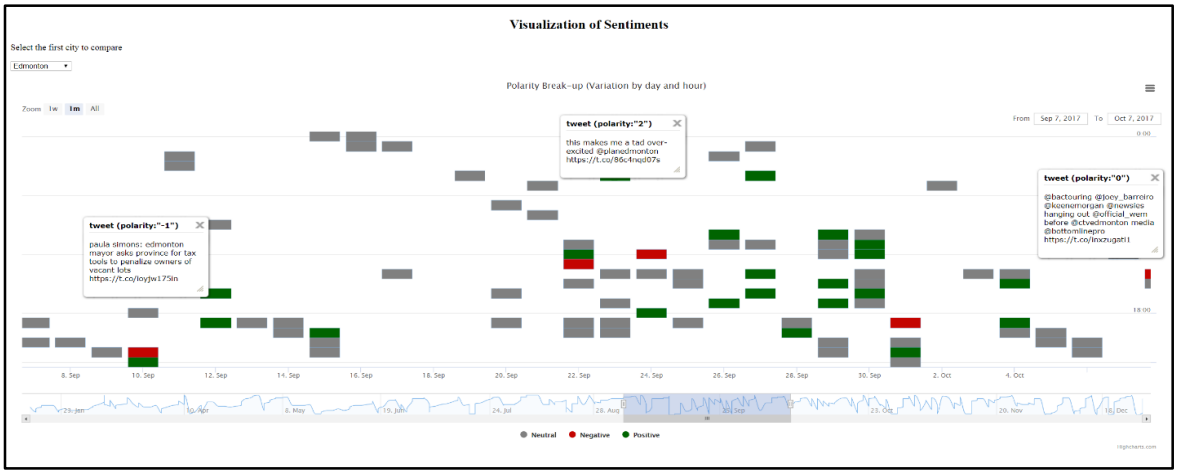


Figure 6.11. Example of Sentiments in Edmonton (One Specific Month) and Topics for a few Select Dates

In conclusion, this research presents visualizations using line charts and column stacked charts for both emotions and sentiments. Sample illustrations of line charts and column stacked charts are included in the discussion on emotions, and sample illustrations of a third type of visualization, using heatmap is presented in the discussion of sentiments. The proposed application offer flexibility to review emotions and sentiments by user control features such as: for pre-determined temporal periods; by manually entering start and end dates; to visualize and compare all the emotions; or to select a few by clicking (toggle buttons) on the specific emotion(s) or sentiment(s). Finally, the proposed visualization gives options to compare different emotions and sentiments between cities or compare the same emotion or sentiment for different time periods within a city using a two-pane setup. Also the same two-pane setup can be used to compare one or more emotions and sentiments between two cities during the same or different temporal period.

Chapter 7

Tweet Categorization and Visualizations

7.1 Overview

Twitter users post a combined total of over 500 million tweets every day worldwide ([7][127]) on a wide range of topics but the majority of users are interested in only a relatively small set of topics and not in every topic. It can be a challenging task for users to find information on events of interest [142] with ease due to sheer volume of tweets as it may cause information overload ([101][139]). Thus categorization of tweets into topic categories is important and will help users to better manage information from tweets on topics that are of interest to them. Further, categorization of tweets into themes/topics coupled with visualization can be a useful strategy in handling a large volume of data. Visualizations of topic categories can be useful in many ways such as an overview of temporal patterns of information flow in specific categories and, any potential outbreak leading to a surge in tweet posting can inform interested users on emerging scenarios. The scatter charts from Highcharts Library were used to create visualizations related to categorization of tweets into topics or themes. The findings from manual assignment and automatic assignment using Google Knowledge Graph and WordNet of topic labels to tweets are discussed in this chapter. Also, a number of screenshots are included to exhibit examples of different visualizations developed for this

research, and more importantly, to compare the manual labelling with WordNet-based labelling of topics to tweets.

7.2 Categorization of Tweets into Topics

As noted in Chapter 3, a total of 400 tweets comprising of 100 tweets from each of the four shortlisted cities i.e., Banff, Calgary, Edmonton and Red Deer were manually categorized into eight pre-determined categories including 'Others' category. The other seven categories include Weather, Jobs, Sports, Entertainment, Health, News, Traffic & Urgent Events. The analysis of the manual categorization reveals the following:

- The majority of tweets were categorized into one category i.e., 73.5% (= 294) tweets out of the 400 tweets were assigned one category labels. There were only 5.25% (= 21) tweets and 0.25% (= 1) tweet were assigned to two and three or more categories respectively (Table 7.1). Though the sample size is small to generalize the finding but the trends suggests that primarily users talks about one topic in a tweet, and comparatively, only a limited number of tweets include multiple topics.

- 21% (= 84) tweets out of 400 tweets were assigned to 'Others' category (Table 7.1). The number of tweets categorized in this category suggests that users post content related to other topics as well and thus, additional categories are required. Based on the analysis, it is recommended that categorization of tweets should be data driven and a set of new categories could be added after reviewing the tweets in the 'Others' category to make categorization more comprehensive and robust. For example, the review of tweets in the 'Others' category revealed that a number of tweets were posted related to 'Rental' information (e.g., Tweet related to Calgary: "*Condo for #rent in Beltline Inner-City SW 1 bedrooms - $$1350.00. Available September 01*

https://t.co/TIPe1RncMs Calgary") or general information (e.g., Tweet related to the Red Deer: "*I'm at City centre stage in Red Deer Alberta https://t.co/fHNRpr1mVq*"). Overall, Red Deer had the highest number of tweets in the 'Others' categories and additional categories such as 'informational' or 'acknowledgement' could capture some of them.

- The analysis at city-level reveals that Banff had the highest number of single topic tweets as compared to other cities. The plausible reason for such high number is that it is a tourist destination and transient population of Banff posts tweets related to various facets of entertainment (e.g., tourism, food, etc.). Thus, the majority of tweets in this city falls under the 'Entertainment' category as compared to other cities (Table 7.1).

- Analysis of categorization at topic level reveals that the 'Entertainment' category was relatively the most popular category. This is primarily due to two factors: a) the inclusion of tweets from Banff in the sample dataset which contributed to a large number of tweets in this category, and; b) the 'Entertainment' category subsumed multiple sub-categories of Food, Festival, Music, Movies and Tourism.

- Analysis at the city-level reveals that after excluding 'Others' category, the 'Jobs' and 'News' categories were more popular categories in Calgary and Edmonton, the two large cities in Alberta, while the 'Jobs' and 'Entertainment' categories were relatively more popular in the relatively smaller city of Red Deer (Table 7.1).

Table 7.1: Assignment of Tweets to Topic Categories

City name	One Category	Two Categories	Three or more Categories	'Others' Category
Topic Labels Assignment – Manual				
Banff	92	5	0	3
Calgary	77	4	0	19
Edmonton	64	11	1	24
Red Deer	61	1	0	38
Total	**294**	**21**	**1**	**84**
Percentage	**73.50%**	**5.25%**	**0.25%**	**21.00%**
Topic Labels Assignment – Using Keywords from Google Knowledge Graph				
Banff	24	20	23	33
Calgary	42	13	16	29
Edmonton	29	22	16	33
Red Deer	41	17	10	32
Total	**136**	**72**	**65**	**127**
Percentage	**34.00%**	**18.00%**	**16.25%**	**31.75%**
Topic Labels Assignment – Using Keywords from WordNet				
Banff	27	38	35	0
Calgary	21	47	31	1
Edmonton	20	40	38	2
Red Deer	6	50	44	0
Total	**74**	**175**	**148**	**3**
Percentage	**18.50%**	**43.75%**	**37.00%**	**0.75%**

7.2.1 *Approaches to Categorization using WordNet and Google Knowledge Graph*

Automated categorization of tweets into topic categories was done using WordNet and Google's Knowledge Graph (GKG). The further analysis of results in Table 7.1 and Table 7.2 showed the following three types of categorizations.

Imperfect Categorization: A total of 60% (240 tweets out of 400) tweets were assigned to the same categories by WordNet as they were assigned manually but in the case of GKG, the score was relatively low at 46% (=184 tweets out of 400 tweets). WordNet was able to achieve a higher score of 60% as compared to 46% of GKG, due to better results from Banff i.e., WordNet was able to categorize 91% of tweets to the same category as the manual category. Though, GKG was able to match only 33% of tweets to the manually assigned tweets. However, there was a limited difference between the WordNet score and GKG score (although not tested for statistical significance) to correctly categorize tweets in the appropriate topic categories for the other three cities. WordNet assigned five extra tweets to the right categories for Edmonton and Red Deer, and this difference was only of two tweets for Calgary.

Over Categorization: A total of 80.75% (=323 tweets out of 400) tweets i.e., nearly four in five tweets were categorized into two or more categories by the use of WordNet, and only 18.5% (=74 tweets out of 400) tweets were categorized into one topic. A total of 34.25% (=137 tweets out of 400) tweets i.e., one in three tweets were categorized into two or more topic categories by the use of GkG. These results are higher when compared to manual categorization whereas 5.5% (=22 tweets out of 400) tweets were categorized into two or more topic categories.

Table 7.2: Comparison of Manual and Automated Approaches for Categorizing Tweets into Topics

Labels (➔) & Approaches (↓)	**Weather**	**Sports**	**Jobs**	**Entertainment**	**Health**	**News**	**Traffic & Urgent Events**	**Others**	***Total***
City: Banff									
Manual (1)	4	10	1	86	1	0	0	3	***105***
WordNet (2)	19	21	48	99	16	9	12	0	***224***
GKG (3)	15	18	24	34	26	26	13	33	***189***
Same Label by (1)&(2)	*1*	*5*	*0*	*85*	*0*	*0*	*0*	*0*	***91***
Same Label by (1)&(3)	*0*	*1*	*1*	*31*	*0*	*0*	*0*	*0*	***33***
City: Calgary									
Manual (1)	2	0	23	17	0	37	6	19	***104***
WordNet (2)	9	11	66	96	21	13	6	1	***223***
GKG (3)	14	7	27	16	17	12	27	29	***149***
Same Label by (1)&(2)	*2*	*0*	*23*	*14*	*0*	*6*	*0*	*0*	***45***
Same Label by (1)&(3)	*1*	*0*	*23*	*5*	*0*	*5*	*3*	*10*	***47***
City: Edmonton									
Manual (1)	6	4	34	10	4	23	8	24	***113***
WordNet (2)	11	8	67	95	24	12	6	2	***225***
GKG (3)	22	4	40	16	14	19	10	33	***158***
Same Label by (1)&(2)	*2*	*0*	*34*	*10*	*2*	*4*	*2*	*2*	***56***
Same Label by (1)&(3)	*1*	*0*	*33*	*4*	*1*	*4*	*3*	*15*	***61***
City: Red Deer									
Manual (1)	7	10	18	21	1	2	4	38	**101**
WordNet (2)	14	16	82	100	10	14	15	0	**251**
GKG (3)	8	6	37	15	13	17	14	32	**142**
Same Label by (1)&(2)	*6*	*1*	*18*	*21*	*0*	*1*	*1*	*0*	**48**
Same Label by (1)&(3)	*1*	*0*	*18*	*4*	*0*	*2*	*4*	*14*	**43**

The average number of categories (including 'Others' category) assigned by WordNet were 2.3 labels per tweets (i.e., a total of 923 categories including 'Others' category were identified for 400 tweets) as compared to GKG which assigned 1.6 labels per tweets (i.e., a total of 638 topic categories including 'Others' category were identified for 400 tweets).

Imposed Categorization: A total of 99.25% of tweets were categorized into one or more categories and only 0.75% (=3 tweets out of 400) tweets were assigned to the 'Others' category by the use of WordNet. These results are in contrast to the categorization done manually and using GKG where 21% (=84 tweets) and 31.75 % (=127) were assigned to 'Others'.

7.3 Visualization of Topic Categorization

Visualization can provide important details while handling a large volume of data. Considering topic visualizations, it can help in identifying range of topics and temporal patterns including popularity (or trendiness) of different topics, and potentially, assist users in filtering information based on their topic of interest. The scatter charts from Highcharts Library were used to create visualizations related to categorization of tweets into topics or themes. These charts were also used in the emotion and sentiment visualizations included in this thesis work.

7.3.1 Manual Topic Categorization

The topic categories (e.g., Weather, Sports, and Jobs) are represented by their respective symbol, which is a combination of shape and colour, are shown in the legend at the bottom of the plotted chart. For example, the 'Weather' category is represented by the circle filled by blue colour and the 'Jobs' category is represented by the square filled with light green colour as shown in Figure 7.1. A sample of 100 tweets each

associated with four cities (i.e., Banff, Calgary, Edmonton and Red Deer) were used to create temporal visualization of topics.

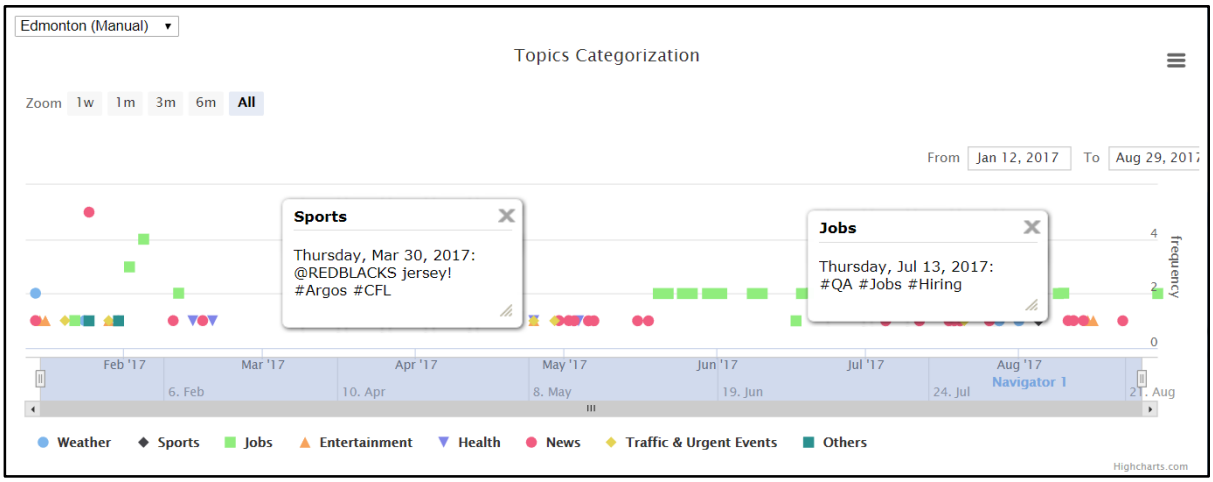


Figure 7.1: Example of Topics Categorization with topics associated

The visualization in Figure 7.1 is showing Edmonton tweets' topics categorized manually by us. The temporal period associated with tweets included in this visualization is from January 12, 2017 to August 29, 2017. The created visualizations have same set of features, as noted in earlier visualization examples in this chapter, including temporal filtering by specifying date range or by using pre-determined temporal window (e.g., one week, one month, and three months), and 'on-click' recall of keywords associated with each plotted data point. Such visualizations are helpful in identifying popular topic categories during a temporal period and the keywords associated with categories in a city.

7.3.2 Comparing Topics Between Cities

The developed visualization application has potential to allow comparison between two cities on various facets such as different temporal period and topic categories. For

example, the visualization in Figure 7.2 presents visualization to compare topic categories (e.g., 'Jobs' vs. 'News') in a city (e.g., Calgary). Such comparative visualization helps in understanding peaks and troughs of topics during particular temporal period. For example, Figure 7.2 shows that there were more tweets in the sample data related to the 'News' category as compared to the 'Jobs' category but there is one temporal window in the month of July when the 'Jobs' related tweets outnumbered the tweets related to the category 'News'.

Figure 7.2: Comparison of Topic Categories in Calgary during Specific Temporal Period

The visualization in Figure 7.3 presents a comparison between two cities, Edmonton and Calgary on multiple topics as they emerged during the three-month temporal period. Such visualizations are useful in comparing overall topic categories, identifying relatively popular topics and inferring pattern variations in topics. For example, the review of visualization in Figure 7.3 suggests that tweets related to the 'Jobs' category is spread over a large temporal window for Edmonton as compared to Calgary in which, it is more concentrated during the month of July. Similarly, tweets related to 'Entertainment' occur more often in Calgary as compared to Edmonton. Thus, such visualizations help in identifying and comparing, for example, the leading (or popular) categories (e.g., 'Jobs' in Edmonton, 'Entertainment' in Calgary) and the least occurring categories (e.g., 'Weather') in the two cities during the same temporal period or different temporal period. This comparison can be further refined by selecting a limited number of categories such as 'Entertainment', 'Traffic & Urgent Events' and 'News', and users can learn about keywords associated with different topic categories which are accessible by 'on-Click' with results displayed in the pop-up block as shown in Figure 7.4.

7.3.3 Comparison Between WordNet and Manual Topic Categorization

As noted in Chapter 3, WordNet and Google Knowledge Graph (GKG) were used to categorize tweets into topics (e.g., Weather, Jobs, News and Helath). Visualizations were also developed for WordNet-based topic categorization including visualizations comparing the outcome of categorization resulting from the automated categorization using WordNet and the manual categorization of sample tweets into topics. Visualization in Figure 7.5 shows such comparison for Edmonton. Such visualization helps in evaluating and inferring the difference in the outcome from the two approaches (i.e., automated vs. manual). WordNet did both 'imperfect categorization' and 'over

categorization' (findings as discussed above). For example, tweet on February 2nd was categorized into the 'Jobs' category using the manual approach but WordNet assigned multiple categories such as 'Jobs' and 'News'. This tweet was manually assigned the 'jobs' category because of keywords such as 'Job' and 'Nursing' but was assigned the additional category of 'News' by the automated approach due to the inclusion of keywords such as 'registered' in the tweet. This example demonstrates an example of 'imperfect categorization' of tweets using automated approach. Also, the automated approach using WordNet did 'over categorization' of tweets by placing tweets into multiple categories. For example, the visualization example below shows a high concentration of the 'Entertainment' category.

Further, a comparative visualization, as shown in Figure 7.5, can help in identifying keywords used to assign topic category to a tweet by different approaches, for example in this visualization example, it is an automated approach (i.e., WordNet) vs. manual approach. For example, the tweets for May 22nd is assigned to the 'Jobs' category by both the approaches but the use of keywords were different. The manual approach used keywords such as 'Cosmetology' and 'Hiring' while the automated approach used keywords such as 'jobs' and 'work'.

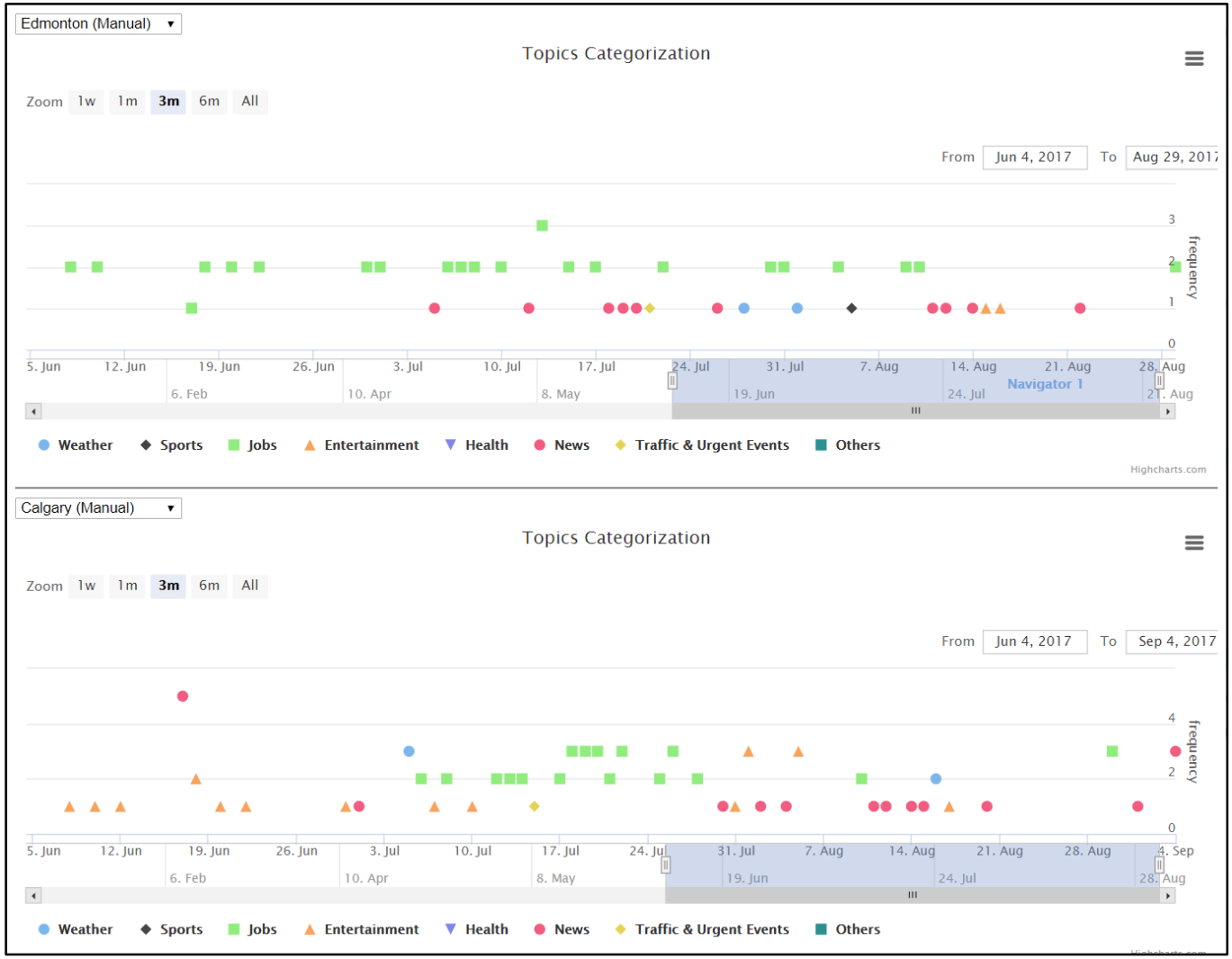


Figure 7.3: Comparison between Edmonton and Calgary During Three-Month Temporal Window

Figure 7.4: Comparison on the basis of three topic categories in Edmonton and Calgary

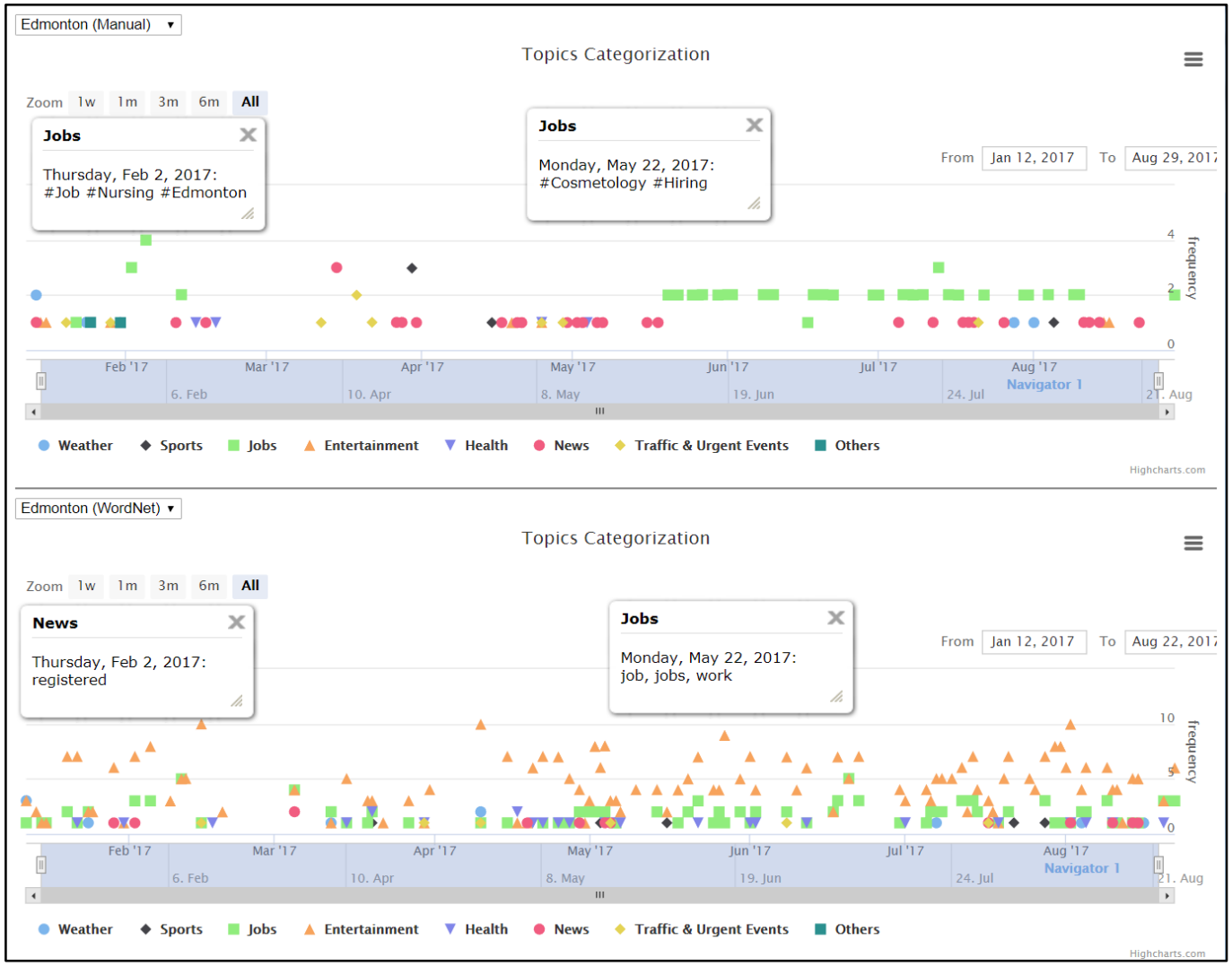


Figure 7.5: Comparison between Manually-Labelled and WordNet-Labelled Topic Categories for Edmonton

Chapter 8

Conclusion and Future Work

8.1 Summary

It is a challenging task to correctly pair a tweet with a location relevant to the content discussed in a tweet. Such challenges arise due to many reasons: limited (or no) explicitly stated location information included in the tweet content; limited document length because of the maximum number of characters allowed in a tweet (originally 140 which was recently changed to 280 characters); varying location granularities such as city, province, or country levels, etc.; location ambiguity (e.g., London in Canada vs London in U.K.); poor location relevant metadata and spelling errors. The identification of accurate location relevant to tweets can be useful to users in many ways. It can aid users in making informed decisions through learning about events happening in a particular location in real-time, as well as learning about users' emotions and sentiments and how they evolve over a period of time in a given location.

This study focused on solving the problem of identifying location (i.e., cities) relevant to tweets and proposed a novel approach called *DigiCities,* which is the digital avatar of real world locations i.e., cities which are represented in digital environments like Twitter by facets such as People, Organizations, and Places (POP). The proposed approach, *DigiCities* uses elements of the POP Framework to create digital profiles of cities from Twitter, which helps in accurately identifying location relevant to given tweets. This approach is based on the assumption that the majority of times, users

would include information about people or organization(s) or places (or a combination of these elements) relevant to a city, when they post tweets in relation to that city. Such information about, or connection to a city, is made by including user-ids and hashtags associated with key people, organizations and places (e.g., @FortEdPark is the user-id for the Fort Edmonton Park).

This study also identifies topics categories, sentiments and emotions and visualizes them as they evolve over a period of time for different geographical locations used in this study. The study also compares true emotions as they emerge from a gold standard dataset with emotions as they emerge from different experimental datasets (discussed in the following paragraphs).

Tweets relevant to eight cities from the Province of Alberta in Canada were used in this research to evaluate the effectiveness of the proposed approach. *DigiCites i.e.,* digital profiles of all these cities were created by identifying relevant user-ids and hashtags. Additionally, this research presents two strategies labelled as *append strategy* and *replace strategy* to implement the POP framework element of *DigiCities*. Statistics were collected when the approach was implemented on tweet datasets related to all the cities included in this research. The collected statistics included the total number of terms replaced in city-related tweets, the terms most commonly replaced in tweets for each city, and the number of times a different city's name was added to another city's tweets.

A total of 4,000 tweets (500 tweets per city) were used in this research. All the tweets for each location were manually identified relevant to each city used in this research; this was the gold standard data to compare and evaluate the accuracy scores from multiple classifications experiments conducted in this research. There were thirty-six

experiments conducted by manipulating and controlling different experimentation facets. These include: the use of three different classification algorithms, kNN, NB and SMO; preprocessing or without preprocessing; and, most importantly with or without implemention of our proposed, *DigiCities* approach. The implementation of *DigiCities* means that the city name was added to the tweet dataset by using replace or append strategy, and not implementation of *DigiCities* means that the city name was not added in the tweet dataset. t-tests were also conducted to evaluate if accuracy scores from the experiments were statistically different or not.

The comparison of accuracy scores suggests that the proposed approach can help in identifying locations relevant to tweets. The accuracy scores for all the three algorithms, kNN, NB and SMO, improved after the implementation of the proposed approach, using the *DigiCities* approach, when compared with the accuracy scores achieved on the baseline dataset. In addition, append strategy performed better than the replace strategy. In algorithms, SMO algorithm was the best among the three chosen algorithms.

This study created a number of visualizations to present topics, emotions and sentiments, and the associated keywords over a selected temporal period (e.g., weekly). Topic categorization was done by Google Knowledge Graph and WordNet. Emotions were identified using algorithm developed by Shahraki and Zaiane [108] and sentiments were detected by using Sentistrength [124], and the keywords were identified based on frequency of occurrence. Visualizations for all eight cities, based on the gold standard data, were created, however, only a few screenshots were included in this thesis. A visual comparison of emotions for the city of Calgary was performed on tweets for the baseline data and the dataset prepared after append strategy. The analysis of the visualization developed on the gold standard data, data post classification on the baseline and append

data showed that the emotions emerging from the append strategy were very close to the emotions which were identified using the gold standard dataset. Thus, the use of our approach, *DigiCities,* when implanted using append strategy, can help in identifying a closer reflection of true emotions in a given location.

8.2 Contributions

This thesis makes the following contributions:

Proposes a novel approach labelled as DigiCities: Identifying location relevant to tweets can be useful for users in multiple ways. For example, users can learn about events in real-time, and about users' topics of discussion, emotions and sentiments in a location during a temporal period. However, these benefits can only be achieved if tweets are accurately paired with the location. Furthermore, in spite of a number of challenges associated with Twitter data in terms of location identification, there is also location related implicit information available in the content of tweets. Thus, this approach will make use of this implicit information embedded in the content of tweets in the form of 'user-ids' and 'hashtags' which are commonly included by users in the tweets. This research proposes a novel approach to identify location relevant to tweets by harnessing the implicit information included in tweets. This *DigiCities* approach harnesses information associated with key People, Organizations and Places (POP) relevant to a location, which have Twitter presence with usersids and associated hashtags. This is a novel approach as this has not been used in prior Twitter research to identify location relevant to tweets. As noted in the findings and the summary section of this chapter, the proposed approach has improved identification of location relevant to tweets.

Proposed two unique strategies - append strategy and replace strategy: One of the challenges in the Twitter-based data is of data sparsity. There is limited, explicit location information in tweets, which further adds to this data sparsity in the context of location identification. This study investigated the use of append strategy and replace strategy to overcome issues such as data sparsity in Twitter data, particularly associated with explicit location information. The findings suggest that in the case of kNN algorithm, location identification is better with the use of append strategy as compared to replace strategy. There was limited to no impact on the other algorithms such as SMO i.e., where either strategy would enhance the location identification relevant to tweets.

Digital Profile of Cities using the POP Framework: This thesis work proposes the POP framework to create the digital profiles of cities. The initial proposed POP framework contains three elements i.e., key People, key Organizations and key Places of a city, and this can be expanded through future research work. The thesis also creates digital profiles of eight cities which can be used in other research work, as is, and also, the profiles can be used to build more comprehensive digital profiles, for example through crowdsourcing.

Harnessing Implicit Information in Tweets: This thesis presents ideas to harness implicit location relevant information embedded in users' ids and hashtags included in tweet content. This information, when combined with other approaches, can be used to disambiguate location correctly.

Automated Topic Categorization: This thesis explored the use of Google Knowledge Graph (GKG) and WordNet for categorizing tweets into topic categories. Both GKG and WordNet showed promising results.

Other Contributions: This study supports the findings of other research such as Hull [43] who noted that stemming may not always be useful, particularly in datasets that are short and informal like Tweets. The findings suggest that when stemming alone was applied, there were not a major change in the accuracy scores. This was also observed in instances when append and replace strategies were implemented. Further, in the context of implementation of append strategy and replace strategy, the findings suggest that both removal of stopwords and stemming can be ignored when using the SMO algorithm, as there was not a major change in the accuracy scores, however, accuracy scores can be improved by removing stopwords when using kNN and NB algorithms. This can be further investigated by using larger sample sizes and using other classification algorithms.

8.3 Limitations

The study has a few limitations and these include:

- *Geographical Biasness and Number of Locations used in this Research*: The study includes only eight cities in Canada; all of these cities were from the Province of Alberta. A diversity of locations, for example cities from other regions of Canada, or from other countries, might impact the current accuracy scores achieved in this research.
- *No Prior Research to Develop City Profiles:* It was challenging to develop and create *DigiCities* i.e., the digital profile of cities as represented on Twitter as there was no prior research work that would provide guidance and established framework to create such profiles.

- *Lack of Geographical Knowledge:* Lack of geographical knowledge created a challenge to develop digital profiles of cities included in this research as they had to be developed manually; it was difficult to identify all the potential members and therefore the digital profiles were not inclusive of all the potential sets of people, organizations and places relevant to a location. Additionally, limited understanding of each city might have impacted the selection of tweets while developing the gold standard data.

- *Tweet Selection Bias*: Tweets for different cities were selected us and therefore, there is a potential researcher's bias in the shortlisting of tweets.

- *Tweet Data Size*: 500 tweets were selected for each city leading to a total of 4,000 tweets for eight cities and a set of 500 tweets for the 'Others' category.

- *Small Sample Selection for Manual Topic Labelling*: Only 100 tweets per city for four cities (equals to a total of 400 tweets) were manually labelled for topic categories which is a small data for generalizing results.

- *No Inter-Coder Reliability*: Tweets were manually labelled by us and thus lacked inter-coder reliability.

- *Twitter-Data Based Pulse of a City*: This approach uses Twitter data, and therefore *Pulse of a City* is primarily expressing Twitter users' perspective. Twitter is not used equally by all age groups and genders. For example, gender-wise distributions of Twitter users include 34% females and 66% males, and about three-fifth (~ 62%) of Twitter users are between age 18 and 49 [7], and thus, *Pulse of a City* will predominantly show emotions, sentiments and topics of its user demographics.

8.4 Conclusion and Future Work

This research demonstrated that identification of location relevant to tweets can be enhanced by using the novel approach, labelled as *DigiCities*. This framework aims to create a linkage between the physical world and the digital world. Physical world locations are represented by entities such as people, organizations and places. In today's digital world, these entities are also present on social media tools such as Twitter through user-ids and hashtags. The use of such digital equivalent representation can help to correctly identify location relevant to tweets. The improvement in location identification is evaluated by using traditional classification algorithms like kNN, NB and SMO.

This thesis also presented examples of how correct identification of location relevant to tweets can help in learning the *Pulse of a City* as reflected through topics, emotions and sentiments. This research identified emotions and sentiments on pre-classified data (i.e., gold standard data) for all the eight cities and topics for four cities used in this research and were visualized by using HighCharts Library. In addition, emotions generated from the pre-classification data were compared with the emotions from the post-classification data prior to and after the implementation of the *DigiCities* approach. The visualization comparison demonstrated that with the help of the *DigiCities* approach, a near to real reflection on the *Pulse of a City* can be discovered.

The categorization of tweets into topic labels using both manual approach and automated approaches which inlcude the use of Google Knowledge Graph (GKG) and WordNet. Both GKG and WordNet showed promising results to categorizing tweets into topics.

There are a number of ways this thesis work can be extended and developed further in the future. First, the novel approach can be tested by increasing the diversity of cities (e.g., including cities from different countries) and by increasing the number of cities in the dataset. Second, currently only three high level elements are included in the POP Framework. The POP framework can be extended by adding other elements in the framework such as local terms used in a geographical location and local images. There is a potential to harness local images to further build the digital profiles of cities. Third, there is a scope to further develop more comprehensive digital profiles by automating the process and also to include seasonal profile terms (such as hashtags or user-ids of yearly occurring events). Fourth, the approach proposed in this thesis work can be used in combination with other approaches (e.g., [44]) for location disambiguation. Fifth, the study could be extended to evaluate the impact of the proposed approach, *Digicities*, when used with other stemming algorithms (e.g., Porter algorithm [97]). Finally, there is a potential to conduct more research work to improve automated categorization using GKG and WordNet individually as well as by combining them. This would entail harnessing the power of GKG and WordNet individually, and then combining their strengths to improve categorization of tweets into topics.

References

[1]. Abdelhaq, H., Sengstock, C., & Gertz, M. (2013). Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment*, *6*(12), 1326-1329.

[2]. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2012, February). Twitter Improves Seasonal Influenza Prediction. In *Healthinf* (pp. 61-70).

[3]. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (pp. 30-38). Association for Computational Linguistics.

[4]. Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., & Jaimes, A. (2013). Sensing trending topics in Twitter. *IEEE Transactions on Multimedia*, *15*(6), 1268-1282. http://dx.doi.org/10.1109/TMM.2013.2265080

[5]. Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering*, *15*(3), 72.

[6]. Armentano M.G., Schiaffino S., Christensen I., & Boato F. (2015) Movies Recommendation Based on Opinion Mining in Twitter. In: Pichardo Lagunas O., Herrera Alcántara O., Arroyo Figueroa G. (eds) Advances in Artificial Intelligence and Its Applications. MICAI 2015. Lecture Notes in Computer Science, 9414, 80 – 91. Springer, Cham

[7]. Aslam, S. (2019, January 6) Twitter by the Numbers: Stats, Demographics & Fun Facts. Retrieved from https://www.omnicoreagency.com/twitter-statistics/

[8]. Bae, Y., & Lee, H. (2012). Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the Association*

for Information Science and Technology, *63*(12), 2521 – 2535. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1002/asi.22768

[9]. Balabantaray, R. C., Mohammad, M., & Sharma, N. (2012). Multi-class twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, *4*(1), 48-53. http://dx.doi.org/10.5120/ijais12-450651

[10]. Barbosa, L., & Feng, J. (2010, August). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd international conference on computational linguistics: posters* (pp. 36-44). Association for Computational Linguistics.

[11]. Batool, R., Khattak, A. M., Maqbool, J., & Lee, S. (2013, June). Precise tweet classification and sentiment analysis. In *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)* (pp. 461-466). IEEE.

[12]. Benkhalifa, M., Mouradi, A., & Bouyakhf, H. (2001). Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization. *International Journal of Intelligent Systems*, *16*(8), 929-947.

[13]. Biernacki, P., & Waldorf, D. (1981). Snowball Sampling: Problems and Techniques of Chain Referral Sampling, *Sociological Methods and Research,* 10(2) (November, 1981), 141 – 163. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1177/004912418101000205

[14]. Bifet, A., & Frank, E. (2010, October). Sentiment knowledge discovery in twitter streaming data. In *International conference on discovery science* (pp. 1-15). Springer, Berlin, Heidelberg.

[15]. Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, *7*(1), 61-70.

[16]. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[17]. Bollen, J., Mao, H., & Pepe, A. (2011a, July). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media.*

[18]. Bollen, J., Mao, H., & Zeng, X. (2011b). Twitter mood predicts the stock market. *Journal of computational science*, *2*(2011), 1–8. http://dx.doi.org/10.1016/j.jocs.2010.12.007

[19]. Boyatzis, R. E. (1998). Transforming qualitative information: Thematic analysis and code development. sage.

[20]. Card, S., Mackinlay, J., & Shneiderman, B. (1999). Readings in information visualization: using vision to think. San Francisco, CA, USA:Morgan Kaufmann.

[21]. Cataldi, M., Di Caro, L., & Schifanella, C. (2010, July). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining* (p. 4). ACM.

[22]. Chang, H. W., Lee, D., Eltaher, M., & Lee, J. (2012, August). @ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (pp. 111-118). IEEE Computer Society. http://dx.doi.org/10.1109/ASONAM.2012.29

[23]. Chen, X., Li, L., Xu, G., Yang, Z., & Kitsuregawa, M. (2012, July). Recommending related microblogs: A comparison between topic and wordnet based approaches. In *Twenty-Sixth AAAI Conference on Artificial Intelligence.*

[24]. Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759-768). ACM.

[25]. Clement, J. (2019). Statista 2019: Most popular social networks worldwide as of July 2019, ranked by number of active users (in millions). Available at

https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Accessed on July 31, 2019.

[26]. Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters* (pp. 241-249). Association for Computational Linguistics.

[27]. Davis Jr, C. A., Pappa, G. L., de Oliveira, D. R. R., & de L. Arcanjo, F. (2011). Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, *15*(6), 735-751. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1111/j.1467-9671.2011.01297.x

[28]. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

[29]. Duong-Trung, N., Schilling, N., & Schmidt-Thieme, L. (2016, October). Near real-time geolocation prediction in twitter streams via matrix factorization based regression. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (1973-1976). ACM.

[30]. Dykov, M. A., & Vorobkalov, P. N. (2013, May). Twitter Trends Detection by Identifying Grammatical Relations. In The Twenty-Sixth International FLAIRS Conference.

[31]. Ekman, P. (1992.) An argument for basic emotions, *Cognition & Emotion*, 6(3-4), 169-200. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1080/02699939208411068

[32]. Elberrichi, Z., Rahmoun, A., & Bentaalah, M. A. (2008). Using WordNet for Text Categorization. *International Arab Journal of Information Technology (IAJIT)*, *5*(1), 16- 24

[33]. Goldsmith J.A., Higgins D., & Soglasnova S. (2001) Automatic Language-Specific Stemming in Information Retrieval. *In: Peters C. (eds) Cross-Language Information Retrieval and Evaluation. CLEF 2000. Lecture Notes in Computer Science*, *2069*, 273-283, Springer, Berlin, Heidelberg.

[34]. Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, *66*(4), 568-578. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1080/00330124.2014.907699

[35]. Gupta, M., Gao, J., Zhai, C., & Han, J. (2012). Predicting future popularity trend of events in microblogging platforms. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1-10.

[36]. Hamdan, H., Béchet, F., & Bellot, P. (2013, June). Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 455-459).

[37]. Han, J. & Kamber, M. (2001). *Data mining: concepts and techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers.

[38]. Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L. E., & Hsu, M. C. (2011, October). Visual sentiment analysis on twitter data streams. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 277-278). IEEE.

[39]. Hasan, M., Rundensteiner, E., & Agu, E. (2014, May). Emotex: Detecting emotions in twitter messages. In the *Proceedings of the 2014 ASE Bigdata/SocialCom/CyberSecurity Conference*, Stanford University.

[40]. Heckathorn, D. D. (2011). Comment: Snowball versus respondent-driven sampling. *Sociological methodology*, *41*(1), 355-366. https://doi.org/10.1111/j.1467-9531.2011.01244.x

[41]. Hewis, J. (2015). Do MRI patients tweet? Thematic analysis of patient tweets about their MRI experience. *Journal of medical imaging and radiation sciences*, *46*(4), 396-402.

[42]. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., & Tsioutsiouliklis, K. (2012, April). Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web* (pp. 769-778). ACM.

[43]. Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, *47*(1), 70-84.

[44]. Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., & Ghazi, D. (2015, April). Detecting and disambiguating locations mentioned in Twitter messages. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 321-332). Springer, Cham. http://dx.doi.org/10.1007/978-3-319-18117-2_24.

[45]. Iosifidis, V., & Ntoutsi, E. (2017, August). Large scale sentiment learning with limited labels. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1823-1832). ACM.

[46]. Irani, D., Webb, S., Pu, C., & Li, K. (2010). Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS).*

[47]. Issac, B., & Jap, W. J. (2009, January). Implementing spam detection using Bayesian and Porter Stemmer keyword stripping approaches. In *TENCON 2009-2009 IEEE Region 10 Conference* (pp. 1-5). IEEE.

[48]. Ji, X., Chun, S. A., & Geller, J. (2013, September). Monitoring public health concerns using twitter sentiment classifications. In *2013 IEEE International Conference on Healthcare Informatics* (pp. 335-344). IEEE.

[49]. Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-*

Volume 1 (pp. 151-160). Association for Computational Linguistics. Retrieved from https://dl-acm-org.login.ezproxy.library.ualberta.ca/citation.cfm?id=2002492

[50]. Joachims, T. (1998). *Making large-scale SVM learning practical* (No. 1998, 28). Technical report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

[51]. Karidi, D. P. (2016, May). From user graph to Topics Graph: Towards twitter followee recommendation based on knowledge graphs. In *2016 IEEE 32nd International Conference on Data Engineering Workshops (ICDEW)* (pp. 121-123). IEEE.

[52]. Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. Neural computation, 13(3), 637-649.

[53]. Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(4), 580-585.

[54]. Khan, M., & Khan, S. S. (2011). Data and information visualization methods, and interactive mechanisms: A survey. *International Journal of Computer Applications*, *34*(1), 1-14.

[55]. Kim, E. K., Seok, J. H., Oh, J. S., Lee, H. W., & Kim, K. H. (2013). Use of hangeul twitter to track and predict human influenza infection. *PLoS one*, *8*(7), e69305. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1371/journal.pone.0069305

[56]. Kindberg, T., Barton, J., Morgan, J., Becker, G., Caswell, D., Debaty, P., Gopal, G., Frid, M., Krishnan, V., Morris, H., Schettino, J., Serra, B. & Spasojevic, M. (2002). People, places, things: Web presence for the real world. *Mobile Networks and Applications*, *7*(5), 365-376. http://dx.doi.org/10.1023/A:1016591616731

[57]. Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg!. In *Fifth International AAAI conference on weblogs and social media*.

[58]. Krouska, A., Troussas, C., & Virvou, M. (2016, July). The effect of preprocessing techniques on Twitter sentiment analysis. In *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)* (pp. 1-5). IEEE.

[59]. Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (pp. 591-600). ACM.

[60]. Lamb, A., Paul, M. J., & Dredze, M. (2013). Separating fact from fear: Tracking flu infections on twitter. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 789-795).

[61]. Larsen, M. E., Boonstra, T. W., Batterham, P. J., O'Dea, B., Paris, C., & Christensen, H. (2015). We feel: mapping emotion on Twitter. *IEEE journal of biomedical and health informatics*, *19*(4), 1246-1252. http://dx.doi.org/10.1109/JBHI.2015.2403839

[62]. Lau, J. H., Collier, N., & Baldwin, T. (2012). On-line trend analysis with topic models twitter trends detection topic model online. *Proceedings of COLING 2012*, 1519-1534.

[63]. Lee, K., Ganti, R. K., Srivatsa, M., & Liu, L. (2014, December). When twitter meets foursquare: tweet location prediction using foursquare. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services* (MOBIQUITOUS '14). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[64]. Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011, December). Twitter trending topic classification. In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 251-258). IEEE. http://dx.doi.org/10.1109/ICDMW.2011.171.

[65]. Lennon, M., Peirce, D. S., Tarry, B. D., & Willett, P. (1981). An evaluation of some conflation algorithms for information retrieval. *Journal of information Science*, *3*(4), 177-183.

[66]. Lewis, C. S. (2001). *The Chronicles of Narnia*. New York, NY, USA: Harper Collins.

[67]. Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *European Conference on Machine Learning*, 4-15, Springer, Berlin, Heidelberg.

[68]. Li, C. H., Yang, J. C., & Park, S. C. (2012). Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet. *Expert Systems with Applications*, *39*(1), 765-772.

[69]. Li, Q., Shah, S., Liu, X., Nourbakhsh, A., & Fang, R. (2016). Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In the *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management,* 2429-2432/ ACM.

[70]. Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012, April). Tedas: A twitter-based event detection and analysis system. In 2012 IEEE 28th International Conference on Data Engineering (pp. 1273-1276). IEEE.

[71]. Li, R., Wang, S., & Chang, K. C. C. (2012). Multiple location profiling for users and relationships from social network and content. *Proceedings of the VLDB Endowment*, *5*(11), 1603-1614.

[72]. Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, *11*(1-2), 22-31.

[73]. Lu, R., & Yang, Q. (2012). Trend analysis of news topics on twitter. *International Journal of Machine Learning and Computing*, *2*(3), 327. http://dx.doi.org/10.7763/IJMLC.2012.V2.139

[74]. MacArthur, A. (2018, November 2). The Real History of Twitter, In Brief: How the micro-messaging wars were won. Retrieved from https://www.lifewire.com/history-of-twitter-3288854

[75]. Mahmud, J., Nichols, J., & Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(3), 47. http://dx.doi.org/10.1145/2528548

[76]. Mathioudakis, M., & Koudas, N. (2010, June). Twittermonitor: trend detection over the twitter stream. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 1155-1158). ACM.

[77]. McGee, J., Caverlee, J., & Cheng, Z. (2013, October). Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 459-468). ACM.

[78]. Meyer, B., Bryan, K., Santos, Y., & Kim, B. (2011, March). TwitterReporter: Breaking News Detection and Visualization through the Geo-Tagged Twitter Network. In *CATA* (pp. 84-89).

[79]. Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

[80]. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, *3*(4), 235-244.

[81]. Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301-326. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1111/coin.12024

[82]. Morstatter, F., Kumar, S., Liu, H., & Maciejewski, R. (2013, August). Understanding twitter data with tweetxplorer. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1482-1485). ACM.

[83]. Naaman, M., Becker, H., & Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, *62*(5), 902-918. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1002/asi.21489

[84]. Nazemi, K., Burkhardt, D., Retz, W., & Kohlhammer, J. (2014, December). Adaptive visualization of social media data for policy modeling. In International Symposium on Visual Computing (pp. 333-344). Springer, Cham.

[85]. Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*(1), 11-33.

[86]. Ozdikis, O., Senkul, P., & Oguztuzun, H. (2012, August). Semantic expansion of tweet contents for enhanced event detection in twitter. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 20-24). IEEE

[87]. Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).

[88]. Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery*, *24*(3), 515-554. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1007/s10618-011-0224-z

[89]. Paradesi, S. M. (2011, March). Geotagging tweets using their content. In *Twenty-Fourth International FLAIRS Conference*.

[90]. Paulheim, H. (2016). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, *8*(3), 489-508.

[91]. Pedersen, T. & Michelizzi, J. (n.d.) WordNet::Similarity. Available at http://maraca.d.umn.edu/cgi-bin/similarity/similarity.cgi. Accessed on July 15, 2019.

[92]. Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May). WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004* (pp. 38-41). Association for Computational Linguistics.

[93]. Petersen, M. G., Iversen, O. S., Krogh, P. G., & Ludvigsen, M. (2004, August). Aesthetic interaction: a pragmatist's aesthetics of interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 269-276). ACM.

[94]. Petrović, S., Osborne, M., & Lavrenko, V. (2010, June). Streaming first story detection with application to twitter. In Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics (pp. 181-189). Association for Computational Linguistics.

[95]. Piao, G., & Breslin, J. G. (2016, October). User modeling on Twitter with WordNet Synsets and DBpedia concepts for personalized recommendations. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 2057-2060). ACM.

[96]. Platt, J. C. (1998) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Technical Report MSR-TR-98-14. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf.

[97]. Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137.

[98]. Rakesh, V., Reddy, C. K., & Singh, D. (2013, August). Location-specific tweet detection and topic summarization in twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1441-1444). ACM.

[99]. Ray, S. K., & Singh, S. (2010, October). Blog content based recommendation framework using WordNet and multiple Ontologies. In *2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM)* (pp. 432-437). IEEE.

[100]. Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012, May). EmpaTweet: Annotating and Detecting Emotions on Twitter. In *LREC* (Vol. 12, pp. 3806-3813).

[101]. Rosa, K. D., Shah, R., Lin, B., Gershman, A., & Frederking, R. (2011). Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*, *63*.

[102]. Rosso, P., Ferretti, E., Jiménez, D., & Vidal, V. (2004). Text categorization and information retrieval using wordnet senses. In *Proceedings of the Second International WordNet Conference—GWC* (pp. 299-304).

[103]. Rybalko, S., & Seltzer, T. (2010). Dialogic communication in 140 characters or less: How Fortune 500 companies engage stakeholders using Twitter. *Public relations revieu*, *36*(4), 336-341.

[104]. Sadilek, A., Kautz, H., & Silenzio, V. (2012, May). Modeling spread of disease from social interactions. In Sixth International AAAI Conference on Weblogs and Social Media.

[105]. Saif, H., He, Y., & Alani, H. (2012, November). Semantic sentiment analysis of twitter. In *International semantic web conference* (pp. 508-524). Springer, Berlin, Heidelberg. https://doi-org.login.ezproxy.library.ualberta.ca/10.1007/978-3-642-35176-1_32

[106]. Samuel, H., Noori, B., Farazi, S., & Zaiane, O. (2018, December). Context Prediction in the Social Web Using Applied Machine Learning: A Study of Canadian Tweeters. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 230-237). IEEE.

[107]. Schofield, A., & Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, *4*, 287-300.

[108]. Shahraki, A. G., & Zaiane, O. R. (2017). Lexical and learning-based emotion mining from text. In Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing.

[109]. Shaikh, S., Ravi, K., Gallicano, T., Brunswick, M., Aleshire, B., El-Tayeby, O. & Levens, S. (2019, July) EmoVis – An Interactive Visualization Tool to Track Emotional Trends During Crisis Events. *In: Ahram T. (eds),* In *Proceedings of the AHFE 2019 International Conference on Human Factors in Artificial Intelligence and Social Computing, the AHFE International Conference on Human Factors, Software, Service and Systems Engineering, and the AHFE International Conference of Human Factors in Energy* (July 24-28, 2019), (pp. 14-24) Washington D.C., USA.

[110]. Shiroi, S., Misue, K., & Tanaka, J. (2012, July). Chronoview: Visualization technique for many temporal data. In *2012 16th International Conference on Information Visualisation* (pp. 112-117). IEEE.

[111]. Shneiderman, B. (2008, June). Extreme visualization: squeezing a billion records into a million pixels. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 3-12). ACM.

[112]. Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS one*, *6*(5), e19467 (1-10). Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1371/journal.pone.0019467

[113]. Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter?, *Journal of Interactive Marketing*, *26*(2), 102-113. https://doi.org/10.1016/j.intmar.2012.01.002

[114]. Smith, K. (2019, January 2) 122 Amazing Social Media Statistics and Facts. Retrieved from https://www.brandwatch.com/blog/96-amazing-social-media-statistics-and-facts/

[115]. Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, *29*(4), 470-479. https://doi.org/10.1016/j.giq.2012.06.005

[116]. Soranaka, K., & Matsushita, M. (2012, November). Relationship between emotional words and emoticons in tweets. In *Technologies and Applications of Artificial Intelligence (TAAI), 2012 Conference on* (pp. 262-265). IEEE.

[117]. Specia, L., & Motta, E. (2007, June). Integrating folksonomies with the semantic web. In *European semantic web conference* (pp. 624-639). Springer, Berlin, Heidelberg.

[118]. Statistics Canada (2018, May 30). Retrieved from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E

[119]. Sul, H. K., Dennis, A. R., & Yuan, L. I. (2014, January). Trading on Twitter: The financial information content of emotion in social media. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (pp. 806-815). IEEE.

[120]. Sutton, J., Spiro, E. S., Johnson, B., Fitzhugh, S., Gibson, B., & Butts, C. T. (2014). Warning tweets: Serial transmission of messages during the warning phase of a disaster event. *Information, Communication & Society*, *17*(6), 765-787.

[121]. Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*, Boston, MA, USA: Pearson Education Inc.

[122]. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1555-1565).

[123]. Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal of mixed methods research*, *1*(1), 77-100.

[124]. Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *65*(1), 163-173.

[125]. Torkildson, M. K., Starbird, K., & Aragon, C. (2014, September). Analysis and visualization of sentiment and emotion on crisis tweets. In Luo Y. (eds), Cooperative Design, Visualization, and Engineering. CDVE 2014. Lecture Notes in Computer Science, vol. 8683, 64-67. Springer, Cham. DOI https://doi.org/10.1007/978-3-319-10831-5_9

[126]. Tsou, M. H. (2011, January). Mapping Cyberspace: Tracking the Spread of Ideas on the Internet. In *Proceedings of the 25th International Cartographic conference*. Retrieved from https://icaci.org/files/documents/ICC_proceedings/ICC2011/Oral%20Presentations%20PDF/D3-Internet,%20web%20services%20and%20web%20mapping/CO-354.pdf

[127]. Twitter Usage Statistics (n.d.). Retrieved from http://www.internetlivestats.com/twitter-statistics/

[128]. Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104-112.

[129]. Valkanova, N., Jorda, S., Tomitsch, M., & Vande Moere, A. (2013, April). Reveal-it!: the impact of a social visualization projection on public awareness and discourse. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3461-3470). ACM.

[130]. Wang, L., Wang, G., & Alexander, C. A. (2015). Big data and visualization: methods, challenges and technology progress. *Digital Technologies*, *1*(1), 33-38.

[131]. Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012, September). Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (pp. 587-592). IEEE.

[132]. Wang, W., Hernandez, I., Newman, D. A., He, J., & Bian, J. (2016). Twitter analysis: Studying US weekly trends in work stress and emotion. *Applied psychology*, *65*(2), 355-378. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1111/apps.12065

[133]. Wang, X., & Cosley, D. (2014, February). Tweetdrops: a visualization to foster awareness and collective learning of sustainability. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 33-36). ACM.

[134]. Warf, B., & Sui, D. (2010). From GIS to neogeography: ontological implications and theories of truth. *Annals of GIS*, *16*(4), 197-209. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1080/19475683.2010.539985

[135]. Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011, October). Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2541-2544). ACM. http://dx.doi.org/10.1145/2063576.2064014

[136]. Xia, C., Hu, J., Zhu, Y., & Naaman, M. (2015, May). What is new in our city? a framework for event extraction using social media posts. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 16-32). Springer, Cham. https://doi-org.login.ezproxy.library.ualberta.ca/10.1007/978-3-319-18038-0_2

[137]. Xu, G., Zhang, Y., & Yi, X. (2008, December). Modelling user behaviour for web recommendation using lda model. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 3, pp. 529-532). IEEE.

[138]. Yoon, S., Elhadad, N., & Bakken, S. (2013). A practical approach for content mining of tweets. *American journal of preventive medicine*, *45*(1), 122-129. https://doi.org/10.1016/j.amepre.2013.02.025

[139]. Zaïane, O. R., & Antonie, M. L. (2002, January). Classifying text documents by associating terms with text categories. In *Australian computer Science communications* (Vol. 24, No. 2, pp. 215-222). Australian Computer Society, Inc.

[140]. Zhang H. & Ling C.X. (2003) A Fundamental Issue of Naive Bayes. In: Xiang Y., Chaib-draa B. (eds) Advances in Artificial Intelligence. Canadian AI 2003.

Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), Vol 2671, 591-595, Springer, Berlin, Heidelberg.

[141]. Zhao, J., Gou, L., Wang, F., & Zhou, M. (2014, October). Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 203-212). IEEE.

[142]. Zhao, Z., Cheng, Z., Hong, L., & Chi, E. H. (2015, May). Improving user topic interest profiles by behavior factorization. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1406-1416). International World Wide Web Conferences Steering Committee.

[143]. Zhou, D., Chen, L., & He, Y. (2015, February). An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[144]. Zhou, X., & Chen, L. (2014). Event detection over twitter social media streams. *The VLDB Journal —The International Journal on Very Large Data Bases*, *23*(3), 381-400. http://dx.doi.org/10.1007/s00778-013-0320-3

[145]. Zubiaga, A., Spina, D., Martinez, R., & Fresno, V. (2015). Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, *66*(3), 462-473. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1002/asi.23186

Appendices

A. Classification Accuracy Scores

A.1 Classification Accuracy Score (Numbers and Percentage) for all the Experiments with 'Others' Category Tweets Included in the Dataset

S. No.	Experiment Name	Correctly Classified		Incorrectly Classified		Total Tweets
		Number	Percent (%)	Number	Percent (%)	
1	NB_B	3146	69.9%	1354	30.1%	**4500**
2	NB_B_SA	3133	69.6%	1367	30.4%	**4500**
3	NB_B_WS	3479	77.3%	1021	22.7%	**4500**
4	NB_B_WS_SA	3418	76.0%	1082	24.0%	**4500**
5	kNN_B	2144	47.6%	2356	52.4%	**4500**
6	kNN_B_SA	2174	48.3%	2326	51.7%	**4500**
7	kNN_B_WS	2647	58.8%	1853	41.2%	**4500**
8	kNN_B_WS_SA	2709	60.2%	1791	39.8%	**4500**
9	SMO_B	3951	87.8%	549	12.2%	**4500**
10	SMO_B_SA	3935	87.4%	565	12.6%	**4500**
11	SMO_B_WS	4010	89.1%	490	10.9%	**4500**
12	SMO_B_WS_SA	3981	88.5%	519	11.5%	**4500**
13	NB_A	3830	85.1%	670	14.9%	**4500**
14	NB_A_SA	3832	85.2%	668	14.8%	**4500**
15	NB_A_WS	4044	89.9%	456	10.1%	**4500**
16	NB_A_WS_SA	4026	89.5%	474	10.5%	**4500**
17	kNN_A	3130	69.6%	1370	30.4%	**4500**
18	kNN_A_SA	3150	70.0%	1350	30.0%	**4500**
19	kNN_A_WS	3737	83.0%	763	17.0%	**4500**
20	kNN_A_WS_SA	3698	82.2%	802	17.8%	**4500**
21	SMO_A	4224	93.9%	276	6.1%	**4500**

22	SMO_A_SA	4227	93.9%	273	6.1%	**4500**
23	SMO_A_WS	4239	94.2%	261	5.8%	**4500**
24	SMO_A_WS_SA	4238	94.2%	262	5.8%	**4500**
25	NB_R	3645	81.0%	855	19.0%	**4500**
26	NB_R_SA	3614	80.3%	886	19.7%	**4500**
27	NB_R_WS	3977	88.4%	523	11.6%	**4500**
28	NB_R_WS_SA	3928	87.3%	572	12.7%	**4500**
29	kNN_R	2523	56.1%	1977	43.9%	**4500**
30	kNN_R_SA	2599	57.8%	1901	42.2%	**4500**
21	kNN_R_WS	3358	74.6%	1142	25.4%	**4500**
32	kNN_R_WS_SA	3167	70.4%	1333	29.6%	**4500**
33	SMO_R	4223	93.8%	277	6.2%	**4500**
34	SMO_R_SA	4228	94.0%	272	6.0%	**4500**
35	SMO_R_WS	4234	94.1%	266	5.9%	**4500**
36	SMO_R_WS_SA	4230	94.0%	270	30.1%	**4500**

A.2 Classification Accuracy Score (Numbers and Percentage) for all the Experiments with 'Others' Category Tweets Excluded from the Dataset

S. No.	Experiment Name	Correctly Classified		Incorrectly Classified		Total Tweets
		Number	Percent (%)	Number	Percent (%)	
1	NB_B_WO	3041	76.03%	959	23.98%	**4000**
2	NB_B_WO_SA	3026	75.65%	974	24.35%	**4000**
3	NB_B_WO_WS	3226	80.65%	774	19.35%	**4000**
4	NB_B_WO_WS_SA	3186	79.65%	814	20.35%	**4000**
5	kNN_B_WO	2165	54.13%	1835	45.88%	**4000**
6	kNN_B_WO_SA	2212	55.30%	1788	44.70%	**4000**
7	kNN_B_WO_WS	2632	65.80%	1368	34.20%	**4000**
8	kNN_B_WO_WS_SA	2503	62.58%	1497	37.43%	**4000**

9	SMO_B_WO	3646	91.15%	354	8.85%	**4000**
10	SMO_B_WO_SA	3649	91.23%	351	8.78%	**4000**
11	SMO_B_WO_WS	3681	92.03%	319	7.98%	**4000**
12	SMO_B_WO_WS_SA	3660	91.50%	340	8.50%	**4000**
13	NB_A_WO	3622	90.55%	378	9.45%	**4000**
14	NB_A_WO_SA	3619	90.48%	381	9.53%	**4000**
15	NB_A_WO_WS	3687	92.18%	313	7.83%	**4000**
16	NB_A_WO_WS_SA	3676	91.90%	324	8.10%	**4000**
17	kNN_A_WO	3144	78.60%	856	21.40%	**4000**
18	kNN_A_WO_SA	3154	78.85%	846	21.15%	**4000**
19	kNN_A_WO_WS	3604	90.10%	396	9.90%	**4000**
20	kNN_A_WO_WS_SA	3528	88.20%	472	11.80%	**4000**
21	SMO_A_WO	3812	95.30%	188	4.70%	**4000**
22	SMO_A_WO_SA	3820	95.50%	180	4.50%	**4000**
23	SMO_A_WO_WS	3834	95.85%	166	4.15%	**4000**
24	SMO_A_WO_WS_SA	3828	95.70%	172	4.30%	**4000**
25	NB_R_WO	3495	87.38%	505	12.63%	**4000**
26	NB_R_WO_SA	3482	87.05%	518	12.95%	**4000**
27	NB_R_WO_WS	3617	90.43%	383	9.58%	**4000**
28	NB_R_WO_WS_SA	3592	89.80%	408	10.20%	**4000**
29	kNN_R_WO	2563	64.08%	1437	35.93%	**4000**
30	kNN_R_WO_SA	2635	65.88%	1365	34.13%	**4000**
21	kNN_R_WO_WS	3289	82.23%	711	17.78%	**4000**
32	kNN_R_WO_WS_SA	3166	79.15%	834	20.85%	**4000**
33	SMO_R_WO	3807	95.18%	193	4.83%	**4000**
34	SMO_R_WO_SA	3807	95.18%	193	4.83%	**4000**
35	SMO_R_WO_WS	3824	95.60%	176	4.40%	**4000**
36	SMO_R_WO_WS_SA	3815	95.38%	185	4.63%	**4000**

Note: 'WO' label in the experiment name indicates that the dataset used in the experiments does not have 'Others' Category Tweets.

B. Statistical Test Results (p-values)

S. No.	Data Types	p-value
1	kNN_B and kNN_A	0.00067
2	kNN_B and kNN_R	0.01134
3	kNN_A and kNN_R	0.00350
4	NB_B and NB_A	0.00017
5	NB_B and NB_R	0.00064
6	NB_A and NB_R	0.00161
7	SMO_B and SMO_A	0.00063
8	SMO_B and SMO_R	0.00071
9	SMO_A and SMO_R	0.82430
10	kNN_B and kNN_B_WS	0.00424
11	kNN_A and kNN_A_WS	0.09883
12	kNN_R and kNN_R_WS	0.00732
13	kNN_A_WS and kNN_R_WS	0.26142
14	NB_B and NB_B_WS	0.00002
15	NB_A and NB_A_WS	0.00194
16	NB_R and NB_R_WS	0.00017
17	NB_A_WS and NB_R_WS	0.03032
18	SMO_B and SMO_B_WS	0.12810
19	SMO_A and SMO_A_WS	0.11022
20	SMO_R and SMO_R_WS	0.17891
21	SMO_A_WS and SMO_R_WS	0.13880

S. No.	Data Types	p-value
22	kNN_B and kNN_B_SA	0.24064
23	kNN_A and kNN_A_SA	0.39675
24	kNN_R and kNN_R_SA	0.03117
25	kNN_A_SA and kNN_R_SA	0.00465
26	NB_B and NB_B_SA	0.45368
27	NB_A and NB_A_SA	0.82430
28	NB_R and NB_R_SA	0.16725
29	NB_A_SA and NB_R_SA	0.00059
30	SMO_B and SMO_B_SA	0.34859
31	SMO_A and SMO_A_SA	0.70513
33	SMO_R and SMO_R_SA	0.55388
34	SMO_A_SA and SMO_R_SA	0.82430
35	kNN_R_WS and kNN_R_WS_SA	0.03318
36	kNN_A_WS and kNN_A_WS_SA	0.79145
37	kNN_B_WS and kNN_B_WS_SA	0.52207
38	kNN_A_WS_SA and kNN_R_WS_SA	0.04632
39	NB_A_WS_SA and NB_R_WS_SA	0.00319
40	NB_B_WS and NB_B_WS_SA	0.00173
41	NB_A_WS and NB_A_WS_SA	0.26666
42	NB_R_WS and NB_R_WS_SA	0.03459
43	SMO_A_WS_SA and SMO_R_WS_SA	0.23744
44	SMO_B_WS and SMO_B_WS_SA	0.28740
45	SMO_A_WS and SMO_A_WS_SA	0.84869
46	SMO_R_WS and SMO_R_WS_SA	0.44681

C. Confusion Matrices – Classification Accuracy Scores

C.1 Confusion Matrix - NaiveBayes_B

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**418**	3	1	6	7	10	53	1	1
cal_B	17	**320**	28	7	6	8	109	1	4
edm_B	21	10	**295**	25	8	13	122	0	6
fm_B	22	10	26	**343**	11	7	78	0	3
leth_B	11	7	46	11	**332**	10	76	3	4
mhat_B	3	9	30	5	14	**360**	69	8	2
oth_B	39	23	13	14	21	19	**361**	3	7
rd_B	13	0	24	11	9	12	67	**362**	2
sta_B	10	5	51	10	6	6	48	9	**355**

C.2 Confusion Matrix - NaiveBayes_A

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**446**	4	1	1	3	3	41	1	0
cal_B	4	**443**	2	3	0	0	45	0	3
edm_B	11	6	**396**	11	0	1	67	0	8
fm_B	10	3	30	**388**	2	2	63	2	0
leth_B	1	0	12	1	**455**	3	27	1	0
mhat_B	6	2	9	1	6	**425**	39	11	1
oth_B	36	3	11	6	4	14	**407**	11	8
rd_B	5	1	0	2	2	9	53	**428**	0
sta_B	4	0	13	0	1	3	36	1	**442**

C.3 Confusion Matrix - NaiveBayes_R

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**446**	2	1	4	4	5	37	1	0
cal_B	7	**427**	5	3	1	1	53	0	3
edm_B	11	4	**375**	17	1	2	71	0	19
fm_B	14	4	31	**380**	5	2	62	1	1
leth_B	2	0	26	2	**424**	4	29	11	2
mhat_B	5	4	21	2	7	**407**	34	17	3
oth_B	45	3	11	10	11	20	**376**	12	12
rd_B	5	0	8	3	3	9	59	**412**	1
sta_B	8	0	35	1	3	4	43	8	**398**

C.4 Confusion Matrix - SMO_B

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**461**	1	0	0	2	6	29	1	0
cal_B	1	**438**	4	0	1	2	54	0	0
edm_B	3	3	**399**	3	2	5	82	1	2
fm_B	2	2	7	**426**	3	2	56	0	2
leth_B	1	3	3	4	**434**	9	43	2	1
mhat_B	3	1	3	3	5	**444**	36	5	0
oth_B	5	2	16	8	3	9	**451**	4	2
rd_B	3	0	0	5	4	6	38	**444**	0
sta_B	1	1	8	1	1	1	30	3	**454**

C.5 Confusion Matrix - SMO_A

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**473**	3	0	1	0	1	21	1	0
cal_B	0	**480**	3	0	0	0	17	0	0
edm_B	1	6	**455**	1	0	0	35	0	2
fm_B	1	4	4	**440**	0	0	51	0	0
leth_B	0	2	0	0	**483**	2	13	0	0
mhat_B	3	4	1	0	3	**463**	24	2	0
oth_B	1	0	0	2	0	1	**493**	2	1
rd_B	0	2	0	0	1	1	31	**465**	0
sta_B	0	1	5	0	0	0	21	1	**472**

C.6 Confusion Matrix - SMO_R

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**474**	2	0	1	0	1	21	1	0
cal_B	0	**480**	2	0	0	0	18	0	0
edm_B	1	7	**454**	1	0	0	35	0	2
fm_B	1	4	8	**437**	0	0	50	0	0
leth_B	0	2	0	0	**483**	2	13	0	0
mhat_B	3	4	1	0	4	**462**	24	2	0
oth_B	1	0	0	1	0	0	**495**	2	1
rd_B	0	2	1	0	1	1	30	**465**	0
sta_B	0	1	5	0	0	0	20	1	**473**

C.7 Confusion Matrix - kNN_B

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**302**	1	64	119	6	5	0	0	3
cal_B	7	**218**	112	131	3	15	3	1	10
edm_B	6	13	**355**	88	2	26	2	1	7
fm_B	3	11	68	**396**	0	13	1	0	8
leth_B	21	9	142	192	**108**	19	1	0	8
mhat_B	15	2	125	159	6	**186**	1	2	4
oth_B	2	8	246	154	3	61	**18**	1	7
rd_B	2	15	82	103	3	13	1	**274**	7
sta_B	0	13	76	106	2	8	0	8	**287**

C.8 Confusion Matrix - kNN_A

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**442**	1	3	38	5	6	1	0	4
cal_B	36	**374**	19	35	3	20	3	3	7
edm_B	7	6	**450**	18	1	5	4	2	7
fm_B	17	11	27	**431**	1	7	2	0	4
leth_B	51	6	51	86	**271**	26	4	1	4
mhat_B	18	4	23	58	4	**388**	3	1	1
oth_B	70	8	107	146	7	109	**45**	0	8
rd_B	23	9	32	75	3	21	3	**330**	4
sta_B	9	3	51	25	1	7	2	3	**399**

C.9 Confusion Matrix - kNN_R

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**348**	2	6	123	4	11	0	0	6
cal_B	4	**294**	24	125	3	37	0	1	12
edm_B	2	13	**443**	20	2	14	0	1	5
fm_B	1	12	27	**448**	0	8	0	0	4
leth_B	8	15	45	244	**145**	41	0	0	2
mhat_B	9	8	26	189	2	**260**	0	2	4
oth_B	1	28	94	217	5	150	**4**	0	1
rd_B	3	21	29	161	4	22	1	**250**	9
sta_B	0	13	67	66	1	14	0	8	**331**

C.10 Confusion Matrix - kNN_B_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**369**	1	1	102	12	10	4	0	1
cal_B	7	**254**	13	139	27	33	19	2	6
edm_B	1	13	**329**	67	8	51	28	1	2
fm_B	3	7	12	**417**	18	18	20	1	4
leth_B	10	8	13	162	**233**	42	24	3	5
mhat_B	5	2	12	178	17	**257**	24	2	3
oth_B	0	10	18	212	33	116	**103**	4	4
rd_B	2	14	11	77	17	24	18	**331**	6
sta_B	0	5	37	65	9	15	8	7	**354**

C.11 Confusion Matrix - kNN_A_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**446**	1	0	9	9	0	33	1	1
cal_B	2	**396**	5	16	43	0	31	2	5
edm_B	2	1	**455**	1	4	0	30	1	6
fm_B	2	2	11	**422**	21	0	41	0	1
leth_B	0	3	3	5	**433**	3	53	0	0
mhat_B	3	2	1	4	45	**409**	35	1	0
oth_B	0	6	8	39	79	0	**364**	3	1
rd_B	1	0	2	5	30	3	50	**406**	3
sta_B	0	0	43	9	12	0	27	3	**406**

C.12 Confusion Matrix - kNN_R_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**432**	2	3	56	0	3	1	0	3
cal_B	2	**433**	10	41	2	0	0	1	11
edm_B	1	10	**453**	29	0	1	2	0	4
fm_B	1	9	14	**463**	0	2	7	0	4
leth_B	0	5	8	81	**401**	4	0	0	1
mhat_B	2	4	13	76	2	**398**	0	2	3
oth_B	0	6	56	329	8	42	**56**	2	1
rd_B	0	13	11	99	1	8	8	**354**	6
sta_B	0	5	59	57	1	2	1	7	**368**

C.13 Confusion Matrix - NaiveBayes_B_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**443**	1	2	7	1	4	40	1	1
cal_B	11	**373**	8	8	1	9	85	1	4
edm_B	9	8	**337**	15	1	12	105	0	13
fm_B	10	15	19	**369**	2	3	78	0	4
leth_B	6	1	38	7	**371**	11	61	3	2
mhat_B	2	3	36	17	3	**378**	56	4	1
oth_B	9	15	21	23	6	14	**403**	5	4
rd_B	3	0	5	9	4	12	71	**396**	0
sta_B	2	2	24	5	3	4	45	6	**409**

C.14 Confusion Matrix - NaiveBayes_A_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**459**	2	0	1	0	0	36	2	0
cal_B	1	**477**	2	2	0	0	16	0	2
edm_B	4	6	**449**	4	0	0	31	0	6
fm_B	4	3	22	**408**	0	1	58	0	4
leth_B	1	0	3	1	**474**	1	19	0	1
mhat_B	1	3	6	1	2	**442**	35	10	0
oth_B	8	2	12	2	1	10	**449**	10	6
rd_B	2	1	0	1	2	9	50	**435**	0
sta_B	1	0	12	0	0	2	32	2	**451**

C.15 Confusion Matrix - NaiveBayes_R_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**464**	1	0	2	0	3	27	3	0
cal_B	4	**475**	1	3	0	0	14	0	3
edm_B	5	8	**434**	9	0	2	27	1	14
fm_B	4	4	29	**401**	0	1	56	1	4
leth_B	1	0	15	1	**461**	3	11	7	1
mhat_B	1	2	15	2	4	**433**	25	17	1
oth_B	9	2	13	13	3	14	**427**	11	8
rd_B	2	0	0	4	2	13	50	**429**	0
sta_B	1	0	17	0	1	2	22	4	**453**

C.16 Confusion Matrix – SMO_B_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**461**	1	0	0	0	4	32	2	0
cal_B	0	**449**	4	0	0	0	47	0	0
edm_B	1	3	**403**	1	1	4	85	0	2
fm_B	2	2	4	**426**	3	1	60	1	1
leth_B	0	1	0	2	**441**	5	50	0	1
mhat_B	6	1	1	0	2	**439**	46	5	0
oth_B	3	0	5	2	0	2	**486**	1	1
rd_B	2	0	0	1	3	3	43	**448**	0
sta_B	1	0	5	0	3	1	32	1	**457**

C.17 Confusion Matrix - SMO_A_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**473**	3	0	0	0	1	22	1	0
cal_B	0	**482**	2	0	0	0	16	0	0
edm_B	1	5	**456**	1	0	0	34	0	3
fm_B	1	4	4	**442**	0	0	49	0	0
leth_B	0	2	0	0	**483**	2	13	0	0
mhat_B	2	4	1	0	2	**460**	30	1	0
oth_B	0	0	0	0	0	0	**497**	1	2
rd_B	0	2	0	0	1	2	23	**472**	0
sta_B	0	1	4	0	0	0	20	1	**474**

C.18 Confusion Matrix - SMO_R_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**474**	2	0	0	0	1	22	1	0
cal_B	0	**482**	2	0	0	0	16	0	0
edm_B	1	6	**454**	2	0	0	34	0	3
fm_B	1	4	6	**440**	0	0	49	0	0
leth_B	0	2	0	0	**483**	2	13	0	0
mhat_B	2	5	1	0	3	**459**	29	1	0
oth_B	0	0	0	0	0	0	**497**	1	2
rd_B	0	2	1	0	1	2	23	**471**	0
sta_B	0	1	4	0	0	0	20	1	**474**

C.19 Confusion Matrix - kNN_B_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**313**	1	58	108	6	4	1	0	9
cal_B	6	**223**	112	129	3	14	0	1	12
edm_B	7	18	**342**	88	2	31	3	1	8
fm_B	2	11	69	**397**	0	10	1	0	10
leth_B	17	10	130	194	**110**	24	2	0	13
mhat_B	11	4	113	161	6	**193**	2	2	8
oth_B	2	18	223	154	3	70	**20**	1	9
rd_B	1	18	80	100	3	12	2	**274**	10
sta_B	16	14	69	81	3	7	0	8	**302**

C.20 Confusion Matrix - kNN_A_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**442**	1	3	38	5	6	1	0	4
cal_B	34	**386**	15	31	3	18	4	2	7
edm_B	7	6	**451**	18	1	5	4	2	6
fm_B	15	11	29	**430**	1	8	2	0	4
leth_B	49	7	54	79	**271**	30	7	0	3
mhat_B	16	4	25	55	4	**392**	2	1	1
oth_B	67	14	101	143	7	105	**56**	0	7
rd_B	21	13	33	72	3	14	4	**336**	4
sta_B	31	4	49	16	2	7	2	3	**386**

C.21 Confusion Matrix - kNN_R_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**367**	2	7	104	4	10	0	0	6
cal_B	2	**310**	22	118	3	30	0	1	14
edm_B	1	14	**443**	20	2	14	0	1	5
fm_B	1	12	29	**446**	0	7	1	0	4
leth_B	8	18	46	233	**155**	39	0	0	1
mhat_B	10	9	28	177	2	**267**	0	2	5
oth_B	2	34	99	212	5	143	**4**	0	1
rd_B	4	23	31	156	4	20	1	**251**	10
sta_B	6	13	66	34	2	15	0	8	**356**

C.22 Confusion Matrix - NaiveBayes_B_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**416**	2	1	11	6	11	51	1	1
cal_B	14	**321**	25	10	6	12	103	1	8
edm_B	20	8	**282**	30	10	12	124	0	14
fm_B	18	11	23	**343**	11	8	80	0	6
leth_B	12	7	43	13	**339**	11	70	3	2
mhat_B	3	10	28	9	11	**358**	67	9	5
oth_B	33	17	13	17	24	19	**364**	3	10
rd_B	14	0	24	13	8	13	67	**359**	2
sta_B	13	7	45	12	7	8	49	8	**351**

C.23 Confusion Matrix - NaiveBayes_A_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**451**	3	1	4	2	5	33	1	0
cal_B	3	**442**	2	3	0	1	46	0	3
edm_B	9	4	**396**	10	0	1	65	0	15
fm_B	11	3	26	**393**	0	2	62	1	2
leth_B	1	0	8	2	**456**	3	29	1	0
mhat_B	5	2	9	2	5	**423**	42	11	1
oth_B	29	2	11	10	5	16	**405**	12	10
rd_B	6	1	0	4	3	8	52	**426**	0
sta_B	4	0	13	0	1	5	36	1	**440**

C.24 Confusion Matrix - NaiveBayes_R_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**448**	2	1	7	4	7	30	1	0
cal_B	5	**430**	3	4	0	3	52	0	3
edm_B	9	4	**357**	19	2	2	78	0	29
fm_B	9	4	29	**379**	4	3	66	2	4
leth_B	2	0	24	3	**423**	4	31	10	3
mhat_B	3	4	23	3	7	**398**	42	15	5
oth_B	38	2	12	13	14	18	**378**	12	13
rd_B	7	0	9	4	4	10	59	**405**	2
sta_B	7	0	37	1	3	5	44	7	**396**

C.25 Confusion Matrix - SMO_B_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**464**	2	0	2	2	5	23	2	0
cal_B	1	**435**	6	0	0	4	50	1	3
edm_B	3	2	**398**	5	2	5	80	1	4
fm_B	3	2	7	**427**	6	1	50	2	2
leth_B	2	4	2	5	**435**	5	44	2	1
mhat_B	10	1	3	3	5	**438**	35	4	1
oth_B	5	3	17	15	5	10	**437**	6	2
rd_B	4	0	0	2	5	5	39	**445**	0
sta_B	3	1	8	1	2	1	26	2	**456**

C.26 Confusion Matrix - SMO_A_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**474**	4	0	1	0	1	19	1	0
cal_B	0	**481**	3	0	0	0	16	0	0
edm_B	1	4	**456**	2	0	0	33	0	4
fm_B	1	4	4	**440**	0	0	50	0	1
leth_B	0	2	0	0	**483**	2	13	0	0
mhat_B	3	4	1	0	3	**463**	23	2	1
oth_B	2	0	0	1	0	1	**493**	2	1
rd_B	1	2	0	0	1	1	25	**470**	0
sta_B	0	0	8	1	0	0	22	2	**467**

C.27 Confusion Matrix - SMO_R_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**476**	2	0	1	0	1	19	1	0
cal_B	0	**482**	2	0	0	0	16	0	0
edm_B	1	7	**455**	2	0	0	32	0	3
fm_B	1	4	7	**437**	0	0	50	0	1
leth_B	0	2	0	0	**483**	2	13	0	0
mhat_B	3	4	1	0	4	**463**	22	2	1
oth_B	2	0	0	1	0	1	**494**	1	1
rd_B	1	2	1	0	1	1	24	**470**	0
sta_B	0	1	6	1	0	0	22	2	**468**

C.28 Confusion Matrix - kNN_B_WS_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**363**	3	2	107	0	1	2	4	18
cal_B	1	**273**	12	168	4	2	6	2	32
edm_B	1	28	**324**	122	3	2	6	2	12
fm_B	2	5	7	**459**	0	2	4	1	20
leth_B	6	16	9	215	**209**	9	3	1	32
mhat_B	2	5	5	173	4	**292**	3	3	13
oth_B	0	36	20	346	2	3	**53**	5	35
rd_B	2	8	6	100	1	2	4	**345**	32
sta_B	3	6	20	60	6	1	3	10	**391**

C.29 Confusion Matrix - kNN_A_WS_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**449**	2	5	26	0	1	15	1	1
cal_B	7	**450**	5	18	2	0	12	1	5
edm_B	4	2	**460**	13	0	0	14	1	6
fm_B	9	5	9	**446**	0	0	28	0	3
leth_B	10	4	5	44	**392**	3	41	0	1
mhat_B	8	4	8	32	2	**431**	13	1	1
oth_B	17	8	35	172	7	1	**252**	2	6
rd_B	6	4	8	43	0	3	24	**409**	3
sta_B	4	0	44	24	4	1	10	4	**409**

C.30 Confusion Matrix - kNN_R_SA_WS

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**407**	2	4	80	0	3	1	0	3
cal_B	2	**414**	9	60	2	0	0	2	11
edm_B	1	9	**459**	25	0	1	2	0	3
fm_B	1	9	24	**462**	0	0	0	0	4
leth_B	1	17	39	99	**332**	12	0	0	0
mhat_B	2	7	18	71	1	**396**	0	2	3
oth_B	0	13	135	291	9	40	**9**	1	2
rd_B	0	18	23	121	0	8	1	**323**	6
sta_B	0	5	66	46	4	6	0	8	**365**

C.31 Confusion Matrix – NaiveBayes_B_WS_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**442**	1	2	6	2	2	42	2	1
cal_B	8	**366**	4	8	1	12	93	1	7
edm_B	6	5	**322**	20	3	12	117	0	15
fm_B	7	16	19	**359**	7	4	79	0	9
leth_B	10	7	38	11	**362**	8	57	3	4
mhat_B	2	6	34	13	4	**372**	64	3	2
oth_B	8	15	19	21	9	14	**401**	6	7
rd_B	3	1	6	7	6	9	75	**393**	0
sta_B	3	4	25	7	6	5	42	7	**401**

C.32 Confusion Matrix - NaiveBayes_A_WS_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**464**	2	0	0	1	1	29	3	0
cal_B	2	**467**	2	3	0	1	23	0	2
edm_B	5	6	**439**	3	0	0	36	0	11
fm_B	2	3	19	**410**	1	1	62	0	2
leth_B	1	0	0	1	**474**	1	22	0	1
mhat_B	1	3	7	2	2	**440**	36	8	1
oth_B	5	2	13	4	1	8	**447**	13	7
rd_B	1	1	0	2	2	7	52	**435**	0
sta_B	0	0	12	0	0	2	34	2	**450**

C.33 Confusion Matrix - NaiveBayes_R_WS_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**465**	1	0	1	1	2	26	4	0
cal_B	3	**465**	1	3	0	1	24	0	3
edm_B	4	7	**417**	8	0	3	33	0	28
fm_B	5	4	29	**399**	3	1	54	1	4
leth_B	1	0	13	1	**460**	3	12	8	2
mhat_B	2	3	19	2	4	**424**	30	15	1
oth_B	9	2	13	12	4	12	**427**	13	8
rd_B	2	0	2	2	3	10	52	**429**	0
sta_B	2	0	18	0	1	2	31	4	**442**

C.34 Confusion Matrix - SMO_B_WS_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**465**	1	0	1	0	3	28	2	0
cal_B	1	**442**	5	0	0	1	46	2	3
edm_B	2	2	**400**	2	1	3	84	1	5
fm_B	5	3	5	**428**	3	1	54	0	1
leth_B	1	1	0	3	**440**	6	47	2	0
mhat_B	8	1	3	0	2	**441**	40	4	1
oth_B	6	1	12	5	1	8	**462**	4	1
rd_B	2	0	1	1	5	4	42	**445**	0
sta_B	3	0	8	0	1	1	27	2	**458**

C.35 Confusion Matrix - SMO_A_WS_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**475**	3	0	0	0	1	20	1	0
cal_B	0	**482**	3	0	0	0	15	0	0
edm_B	1	6	**456**	1	0	0	32	0	4
fm_B	1	4	4	**441**	0	0	49	0	1
leth_B	0	2	0	0	**485**	1	12	0	0
mhat_B	2	4	1	0	3	**460**	28	1	1
oth_B	1	0	0	0	0	1	**494**	2	2
rd_B	1	2	0	1	1	1	21	**473**	0
sta_B	1	1	4	0	0	0	21	1	**472**

C.36 Confusion Matrix - SMO_R_WS_SA

	bnf_B	cal_B	edm_B	fm_B	leth_B	mhat_B	oth_B	rd_B	sta_B
bnf_B	**476**	2	0	0	0	1	20	1	0
cal_B	0	**483**	2	0	0	0	15	0	0
edm_B	1	8	**452**	1	0	0	33	0	5
fm_B	1	4	7	**438**	0	0	49	0	1
leth_B	0	2	0	0	**484**	2	12	0	0
mhat_B	2	5	1	0	5	**457**	28	1	1
oth_B	1	0	0	0	0	1	**496**	0	2
rd_B	1	2	1	1	1	1	21	**472**	0
sta_B	1	1	4	0	0	0	21	1	**472**

D. Tweet Examples (Relevant to Table 3.4)

Tweet (T1): 1/10/2017, Hey Banff! Forty Creek masterclass later this aft @parkdistillery Can't wait! #wearefortycreek https://t.co/W0rgzByyAj-re10

Tweet (T2): 1/10/2017, My afternoon is looking pretty magical... @Banff National Park https://t.co/vvPldHZvo0

Tweet (T3): 1/10/2017, Its officially on! The new season begins with todays opening at Norquay. See yall out there! https://t.co/kxKOgtde82-pu68eob2a78

Tweet (T4): 15/11/2017, Early season link up on Cascade #banff #alberta game on winter18 @Banff_Squirrel https://t.co/jYYPnqPwn2-ks80

Tweet (T5): 15/11/2017, Watching the first rays of sunshine hitting mountains never gets old. Moraine Lake Banff https://t.co/rFH5aPRTpA-re10