

LOCATION BASED POPULARITY ANALYSIS OF TWITTER DATA

By

S.M. Rashel Rana

Bachelor of Science in Computer Science and Engineering

DIU, Bangladesh, 2004

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Science

in the program of Computer Science

Toronto, Canada, 2015

©S.M. Rashel Rana 2015

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

LOCATION BASED POPULARITY ANALYSIS OF TWITTER DATA

S.M. Rashel Rana

Master of Science in Computer Science

Ryerson University, 2015

Abstract

This research aims to analyze location based twitter data to measure the popularity of the products/persons or any given user parameter. For this purpose this work has integrated sentimental analysis, location based system and ontology. An application with a novel user interface has been developed to search and visualize the data on Google map. This research work uses publicly available and location enabled twitter data. This work also has the capability to process tweet data without user's locations.

The main contribution of this research is the integration of sentimental analysis on location based Twitter data. Another significant contribution is the development of a novel user interface, which allows the user to search on a map interactively with multi-focusing features on Google map. This integrated sentiment analysis work efficiently performs location based popularity scaling on products, persons, brands or any given topic.

Acknowledgements

I would like to express my utmost gratitude to my supervisor Dr. Abdolreza Abhari for the encouragement, dedication and periods of time he provided during the past two years of my graduate studies. Dr. Abhari always found time for me from his office schedule which allowed me to improve my thesis as well as provide me with all the required support. Working under the supervision of Dr. Abhari was a good learning experience for me, which allowed me to enhance my skills and expand my proficiency as a computer scientist. This thesis would not be possible to complete without the help of Dr. Abhari.

I would also like to acknowledge the faculty members of the department of Computer science at Ryerson University. The knowledge and experience of the faculty has allowed me to strengthen my insight into computer science and expand my knowledge base.

I would also like to express my gratitude to the staff members of the Computer Science department at Ryerson University as well as graduate peers for their consistent support over the past two years of our study.

Lastly I am thankful for the inspiration, patience and support of my family, for without them none of this would be possible. I would not have been able to achieve my goal and grasp my dream in working in the field I'm interested in without them. No sum of appreciation would be enough to express my gratitude, love and thankfulness towards my family.

Dedication

To my Parents and Family

Table of Contents

Author's Declaration.....	II
Abstract	III
Acknowledgements.....	IV
List of Figures.....	VIII
List of Algorithms.....	X
List of Abbreviations.....	XI

Chapter 1

Introduction.....	1
1.1 Motivation.....	1
1.2 Problem Statement	2
1.3 Methodology.....	3
1.4 Application and Practical Usage.....	4
1.5 Research Challenges and Contribution.....	4
1.6 Thesis Outline.....	5

Chapter 2

2.1 Background Information.....	6
2.1.1 Twitter.....	6
2.1.2 Importance of Information and Sentiment Analysis.....	6
2.1.3 Data Mining.....	7
2.1.4 Popularity and Popularity Scaling.....	8
2.1.5 Location-based analysis with semantic web.....	9
2.1.6 Mining Permission Request Patterns	9
2.1.7 Geographical Location	10
2.1.8 Semantic Web Search	11
2.1.9 Entity Recognition	11
2.1.10 Entity Monitoring.....	12
2.1.11 Ontology Based Text mining.....	12
2.2 Related Work.....	13
2.2.1 Supervised Approach for Twitter Sentiment	13
2.2.2 Combining strengths, emotions and polarities	13
2.2.3 Twitter Sentiment Analysis for Election Cycle.....	14
2.2.4 Twitter Sentiment by Modelling Public Mood and Emotion.....	15
2.2.5 Mining Social Media with Social Theories: A Survey.....	16
2.2.6 Frequent pattern.....	17

Chapter 3

Material and Methods	18
3.1 System Architecture.....	18
3.2 Google Maps API	20
3.3 Location Selection.....	20
3.4 Semantic Search.....	21
3.5 Twitter API.....	25
3.6 Data Crawler.....	26
3.7 Text Processor.....	28
3.8 Data Cleaning.....	29
3.9 Word Parsing.....	30
3.10 Sentiment Dictionary.....	33
3.11 Word Matching and Final Score Calculation.....	34
3.12 Data Visualization.....	36

Chapter 4

Experiments & Results.....	37
4.1 Experiment Setup and Design	37
4.2 Experiment sets for System Scaling	37
4.2.1 Single product in Single Location	38
4.2.2 Single product in Multiple Locations.....	41
4.2.3 Multiple products in Single Location.....	44
4.2.3.1 Data Validation.....	46
4.2.4 Multiple products in Multiple Locations.....	48

Chapter 5

Conclusion & Future work.....	52
5.1 Conclusion.....	52
5.2 Contributions.....	53
5.3 Future Work.....	54

Chapter 6

References.....	55
------------------------	-----------

List of Figures

Fig.1 Emotions Ontology sample with sub-categories	12
Fig. 2 System Architecture.....	19
Fig. 3 System interface with Google Maps API, Location Selection	20
Fig. 4: Searching a keyword along with ontology	22
Fig. 5 Twitter location on focal area on Google map.....	24
Fig. 6 Multiple Zone Selection on the Globe.....	25
Fig. 7 Sample Tweet about election candidate Olivia Chow	29
Fig. 8 REGXP_INSTR function working pattern	31
Fig. 9 Word Array Extraction from Tweet	32
Fig.10 Developed Word Array and Lexicon Dictionary.....	34
Fig. 11 Score Calculation using Word Array and Lexicon Dictionary	35
Fig.12 Tag Cloud for #TOpoli related Tweet.....	36
Fig.13 Line Graph for the Popularity of iPhone in Toronto for 100 Tweet	38
Fig.14 Cumulative Line Graph for the Popularity of iPhone in Toronto based on 100 Tweet....	39
Fig.15 Bar Chart and Pie Chart for 100 Tweet about iPhone popularity in Toronto	39
Fig.16 Line Graph for the Popularity of iPhone in Toronto for 10K Tweet.....	40
Fig.17 Cumulative Line Graph for the Popularity of iPhone in Toronto based on 100 Tweet....	40
Fig.18 Bar Chart and Pie Chart for iPhone popularity in Toronto	41
Fig.19 Toronto VS New York Popularity for iPhone based on 100 Tweet.....	42
Fig. 20 Cumulative popularity line graph on iPhone popularity in Toronto VS New York	42
Fig.21 Cumulative line graph on iPhone popularity in Toronto vs New York.....	44
Fig.22 Pie Chart for iPhone Popularity Toronto VS New York on (10,000 Tweet).....	44
Fig.23 Line graph for popularity scaling for the Election base on 100 Tweet.....	44
Fig.24 Cumulative line graph for popularity scaling for the Election base on 100 Tweet.....	45
Fig.25 Cumulative Line graph for popularity scaling for the Election base on 850 Tweet.....	46
Fig.26 Pie Chart for the Predicted Election Result VS Real Election Result.....	47
Fig.27 Bar chart for Weekly Election Trend base on 985+ Tweets.....	47
Fig.28 Cumulative Line graph for Weekly Election Trend base on 980+ Tweets.....	48
Fig.29 Popularity scaling for Toyota and Honda for 100 Random Tweet in Toronto and NY ...	49
Fig.30 Popularity scaling of Toyota VS Honda in Toronto & New York	50

Fig.31 Popularity scaling Bar chart for brands on different locations	50
Fig.32 Popularity scaling for brands over the Locations	51
Fig.33 Popularity scaling for Brands over the Targeted market	51

List of Algorithms

Algorithm 1: Location Based Searching	21
Algorithm 2: Semantic Search.....	22
Algorithm 3: Data Crawler.....	27
Algorithm 4: Text Processor.....	28
Algorithm 5: Data cleaning	29
Algorithm 6: Word parsing	31
Algorithm 7: Word Matching Strategy.....	34

List of Abbreviations

API	Application programming interface
Acc	Accuracy
Avg	Average
GPS	Global Positioning System
XML	Extensible Markup Language
Web	Ontology Language
NLP	natural language processing
SQL	Structured Query Language
SVM	Application programming interface
SN	Social Network
SNS	Social Network Site
UGC	user generated contents
GPS	Global Positioning System
XML	Extensible Markup Language
Web	Ontology Language
NLP	Natural language processing
SQL	Structured Query Language
SVM	Support Vector Machines

Chapter 1

Introduction

The recent explosion of social networks is so vast that it has become a very important platform for communication on the web. It is gaining popularity every day. By using micro-blogging services, users post messages about their daily life and initiate discussions by sharing personal opinions and emotions on different topics. These can range from something as simple as some products, events and services to more complex issues that deal with economic issues, problems, interests, culture, politics, religions, diseases, epidemic, food crisis, and famine and so on. The topics of Twitter discussions are limitless. According to C. Aggarwal [1] “The richness of this network provides unprecedented opportunities for data analytics in the context of social networks [1]”. The vast amount of available behavioral data in online social networks may give an opportunity for knowledge discovery. Data mining techniques may detect implicit or hidden patterns within a social networking site. This technique provides feedback to sense user sentiments for purchasing behavior, identification of social groups and understanding the hidden trends in network evolution. This research focuses on products popularity and the popularity of election candidates.

1.1 Motivation

According to M. Russell [2], the Web’s ongoing evolution is an important step forward because they provide an effective mechanism for embedding “smarter data” into web pages and are easy for content authors to implement. The Web 2.0 has changed the way of communication on the web. Internet users are no longer passive consumers. Using the Social networks (SNs) they have become active participants by connecting, producing and sharing information, experiences and opinions with each other. This vast number of communication messages express sentiments and different types of influences which attracts the attention of the information extraction among the research community. These user communications are not just only

messages but a useful source of valuable information, as it contains personal opinions, emotions and expressions about different topics and entities. Public opinions extracted in the form of trends are interesting for researchers, sociologists, news reporters, marketing professionals and opinion tracking companies. The number of information processing companies is increasing for online data analysis. These companies are interested in fast discovery of trends that are extracted from social media platforms. Therefore, an efficient and capable tool for automatic detection and monitoring of topics is needed, which has influenced the research in text analysis.

Information retrieval and trend detection from social media has to ensure the deeper insight from the user generated contents (UGC). This research is only interested in online social network data. As a data source this project preferred Twitter over Facebook (FB). Facebook has constricted privacy policies which limits the scope of data extraction. Facebook displays personal preferences through visual orientation. On the other hand, Twitter is more text oriented with general sentiment. Twitter has released, Application Programming Interface (API) to collect the public data significantly earlier than FB. Twitter API can deliver better insights with easier access to the information extraction. According to Barbosa and J. Feng, this is possible because Twitter is one of the largest SN with more than a billion tweets posted every day. These are the main reasons to select Twitter as a data source in this research work.

1.2 Problem Statement

This thesis aims to measure location based popularity of any topic (keywords, products, persons, event etc.) by analyzing public statements. This is possible by analyzing messages that contain general sentiment. The first step is to identify the tweet-location. Targeting location can be identified by users' choice or the density calculation on the globe. Tweet-Marker-Density can be visualized by using the developed application for this research. This application software (implemented in this work) not only identifies the location but also allows the data to be visualized. The second step is to identify a big data source, which can deliver real-time or relatively new data. As mentioned before, this research work is using Twitter-API as a feeding data source. Once the location is identified and tweet data is captured, this project will undertake

detailed data analysis. Each single tweet message will be processed and analyzed separately to extract all the useful information. Generally Twitter message contains hash-tags (marked with the character #), white spaces, punctuations (!,;:- ({[]})|/ ' "" ---),web addresses and special characters ('@%!~^&^â, °Å,€~Å'-â,,¢?'). For the pre-processing, it is required to undergo data-cleaning process by removing these elements. As a result the process will generate more concise output (tweet string), which is easier and faster to parse. Afterwards, the smaller filtered message will be parsed and stored into arrays of words to be compared with the sentiment dictionary to score as positive, negative or neutral.

1.3 Methodology

The solution for problem statement in section 1.2 consists of a few tools, components, techniques and methods. The detailed information and the process of the implementation will be described in Chapter 3. Only a brief overview of the proposed solution is summarized here. First of all the application user (developed for this research) will draw target location/locations on Google map by using described application. In the next step, user will search for particular topic or different keywords. As a result, the related information will be retrieved from Twitter using Twitter Streaming-API and visualized on Google maps (API). Afterwards, the tweet information will be written on a local database containing user information (tweet sender), longitude, latitude, generation time, user tweet (text message) and the tweet-location. By nature tweets are noisy and unstructured. Several logical processes (will be explained in CH-3) are needed to be perform to clean these data. Let us consider the following tweet, “*@Ryerson_Alumni: What Every #Entrepreneur Needs to Know - Free @ChangSchool event w @RBC at #Ryerson <http://t.co/onPUy54Rd7>*”. The data cleaning process will remove hashtag (#), @, URL and other special characters. Finally it will generate the following: “**Ryerson Alumni What Every Entrepreneur Needs to Know Free ChangSchool event RBC at Ryerson**”. Following this step, the cleaned *tweet string* will be processed for parsing and as a result a word array will be constructed. For basic communication in any language a person needs ~2000 words [3]. This work has developed a ranking dictionary by using Sentiword.net. SentiWord.Net is widely used popular lexical resource for sentimental analysis [4]. This dictionary has over 146,600 words, which is more than adequate to score a sentence efficiently. The word array will be compared

with the developed dictionary and a score will be generated for each tweet. Each extracted (parsed from the tweet) word has an individual weight (score) in the dictionary and the tweet score is the sum of the values of the extracted words of the tweet.

1.4 Application and Practical Usage

Practically a single user message or tweet can play a very little role to encapsulate the public opinion. Only gathering and analyzing a large number of tweets can generate a relevant meaningful expression of public opinion. Once the considerable amount of data is gathered, it has to undergo the data cleaning process to perform the sentiment analysis and as a result it will generate some useful output. These output can be used for product market line, stock market forecasting, business promotion, user feedback, post event decision making system, election prediction etc.

1.5 Research Challenges and Contributions

Sentiment analysis is an active research area and many researchers have performed this research. Relatively low number of location based research has been performed by the research community. This research has integrated sentimental analysis and location based system along with ontology. This work has resolved some challenges:

- (1) Application development and visualization of location based data
- (2) Ontology based searching integration
- (3) Location based and Location independent data extraction
- (4) Sentiment analysis on the extracted data

This Research work can perform popularity scaling on different products but is not limited to scope of finding the popularity of products. It can also compare a person's popularity or any user defined object. Popularity scaling can be done in several ways:

- (1) Single product in single location
- (2) Single product in multiple locations
- (3) Multiple products in single location
- (4) Multiple products in multiple locations

This research have done experiment of every mentioned scope. Firstly, this thesis work has proposed a location based popularity measurement system by using sentiment analysis. Secondly, it has integrated semantic search approach. Thirdly, this work has designed, developed and implemented an efficient application for data crawling, capturing and processing. The usage of interactive GUI and real-time visualization makes it more sensible and enables the popularity scaling system to act smoothly with the experimental approach. This research has collected Twitter data for a three months' timeframe. This data consists of a wide snapshot of different users at different timestamps.

1.6 Thesis Outline

Chapter 1 describes the Introduction which includes motivation, problem statement, solution approach and application. Related information has been discussed for the potential goals of the proposed problem.

Chapter 2 briefly presents background information. Related key concepts are discussed only, which are pertinent to the proper understanding of the rest of the thesis. Background information is not comprehensive because of broad research area.

Chapter 3 describes the materials and methods of the proposed solution, related issues and components. Detailed explanations have been provided on System Architecture, Algorithms, Google Maps API, Location Selection, Semantic Search, Sentiment Classification, Twitter API, Data Crawler, Text Processor, Data Cleaning, Word Parsing, Sentiment Dictionary, Word Matching and the Final score calculation process.

Chapter 4 explains experiment, data set, sentiment analysis, graphical representation with score calculation and results of the proposed system. It discusses about some real life problems, performs the experiments and proves its adaptability.

Finally Chapter 5 discusses conclusions, contributions and future developments of this research work.

Chapter 2

2.1 Background Information

2.1.1 Twitter

Twitter is an online social media website. Twitter allows its users to communicate through micro-blogging services and opens a way to share their feelings, opinions or messages to other users over the internet. Twitter was developed in 2006. Currently Twitter is one of the top three popular online social media websites [5]. It provides micro-blog services to write messages up to 140 characters at one time, which are typically not more than 30 words. As of September 2013, Twitter has 645 million users and who produced about 300 billion tweets. More than 143,199 Tweets are tweeted (i.e. transmitted or delivered) per second [6].

Twitter uses several character symbols for performing operations. For instance, '@USER_ID' transmits direct messages, 'RT' for performing re-tweet. To enforce a category or a topic discussion '#' (hash tag) is widely used. Twitter itself also suggests its users to use hash tag for putting extra wastage of the words. Approximately 98% users do not have enough followers [7]. Some of the tweeter users have millions of followers but they are very few in number (in year 2010) [8]. These users are different kind of media-stars, politicians or popular news web sites (CNN, newspaper websites etc.). Follower is very important to keep the flow of tweeting. A tweet user must have at least 10 followers for a regular flow, otherwise they will lose interest of tweeting [9]. Twitter has a very high number of users and many new users are joining every day. Location enabled tweet can play an important role for business [10]. Benjamin and Krzysztof mentioned that the geographic-location is a key component for information retrieval on the web for recommendation systems especially in mobile computing, social networks and place-based integration.

2.1.2 Importance of Information and Sentiment Analysis

Information is very important and valuable for researchers, processes, methods and organizations. It is the key component for any decision making process. Proper information is sometimes as valuable as gold. For example, proper information and public opinion, about the

stock market, may save investors from millions of dollars of bad investments. Information about weather forecasting may lead to safe landing of aircrafts or ensure smooth sea travels.

Sentiment is an attitude, feeling, emotion or opinion toward something. Sentiment can be comprehended with hearing, sight, touch, smell and taste. Sentiment analysis technique is used for the extraction and determination of better insights of natural language. Sentiment analysis can extract sentiment in three different forms: positive, negative and neutral. Sentiment analysis can be categorized in the document, sentence and word level. This research is using “word level” sentiment analysis. Word level sentiment analysis uses a dictionary to compare each word and returns positive, negative or neutral value of that word in a sentence. This thesis is using the social networking website, Twitter, as a data source.

2.1.3 Data Mining

Data mining is a process of identification or extraction of hidden information from large volume of information. Data mining process discovers useful patterns that are not necessarily found by merely querying or processing data or metadata in the data warehouse. To be practically useful, data mining must be carried out efficiently on large files and databases. Although some data mining features are being provided in RDBMSs, data mining is not well-integrated with database management systems [11]. Data mining includes following different sub-areas.

Association rules: It is a relation among the objects inside large volume of data. This rule defines the correlation, such as, if variable-A happens, what is the probability that variable-B also happens. For instance: (i) when a customer purchases some toys, he is likely to buy some baby food; (ii) if a patient has fever, he is likely to have some muscle pain or headache. Let us consider an experiment on a customer’s purchase behavior. Here prediction will show how certain attributes or behavior within the data normally behave and continue in future. Secondly for identification of a particular trend, data patterns can be used to show that the occurrence of an item/ event/ activity may create and influence some other activity. Thirdly the classification can split into various categories depending on different parameters. One of the major technologies in data mining involves the discovery of association rules. For *Association rules* there are several usages, such as (i) Market-Basket model, support, and confidence, (ii) Apriori algorithm for finding frequent (large) item sets, (iii) Sampling algorithm, (iv) Frequent-Pattern (FP) Tree and

FP-Growth algorithm for finding frequent item sets, (v) Partition algorithm, (vi) Association rules among hierarchies, (vii) Multidimensional associations, and (viii) Negative associations [11, 27].

Classification hierarchies: It is a system where objects or entities appeared in a form of hierarchical relation. It constructs a system by a classification relation. For instance, (i) a group of students may be divided into four categories: hardworking, average working, somehow working, and not working; depending on their class time behavior. (ii) The health deterioration can be classified on the diet behavior like vegetarian, fish consumer and red meat consumer. *Classification* is the process of learning a model that describes different classes of data. For constructing a decision tree from a training data set, decision Tree Induction Algorithm is widely used.

Patterns discovery in sequence: Sequential patterns over time in a series is an interesting research area in data mining. It can be visible on daily basis, e.g. on amount of sales or in product distribution. For instance, (i) predicting amounts of sales of new product based on advertising expenditure. (ii) Two products may show the same selling pattern in warm season but a different pattern in cold season. (iii) If a customer buys a car, he or she has to buy insurance within 1 week.

Clustering: It is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). For instance, (i) identification of areas of similar land use in an earth observation database. (ii) identifying groups of motor insurance policy holders with a higher than average claim cost. (iii) earthquake studies have observed earthquake epicenters clustered along continental fault lines. *K-Means* algorithm is a very popular tool for clustering.

2.1.4 Popularity and Popularity Scaling

Popularity is a state or a condition when many people like certain things. This can be demonstrated as the favor of general people, a group of people or public. In most of the cases, it is a social phenomenon. Popularity can be increased and decreased over time, which is explained by ‘scaling’. By the Geometric definition, Scaling is a linear transformation of the object in

increasing or decreasing factor. It is a system that defines the standards of measurement. Popularity scaling quantifies the phenomenon in social measure. The popularity of different products or brands can be identified by scaling the consumer opinion or likeness. The social influence can be measured and scaled by analyzing the popularity scaling process. Josep et al. [12] argued that the scaling of popularity is a unique challenge in terms of identification, management and maintenance.

2.1.5 Location-based analysis with semantic web

Location-based system and services have a great influence in our daily life. People experience these influences through the usage of smartphone and location enabled electronic gadgets. Widely used services through these devices includes weather forecasts checking, traffic updates, locations of nearby gas stations, restaurants etc. and sometime as a GPS device. Celino et al. [13] have developed an android application BOTTARI that processes social media content to identify people's point of interest (POI) and visualize it on google map. The system has integrated semantic analysis for identifying people's influence through location based microblog postings. They have used machine learning algorithm SVM, rule base approach for language processing and a kernel syllable for sentiment elicitation. This system can process only single location based objects but our system can process multiple user defined objects on targeted locations.

2.1.6 Mining Permission Request Patterns

Mobile applications are very popular now-a-days and have a fast paced growth, especially with open development platform. Hundreds of thousands of third-party applications are available for different platforms on the web. Normal trend for end users is to install these third-party apps with couple of clicks. With hundreds of choices, end users often get confused about different privacy policies. These applications can access different information of the user resources like hardware, profile info, likes, posting, ads and information of other social networking sites as well. The framework of applications is driven by permission systems to control the application privileges. The access of the application about the user's privacy and security is determined by specific permission request. Many of the end users are not aware of the

permission warnings or do not pay attention to this matter. Most of the users do not know about permission combinations of applications. Primary goal here is to simplify permission systems followed by statistical methods. Secondly, development of an identical common pattern for permission request system which will not allow the predominant patterns for user security. Reputed applications normally responds well for permission request patterns but poor (reputation, normally <10 ratings) applications deflects from those patterns.

The third-party application markets limit mobile applications access, private information and resources to users by system permission. The official Android Market stores Android applications user ratings and reviews with other technical information regarding the apps. A research study on popular 1,100 apps have set-out that these applications take a number of hardware permissions from the users but most of them remain unused. Some of the permissions' categories are not related with the apps at all [14]. Frank et al. have proposed to develop a white-listing approach, which will warn about the safe permission patterns before the installation of any mobile applications. The authors expect that the successful deployment of white-listing-approach can elevate the pattern mining system if it is synchronized with human review.

2.1.7 Geographical Location

Geographical location is a physical position on the Earth's surface. There are numbers of ways to obtain Geo-location information. Such as user's latitude and longitude, IP / MAC address, RFID, Wi-Fi access points and GPS coordinates. But it is typically identified by latitude & longitude coordinates. For example *Toronto* is identified by (43°4259N, 79°2026W). When it comes to the case of sensor based information retrieval and comparison, it is very important to identify the geo-location. Geo-Location of a tweet can be extracted (if exists) from Twitter API by passing user defined location (latitude and longitude).

Limitations of Geo-Location: IP based geo-location may not be correct because of incorrect location association in the Geo-database. Sometime it may be scaled on broad geographic area. Many addresses are associated at city level, but there are still some new addresses for which longitude and latitude might not appear because of their nonexistence in Google Geo-database.

2.1.8 Semantic Web Search

The Semantic Web consists of two things: Common formats for integration and combination of data drawn from diverse sources [15]. It is also about language for recording how the data relates to real world objects. Semantic web can be treated as the extension of present web. Extensible Markup Language (XML) is the main technology for the semantic web. Web Ontology Language (OWL) is used for graphically expressed information (Ontologies). The combination of these powerful tools will allow the creation of search engines and will retrieve related search content.

2.1.9 Entity Recognition

Entity recognition is a natural language processing (NLP) job. This is designed to extract and identify the entities about objects names, geo-locations, company or organization, etc. from any data source, which is expressed in natural language. Finding the relevant information from systems is very important. Entities can be linked to open data to search and extract more related information. Just only the keyword may not be enough for searching in many cases. Moreover some entity names may be ambiguous. Ambiguous entities consist of at least two specific words to express that entity, where each specific word is different. For example, ‘Rob Ford’ is not an ambiguous entity. If we reduce it to ‘Ford’, it can bring ambiguous result, such as **‘Henry Ford’**, **‘Ben Ford’** and **‘David Robert Ford’**. There are several factors that result in disambiguation, but domain-specific supervised entity recognition approach is the most popular (with string parameter) approach, which can minimize disambiguation. Specific domain can extract better results but the system produces lower number of data with poor performance.

2.1.10 Entity Monitoring

Entity monitoring technique can be applied on company, products or personnel by sentiment analysis. Let us consider Starbucks has introduced a new drink for 2 weeks. Starbucks may want to know how people are reacting to the new product. More sales does not always prove that the product is good, especially on a promotional short timeframe. This kind of observation (i.e. entity monitoring) can help the company to do a quick change/amendment on their promotional product. Sometimes company can retrieve the reaction of people about the user’s desired change. The observations can be stored in a database along with date for further analysis

of products which may determine hidden trends of business over time by applying entity monitoring.

2.1.11 Ontology based Text Mining

Ontology is a metaphysical nature of being, existence, becoming or reality, along with the basic categories of being relations. In the wide sense human ontology is to investigate the relation relying on institutional, social, and technological conventions representing an intellectual activity. Sam et al. have shown that the emotional-ontology from social networking sites combined with product-ontology can analyze the customer behavior [15]. For Ontology-based Text-Mining, unstructured data inside the tweet are analyzed for a particular topic and extracted keywords. Then the ontology of the particular topic and emotions from text mining are matched to discover behavior of the consumer in the market. For example let us consider the class-sentiment of “happiness”. The ontology subclass of “happiness” has been shown in Fig. 1.

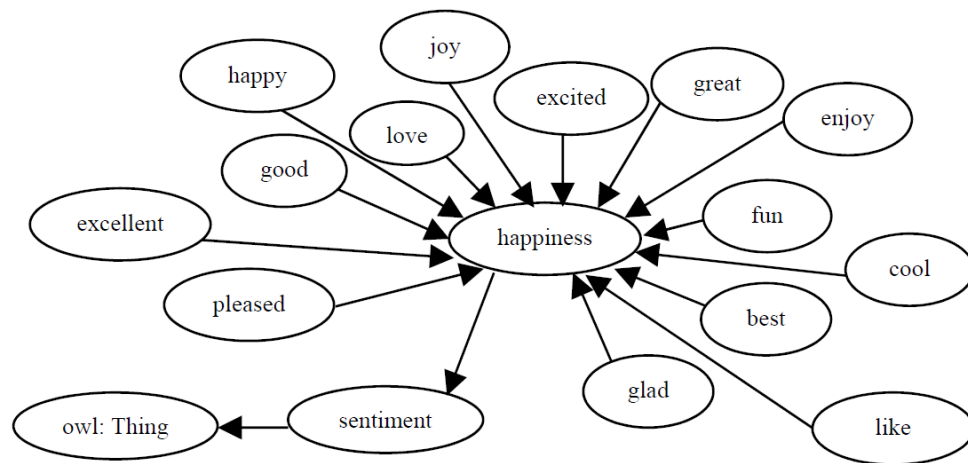


Fig. 1 Emotions Ontology sample with sub-categories [15]

2.2 Related Work

2.2.1 Supervised Approach for Twitter Sentiment

Sentiment analysis represents the use of Natural Language Processing [16]. Efthymios et al. have mentioned that sentiment analysis can be classified in different levels (documents, sentences, word, phrases). Twitter sentiment analysis is not an easy job as many informal languages and specialized characters set are used in tweets. The usage of part-of-speech features is an important research area. Many researchers have tried to explore this usage but results remain inconclusive. Emoticons are common feature of micro-blogging data, but researchers found it less useful for sentimental resource development. Significant numbers of researchers have investigated multiple ways of automatic training data collection. Most of the efforts ended up with polarity classification. Efthymios et al. have investigated linguistic features utility to detect the sentimental weight for Twitter messages [16]. They have analyzed suitability of “automatic part-of-speech tags” and “sentiment lexicons” features for sentiment analysis process. They tried to analyze automatic part-of-speech and sentiment lexicons to examine whether it is suitable for sentiment analysis. User expresses their thoughts and opinion through tweet, however that does not necessarily mean that the user talks about something specific. The user can talk about anything or rather everything. So it was very challenging to build a system that can mine Twitter sentiment (based on any given topic). For development and training purposes the authors have experimented through the use of hash-tagged data set (HASH) and the emoticon data set (EMOT) from <http://twittersentiment.appspot.com>. Through their experiments they found out that the use of "part-of-speech features" might not be a good fit for sentimental analysis, especially for micro blogging domain. Finally, their experiments concluded that the abbreviations, polarity and emoticons were clearly useful for sentiment analysis of micro blogging services.

2.2.2 Combining strengths, emotions and polarities

For web mining community twitter sentiment analysis has an increasing interest. People express their sentiment on many different topics with variable strength and intensities. For sentiment analysis a numbers of methods and lexical resources are available. Felipe, Marcelo and Barbara have proposed to combine opinion strength, emotion and polarity indicator to generate

improved sentimental analysis of polarity and subjectivity. Manual classification for opinion mining is an unfeasible effort at human scale [17], although several methods have been proposed for automatic opinion mining. Supervised classification use polarity estimation and for unsupervised approach lexicon resources are used widely. Emotions are generally expressed by JOY, SURPRISE, TRUST, ANGER, FEAR, SADNESS, DISGUST and ANTICIPATION. Polarity methods extract positive, negative and neutral information from a passage. Emotion methods extract any emotional expression or a person's mood from a text passage. Strength methods provide intensity levels depending on a certain sentiment dimension either in polarity or in an emotional format. Basically these methods return numerical value which demonstrates intensity from a text passage. Felipe et al. proposed to improve sentiment analysis by using supervised learning algorithms. They have considered two major classifications: Subjectivity and Polarity [17]. They have presented an approach with the combination of several lexical resources for sentiment classification. This novel approach validates (classifier's gains) a significant improvement with 5% better accuracy over single-methods. The proposed method is suitable for word-based representation (unigrams or n-grams). The feature does not directly depend on the vocabulary-size and reduce considerable dimensionality. In this representation, several learning algorithms found efficient low-dimensional features. Classification result is significantly dependent on datasets and output is variable. Felipe et al. mentioned that the manual-classification is a subjective task and result is dependent on evaluator's perceptions. That is why this fact is a warning call on any bold conclusions of sentiment classification based on inadequate evidence. The suitability checking for training-dataset is very important and should be measured beforehand. It should be checked whether the training examples are capable of capturing the sentiment diversity of specified domain.

2.2.3 Twitter Sentiment Analysis for Election Cycle

Hao et al. described a real-time system for analyzing public sentiment about U.S. presidential election 2012 using Twitter data [18]. Twitter is considered as a central site for expressing public opinions where people express their political views about candidates and different parties [18]. Almost instantly emerging news or events spread like explosion on Twitter. This explosion introduces the opportunity of sentiment analysis for public emotion about electoral forecast. This kind of research work can explore the effect of public opinion for

any targeted event. Traditional system takes a couple of days or weeks to generate a forecast for election which is expensive. But sentiment analysis research can deliver the result instantly and can produce prediction continuously. This result helps the public and media to think in a timely manner, even the politicians reschedule their activities based on the result. Hao et al. mentioned that past studies on social networks focusing on political sentiment was in a very limited scale. It was somehow post-hoc or dragged out on small sample or limited static samples. To resolve these issues, they had built sentiment model with a unique infrastructure which could analyze Twitter based public sentiment in real-time. They performed a successful experiment on the 2012 U.S. presidential Election. The effort was to scale political sentiment by bringing the social science along with advanced information technology. Finally by understanding traditional statistical sentiment modelling and mixing it up with real-time data processing their approach developed a new way of scaling the political practices. Their infrastructure was on IBM's Info Sphere Streams platform (IBM, 2012) and the standard Natural language Processing (NLP) practices have been used (text is tokenized used for later processing). The data processing infrastructure was in real-time and the statistical-sentiment-model was used to evaluate twitter public sentiment. This particular architecture was capable of using a generic model to unfolding the emerging political events.

2.2.4 Twitter Sentiment by Modelling Public Mood and Emotion

Johan et al. have observed that public mood has a significant and immediate effect on the social and political events. Public mood has various dimensions and it highly affects the cultural and economic events [19]. The authors assume that large scale mood-analysis might bring a solid platform of collective emotive trends model. This model might generate predictive value for social and economic indicators. They have proposed a system to explore how patterns of public mood relate to social and economic indicators. Two of the major components that have been used for the experiment are (a) macroscopic socio-cultural events and (b) public's mood derived from six-dimensional psychometric instrument. The objective of the authors' analysis was to identify a relationship between public mood and events (social, economic and other major), media and culture. The experimental process measured the sentiment of tweets by using the Profile of Mood States [19]. Afterwards the process compared the results with the events in

specified timeline. Johan et al. found a significant correlation among the events (social, political, cultural and economic) and public mood range with various dimensions. Sentiment analysis by syntactic method can bring efficient result without a training or machine learning system [19]. Machine-learning sentiment analysis generates more accurate results for classification only by processing significantly larger dataset (testing and training). But syntactic method faces great difficulties for this kind of approach. Explosion of public mood escalates the sentimental affect across the social networks. So dynamics of public sentiment highly affects the social network services.

2.2.5 Social Theories to Mine Social Media

Currently SNSs have an unquestionable popularity as it is growing day by day, which inspires more and more people to join online media activities [20]. As a result, it keeps producing higher volume of data day by day. These data is linked, noisy, incomplete and unstructured with a very high volume. That is why it differs from traditional data mining. As a result, it becomes a new research area as social media mining. Social theories are helpful for explaining the social phenomena from social sciences. But the scales of social theories are different than the properties of social media data [20]. It is a fundamental question whether the social theories can process social media network data. Advanced computational technology has new tools and techniques to verify large-scale social theories on social media. Jiliang et al. have reviewed various key-approaches of social theories for mining aspects in social media network to find answers to the raised questions. Many social theories are discovered from social sciences that explain social phenomena such as Homophily-theory and Balance theory [20].

Feature engineering (an automatic process to identify the input (X's) in machine learning process) may take a considerable research time if the social media data is not understood and/or organized properly. This process is successful for many realistic cases. The authors proposed that engagement of advanced computational skills and social science theory could mutually benefit to accelerate the social media mining tasks. Social co-relation theory, status theory and balance theory are the key social theories for data mining. Various methods and application of feature engineering, constraint generation and objective identification can be used for better categorization. Finally the combination of these tools can discover hidden social theories for advanced social media mining.

2.2.6 Sequential Frequent Pattern

The reoccurrences of an event in a series is known as Sequential pattern. Sequential pattern in data mining field is a well-known scenario. Sequential pattern normally occurs within a framework followed by some time series. Sequential pattern mining is a difficult problem as it has to be generated or examined combinatorial way on an explosive number of intermediate subsequences [21]. Jian et al. has proposed PrefixSpan approach where a growing sequential pattern is projected repeatedly on smaller sequence array. The benefit of this process is user-specified-constraints, which is used on projection-based methodology for sequential patterns mining. Pattern-growth approach can be extended for efficient frequent patterns identification as well. SPADE, GSP, Free-Span [21] are widely used algorithms for frequent pattern identification. Jian et al. have performed sequential patterns mining on a large database with a systematic matching and found that the PrefixSpan approach is efficient and scalable. They have performed comprehensive performance study and found that the approach reduced the number of physical database projections. This process takes less effort for candidate subsequence generation and it is the fastest among all the algorithms.

Efthymios K et al. [16] have worked on “Twitter Sentiment Analysis: the Good, the Bad and the OMG!” which is a similar to this work but they have used supervised approach which is different than this work. Hao Wang et al. [18] have worked on “A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle”, which is similar to this work. These works were helpful to develop some initial ideas of sentiment analysis but none was identical as the scope is location based sentiment analysis.

Chapter 3

Material and Methods

The SNS phenomenon has a remarkable impact on our daily life. Thom et al. have mentioned that the process of analyzing social blogging services message streams is a challenging task because of enormous document production per day [22]. The research objective of this thesis is to identify possible ways that can lead us to extract knowledge through the sentiment analysis of location based online social networks. When looking at patterns in large datasets, a popular choice is the utilization of Twitter [23]. Geo-tagged Twitter messages have been used as our primary data source but the data (that our application can analyze) is not limited to geo-tagged tweet only. Ontology-based query can be passed through the developed application. Then the developed application can search by tweet content, hash tags, username or keywords. Single/multiple focuses can be set on Google map to extract data from that area in a real-time manner. The tweet data can be stored in a database and can be compared the time frame with the real-time data. A geographical presentation can also take place on the Google Map for geo-tagged tweets, which can allow the user to have a clear understanding of the tweet's location. Our interactive GUI displays the tweets, its user and URLs (if there are any).

3.1 System Architecture

This section will explain the architecture of our system along with the main components. The whole process is divided into four sub modules. The process begins with the data mining application, score generation, data visualization and finally data storing modules. There will be a discussion on each component with a detailed reasoning and set of algorithms as well as why that method has been used. Fig. 2 presents the system architecture of the proposed project.

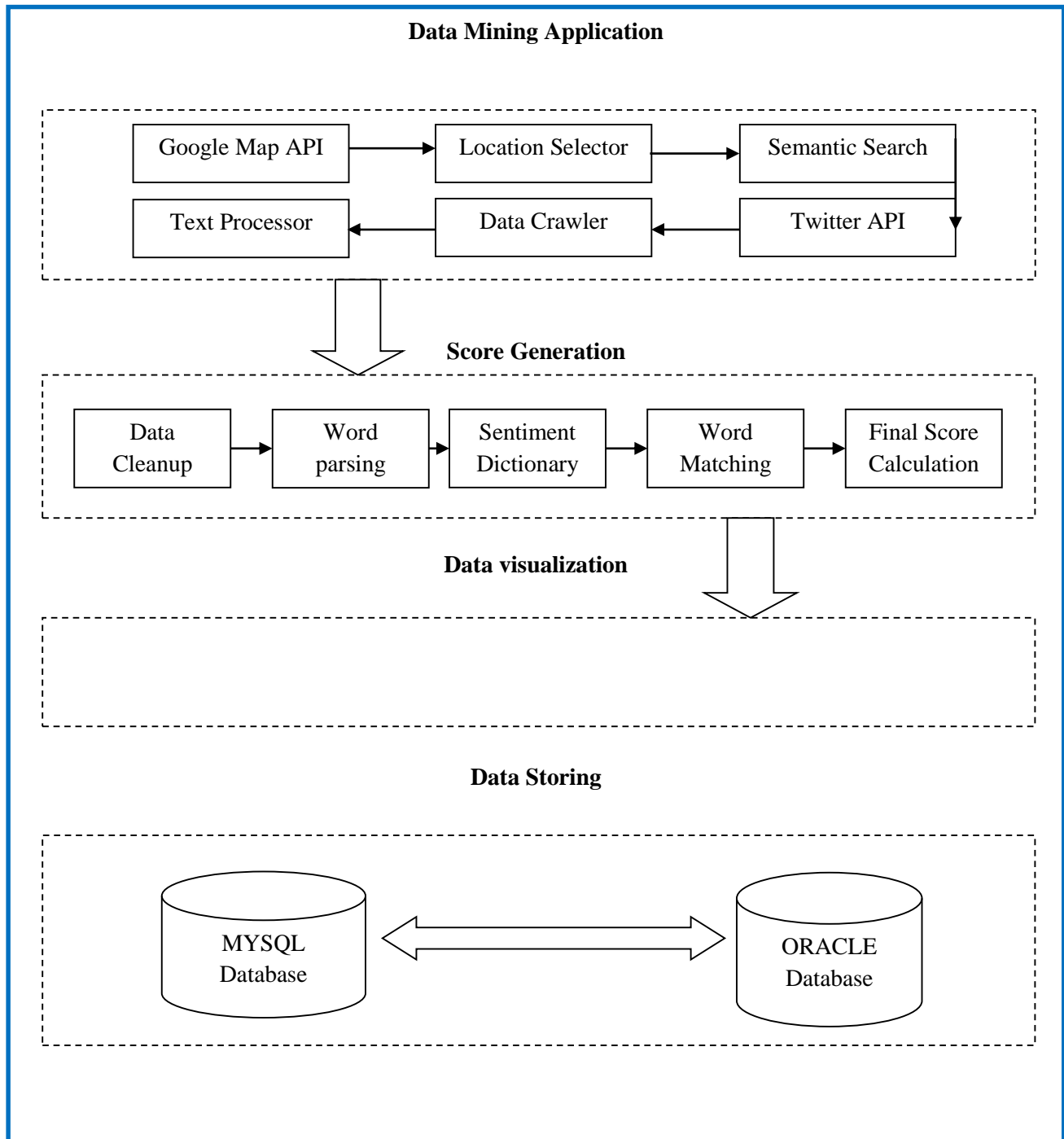


Fig 2: System Architecture

3.2 Google Maps API

Google Maps is a web mapping service that offers satellite images, street view and maps through Google Maps API, released in June 2005. This API allows users to embed robust functionality to explore the world. It also allows users to identify map locations with custom markers. Google Maps API is a free service for private or commercial use. The proposed system has integrated Google API for location selection, data interaction, and visualization. Once the application is run, the very first interface will be with Google Map. If we know the location we can search tweets from the region. Even if we do not know the location we can simply search tweets and identify the targeted location by density of the markers. Fig.3 shows our system interface with Google Maps API.

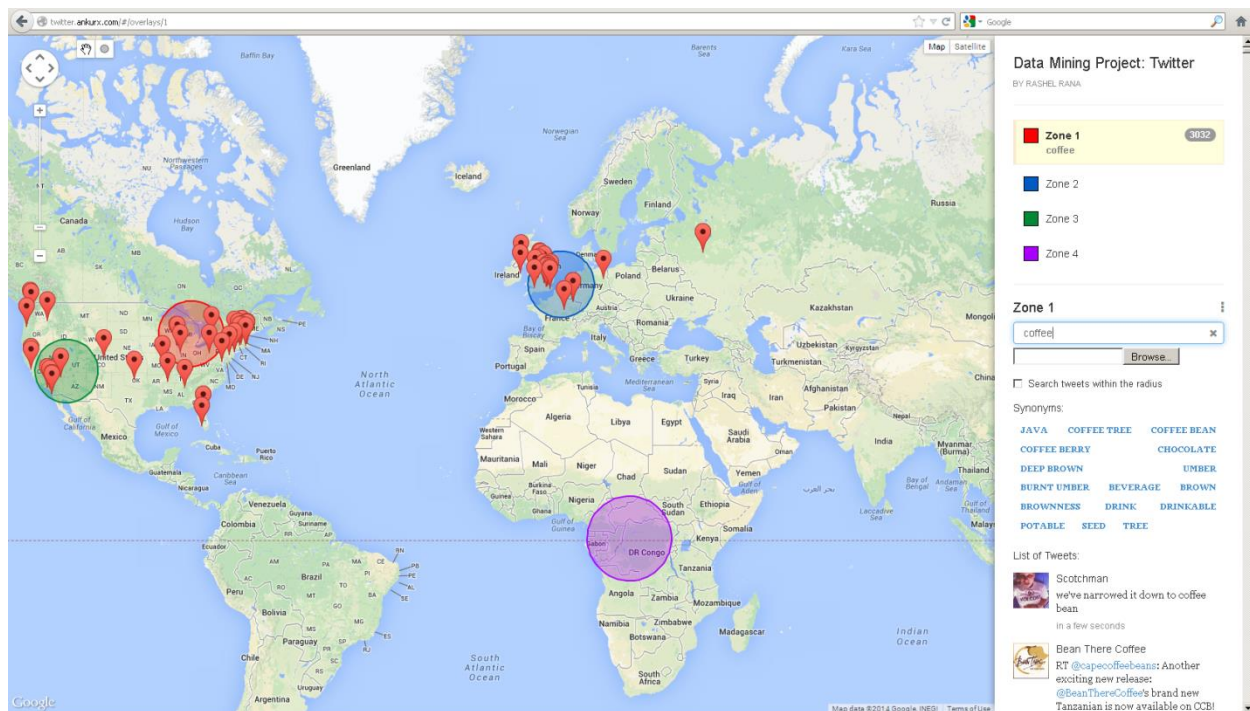


Fig 3: System interface with Google Maps API, Location Selection

3.3 Location Selection

The second step of our application is location selection. Though the application can extract location independent twitter data, it is more interested in location-based data. It is needed to draw single /multiple locations on Google map API to perform location based search. Once the location is selected on Google Maps, it essentially identifies longitude, latitude and diameter,

which is 50km by default. Four different location-circles has been selected on Google map in Fig. 3. Tweet location's system has some limitations as they are not always correct (as user may enter wrong location information while tweeting). Algorithm 1 describes the process of location-based searching.

Algorithm 1: Location Based Searching

Input: Tweet Location/s (longitude, latitude, diameter)

Output: Tweets in that location

1. Begin
2. v-Location $L(x)$ = Point New Location
/*Draw a new location on Google Map and put in location variable*/
3. v-Longitude $LN(x)$ = $LN(x)$
4. v-Latitude $LA(y)$ = $LA(y)$
5. v-Diameter = 50KM;
/*default initial value*/
6. For each keyword in $LN(x)$, $LA(y)$, v-Diameter
7. {
8. Search Tweet in that Location
9. }
10. Return Tweets
11. End;

3.4 Semantic Search


In the third phase of our application, users can search tweets either by keyword, hash-tag (#) or from text file. After performing the search it will generate data visualization instantly on the map with markers. For an experiment, the keyword “COFFEE” was searched and instantly the application pointed the markers of the locations of the users talking about coffee around the globe along with their tweet. Now COFFEE is related with *coffee tree, coffee bean, coffee berry, beverage, drink, drinkable, chocolate, deep brown, umber, burnt umber, brown, brownness, potable, seed, and tree*. Once the search is conducted with the keyword, the word and related ontology will generate the query. Finally it will pull up related tweet information and visualize on the globe using Google Map. For ontology keywords this thesis have used WordNet v3.1 search. Fig. 4 shows the tags on the Google Map. Algorithm 2 describes the process of semantic search.

Data Mining Project: Twitter

BY RASHEL RANA

 **Zone 1** 3032

 Zone 2

 Zone 3

 Zone 4

Zone 1



☐ Search tweets within the radius

Synonyms:

JAVA COFFEE TREE COFFEE BEAN
COFFEE BERRY CHOCOLATE
DEEP BROWN UMBER
BURNT UMBER BEVERAGE BROWN
BROWNNESS DRINK DRINKABLE
POTABLE SEED TREE

List of Tweets:



Scotchman
we've narrowed it down to coffee
bean
in a few seconds



Bean There Coffee
RT @capecoffeebeans: Another
exciting new release:
@BeanThereCoffee's brand new
Tanzanian is now available on OCB!
<http://t.co/Q6z02InoOB> h...

Fig.4 searching a keyword (Coffee) along with Ontology

Algorithm 2: Semantic Search

Input: $K = \{k_1, k_2, k_3, \dots, k_n\}$ Where k_i is the user delivered search keyword

Output: A set of Ontology Keywords Matrix

1. Begin New Search
2. VAR-Matrix = Build empty matrix[]
3. Set Keyword K(x) [N] = New K(x)[n]
4. For each keyword in K(x)
5. {
6. VAR-Ontology=Call WordNet API (K_(x))
7. /* Search Ontology for the K(x)[n] into variable VAR-Ontology */
- VAR-Matrix = VAR-Matrix + VAR-Ontology;
8. }
9. Return VAR-Matrix
10. End Search;

The visualization includes the tweet along with link of the user (URL). This application allows non-geo-tagged tweets by unmarking “*Show tweets with location only*” checkbox (Fig.5). Now all the tweets can be extracted and analyzed to identify trends, likenesses, opinions, effects etc. The easiest way to visualize the trend is to use tag-cloud/info-graphics which allows to understand better insight of the text data.

Extracting the user statement about a particular topic is not an easy job. Keyword or Ontology searches may result in too much irrelevant information that can be analyzed and filtered to extract more sophisticated results. Let us consider the extraction of user opinion about 'Product X' - do they like it, or not? Users may express the same thing in different ways. i.e. “*I like this product*”, “*I love the product*” or “*I wish to have this product*”. To manage these expressions sentiment analysis is applied to the related products. Semantic analysis on the Twitter data can generate polarity by determining the positive and negative feedback on the given product or topic.

To focus particular areas on the Globe, simply the areas are pointed on Google map and the application is guided to extract information from that specific area. Our system gets location data in two ways. One is from geo-tagged location which is typically longitude and latitude. Another is from users profile location data.

The application converts the location into longitude and latitude coordinates and match it with the selected area. For example if a user location is “SAN_FRANCISCO”, the application will convert it into (-122.75, 36.8, -121.75, 37.8). Finally it is translated and used as functional parameter. Fig. 5 shows tweets in a selected focal area.

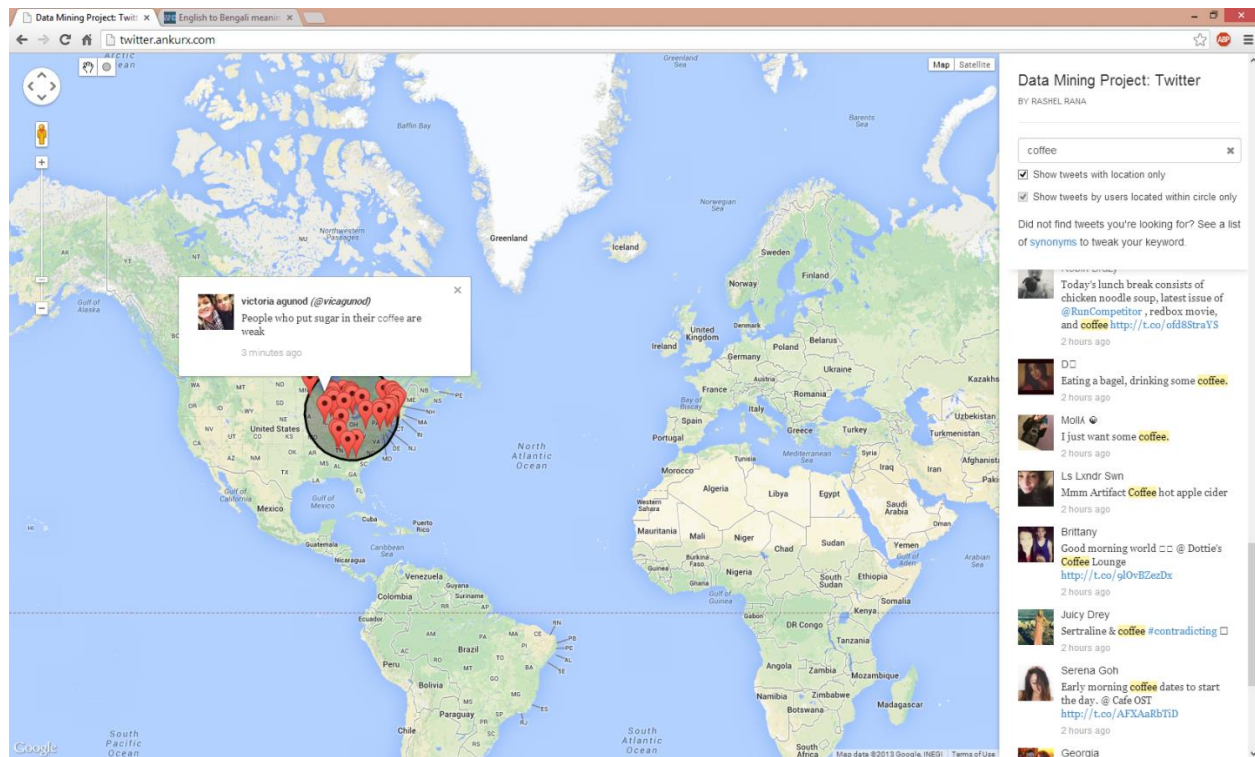


Fig. 5 Twitter location on focal area on Google map

Depending on the result, this application can focus on multiple sites for a timeframe to discover data analysis in multiple ways. For example, if a marketing theme is considered to market a particular product/products, the system can generate an idea of the public reaction after analyzing the Twitter data related to that product/products. Focusing on multiple points generates a contextual feedback of situation comparison. As a result the system can identify the positive and the negative poles of feedback in the targeted tweet-conversation. Once the information about success or defects are known, necessary steps may be taken for the product market line.

Let us consider the goal to promote the iPhone in Europe and America. Now assume some advertisement or promotion has been done on the targeted market. Now the system wants to show, which area has more promotional effect. Assume the observation found that the American zone has less response (promotional) than the European zone. But the expectation (targeted business) was higher for the American market. Now the inspection has to go into further details for the information retrieval. Afterwards, the captured historical information about the zones are compared with recent result. Then necessary steps can be taken in regarding to the

product promotion and how the market can boost up the business. Fig. 6 Multiple Zone (area) shows multiple focus selection on the Google Map.

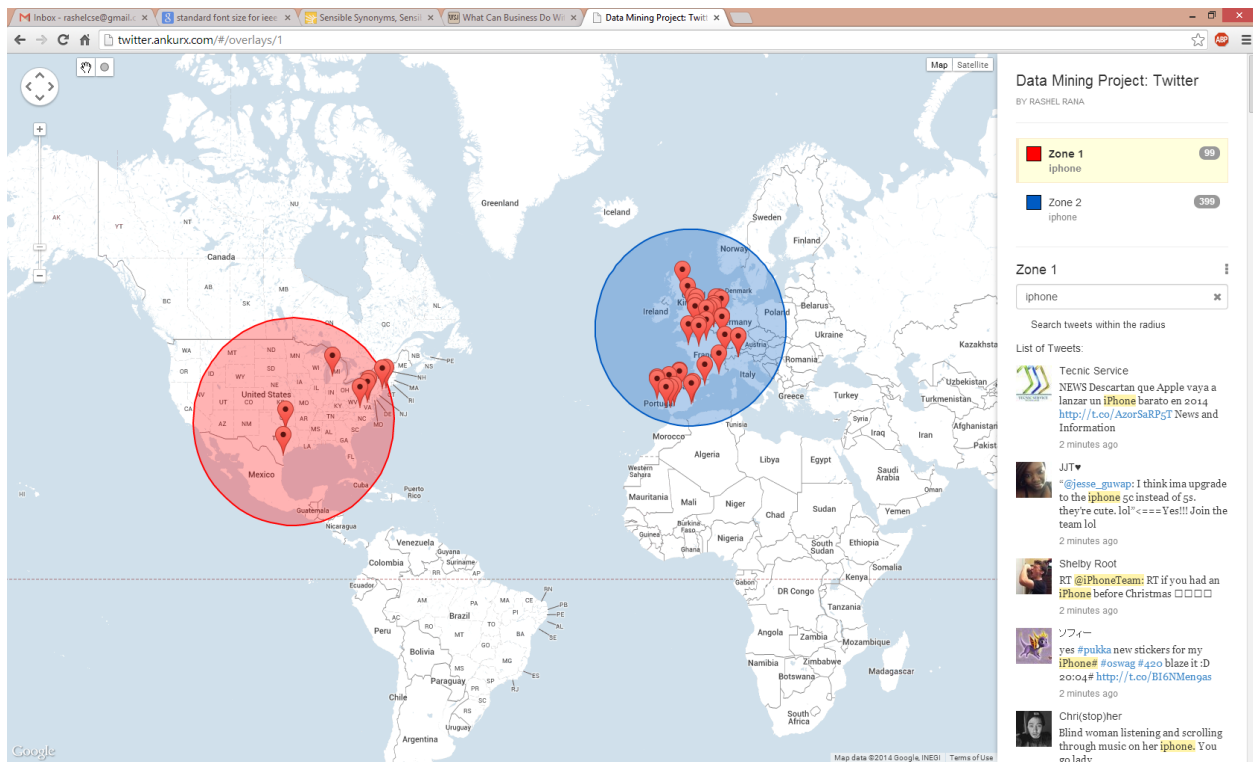


Fig. 6 Multiple Zone Selection on the Globe

This is one of the aspects of many other business perspectives. The location selection and focusing process are dynamic and scalable to a small extent. The application can be utilized in the same way to simulate the spread of Flu Virus in the European zone and the American zone. Other popular experimental topics such as election prediction, movie promotions, stock market movements, natural disasters etc. can also be simulated, visualized and analyzed. An external text file (comma separated keywords) can be fed to the application which may serve customized searches.

3.5 Twitter API

Twitter allows users to interact with its data using Twitter APIs. Twitter has three API (Rest, Search and Streaming). “The Rest API provides simple interfaces for most Twitter functionality. The Twitter Search API is part of Twitter's v1.1 REST API. It allows queries

against the indices of recent or popular Tweets and behaves similarity to, but not exactly like the Search feature available in Twitter mobile or web clients, such as Twitter.com search. Before getting involved, it's important to know that the Search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results. If you want to match for completeness you should consider using a Streaming API instead. The Streaming API is a family of powerful real-time APIs for Tweets and other social events” [24].

For efficient data extraction server side programming/scripting language such as PHP, RUBY or PYTHON is needed. An Oath-authentication is required to access data from twitter. Four authentication parameters (*consumer key*, *consumer secret*, *Oauth_token*, *Oauth_secret*) are needed to enable API call. These keys look like following:

<i>Consumer key</i>	<i>xvzIevFS4wEEPTGEFPHBog</i>
<i>Consumer secret</i>	<i>L8qq9PZyRg6ieKGEKhZolGC0vJWLw8iEJ88DRdyOg</i>
<i>OAUTH_TOKEN</i>	<i>5501815334-YG8YkYD8tjFLrT3zw61GmhT1uS8iqvzQxojsu49</i>
<i>OAUTH_SECRET</i>	<i>bfK5P3kZVsfOwgNrM0Gjab56BhLDRiYgMhGRAzxIRry5b</i>

Each application has different Consumer-key, Consumer-secret, Oauth_token and Oauth_secret. Once request is sent to twitter API with proper authentication it will deliver results in JSON (JavaScript Object Notation) format. JSON format is very popular and can easily be read by any program.

3.6 Data Crawler

Data Crawler is one of the most important part of the application. It collects related tweet data from API for our proposed system. It can simply be considered as a raw material collector and deliverer. This data will be used for the experiments which will generate score to explore product's (or any given parameter) popularity. It defines and manages the parameter to be passed to Twitter API and retrieve the related data from Twitter database. The crawler needs location, keywords, or parameter file and ontology as an input parameter to call Twitter API. The crawler will receive data and deliver it to text processor, which will be stored in MYSQL database. The crawler can run in single/multiple machine in single session or multiple sessions simultaneously. As the crawler works online, the system can run in fully online distributed mode. Online data storage helps to store and process data very easily with minimal setup. Twitter does not allow

direct access to its database for any external application. So the proposed system cannot reside on Twitter database. This is the main reason to save data in a personal database. This will ensure data availability and usage offline. Algorithm 3 describes the steps of Data-Crawler.

Algorithm 3: Data Crawler

Input: Location/Locations and Keywords/Keywords

Output: Related Tweet for keywords in any targeted location

1. Begin Process
2. VAR-LocalBuffer = \emptyset
 /* initialize the variable buffer with zero */
 /* Ontology is similarity or similar words */
3. VAR-Crawler = Load matrix (search Keywords $K[x]$ (N) with Ontology)
 /* initialize data crawling process with the Search tweet-keyword with its Ontology */
4. CreateSearchKeyList $K[x]$ (VAR-Crawler)
5. VAR-Queue[n] = LoadKeyList ()
6. While (VAR-Queue[n] is not null)
7. {
8. Call Twitter API (Consumer key, consumer secret, Oauth_token, Oauth_secret,
 Queue [n])
9. VAR-LocalBuffer = Returned result from Twitter API
 /*store document for later processing*/
10. If tweets Contains (VAR-Queue[n]) then
11. {
12. Send tweets to Text Processor
13. } End If
14. }End While
15. End process

3.7 Text Processor

Once the data crawler has completed its job the Text Processor module will be triggered. The text processor will examine the data and decide what actions are needed to perform on the data and how to store and manage it. Then Text processor will send the data to the Score Generation Unit. First step here is to clean up the data to get rid of incorrect, inaccurate, incomplete, irrelevant or duplicate data. The process is performed in several steps. Very first step is to remove white spaces, special characters, URL link, numerical digits etc. Algorithm 4 describes steps for Text Processor.

Algorithm 4: Text Processor

Input: Tweet from Data Crawler

Output: Partially processed string

1. Begin
2. VAR-LocalBuffer L[n] = Load Tweets ()
3. While (VAR-LocalBuffer is not NULL)
4. {
 Trim [VAR-LocalBuffer] 20 character and CheckForDuplicate
 If FoundDuplicate
 Then Remove the Tweets
 Else Store Tweets
}
5. End

3.8 Data Cleaning

In general user tweets have no structure and are mostly written in informal language with many short term words and spelling mistakes. It is very difficult to categorize the variations within tweets by specific domain. Twitter sentiment analysis is a difficult job. Fig. 7 presents a sample Tweet with ID, Screen name, user name, tweet text, and tweet's location by longitude and latitude coordinates with generation time.

ID:	462968139349647360,
Screen Name:	"リザードン▶Ⓢ",
User name:	"THECharizardBoy",
Text:	"RT @nationalpost: Toronto city councillor Adam Vaughan wins Liberal nomination in Ⓢ — Olivia Chow's old riding http://t.co/33tJyXLRKh ",
Longitude:	"43.64842391",
Latitude:	"-79.37536382",
Date:	"Sun May 04 14:53:00 +0000 2014"

Fig. 7 Sample Tweet about election candidate Olivia Chow

Text processor will deliver noisy data to the data cleaning process module. Data cleaning process is performed exclusively on tweet text. Data cleaning process involves the following six steps: First any web link or web references are eliminated. Second any existing whitespaces are removed. Third any special symbol or characters are removed. Fourth the punctuations are removed. These include “*Full Stop/Period - Comma - Semi-colon - Hyphen - Dash - Apostrophe - Question Mark - Exclamation Mark - Slash - Backslash - Quotation Marks - Underline - Underscore - Round Brackets - Square Brackets - Ellipsis - Punctuation Song*”. Fifth step involve the removal of hash (#) symbol. In the final step all the numerical digits are removed. Finally the data cleaning process will rebuilt the string and will deliver a clean output string.

Algorithm 5: Data cleaning

Input: Partially processed Tweet array from Text Processor

Output: Cleaned Tweet string

1. Begin Cleaning
2. VAR-TextFile[n] = PartiallyCleanedTweetArray[n]
3. VAR-FinalTextFile[n]=NULL;
4. {
5. For each value in VAR-TextFile [n]
6. {
- /*Remove http:// header from source string*/


```

7.      VAR-TextFile[n] =Decode (substitution ((VAR-TextFile[n]) 'http ://')
          /*Remove white space*/
8.      VAR-TextFile[n] =Remove whitespace (VAR-TextFile[n])
          /* Remove Special character, smile like '@%!ÃÂçâ, °Å, €~Å'–â,,ç?' */
9.      VAR-TextFile[n] =
          RemoveSpecialCharecter (Translate replace (VAR-TextFile[n]))
          /*Remove punctuation. !, ;:- ({}))|\ ' "" --- that already exists in the string
          anonymously*/
10.     VAR-TextFile[n] =Remove punctuation (VAR-TextFile[n])
          /*Remove the Hash Symbol*/
11.     VAR-TextFile[n] =RemoveHash # (VAR-TextFile[n])
          /* Remove '#1234567890' these character and replace with ';' as separator*/
12.     VAR-TextFile[n] =Translate (VAR-TextFile[n])
          /* Now remove the separator',' that was added at the last position of 'translated
          Value'*/
13.     VAR-TextFile[n] =Reverse (VAR-TextFile[n])
14.     VAR-TextFile[n] =CutFirstChar(VAR-TextFile[n])
15.     VAR-TextFile[n] =Reverse (VAR-TextFile[n])
16.     }
17. BuildArray VAR-FinalTextFile[VAR-TextFile[n]]
18. Return VAR-FinalTextFile
19. End Cleaning

```

3.9 Word Parsing

Word parsing is a process to split a sentence or text into parts by analyzing logical syntactic components followed by the formal grammatical rules of the English Language. Lexical based text analysis is very essential for natural language or computer language processing. Computational linguistics is slightly different than the regular grammatical parsing. Traditional grammatical sentence parsing usually emphasizes on subject and predicate depending on the exact meaning of a sentence. But a computational word parsing transforms a sentence or text into its constituents resulting in a parse tree of syntactic relation.

Text: "RT-@nationalpost: Toronto city **councillor** Adam Vaughan wins Liberal nomination in ~~Q~~ ~~z~~ — Olivia Chow's old riding <http://t.co/33tJyXLRKh>",

This may contain semantic information as well. In our application this process will take a user tweet as an input and will generate a parse tree as an output. This output is transmitted to the word matching function to generate the tweet words weighted value (score) from sentiment dictionary. Oracle SQL built in function *REGEXP_INSTR* has been used for building the tree array. *REGEXP_INSTR* is an extension of the *INSTR* which allows searching a user defined expression pattern [25]. Fig.8 shows working pattern of the *REGEXP_INSTR* function.

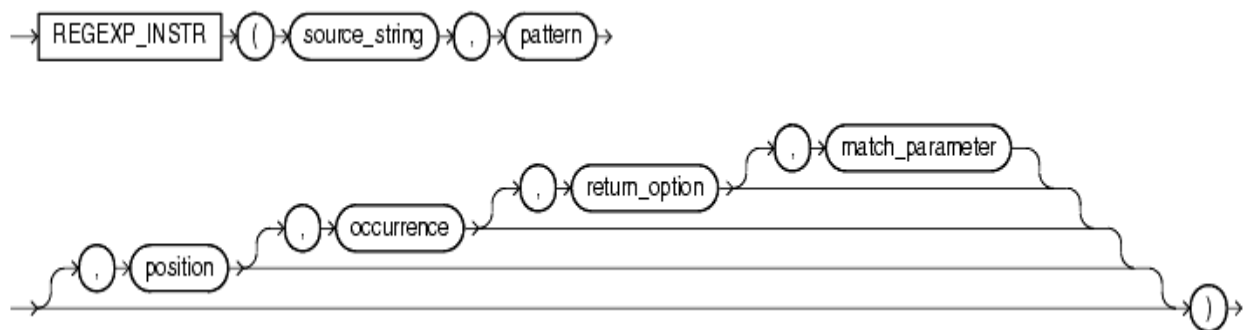


Fig. 8 REGEXP_INSTR function working pattern [25]

Algorithm 6: Word parsing

Input: Cleaned Tweet String

Output: Word Array

1. Function WORD-PARSE (Tweet String, words)
2. Begin WordParsing
3. Declare Variables
4. VAR-Word-Status Integer
5. TMP-WRD varchar(300)
6. New-Words Varchar(300)
7. X_Cur Cursor
8. Fetch TweetString FROM the temporary_table into X_Cur
9. OPEN X_Cur Cursor to Read the Tweet string
10. SET Beginning-Status = 0
11. While Beginning-Status = 0

```

12.      {      Begin Decomposition (X_Cur[n])
          /*set a counter and put words into temporary buffer and increase it*/
13.      For i=1 to N
14.      {      TMP-WRD = TMP-WRD + Whitespace      }
15.      }
          /* for each character-index of whitespaces inside the tweet string*/
16.      For i=1 to N {
17.      Set TMP-WRD = Sub-String (TMP-WRD, ' ', (TMP-WRD+1),
                                   Length of TMP-WRD)
18.      Insert TMP-WRD into New-Words
19.      }End Loop
20. Close the Cursor
21. Return New-Words
22. End Function

```

The word parsing process will generate a word array similar to following result presented in Fig.9.

toronto
city
councillor
adam
vaughan
wins
liberal
nomination
in
olivia
chows
old
riding

Fig.9 Word Array Extraction from Tweet

3.10 Sentiment Dictionary

Sentiment analysis is basically a classification of the polarity for sentences. Classification is a process of separating or filtering different groups or things into classes. It can be done by pattern matching or applying functions for closest match. Sentiment classification can be done mainly in two ways:

- (i) Unsupervised Method: Lexicon based Semantic Classification of text
- (ii) Supervised Method : Principally Machine learning algorithms which are
 - (a) Support Vector Machines (SVM)
 - (b) Naive Bayes
 - (c) Markov Blanket Classifier
 - (d) Maximum Entropy Classifier

This thesis is using Lexicon based Semantic Classification method (unsupervised method). Machine learning approach can sometime generate more accurate results for specific domains. Semantic classification process assigns the extracted sentiment words found from sentences by matching corresponding words and polarity calculations. This work has developed a lexicon dictionary by using SentiWord.Net.

Domain independent very large features are really great content of SentiWord.Net. Semantic classification approach is not dependent on prior trainings. This dictionary is user friendly, customizable and constructed with sentiment and objectivity. This is why it can generate better results for any domain. The advantages of Twitter text analysis is that it is short and easy to clean and classify. Fig 10 displays word-array and sample lexicon dictionary containing Lexicon, Positive value and Negative value.

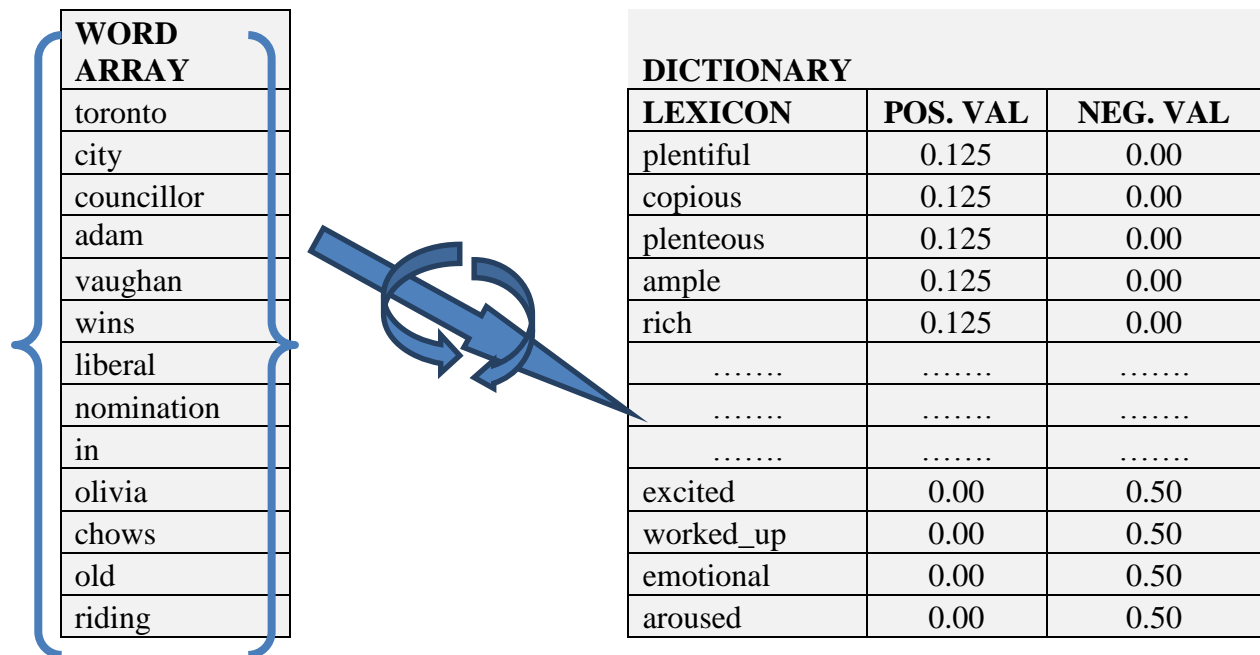


Fig 10. Developed Word Array and Lexicon Dictionary

3.11 Word Matching and Final Score Calculation

After the generation of cleaned tweet text content and word array the system will move ahead for word matching function. Each word of the array will be compared with lexicon dictionary through loop. If there is a match, the value will be returned for the word otherwise it will return zero. For each word, value is stored in local buffer and finally a summarized score for a tweet text is returned as an output. So the final score for the tweet in fig. 9 is $(0+0+0+0+0+0.25+.625+0+0+0+0+0.5+0) = +1.375$. Hence the result is a positive pole tweet. This way we calculate every individual tweet on particular topic and display it in different graph and chart format. Algorithm 7 presents different steps involved in the score calculation process.

Algorithm 7: Score Generation

Input: Input Cleaned Twitter Message String

Output: Tweet Score

1. Begin Score Generation
2. Variable $T[\text{array}] = 0$, $T\text{-word}[\text{array}] = 0$, $T\text{-count} = 0$, $T\text{-word}[x]=0$, $T\text{-score}=0$, $T\text{-Final-Score}=0$
3. Function Calculate-Match ()

```

4. {
5.     T[array] = Get tweet string of users from the dataset
6.     T-word [array] = Parse words from the tweet string
7.     T-count =Count number of words in the tweet string
8.     For (int x = 0; x <= T-count; x++)
9.     {
10.         T-word[x] = Match X with dictionary
11.         Return word score into T-score
12.     }
13.     T-Final-Score = Sum (T-score)
14.     Return T-Final-Score
15. }
16. End Score Generation

```

Fig. 11 shows how the final score calculation is calculated using word array and lexicon dictionary

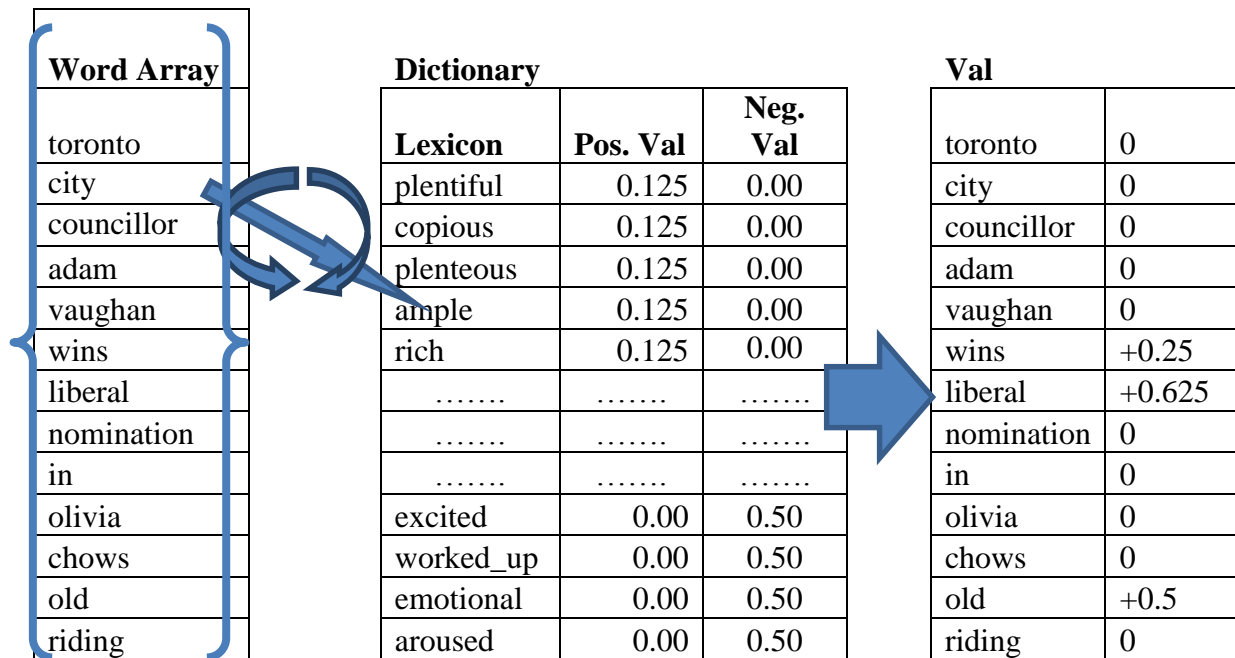


Fig 11. Score Calculation using Word Array and Lexicon Dictionary

3.12 Data Visualization:

#TOpoli is a very popular hash-tag which has come to attention in 2010 during Toronto mayoral election. City Hall, city photographs, political affairs, and election controversy have been broadly discussed using this has-tag. Since it surfaced, it became most popular online platform for political discussion [26]. This research has used tag-cloud to get better insights about the gathered tweets. URL has been used <http://www.wordle.net/> to generate the tag cloud and Fig.12 represents tag-cloud about #TOpoli from Twitter data for Oct 03 2014.

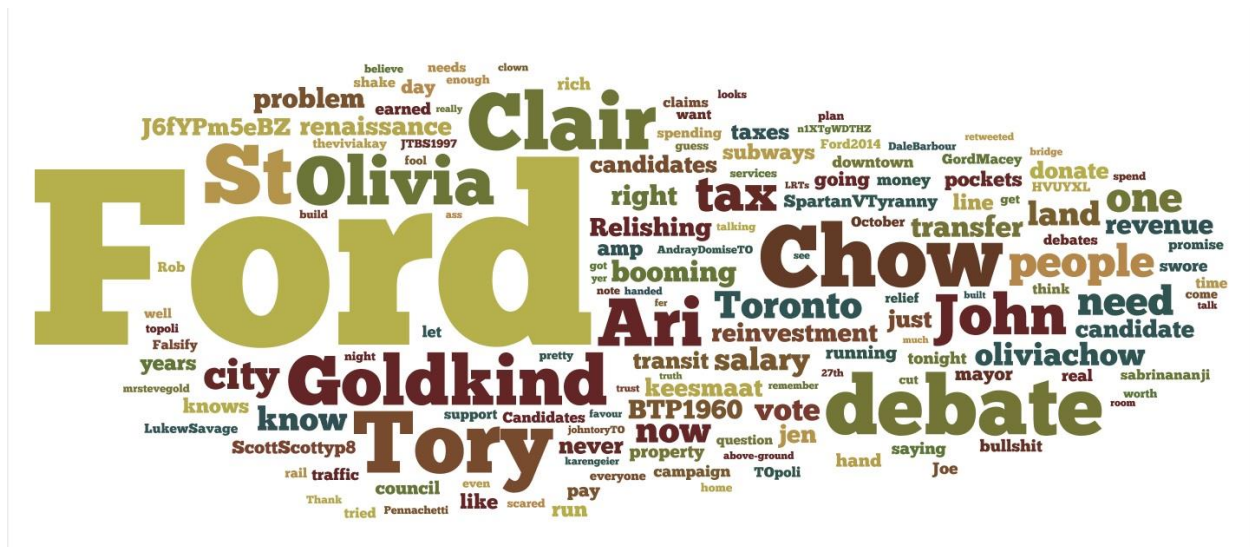


Fig. 12 Tag Cloud for #TOpoli related Tweet.

Wordle.net counts all the words from input text and rank it. Then it analyzes the ranking and then graphically presents them as their count size. Larger the count, greater the size is. In the Fig. 12, “Ford” is the biggest word and it means people are talking more (good or bad) about him. It simply generates an idea of the internal contents ranking through graphical visualization.

Chapter 4

Experiments and Results

This chapter discusses more details of the effectiveness and performance of the proposed system. The discussion begins with experiment setup and design, and follows on to cover data collection and implementation. Relevant results are also discussed and presented for a better understanding.

The user interface of the data crawler was designed and developed with java script. PHP was used to link between Google API and Twitter API. MS SQL Server and Oracle Database were allowed to run in the background to store, manage and process different factors. Data cleaning and text processing was done by Oracle stored procedures and functions. R software was used for relevant histogram drawings and graphical calculations.

Disclaimer: Twitter data (crawled by the application for this research) is only for the purpose of this research and not distributable to the public.

4.1 Experiment Setup and Design

This research is performing the experiment to measure the popularity of our proposed system and the data collection spanned over couple of months. The whole system was divided in multiple modules and was implemented in different time frames. Twitter does not allow accessing its database directly; we had to collect the data with a data crawler and stored it in a local database. So the beginning of our research implementation, the very first module coded was a data crawler which was scheduled to run to collect the data set on given individuals. Since data collection was a time consuming process other modules were developed as data crawler was working on the background. Once we had adequate data, the system was online to produce final result.

4.2 Experiment sets for System Scaling

It is already mentioned that the system is using Twitter as a data source. In the Chapter1 this thesis has discussed the scaling method. Popularity scaling can be performed in several

ways. This research has used the term “PRODUCT”, as a user variable and it can be replaced with any user defined parameter. This work have performed the experiments in the following stepwise categories:

- 4.2.1 Single product in Single Location
- 4.2.2 Single product in Multiple Location (iPhone in Toronto & New York)
- 4.2.3 Multiple product in Single Location (Election Data)
- 4.2.4 Multiple product in Multiple Location

4.2.1. Single product in Single Location.

Experimental Environment

Single product: **iPhone**

Single location: **Toronto.**

The primary scope of the research is to scale the popularity of iPhone in Toronto area (Geolocation = “Toronto”). For this process, system has collected raw iPhone data over a period of three months (6th March 2014 to 10th May 2014) from Toronto by developing a data crawler. After processing the data we have generated a line graph for the selected tweets. For better understanding this working is presenting the scenario from first 100 Tweets (starting from the latest date) in the Fig. 13. In all of the experiments (of section 4.2.4) tweet number is a sequence number (organized from newest to oldest according to tweet posting timeline); this number is generated by the data extraction process (tool: developed data crawler) and these sequence number cannot be reshuffled. In this graph Number of tweets (100) is presented in X axis and popularity score is presented on Y axis. Popularity score is the sentimental weight of each tweet and for 100 tweets.

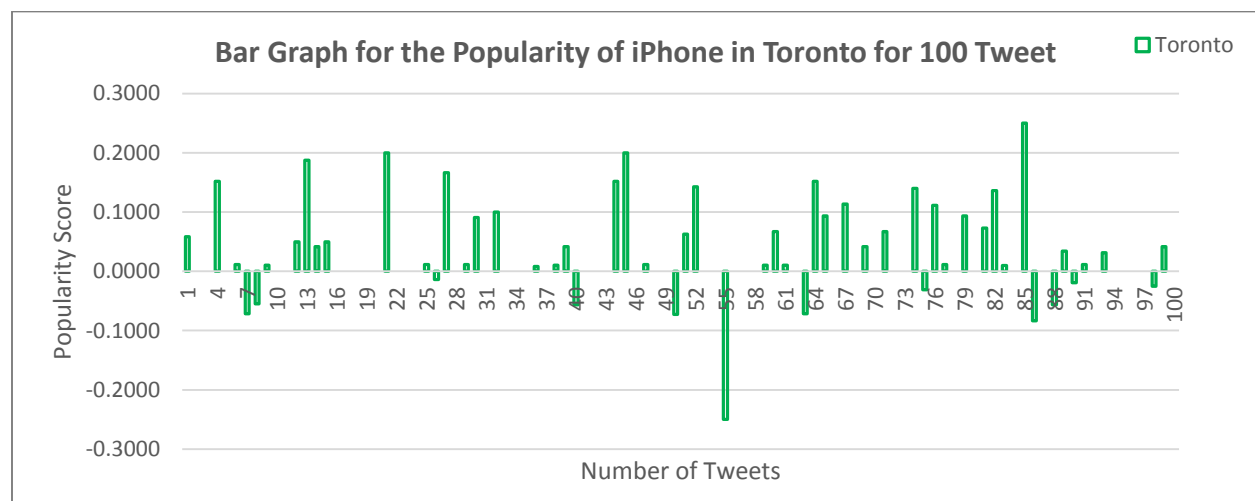


Fig. 13 Line Graph for the Popularity of iPhone in Toronto for 100 Tweet

If the same data is presented as a cumulative line graph the system generates following graph as shown in Fig. 14.

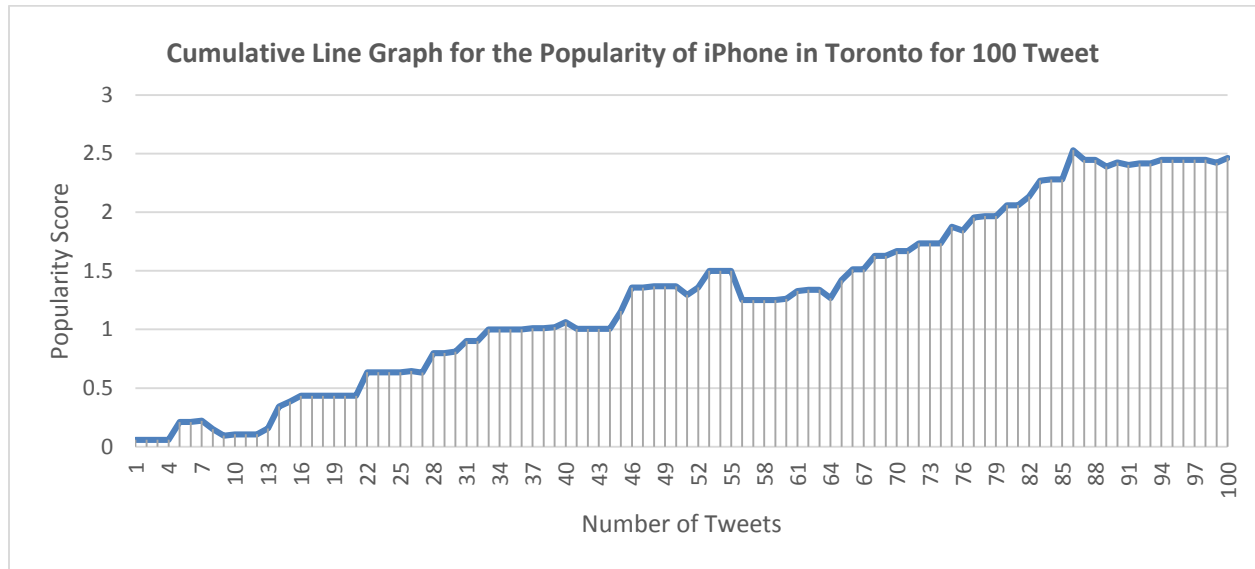


Fig.14 Cumulative Line Graph for the Popularity of iPhone in Toronto based on 100 Tweet

In figure 15 this work is presenting both a bar graph and a pie chart for the popularity of iPhone from 100 tweets within Toronto. Among the tweets the system found 43% positive sentiment, 13% negative sentiment and 44% neutral/zero score.

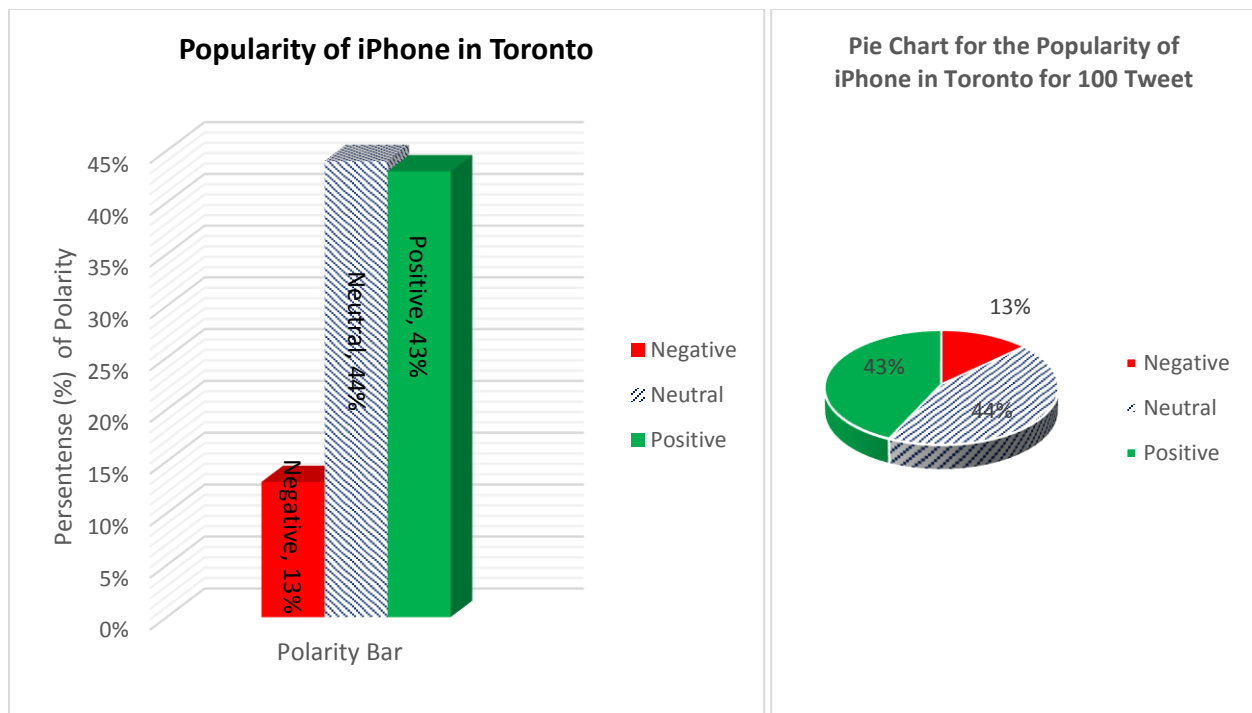


Fig.15 Bar Chart and Pie Chart based on 100 Tweet about iPhone popularity in Toronto

Afterwards this research work has conducted the experiment under the same conditions with 10,000 Tweets (Data collected 6th March 2014 to 10th May 2014) and generated the following data in fig 16.

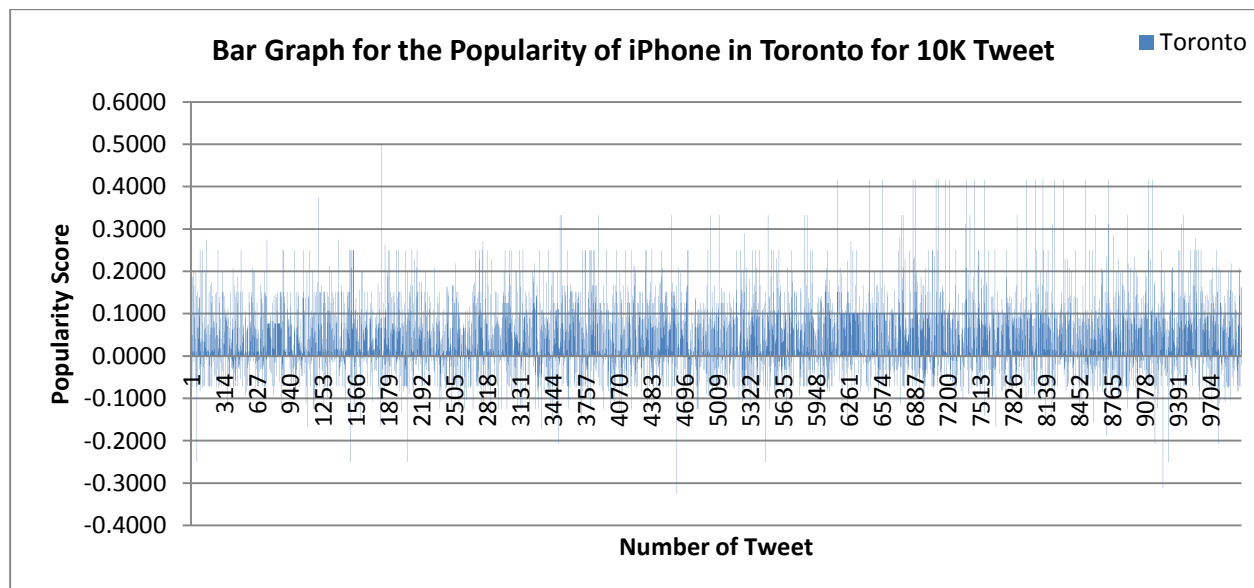


Fig.16 Line Graph for the Popularity of iPhone in Toronto from 10K Tweet

If the data is presented as a cumulative line graph the system generates following graph as Fig 17.

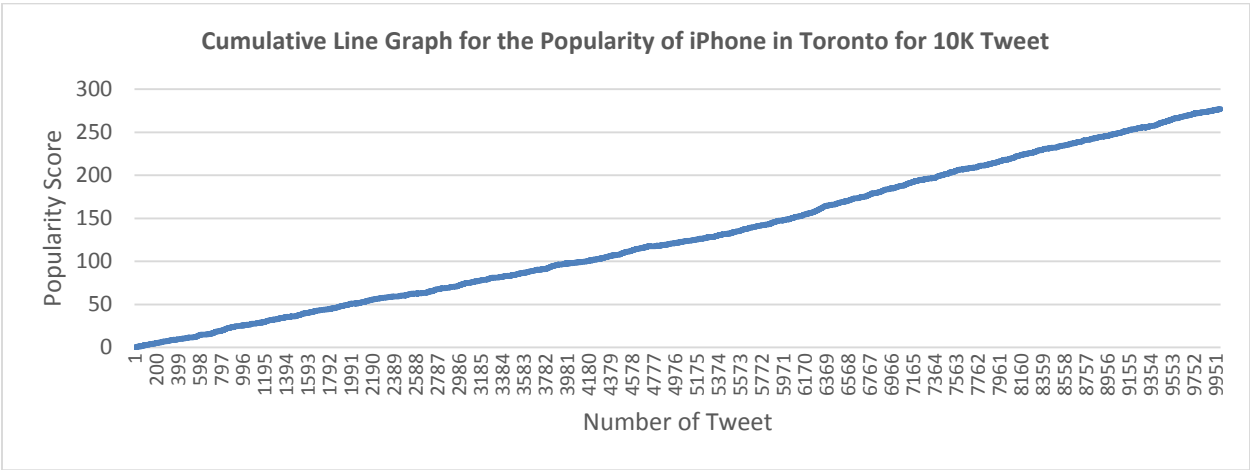


Fig.17 Cumulative Line Graph for the Popularity of iPhone in Toronto based on 100 Tweet

Now the thesis work aims to do polarity analysis on collected tweets. From 10,000 tweets, the thesis work discovered 4555 positive, 5445 neutral and 1516 negative responses which have been presented in the Fig 18 Bar Chart. The Pie Chart represents 40% positive, 47% tweet neutral and 13% tweet negative tweets.

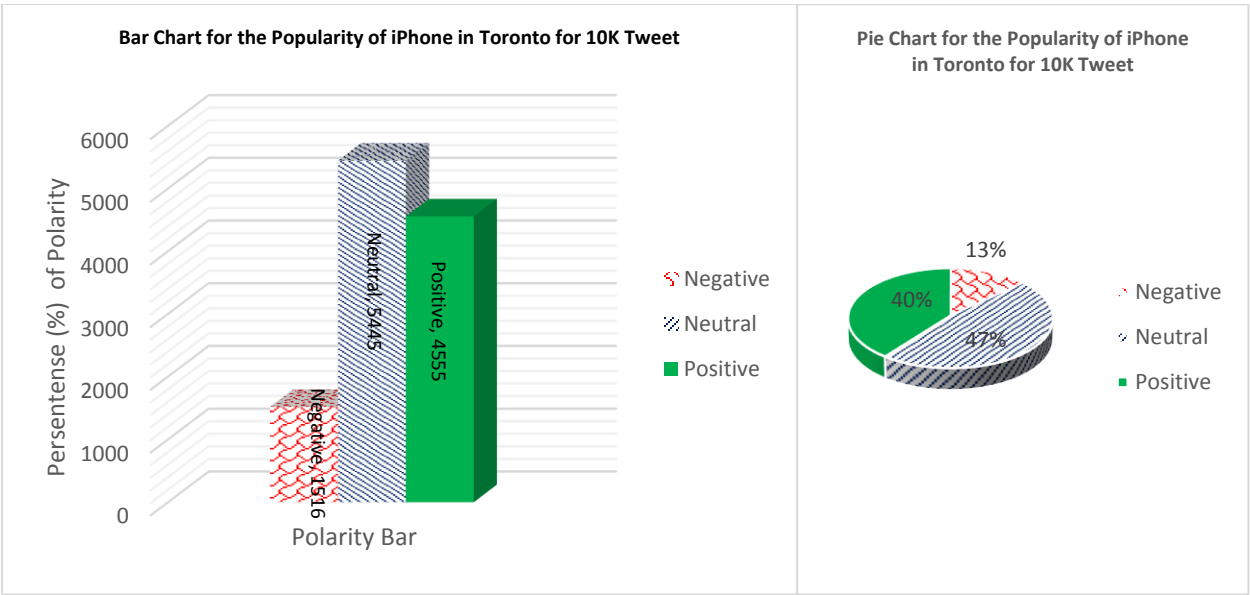


Fig.18 Bar Chart and Pie Chart for iPhone popularity in Toronto based on 10,000 Tweet

4.2.2. Single product in Multiple Location

Experimental Environment

Single product: **iPhone**

Multiple locations: **Toronto and New York**

The research scope of this part of thesis was to scale the popularity of iPhone in **Toronto** and **New York** region. Similar to section 4.1.1 the system has collected iPhone raw data through the data crawler in Toronto and New York region over a period of three months (6th March 2014 to 10th May 2014). The system had processed same number of Tweets from each region. Multi-level data cleaning, processing and analyzing generated the popularity score to plot the column chart. In all of the experiments (of section 4.2.2) tweet number is a sequence number (organized from newest to oldest according to tweet posting timeline); this number is generated by the data extraction process (tool: developed data crawler) and these sequence number cannot be reshuffled. For easier visualization and comprehension of the data, we are presenting the scenario for first 100 Tweets (of totally collected tweets, starting from the latest date). The Solid line presents popularity of Toronto and dotted line represents popularity of New York.

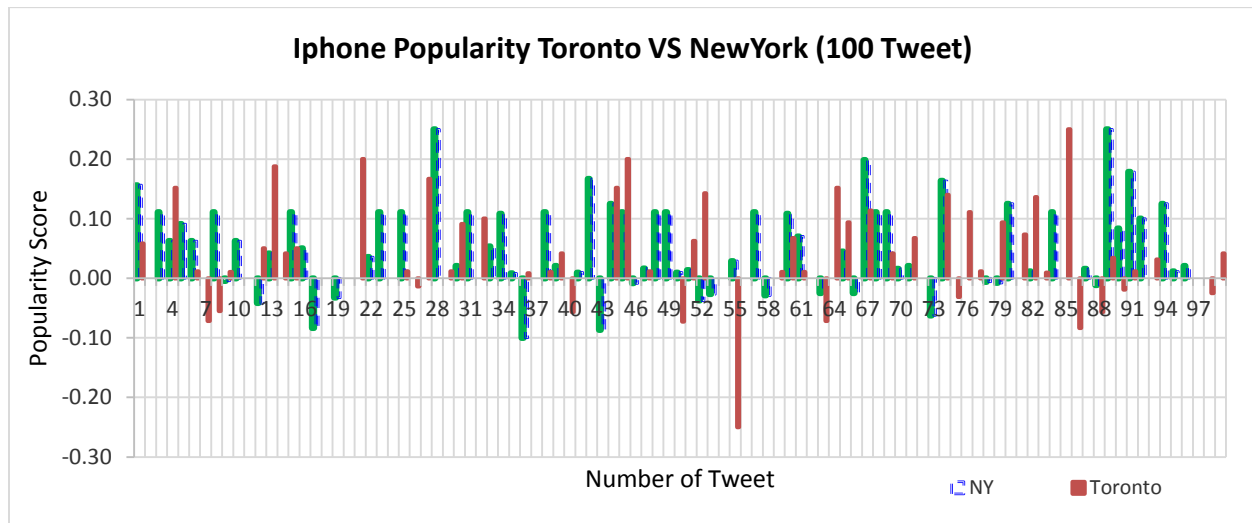


Fig.19 Toronto VS New York Popularity for iPhone based on 100 Tweets

As it may be difficult to discover a trend from line graphs with random values, the research presents cumulative line graph. Fig. 20 represents a cumulative line graph of the popularity of iPhones in Toronto VS New York (6th March 2014 to 10th May 2014). Cumulative popularity

score is the addition of sentimental weight of continuous tweets. The Red line and the Blue line represent cumulative popularity of iPhone in Toronto and New York, respectively. From the graph we can easily discern that iPhone has more popularity in Toronto than in New York.

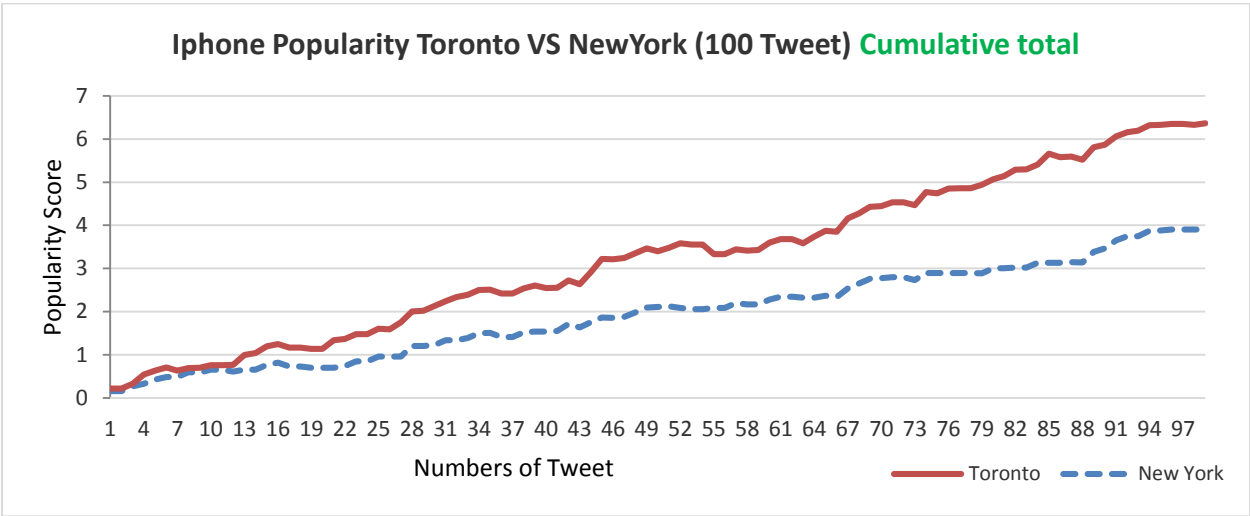


Fig.20 Cumulative popularity line graph for iPhone in Toronto VS New York

It may be visually aesthetic to display smaller number of data which might not be efficient to represent accurate scenario. So we have performed the cumulative line graph process for 10,000 Tweets and have generated the following results. Fig. 21 is representing a cumulative line graph on iPhone popularity in Toronto vs New York

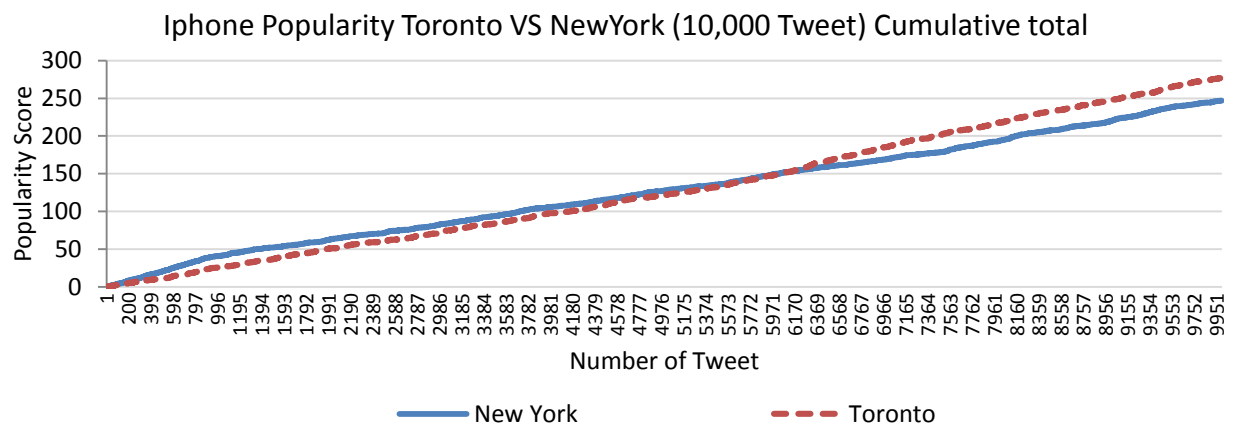


Fig. 21 Cumulative line graph on iPhone popularity in Toronto vs New York based on 10,000 Tweet

From Fig.20 the system presents that until 5345 tweets, the popularity is dominant in New York but after 6347 tweet Toronto got better score. So we found a better insight of the scaling process along with the final result, that the iPhone is more popular in Toronto.

Now the thesis work aims to scale the popularity of iPhone between Toronto and New York. In the cumulative line graph (Fig.21) the system represents that Toronto has a better score but the exact scale is not defined. This percentage of popularity can be scaled by a Pie Chart (Fig.22).

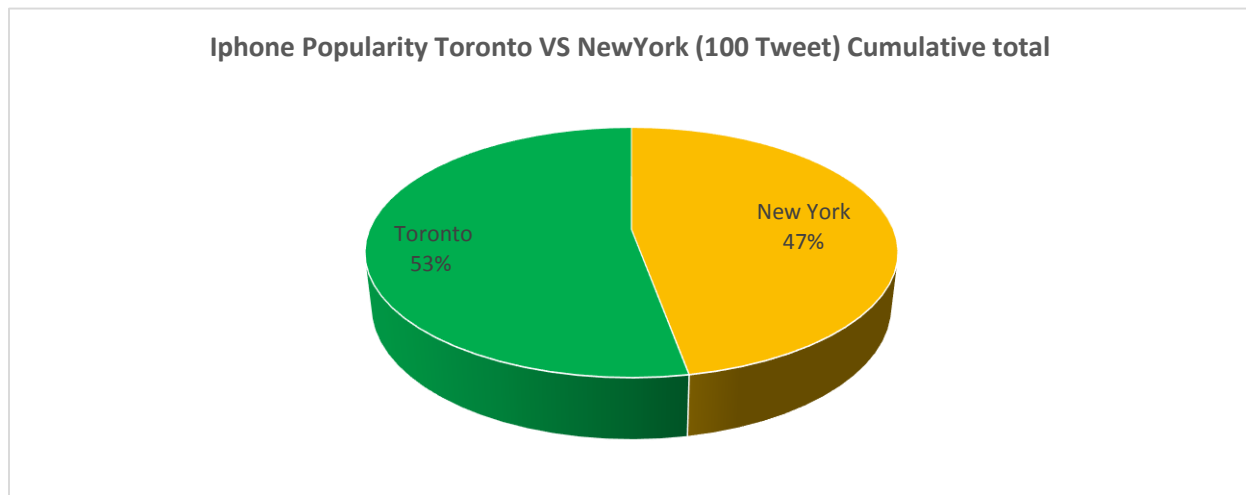


Fig. 22 Pie Chart for iPhone Popularity Toronto VS New York on (10,000 Tweet)

From the pie chart we identify that iPhone-Toronto has 53% popularity and iPhone-New York has 47% popularity. It is concluded that iPhone is more popular in Toronto than New York.

4.2.3. Multiple products in Single Location

Experimental Environment

Multiple products are:

Person1: **Olivia Chow**

Person2: **John Tory**

Person3: **Rob Ford**

Single location: **Toronto**

In this section the research scope was to scale the popularity, of the mayoral candidates of Toronto. Olivia Chow, John Tory and Rob Ford are the leading candidates for the mayoral election. Raw data was collected through the developed crawler in Toronto over a period of three months (25th March 2014 to 9th May 2014). In this experiment we have taken equal number of

tweets for each candidate. The collected data has been passed through multi-level cleaning, processing and analyzing to generate popularity score. Finally a line graph has been plotted by processing first 100 Tweet (starting from the latest date) for each candidate. A lower number of tweets were selected for better understanding any trends in popularity. Election prediction has been plotted in the line graph fig. 23. In all of the experiments (of section 4.2.3) tweet number is a sequence number (organized from newest to oldest according to tweet posting timeline); this number is generated by the data extraction process (tool: developed data crawler) and these sequence number cannot be reshuffled.

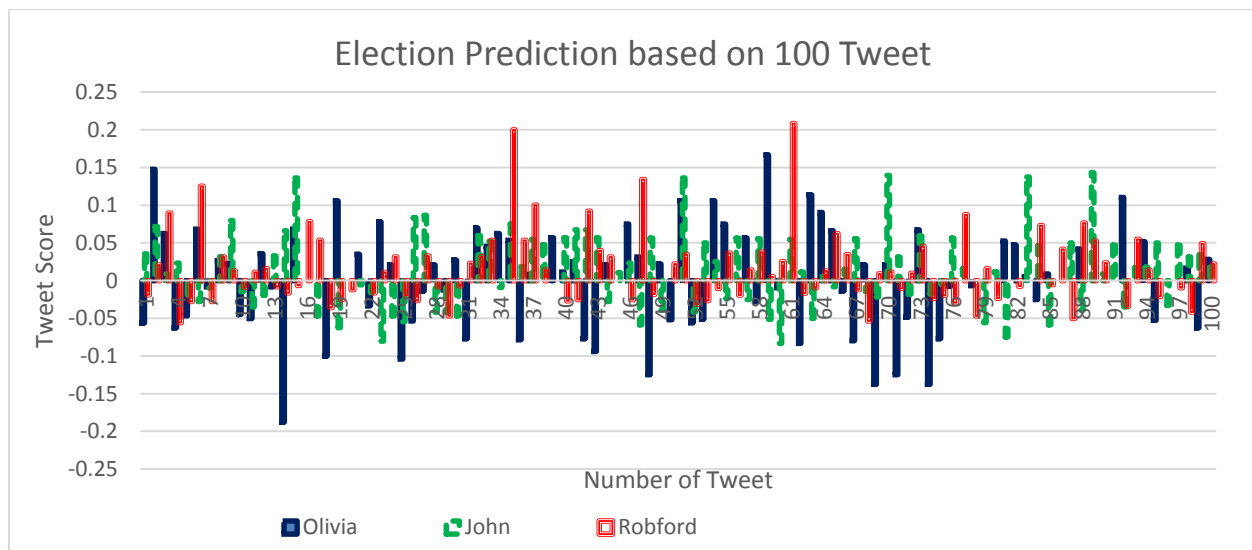


Fig. 23 Line graph for popularity scaling for the Election base on 100 Tweet.

To draw the cumulative line graph the system have used the same data set of 100 Tweets (25th March 2014 to 9th May 2014, starting from the latest date) and have generated the graph in Fig 24 to show cumulative flow of popularity score for the election prediction. As mentioned earlier the graph was not clear enough to provide a trend in popularity. Which is resolved in this graph because lines are not overlapping each other randomly and giving a flow of understandable results. Though this cumulative graph is a good enough representation of the 100 Tweet data we may still not get a much more realistic situation on ground without processing higher number of tweets.

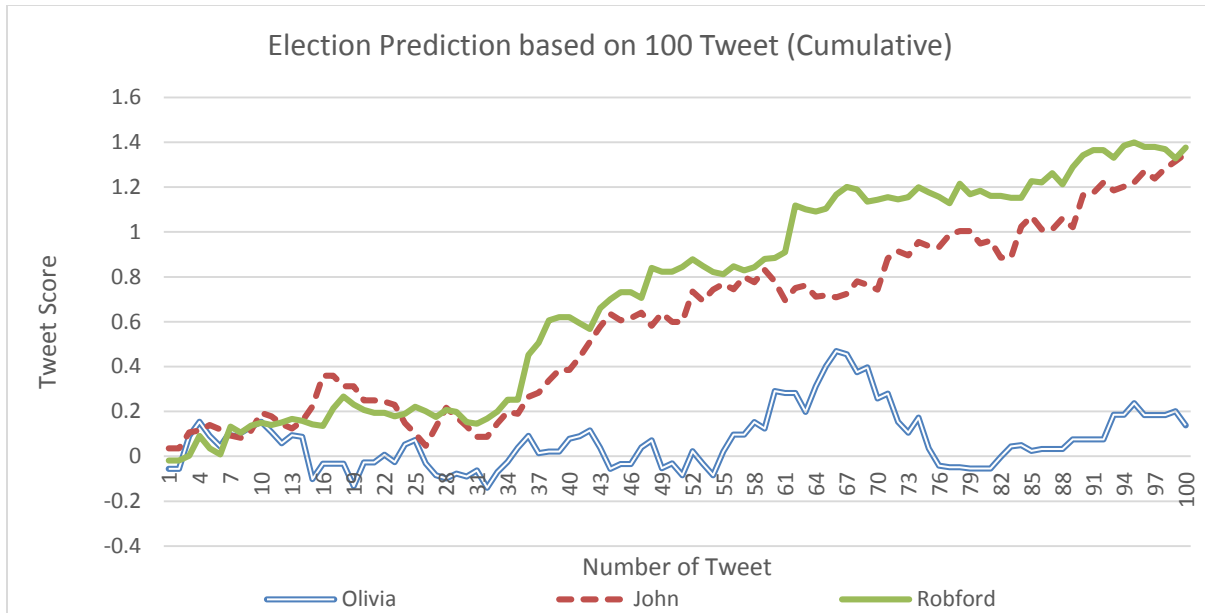


Fig. 24 Cumulative line graph for popularity scaling for the Election base on 100 Tweet.

It is know that more data can generate more accurate results. The thesis work wanted to forecast the trend for the mayoral election. That's why the system have processed 850 tweets (25th March 2014 to 9th May 2014, starting from the latest date) for each candidate and plotted line graph as fig.25. The analysis identify that John Tory is carrying the best score which is represented by broken lines (fig.25).

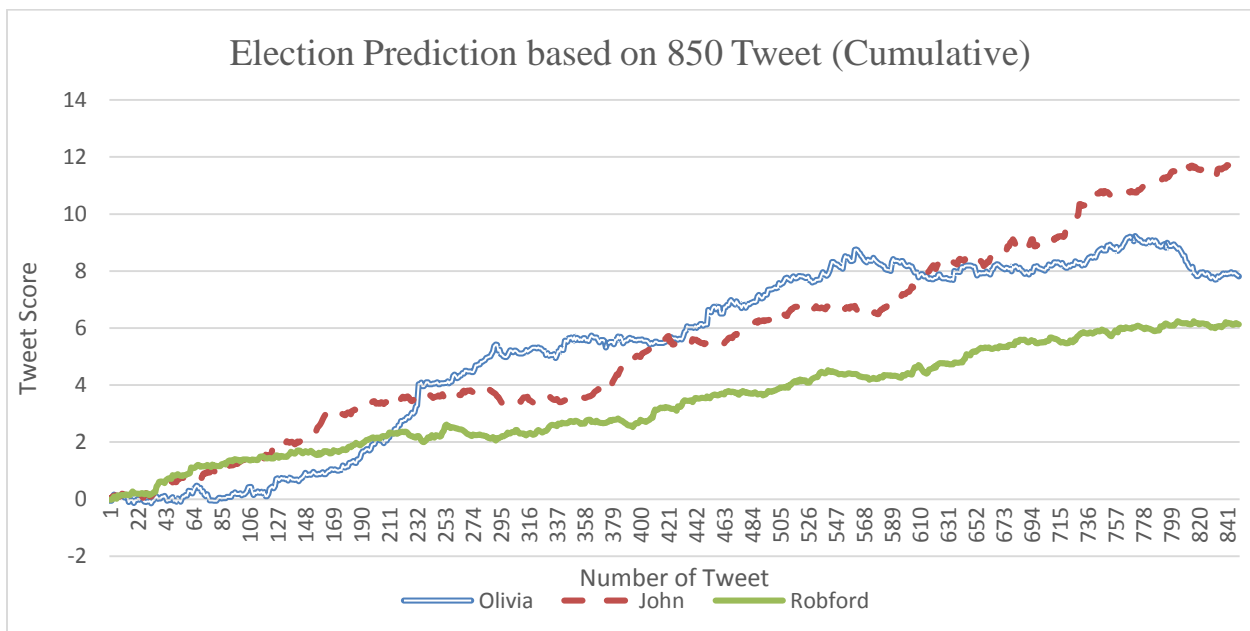


Fig.25 Cumulative Line graph for popularity scaling for the Election base on 850 Tweet.

Second highest score is carried by Olivia Chow which is presented by blue Line. Finally Rob Ford is carrying the lowest score which is presented by double line.

4.2.3.1 Data Validation

The actual score percentage of candidates, can be displayed by using pie chart. In Fig. 26 this research presents a Pie chart for the candidates cumulative score based on 850 tweets (starting from the latest date). Result shows that Olivia had 29% popularity, John had 45% popularity and Rob ford had 22% popularity. Real Election Result shows Tory 40.3%, Doug Ford 33.7% and Olivia 23.1%. Election Result Source: <http://news.nationalpost.com/2014/10/27/toronto-election-2014-live-results-news-and-commentary-on-the-mayoral-race/>

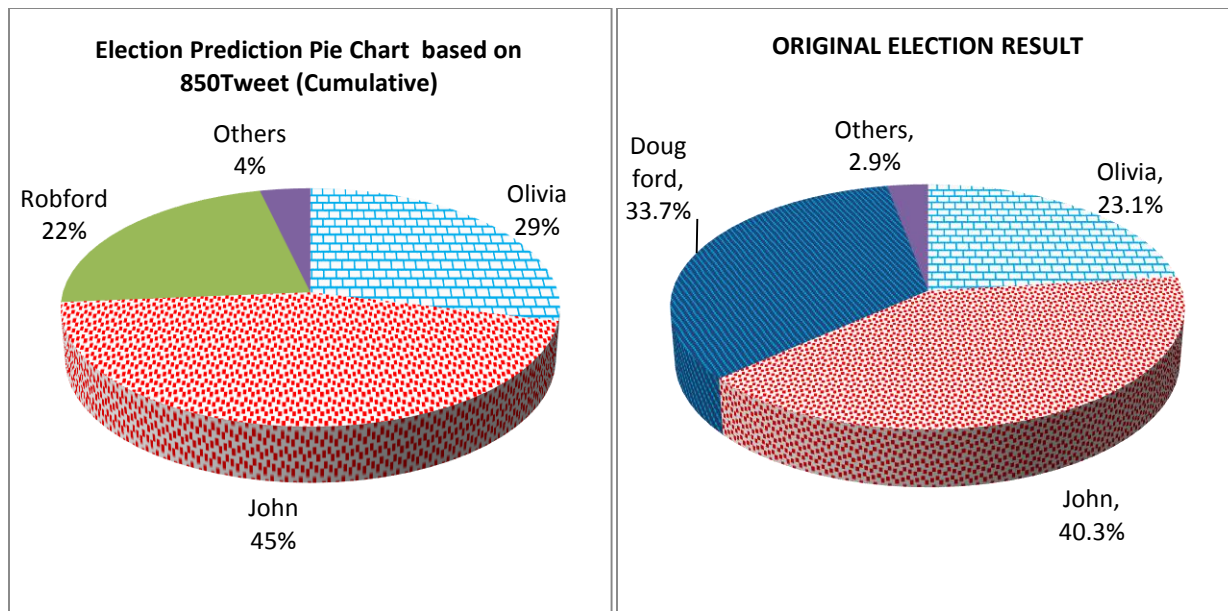


Fig. 26 Pie Chart for the Predicted Election Result VS Real Election Result

This research performed the result calculation on March- May 2014 Data, when Dough Ford was not on the scene but still it generated much closed prediction. Hence the research result is valid.

Popularity of candidate is not always static. Weekly election trend might bring interactive visualization of popularity. For this experiment, data was collected over a week of time Sep18th to Sep 25th 2014. After processing the tweets we have generated a bar chart presented in Fig27.

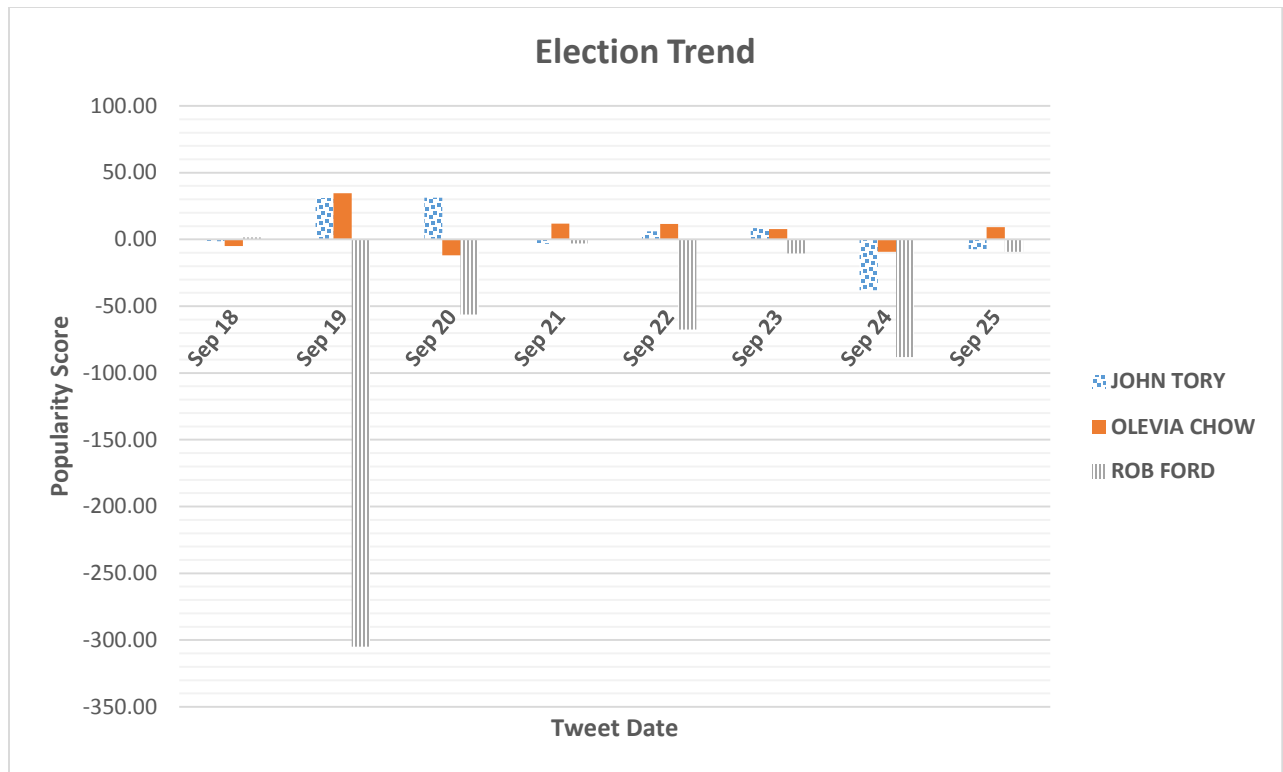


Fig. 27 Bar chart for Weekly Election Trend base on 985+ Tweets.

Now we want to visualize the same dataset (Sep18th to Sep 25th 2014) by continuous line graph.
 IF we draw cumulative line graph we get image like Fig.28

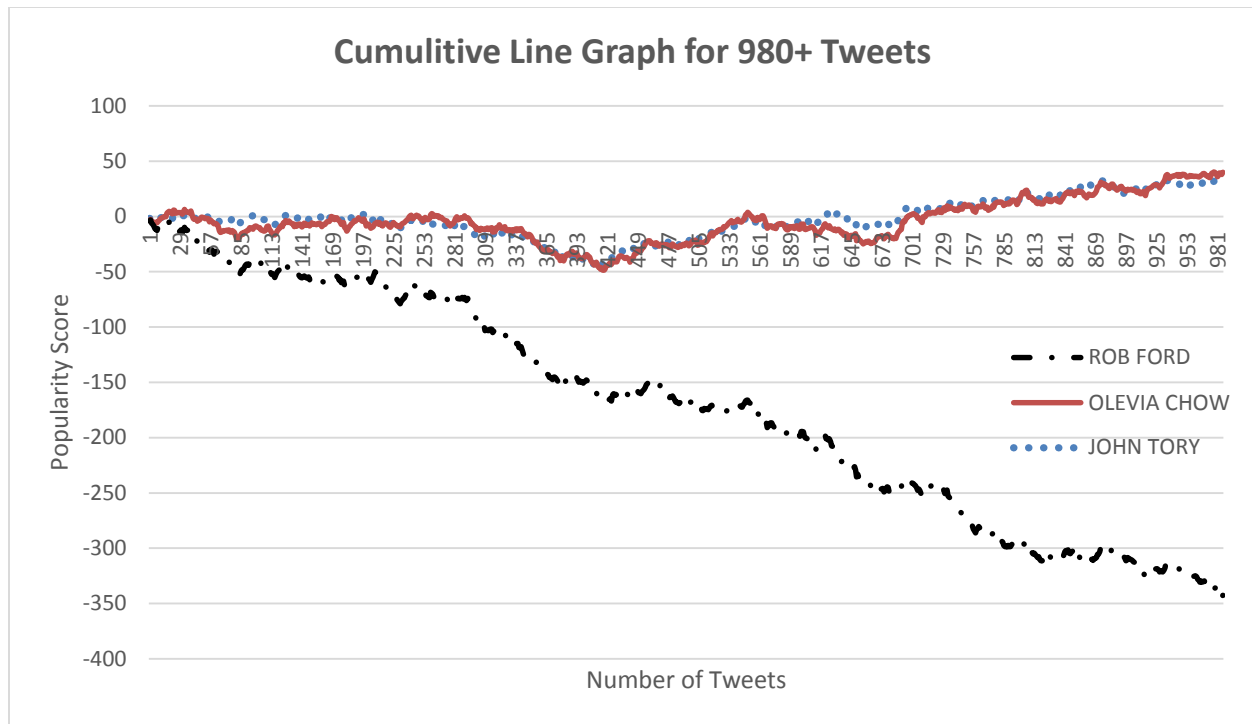


Fig. 28 Cumulative Line graph for Weekly Election Trend base on 980+ Tweets.

4.2.4. Multiple products in Multiple Locations

Experimental Environment

Multiple products are:

Product1: **Honda**

Product2: **Toyota**

Multiple locations: Location 1: **Toronto**

Location 2: **New York**

In this research experiment we wanted to scale the popularity of **Toyota** and **Honda** in two cities, **Toronto** and **New York**. Since both of Toyota and Honda have multiple different models. Thesis work has searched the following models for both brands. Raw data was collected (18th July 2014 to 23rd July 2014) by running our data crawler in Toronto and New York separately. In this experiment we processed equal number of Tweets for each brand from each location. After applying many different cleaning algorithms and logical processes the thesis has prepared the final score of Honda and Toyota data in fig 29. In all of the experiments (of section 4.2.4)

Tweet number is a sequence number (organized from newest to oldest according to tweet posting timeline); this number is generated by the data extraction process (tool: developed data crawler) and these sequence number cannot be reshuffled.

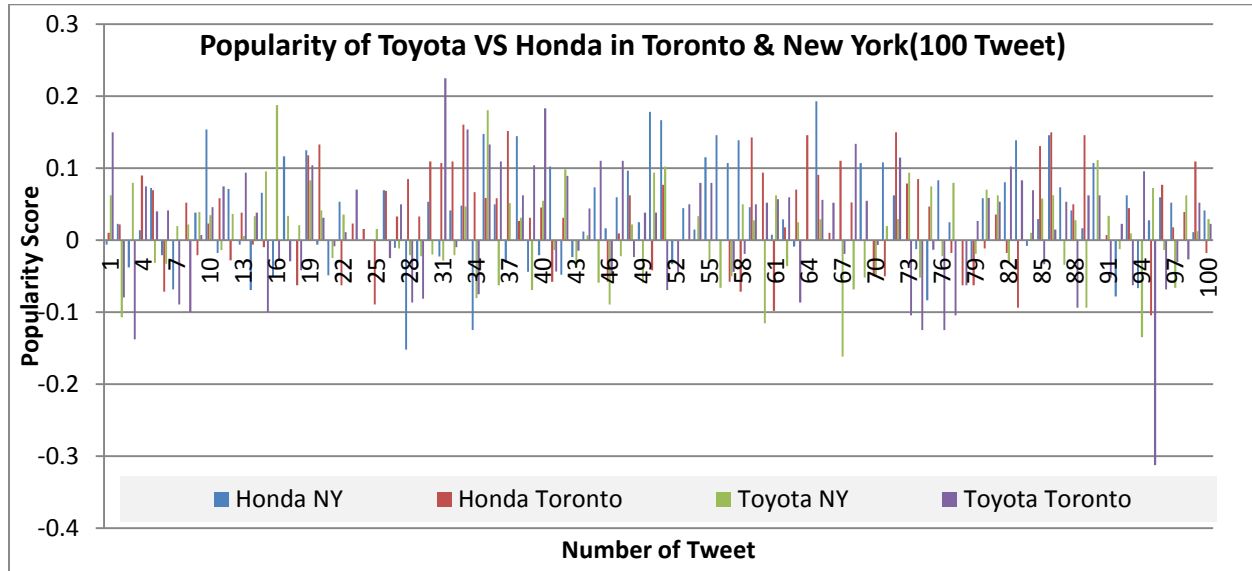


Fig.29 Popularity scaling for Toyota and Honda based on 100 Random Tweet in Toronto and NY area

For random value line graph it is very difficult to discern a trend especially when the same value occurs frequently or the values are very close. So to get a clear idea about the internal trend we have plotted cumulative line graph (shown in fig 30) for 500 tweets of each Toyota and Honda from both Toronto and New York (18th July 2014 to 23rd July 2014). If we analyze the graph we can easily say Honda in NY has higher popularity than Honda-Toronto, Toyota-NY and Toyota Toronto. In the reverse order we can say Toyota-NY has the lowest popularity among the compared brands and locations.

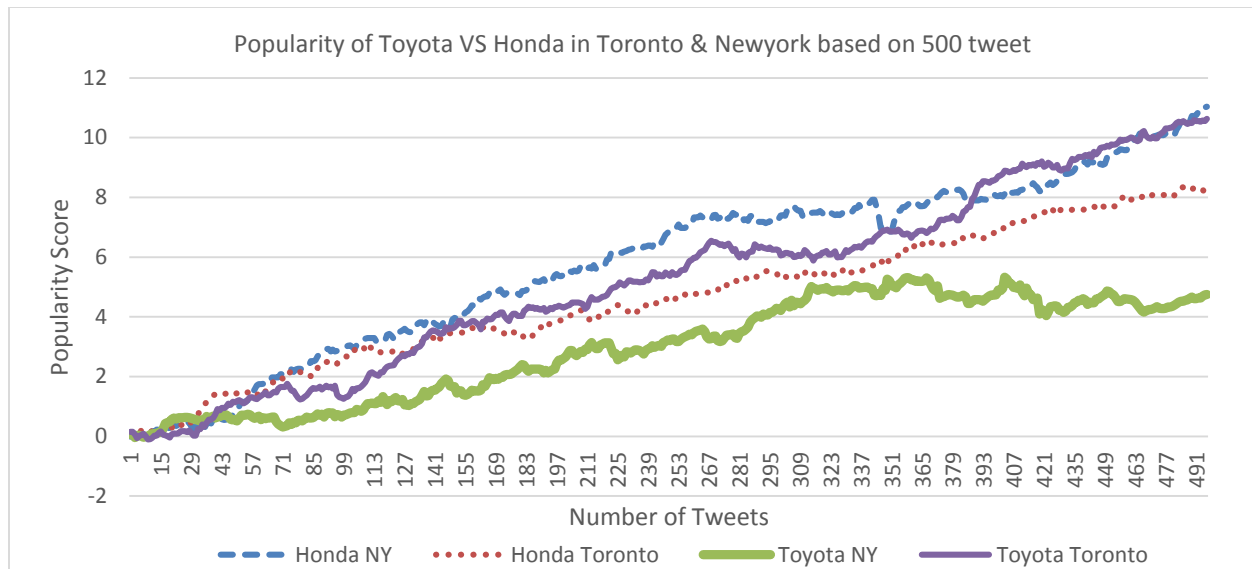


Fig.30 Popularity scaling of Toyota VS Honda in Toronto & New York

System has plotted a Bar graph on cumulative distribution (18th July 2014 to 23rd July 2014) and had generated the following graph as shown in fig.31. Here we can better observe individual brand score on different regions.

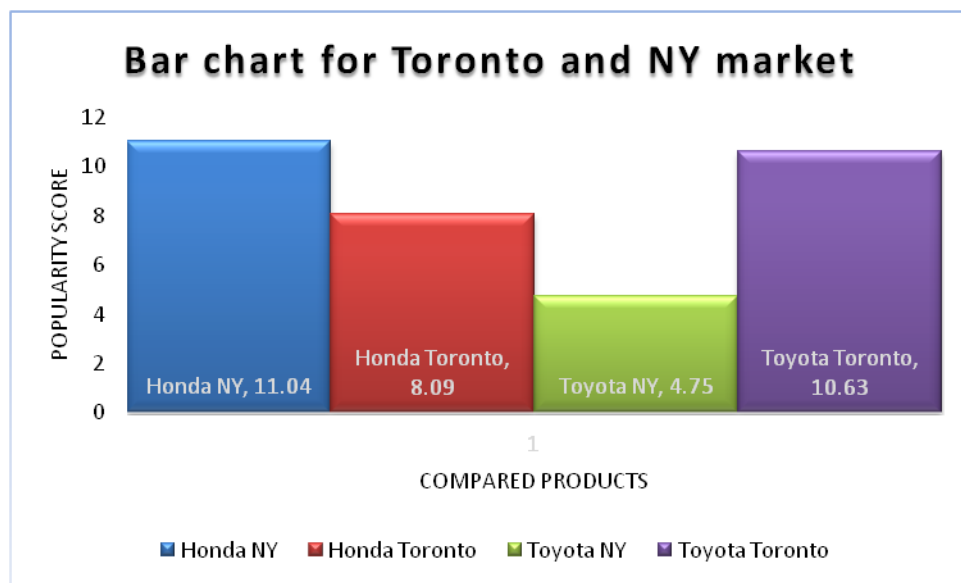


Fig.31 Popularity scaling Bar chart for brands on different locations

The pie chart can give a better visualization of the result of the popularity ratio. System has conducted two experiments for both brands (with data 18th July 2014 to 23rd July 2014). Result

shows (in fig 32) that Toyota has 69% popularity in Toronto whereas 31% popularity in New York. In case of Honda it has 58 % popularity in New York and Toronto has 42% popularity.

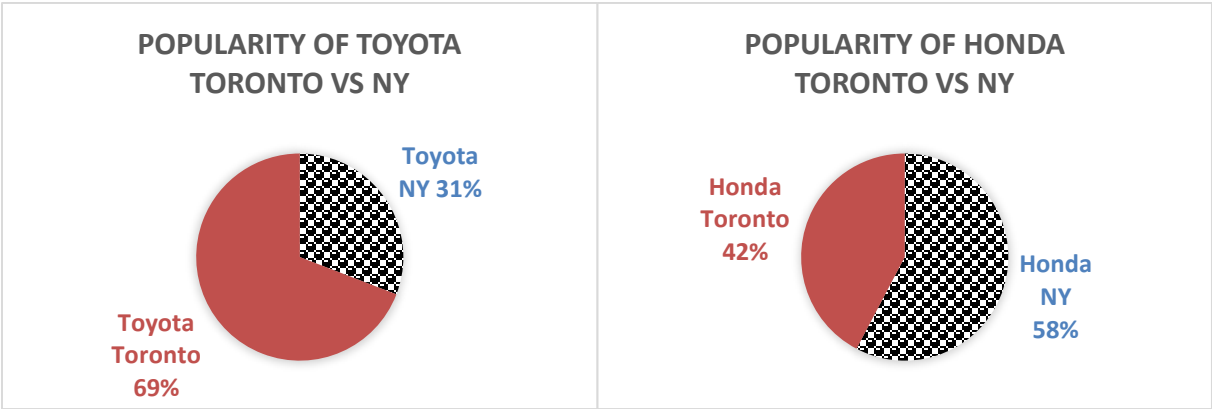


Fig.32 Popularity scaling for brands over the Locations

Finally Fig.33 shows a pie chart considering the Toronto and New York as a targeted market to identify the popularity ratio. It shows Honda-NY has the highest popularity of 32% then second highest Toyota-Toronto has 31% popularity. Honda-Toronto is in the third position with 23% score and Toyota-NY has the lowest score of 14 % market popularity.

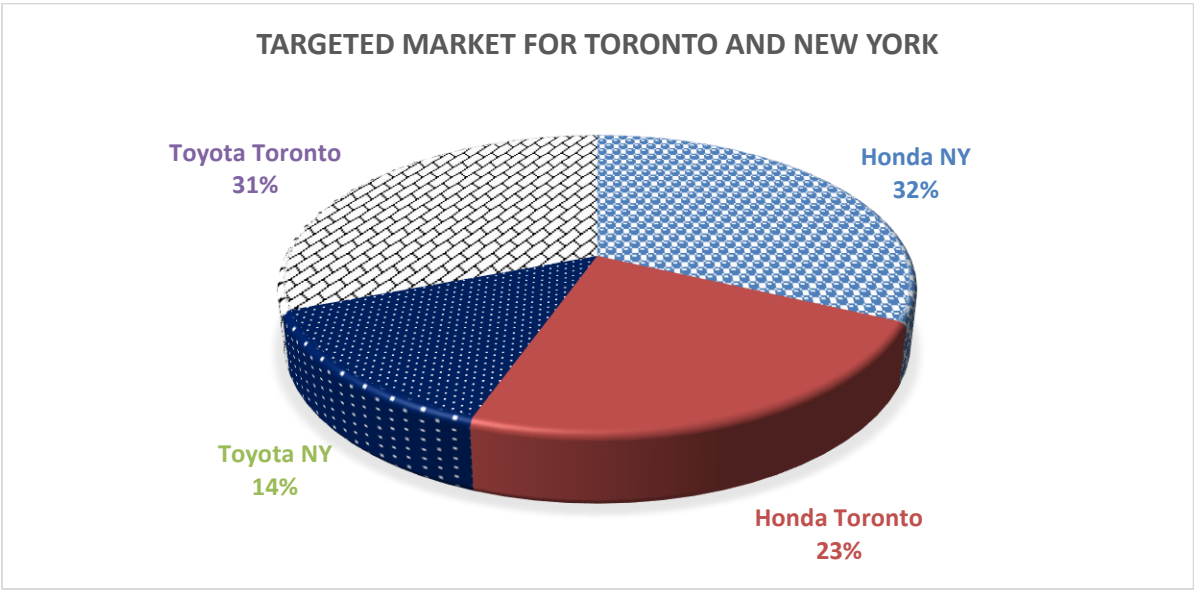


Fig.33 Popularity scaling for Brands over the Targeted market

Chapter 5

Conclusion and Future work

5.1 CONCLUSION

The emergence of social networking sites brings new ideas for data mining. People are sharing their opinions on many topics through microblogging services. Twitter allows to access and analyze these opinionative messages. The purpose of this work was to scale location base popularity for different products or services and use the output for real-life business aspects. This research utilized data mining techniques to collect location based data from twitter, which undergone different processes to generate popularity scaling. These processes consisted of data mining, extraction and visualization, all of which were challenging tasks. Multiple dimension of business aspects have been investigated, analyzed and visualized. An interactive application has been developed for social media data mining and visualization. Ontology technique was used for semantic search protocol to extract better data and relationships. Location based scaling can be applied to different analysis such as product marketing, sentiment analysis, event detection, trend identification, election forecasting, health-related information retrieval and epidemic identification.

This research has developed a user-friendly interface with Google Map. Users can draw single /multiple locations on Google map API to perform location based search. The application can extract location independent (tweet has no location) twitter data. It allows to perform semantic search for tweets either by keyword, hash-tag (#) and text file. The application sends request to Twitter API through Data Crawler. The Crawler collects related tweet data and generates data visualization instantly on the map with google markers. The Text processor examines the data to decide what actions are needed to perform, store and manage the data. Data cleaning process cleans the no-structured format tweet content written in informal language. Word Parsing process splits the cleaned tweet string into parts and develops an array of words. Word-Matching function compares each word with lexicon dictionary which returns numerical

values. Score Calculation process adds up the returned values by word-matching function and generates the final score for a tweet string.

5.2 CONTRIBUTIONS

Sentiment analysis or opinion mining is one of the most active research areas in NLP. It is a popular method that has been used over the past few decades for data, web and text mining. Sentiment analysis still has a growing demand to meet business and social demands. Many researchers have worked for sentiment analysis on social media. Other researchers have worked on location based system.

But the main contribution of this research is the integration of sentimental analysis on location based data from Twitter, which allows to scale popularity of a wide range of topics. There are limited works on location based popularity analysis on Twitter data. The main features of this thesis consist of the following things:

- The research has developed an application that employed a novel user interface to search and visualize the data on google map. This application allows selecting the locations of Twitter users on the globe graphically or by given parameters.
- The research work allows the use of sentiment analysis on the collected tweet data by performing location based popularity scaling on products, persons, brands or any given topic.

The user-friendly interface and lightweight application brings out a new way for efficient information search retrieval and visualization of results which opens a door for looming up various research opportunities. This research explores the raw-data-mining technique from social media networks and has performed different processes to generate useful results. Couple of real-life experiments have been performed and discussed to present results in different dimensions for popularity (of products, persons, brands etc.) scaling. Through a step by step process, it has been proved that the research is capable of performing location based scaling on different products

very efficiently. Hence this research is successful for performing its proposed goal of location based popularity scaling and furthermore, it is not limited to location based data.

5.3 FUTURE WORK

Currently the scope of this research and its application is limited to use Twitter as a data source. In future this research aims to develop a framework and integrate all the popular social networking sites like Facebook, Google+, hi5 and YouTube. In this way the system can process more complex situations and generate a realistic output for different business aspects.

Many researches have shown that the usage of emoticons may generate better insight for microblog processing in some cases. In future work this research aims to adopt emoticons to enhance the insight for sentiment processing.

It is unfortunate that location enabled tweets are currently sparse. This is why gathering high volume location-enabled data is a big challenge. Twitter users generally do not share their geo-location. Twitter API also does not allow the access of location related information such as server origin for user privacy. To overcome this problem, future work aims to develop and integrate an algorithm content based approach to determine user's location. This algorithm will examine user's profile, geographic region, hierarchical location, home location, travel location, city location, time zone, zip code or postal code to predict the user's location.

This research has used an unsupervised method (Lexical based analyzer) for the experiment. For some of the cases (proposed problems), supervised method may generate better results. In future this research will adopt both supervised and unsupervised method at the same time. Successful implementation of proposed system will be efficient enough to compare both outputs and will also yield better results.

References

- [1] C. Aggarwal, “An Introduction to Social Network Data Analytics in Social Network Data Analytics”, New York: Kluwer Academic, January, 2011, pp. 1-14
- [2] M. Russell, “Micro formats: Semantic Markup and Common Sense Collide,” in Mining the Social Web, 1st ed , California , USA: O’Reilly Media, Inc., 2011, pp. 19
- [3] English Vocabulary, <http://www.talkenglish.com/Vocabulary/english-vocabulary.aspx>, Dec, 2014
- [4] SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, [www.http://sentiwordnet.isti.cnr.it/](http://sentiwordnet.isti.cnr.it/)
- [5] L. Barbosa and J. Feng, “Robust sentiment detection on Twitter from biased and noisy data,” 23rd International Conference on Computational Linguistics, Beijing, China, 2010, pp. 36-44
- [6] Digital Marketing Ramblings <http://expandedramblings.com/>
- [7] B. Huber, D. M. Romero and F. Wu, “Social networks that matter: Twitter under the microscope,” website, 2008. [Online] Available: <http://arxiv.org/abs/0812.1045>
- [8] H. Kwak , C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?” in Proc. Int. World Wide Web Conf. ,San Francisco,2010, pp. 591–600.
- [9] N. Jamil and A. Alhadi, “A Collaborative Names Recommendation in the Twitter Environment based on Location,” in Semantic Technology and Information Retrieval (STAIR), International Conference, Putrajaya, Malaysia, 2011, pp. 119 – 124
- [10] A. Benjamin and J. Krzysztof, “On the Geo-Indicativeness of Non-Georeferenced Text,” in 6th International AAAI Conference on Weblogs and Social Media, Dublin, Ireland 2012, pp. 375-37.
- [11] R. Elmasri and S. Navathe, “Data Mining Concepts,” in Fundamentals of Database Systems, 6th ed., Massachusetts, USA: Addison-Wesley Longman Publishing Co., 2011, pp. 1035-1064
- [12] M. Josep, S. Georgos, E. Vijay, “Scaling Online Social Networks without Pains”, in Proc of NETDB, Montana, USA, 2009, pp 104-112
- [13] I. Celino, D. Daniele, E. D. Valle, B. Marco, Y. Huang, L. Tony, K. Seon-Ho, and T.Volker, “Bottari: Location based social media analysis with semantic web,” International Semantic Web Conference (ISWC), Bonn, Germany 2011.

- [14] M. Frank, B. Dong, and A.P. Felt, "Mining Permission Request Patterns from Android and Facebook Applications," in 12th Industrial Conference on Data Mining International Conference, Brussels, 2012, pp. 870 - 875
- [15] K. Sam and R. Chatwin , "Ontology-based text-mining model for social network analysis," in 2012 IEEE International Conference on Management of Innovation &Technology, Bali, Indonesia 2012, pp. 226 – 231
- [16] K. Efthymios, W. Theresa, and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!" in Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain on 2011, pp. 538 – 541
- [17] B. Felipe , M. Marcelo, and P. Barbara, "Combining strengths, emotions and polarities for boosting Twitter sentiment analysis," in Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM New York, 2013, pp. 202-207
- [18] W. Haoet , C. Dogan, K. Abe, B. François, and N. Shrikanth , "A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle," in 12th Proceedings of the ACL System Demonstrations, Stroudsburg, USA, 2012, pp. 115-120
- [19] B. Johan, M. Huina, and P. Alberto, "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena," in Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, July, 2011, pp. 451-453
- [20] T. Jiliang, Y. Chang, and L. Huan, "Mining Social Media with Social Theories: A Survey," in ACM SIGKDD Explorations Newsletter, New York, 2013, pp. 20-29
- [21] P. Jian , H. Jiawei, W. Jianyong, P. Helen, C. Qiming, and D. Umeshwar, "Mining Sequential Patterns by Pattern-Growth:The PrefixSpan Approach," in Sixth Special Interest Group on Knowledge Discovery and Data Mining, ACM New York, NY,2000, pp. 355-359
- [22] D. Thomet , H. Bosch, S. Koch, M. Worner, and T. Ertl, "Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages," in IEEE Pacific Visualization Symposium, Songdo, Korea, 2012, pp. 41- 48
- [23] P. Dudas, "Cooperative, Dynamic Twitter Parsing and Visualization for Dark Network Analysis", in IEEE Network Science Workshop, 2013, West Point, NY, pp. 172 – 176
- [24] Twitter Documentation <https://dev.twitter.com/docs>
- [25] Oracle Corporation Documentation on REGXP_INSTR function http://docs.oracle.com/cd/B12037_01/server.101/b10759/functions114.htm#i1239887

- [26] TOpoli Events Community <http://topoli.ca/>
- [27] C. Charu, A. Mansurul, and A. Mohammad, “Frequent Pattern Mining Algorithms: A Survey,” In *Frequent Pattern Mining*, Springer International Publishing, Switzerland, 2014, pp. 19-64
- [28] G. Pollyanna, A. Matheus, B. Fabrício, and C. Meeyoung, “Comparing and combining sentiment analysis methods”, in *ACM '13 Conference on Online Social Networks*, Boston, USA, 2013, pp. 27-38
- [29] P. Shankar, H. Yun-Wu, P. Castro, B. Nath, and L. Iftode, “Crowds replace Experts: Building Better Location-based Services using Mobile Social Network Interactions,” in *IEEE Pervasive Computing and Communications*, 2012, Lugano, Switzerland, 2012, pp. 20 – 29
- [30] W3C Semantic Web Activity <http://www.w3.org/2001/sw/>
- [31] N. Tatti and B. Cule, “Mining Closed Strict Episodes,” in *IEEE 10th International Conference on Data Mining*, Sydney, 2010, pp. 501 – 510
- [32] A. Mazumder, A. Das, K. Nyunsu, S. Gokalp, A. Sen, and H. Davulcu, H. “Spatio-temporal Signal Recovery from Political Tweets in Indonesia”, in *International Conference on Social Computing (SocialCom)*, 2013, Alexandria, VA, pp.280 - 287
- [33] I. Celino, D. Daniele, E. D. Valle, B. Marco, Y. Huang, L. Tony, K. Seon-Ho, and T. Volker, “Towards BOTTARI: Using Stream Reasoning to Make Sense of Location-Based Micro-Posts.” In *Garcia-Castro, ESWC 2011 Workshops*, Springer, Heidelberg, 2011, pp. 80-87
- [34] G. Andrienko, N. Andrienko, D. Keim, M. MacEachren, and S. Wrobel, “Challenging problems of geospatial visual analytics,” *Journal of Visual Languages & Computing*, 22(4):251 – 256, 2011.
- [35] S. Koch, H. Bosch, M. Giereth, and T. Ertl, “Iterative integration of visual insights during scalable patent search and analysis,” *IEEE Visualization and Computer Graphics*, CA, USA, 2011, pp. 557 –569
- [36] J. Xiang, S.A. Chun ; J. Geller, “Monitoring Public Health Concerns Using Twitter Sentiment Classifications,” *Healthcare Informatics (ICHI)*, 2013 *IEEE International Conference*, Philadelphia, PA ,2013 pp.335 - 344
- [37] B. Trevisan, D. Erasme, EM. Jakobs, “Web comment-based trend analysis on deep geothermal energy,” *IEEE International Professional Communication Conference (IPCC)*, 2013, Vancouver, BC, 2013, pp. 1-8

- [38] K. Mouthami, N.Devi, and V. Bhaskaran, "Sentiment analysis and classification based on textual reviews," in International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, 2013, pp.271 - 276
- [39] C. Keke, S. Spangler, C. Ying, and L. Zhang, "Leveraging Sentiment Analysis for Topic Detection," IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Sydney, NSW, 2008, pp. 265 – 271
- [40] F. Neri, C. Aliprandi, F. Capeci, and M. Cuadros, "Sentiment Analysis on Social Media," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Istanbul, Turkey, 2012, pp. 919 – 926
- [41] J. Fiaidhi, O. Mohammed, S. Mohammed, and S. Fong, "Opinion mining over twitter space: Classifying tweets programmatically using the R approach," Seventh International Conference on Digital Information Management (ICDIM), Macau, China, 2012, pp.313 – 319
- [42] T. Jie and A.C.M. Fong, "Sentiment diffusion in large scale social networks," in IEEE International Conference on Consumer Electronics (ICCE), 2013 , Las Vegas, NV 2013, pp. 244 - 245
- [43] R. Nithish, S. Sabarish, N. Kishen, and A. Abirami, "An ontology based sentiment analysis for mobile products using tweets," Fifth International Conference on Advanced Computing (ICoAC), Chennai, India,2013, pp. 342 – 347
- [44] W. Wei and J. Gulla, "Sentiment analysis in a hybrid hierarchical classification process," Seventh International Conference on Digital Information Management (ICDIM), Macau, China, 2012, pp. 47 – 55
- [45] J. Polpinij and K. Ghose, "An Ontology-Based Sentiment Classification Methodology for Online Consumer Reviews," IEEE/WIC/ACM International Conference on in Web Intelligence and Intelligent Agent Technology (WI-IAT), Sydney, NSW , 2008, pp.518 – 524
- [46] S. Jamali and H. Rangwala, "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis", in International Conference on Web Information Systems and Mining (WISM), Shanghai, 2009, pp.32 – 38
- [47] H. Ming, C. Rohrdantz, H. Janetzko, and U. Dayal, "Visual sentiment analysis on twitter data streams," in IEEE Conference on Visual Analytics Science and Technology (VAST), Providence, RI, 2011, pp. 277 - 278