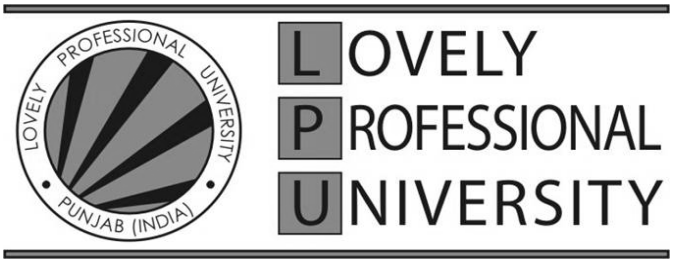

A novel approach of sentiment analysis for social media

A Dissertation Proposal

submitted

By

Hargobind Singh

to

Department of Computer Science & Technology

In partial fulfilment of the Requirement for

the Award of the Degree of

Master of Technology

in Computer Science Technology

Under the guidance of

Mr. Amritpal Singh
(17673)

(November 2017)

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0058

Supervisor Name : Amritpal Singh **UID :** 17673 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Hargobind Singh	11607946	2016	K1637	8568890137

SPECIALIZATION AREA : Database Systems **Supervisor Signature:** _____

PROPOSED TOPIC : A Novel Approach of Sentiment Analysis for social media

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.00
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.00
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	6.67
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.33
5	Social Applicability: Project work intends to solve a practical problem.	7.33
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.00

PAC Committee Members		
PAC Member 1 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 2 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 4 Name: Dr. Pooja Gupta	UID: 19580	Recommended (Y/N): Yes
PAC Member 5 Name: Kamlesh Lakhwani	UID: 20980	Recommended (Y/N): NA
PAC Member 6 Name: Dr.Priyanka Chawla	UID: 22046	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): NA

Final Topic Approved by PAC: A Novel Approach of Sentiment Analysis for social media

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 04 Nov 2017

Abstract

As we know that Social media has become an important part of today's generation. People try to post most of aspect of their life on social media (Facebook, Twitter etc.). Thus, results produced from mining the social media data are very effective in understanding current trends. Trends generated via social media sites are very helpful for different kind of business, launching new products etc. There is one another emerging field where use of social media data become very important that is: understanding social and political dynamic. Basically, our research work deals with introducing research methodology that will be based on sentiment analysis and topic distribution curves to get new insights of social and political dynamic.

CERTIFICATE

This is to certify that **Hargobind Singh** has completed dissertation proposal titled **A novel approach of sentiment analysis for social media** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma.

Date: **Signature of Advisor**

Name: Amritpal Singh

ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of dissertation. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the thesis work. I am sincerely grateful to them for their truthful and illuminating views on many issues related to this research.

I express my sincere thanks to my guide Amritpal Singh for his invaluable assistance, motivation, guidance and encouragement without which this research work will be a dream. In spite of his busy schedule, he was always there to iron out the difficulties which kept on arising at regular intervals.

I am grateful to our **Lovely Professional Universit**y for providing me with an opportunity to undertake this research topic in this university and providing us with all the facilities.

I am highly thankful to my family and friends for their active support, valuable time and advice, whole hearted guidance, sincere cooperation and pains-taking involvement during the study. Lastly, I thankful to all those, particularly the various friends, who have been instrumental in creating proper, healthy and conductive environment and including new and fresh innovative ideas during the project, without their help, it would have been extremely difficult to complete dissertation 1 within time.

DECLARATION

I hereby declare that the dissertation proposal entitled **A novel approach of sentiment analysis for social media** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: _____

Investigator

Regn. No._____

Table of Contents

List of Tables

List of Figures

Chapter 1
INTRODUCTION

Now a days one word is very common in database domain, that is mining. The term data mining is well known to almost all the organizations which deal with the data. Data mining is the process of extracting the hidden pattern or knowledge from a large data set. Data mining around the world is recognized with different synonyms such as- knowledge discovery in databases(KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

1.1 Data Mining Process

Complete knowledge discovery process involves multiple steps-

- Data integration
- Data cleaning
- Data selection
- Data mining
- Pattern evaluation
- Knowledge representation

Initially different resources are recognized, which can be used as a relevant data source according to requirement of analysis task. Further the process of data integration from multiple site is carried out which results in single database (coherent database). After integration the next challenge is to clean the data, which involves removal of unwanted, noisy data and handling the incomplete and inconsistent data. Further for extraction of task relevant data, specific part of cleaned data is selected on which different data mining operations are carried out. Data mining process is comprised of intelligent methods that are applied to extract the data patterns. Further as a data analyst next task is to evaluate the data patterns and identify the interesting patterns. After identifying the interesting patterns, data visualization and knowledge representation techniques are applied to present the most refined form of knowledge to the end user, which is capable enough to provide a support in decision making process. In short, we can say that the data mining process

results into interesting patterns/knowledge which summaries all the large data set. Figure 1.1 explains the mining process.

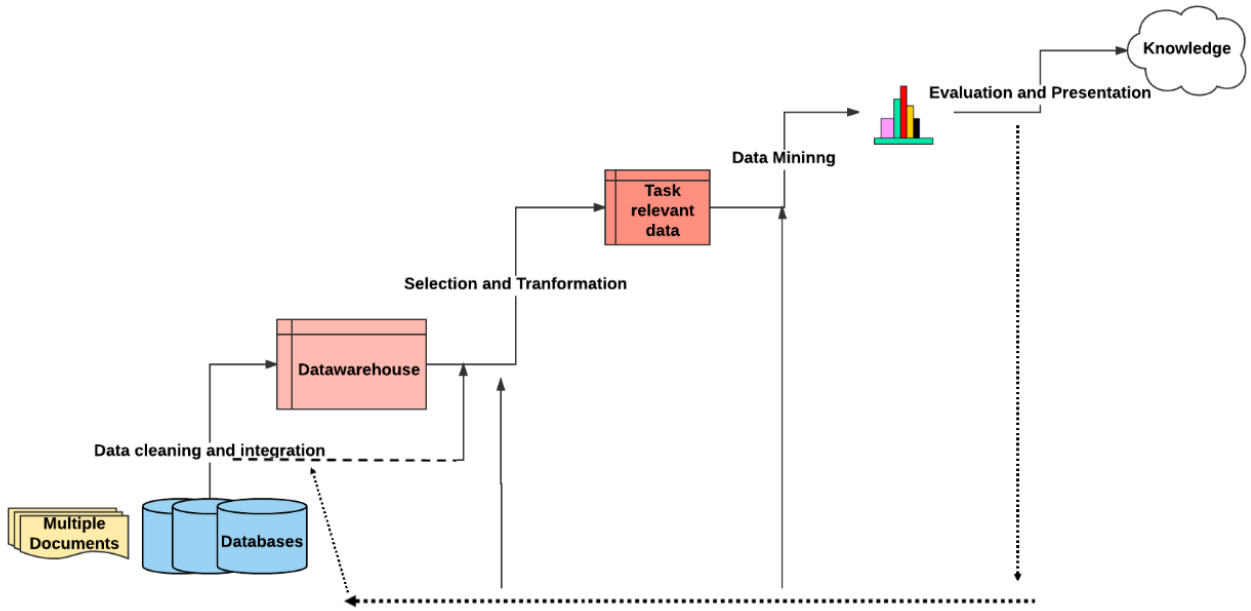

Figure 1.1: Knowledge discovery process.

1.2 Fields and applications of Data mining

Nowadays all the business ventures are spending a healthy amount of money for the purpose of safeguarding data and data analysis. Data analysis is helpful in finding hidden patterns and preparing strategies against predicted risks. Data analysis field is growing very rapidly due to its large acceptance by the financially sound organization. Following are the few fields where data mining applications are very common nowadays-

- **Market analysis and management**- In this field data mining is used for finding target customers, customer relationship management, market basket analysis, market segmentation etc.
- **Risk analysis and management-** Under this field data mining is used for forecasting outcomes, customer retention, quality control (in respect to product and services) etc.
- **Fraud detection and mining unusual pattern-** Data mining with integration with certain clustering algorithm can be used to construct a model which is capable perform the outlier analysis (identify the fraud).

- **Corporate analysis-** Corporates use the data mining applications for the purpose of finance planning and asset evaluation, resource planning, monitor competitors and market directions.
- **Education-** Educational data mining (EDM) refers to research that deals with finding new techniques and methods to process data gathered from different educational sources. There are basically two objective of educational data mining-
    1. Academic objective (to find out information that can be use to uplift the level of education practice).
    2. Administrative object (finding out patterns that lead to better management).
- **Data mining in field of agriculture-** The agriculture is backbone of our countries economy and it is the field having potential of many new research. The agriculture sector contributes 17% in countries total GDP. Technological advancement in this field results provides direct benefits to countries economy. Possible technology requirements are: -
    1. Decision Support System to decide type of crop to be cultivated.
    2. Rainfall Prediction System (with higher accuracy).
    3. Market Price Prediction.
    4. Disease identification system etc**.**

1.3 Data mining: a junction of multiple disciplines

Field of data mining is not only limited to a particular technology or discipline; however, it is comprised of multiple disciplines (confluence of multiple disciplines) as it adopts techniques from multiple domains. Following are example of those domains:

- Database systems and technology
- Machine learning
- Pattern recognition
- Statistics
- Visualizations methods (i.e. Graphs, plots etc.)
- Data warehouse
- Information retrieval
- High performance computing

1.4 Sentiment Analysis (sub domain of data mining)

Nowadays in field of computer science and technology the hottest research area is sentiment analysis. Sentiment analysis can be defined as technique/method of identifying the view of people, given in the form of text regarding to a specific object (event, individual, decision, change etc.).

Other synonyms of sentiment analysis are opinion mining, confidence analysis, people attitude towards an object, deriving opinion etc. The main reason behind the popularity of sentiment analysis it gives us overview of wide spread public opinion/thinking related to a topic. Basically, sentiment associated to a particular object is categorized in one of the following category:

- Positive
- Negative
- Neutral

Sentiment analysis is used at multilevel, it can be used to identify sentiment hidden in a document, to be more précised the analysis can be used to calculate the sentiment associated with each paragraph or may be each line.

The basic method adopted to identify the overall sentiment score is tokenization of each sentence in document into words, further each word is categorized into positive, negative and neutral words. In next step further, these words are categorized according to associated impact (for example: extremely happy is having more impact than happy), finally summation of numbers of positive and negative word is done to find out overall sentiment score.

The major challenges in sentiment analysis are:

- Multilingual
- Sarcasm
- Emoticons handling
- Natural language processing overheads

To increase the accuracy and overcome challenges of sentiment analysis, there are advance sentiment analysis mechanism which incorporate the sarcasm and emoticons handling technique. Moreover, nowadays natural language processing engines/software's (such as OpenNLP by Apache) are also used in the process of sematic analysis.

Example: I am very happy with India's win in cricket match.

if we tokenize the given word there are two positive words: happy and win, and there is no negative word, so it will produce overall positive sentiment score

1.5 Social media data mining

As we know that Social media has become an important part of today's generation. People try to post their daily routine on social media. The two most popular social networking sites are:

- Facebook
- Twitter

All over the world 328 million monthly active twitter user send on an average every 500 million tweets per day. On the other hand, there are over 2 billion active Facebook users, which spend on an average of 35 minutes per day on the Facebook. Now we have clear idea that these social networking sites target a large number of people around the globe. The Figure 1.2 present some more interesting facts related to Facebook and Twitter.

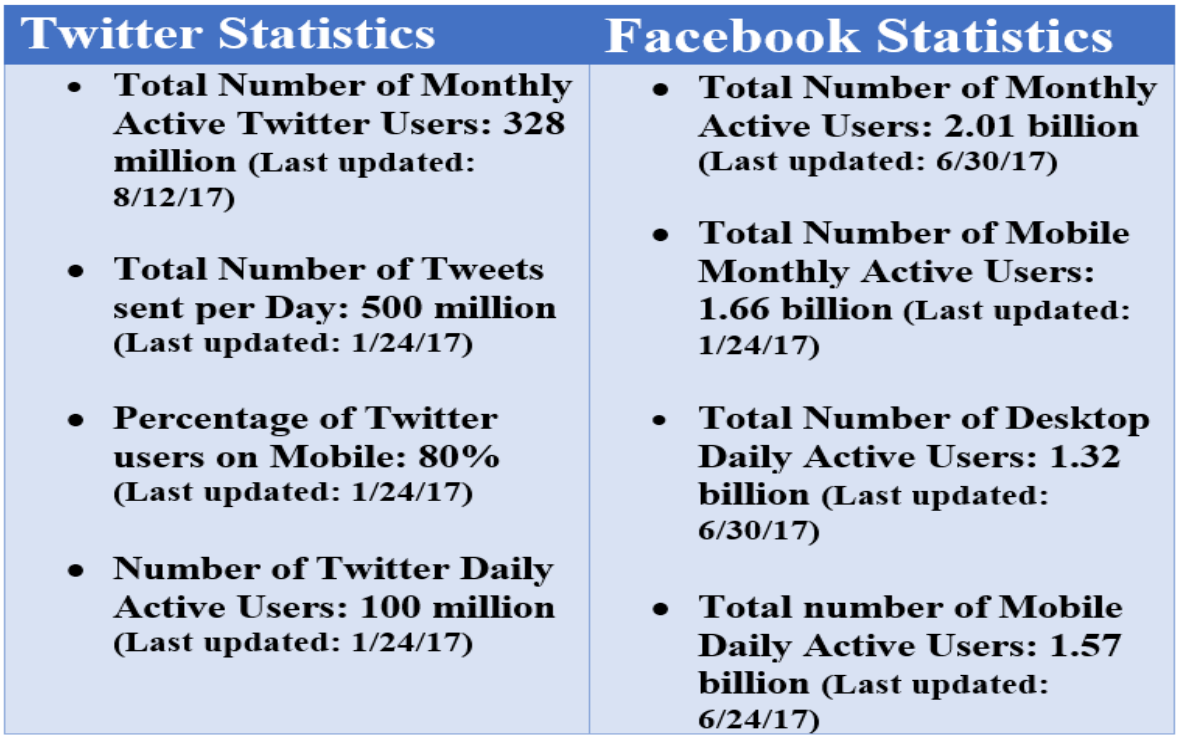

Figure 1.2: Representing statistics related to Twitter and Facebook

No doubt, that data analysts are using social media as the primary source of data. Due to availability of supporting techniques (Facebook graph API, Facebook app, Twitter app, various third-party tool etc.), extraction of this data has also become much easier than ever before. And, results produced from mining the social media data are very effective in understanding current trends.

The analysis of this big social data is useful in various fields such as:

- Sociology
- Politics
- Psychology
- Education
- commercial area

Following are few applications for which social media data mining is very common:

- Finding Popular trend among people
- For feedback of products
- Understanding current social status of society
- Finding Target customer
- Finding student problems etc.

Chapter 2
REVIEW OF LITERATURE

In last few years there is lot of researchers has worked in field of social media data mining. This section will summaries the work of few of them which is reviewed for the proposed research methodology which is described later in this report. Social media is very helpful in understanding the political and social dynamics, this thing is proved from the work of following researchers.

Saud Alashri et al[1] has gathered the data associated with the candidates of 2016 US president election from the social networking sites and presented the different of opinions among the top five presidential candidates: Hillary R. Clinton (Clinton),Donald Trump (Trump), Bernie Sanders (Sanders), Ted Cruz(Cruz), and John R. Kasich (Kasich). The **Figure 2.1** given below shows the distribution of topic among different candidates.

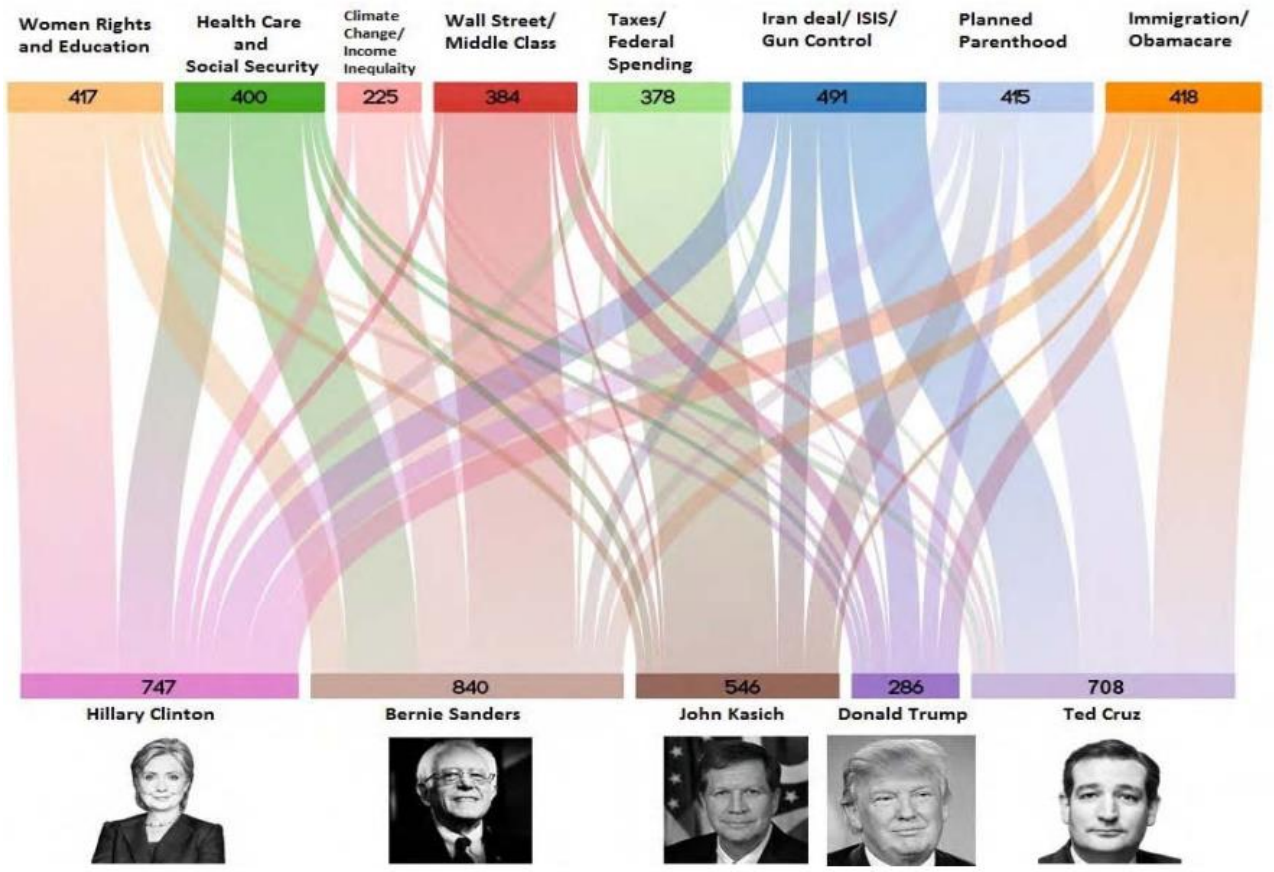

Figure 2.1: Distribution of topic per candidates [1]

The distribution shown in **Figure 2.1** is result of implementation Latent Dirichlet Allocation (LDA) on all candidate's posts.

Further in their paper they presented the people sentiments associated with candidates at different time period through the help of Sentiment and Comments Volume Curves.

Moreover, in their work they have performed Sentiment Analysis to identify people support to the different issues/topics raised by each presidential candidates. And, to perform sentiment analysis authors utilized Stanford CoreNLP as an open-source tool for calculating sentiment score

Another related work done by Anurag P. Jain and Mr. Vijay D. Katkar [2]in which they extracted the data related the political tweets using Twitter API v 1.1. This paper describe the mechanism that can be used to predict overall sentiment inclination of people towards political issues and situations. In the defined methodology two data sets are prepared after preprocessing the raw tweets: training data set of tweets and testing data set of tweets. Further, authors performed the sentiment analysis to build a model to categories the tweet into one of the following category positive, negative and neutral. The complete methodology adopted by authors is explained in Figure 2.2.

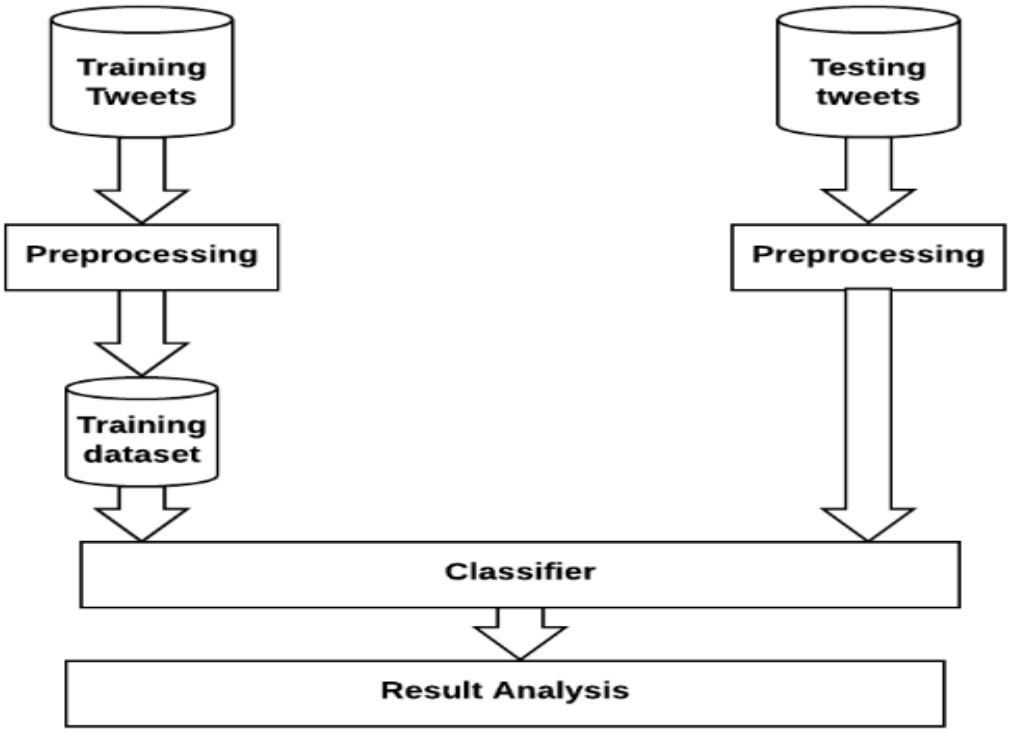

Figure 2.2: Methodology [2]

Basically, single tweet is divided into words, after splitting; polarity of words is calculated from SentiWordNet 3.0.0. Further the accuracy is compared of different classifier: 1)k-nearest neighbour 2)Random Forest 3)Naive Baysin 4)Baysnet. The Figure 2.3 given below is representing the result of comparison.

Accuracies of Machine learner

With stop words

Algorithm	Accuracy
k-nearest neighbour	99.6456%
RandomForest	99.0373%
BaysNet	75.0695%
NaivBays	60.3159

After removal of stop words

Algorithm	Accuracy
k-nearest neighbour	96.6398%
RandomForest	65.6681%
BaysNet	48.9579%
NaivBays	60.3159

Figure 2.3: Comparison of accuracies of different machine learner [2]

Antonio Teixeira and Raul M.S. Laureano [3] are also working on sentiment analysis on the basis the data from Facebook. The main purpose of authors behind this paper is to explain the process of Facebook data extraction, data preparation and sentiment analysis (using open source tools). Their project is using Facebook graph API and java API for fetching the data. The methodology explained in this paper is comprised of five major steps:

- Data collection
- Data preparation
- Sentiment detection
- Sentiment classification
- Presentation of output

In this paper author explained basically data preparation process. To be more précised in data preparation authors has explained how to tackle the emoticons present in the posts/text. In the authors work Facebook data is maintained with the help of three tables:

- Facebook posts
- Facebook comments
- Facebook reaction

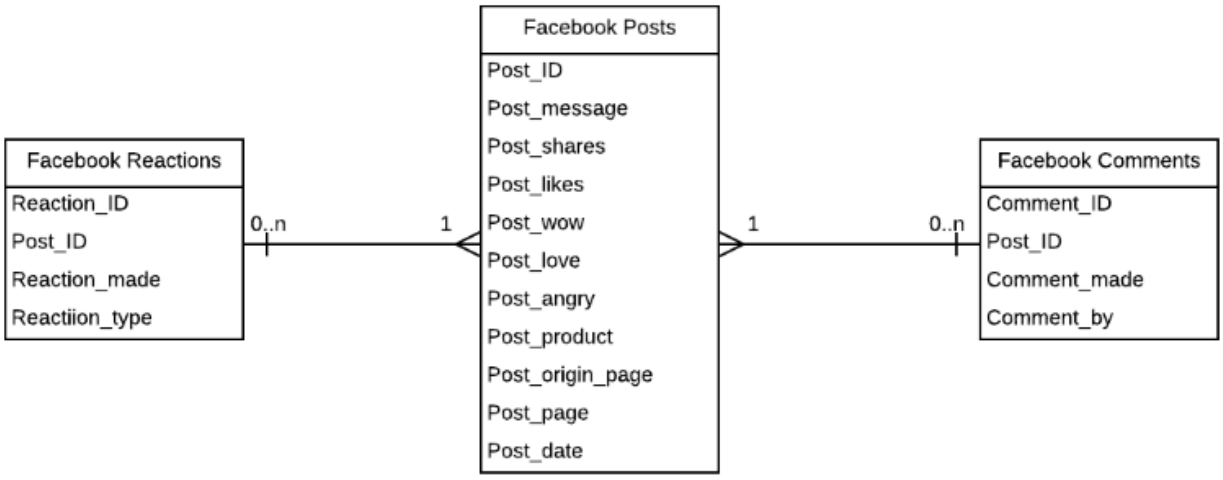

Figure 2.4: Relational database [3]

Further in their future work they are planning to use an open source tool (such as Knime, Hive, Pentaho Community Edition and R) for the purpose of semantic analysis.

Jerome Treboux et al [4] has presented an another work in which they have extracted the data from Flickr and Instagram. Basically, author has performed two tasks:

- Developed a prototype that assist in understanding the need of tourist.
- Analysis of customers detail that support in market decision making.

They have presented the data mining tools and prototype visualizations that are developed for marketing specialists in two different use cases: a Swiss regional tourism promotion agency (Valais Wallis Promotion) and a Swiss online discount retailer (Qoqa.ch). Design and development methodology, they defined for the project is based on CRISP-DM (Cross-Industry Standard Process for Data Mining). In the methodology author explained 7 steps:

- Identify the actors
- Identify the goal
- Define the pre-conditions
- Define the post condition
- Describe the main flow
- Describe the exception
- Describe the alternate flow

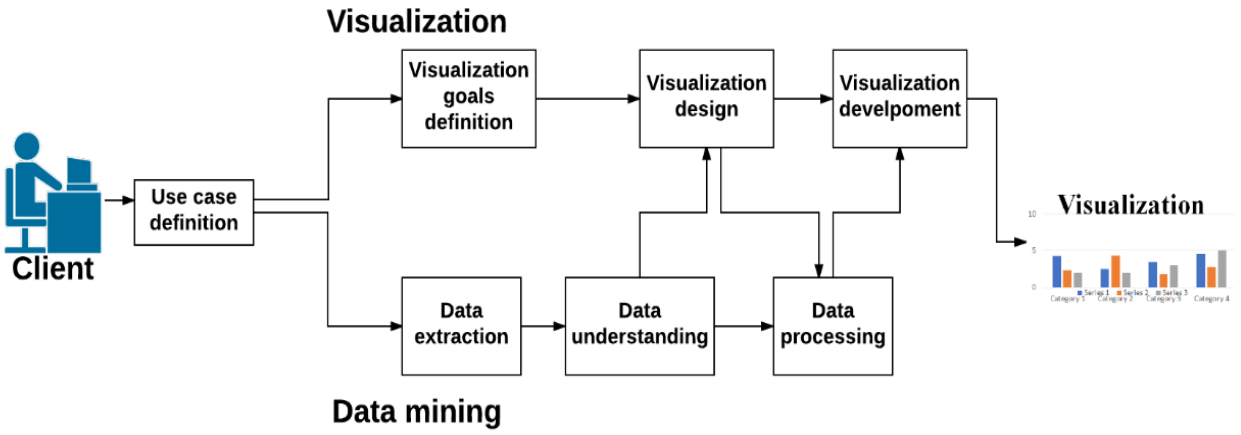

Figure 2.5: Methodology based on CRISP_DM with additional visualization branch [4]

Michalis Korakakis et al [5] has presented a survey paper in which they provided an overview research work on the popular politics domain within the framework of the Twitter social network.

In survey paper authors explained the different research works regarding topics: predicting the election results, sentiment analysis in political topics (opinion polls) and computational social science(Behavior analysis, Social interaction, Identifying social influence) .They basically focus on identifying the different work performed recently in the field of politics collaboration with the social media. Further they defined the pros and cons of each work.

Despoina Antonakaki et al [6] presented a paper in which they covered two events of Greek: Referendum and Election of 2015. The main motivation behind research was to use natural language analysis techniques (sarcasm detection, sentiment analysis, advanced lexicon and entity detection) on the data extracted from Twitter to reveal hidden patterns or content which is hard to find and relations for political domain, and they explored the dataset from related two specific events: referendum and legislative election. By applying natural language processing methods on both datasets, author explored the results in combination, explaining the details of user interaction that is shared among both political parties. For analysis purpose two distinct datasets are formed by extracting the twitter data. First dataset includes all the tweets that contain #dimopsifisma (Greek word that means referendum) and #greferendum from 25th June 2015 to 5th July 2015 (301,000 total tweets). Second dataset includes all the tweets that contain #ekloges and #ekloges2 (ekloges a Greek word that means election) from in September 2015 (182,000 total tweets).

After constructing the twitter corpus, they performed:

- entity identification
- sentiment analysis
- sarcasm detection

They presented final result with the help of different plots representing: Volume analysis, entities co-occurrence, sentiment, sarcasm and hash tags.

Rincy Jose and Varghese S Chooralil [7] proposed an approach for performing sentiment analysis of twitter messages based on lexical resources SentiWordNet and WordNet along with Word Sense Disambiguation. For better accuracy they also implemented the negation handling in the preprocessing step. They have presented their methodology including three steps: Data acquisition, Preprocessing and Sentiment classification (complete methodology is shown in **Figure 2.6**).

And they have implemented this work on 2015 Delhi election. And they have implemented this work on 2015 Delhi election. They extracted the tweet related to Arvind Kejriwal and Kiran Bedi for 3 weeks during the Delhi election days. Further Sentiment analysis performed on the extracted tweets using sentiment lexicons SentiWordNet and WordNet.

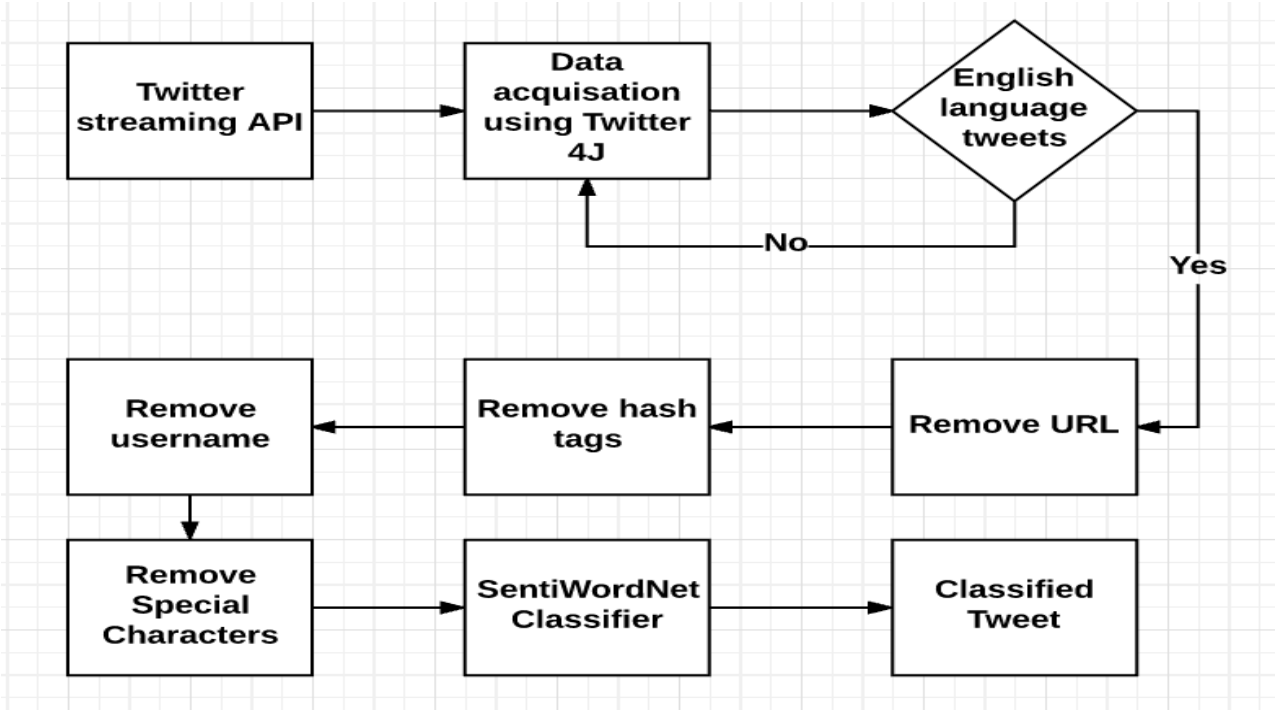

Figure 2.6: Proposed methodology [7]

Pete Burnap et al [8] proposed an paper in which they forecasted the outcome of 2015 UK general election based on Twitter data. Data is extracted from Twitter streaming API (tweets are selected on the basis of leaders and parties name). After fetching the data, further a third party tool (developed by Thelwall et al in 2010) was applied to perform the sentiment analysis, which allocates each word in string from -5(extremely negative) to +5 (extremely positive).

Anastasia Giachanou et al [9] proposed a paper in which they tracked the sentiment toward different entities and detect the sentiment spikes and detect the corresponding reason for the spike (spike in respect to graphs). They have adopted the approach which combines the LDA model with the relative entropy. The proposed methodology is comprised of 3 major steps:

- Identifying most important sentiment spikes
- Identifying Reason (topics discussed) that caused sentiment spikes
- Ranking the extracted topics

Vadim Kangan et al [10] has proposed a paper in which they have explained the methodology by using which they have successfully forecasted 2013 Pakistan election and 2014 India Election. They have used the sentiment scores and sentiment diffusion models for forecasting. They Presented the sentiment score by the graphical representation as shown in **Figure 2.7**.

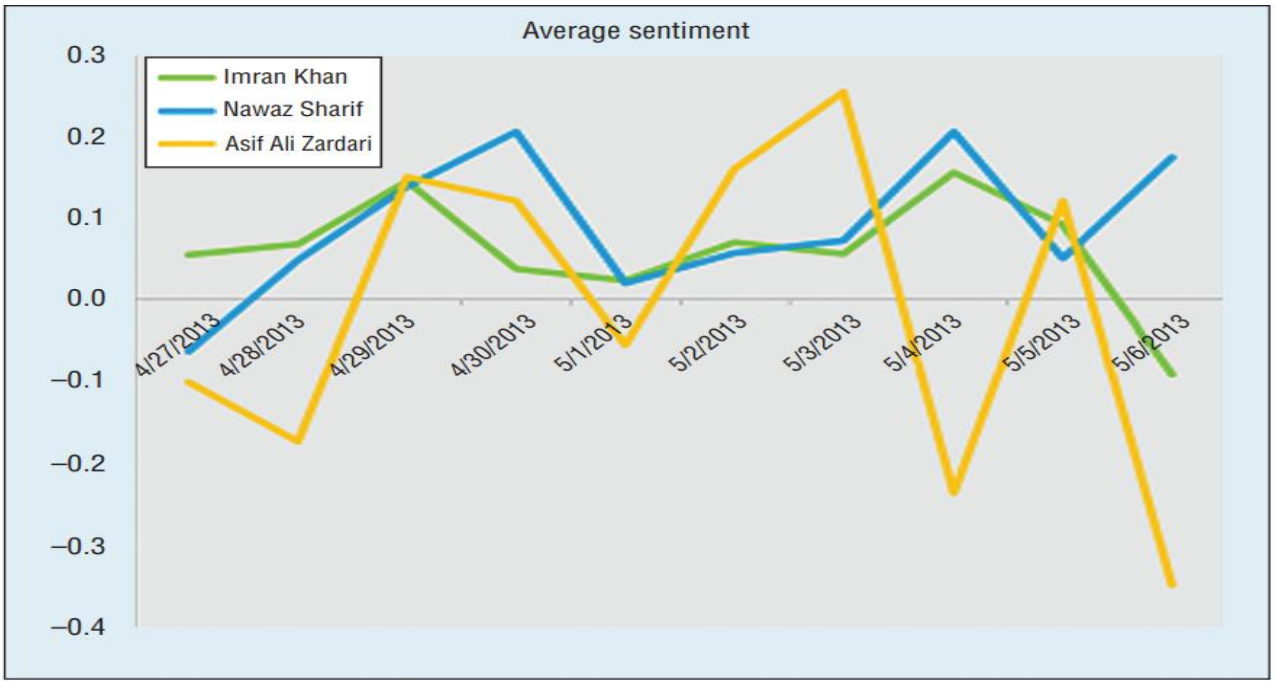

Figure 2.7: Graph representing sentiment score [10]

Sven Rill et al [11] proposed a system known as PoliTwi, which was designed for the purpose of identification of emerging political topics by using Twitter data faster than other standard information channel. In this research work authors have collected around 4,000,000 tweets during parliament election 2013 in Germany from April until September, and they successfully identified top 10 topics.

The system implementation is comprised of different modules: Data selection, Preprocessing, Analysis and Presentation as shown in figure given below.

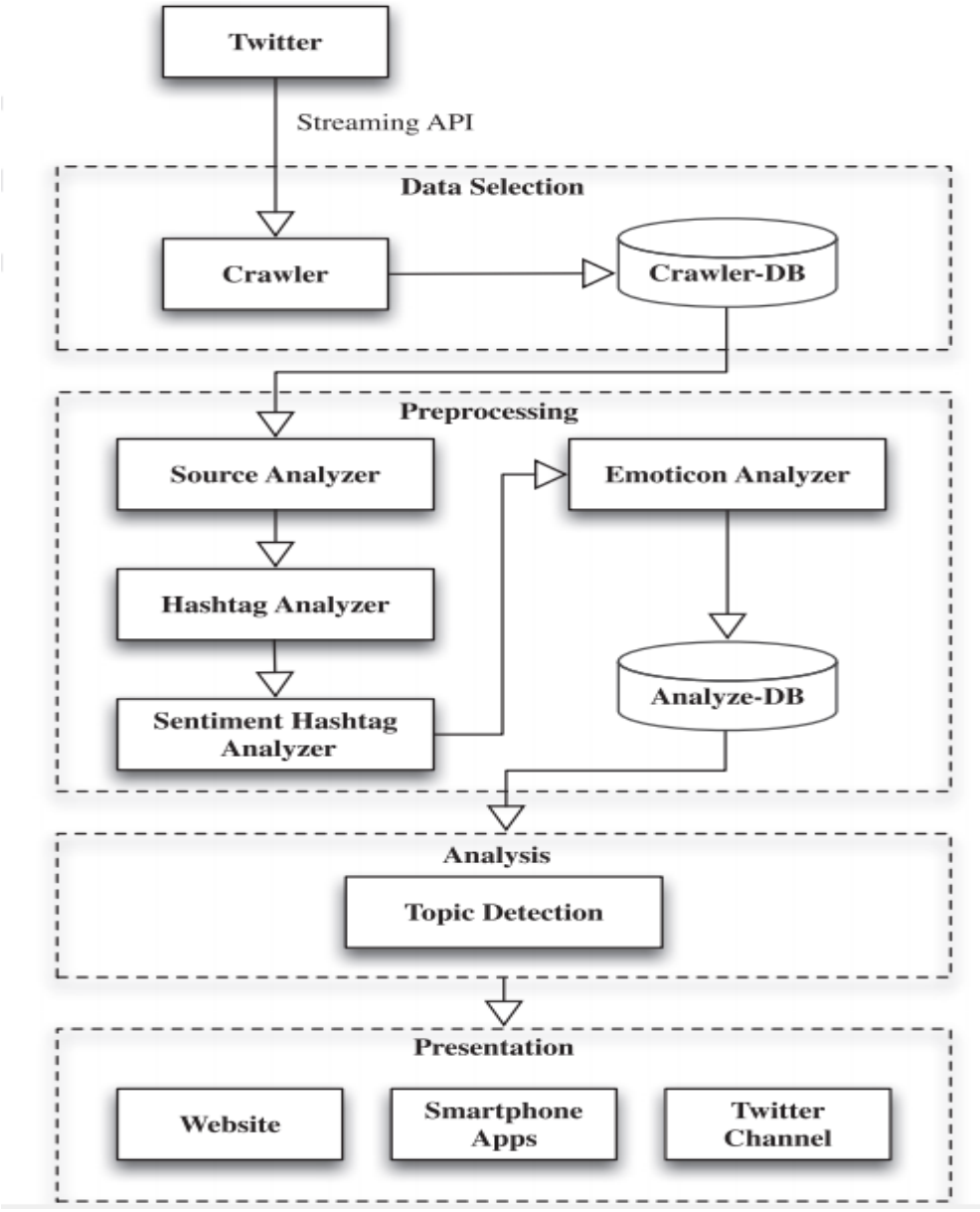

Figure 2.8: System overview of PoliTwi [11]

Zhaoxia wang et al [12] proposed a study describing the existing method for anomaly detection as well as sentiment analysis methods. Moreover, authors had surveyed to highlight the limitations and challenges in semantic analysis. And, to tackle the identified challenges a new methodology based on enhanced sentiment classification is also purposed by the authors. Further the methodology is applied to perform anomaly detection by performing sentiment analysis on the social media data. In this paper data is collected from Twitter with the help of twitter API. While extracting the tweets location constraining geo codes are used to fetch tweets only of Singapore origin. Further for the sentiment analysis lexicon based classifier is used. And, sentiment toward entity is classified into one of the following category:

- Positive
- Negative
- Neutral
- Ambivalence (When both positive and negative sentiments present)

Further authors have explained the method to detect the possible risks on the basis of people sentiment graph as shown in Figure 2.9.

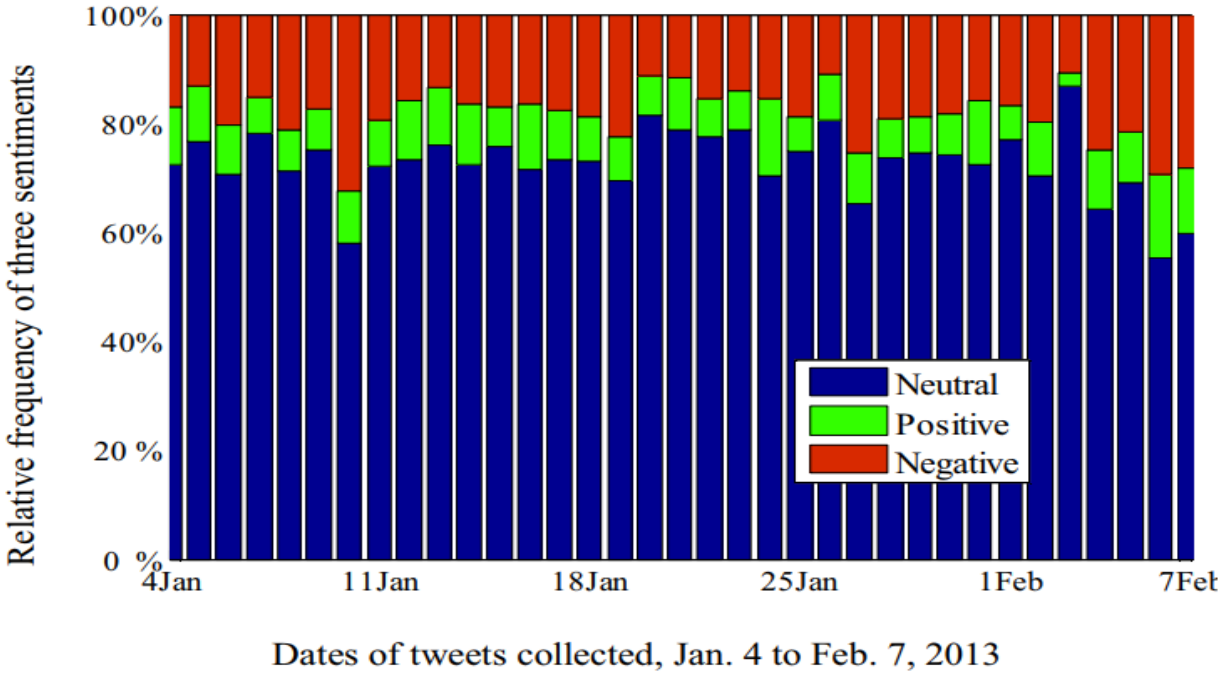

Figure 2.9: Sentiment bar graph [12]

D.M.E.M Hussein et al[13] presented survey paper in which they tried to identify the major challenges in sentiment analysis. The authors taken into account 47 previous researches. Basically, research is based on two comparisons:

- First comparison discusses the relationship between the sentiment analysis challenges and review structure.
- Second comparison examines a significance of solving the sentiment challenges to improve accuracy.

Some of the challenges in different researches are: negation handling, domain dependence, bipolar words, huge lexicon, spam and fake detection, NLP overheads etc.

Monisha Kanakaraj and R.M.R Guddeti[14] proposed a natural language based approach to enhance the sentiment analysis by adding the semantics in the feature vectors. This paper involves analyzing the dynamic of the society on a particular topic from Social networking site Twitter. The main idea proposed in the paper is to increase the accuracy of classification by including Natural Language Processing Techniques (NLP) especially Synsets and Word Sense Disambiguation with ensemble methods for classification. The complete implementation of authors work is presented in Figure 2.10.

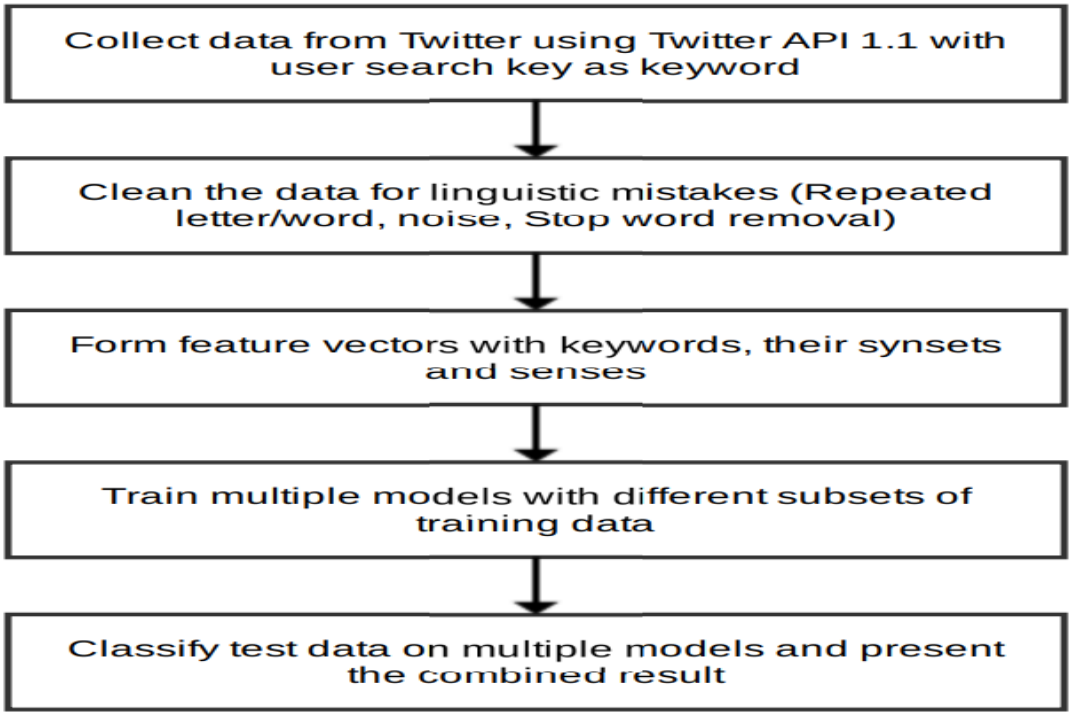

Figure 2.10: Flow chart of proposed system [14]

Work	Task	Methods/Tools	Dataset
[1]	Discover topics, trends and reaction related to 5 Presidential Candidate of 2016 US election.	Latent Dirichlet Allocation (LDA), Stanford Core NLP and Wavelet Transformation.	9,700 Facebook posts And 1,20,50,595 comments.
[2]	Sentiment Analysis on the basis of tweets and comparison of different classifiers.	K-Nearest neighbor, Random Forest, BaysNet, Naive Bays and SentiwordNet dictionary.	21,02,52 tweets about political leaders.
[3]	Personalized Tool to extract data from Facebook.	Java based application and Facebook graph API.	First set of data contain 46 posts,217 comments and 15,994 reactions
[4]	Develop Prototype for Marketing Decision.	CRISP-DM (Cross-Industry Standard Process for Data Mining).	Clean Corpus of 1,47,900 images from Flickr and Instagram.
[6]	Entity detection, Sentiment Analysis and Sarcasm detection on the basis of tweets related to greek referendum and election.	Sentistrength and SVM (support Vector machine)	301000 tweets (#dimopsifism) and 182000 tweets(#ekologes)
[7]	Sentiment Analysis of tweets about Arvind Kejrival during 2015 Delhi election.	SentiWordNet, WordNet and Word sense Disambiguation.	12000 tweets about Arvind Kejrival.
[8]	predict 2015 UK election	sentiment analysis, prior, elections' information	1,38,99,073 tweets
[9]	Identify sentiment spikes	LDA algorithm and KL divergence	10,76,732 + 12,65,001 + 13,69,756 tweets
[10]	predict 2013 Pakistani and 2014 India elections	AVA algorithm adaptation	31 topics, TIEN, IET-DB
[11]	Detect emerging political topics	Calculation of Topic Value	4M tweets
[12]	Anomaly detection via sentiment analysis	Lexicon based classifier	Tweets from Jan 4 to Feb 7, 2013 (Singapore origin)

Table 1: Summary of literature review.

Chapter 3
SCOPE OF THE STUDY

Most of the people around the world has their account on social networking sites (such as Facebook, Twitter etc.). Moreover, people spend a healthy amount of time on these social sites to upload their daily routine. Taking these facts into account, researchers have researched a lot utilizing the data generated on these social networking sites and developing various applications for different fields. In commercial field social networking data mining is done, as it is helpful in:

- Product/services planning
- Trend analysis
- Advertisement
- Feedback for products and services
- Customer relationship management etc.

Similarly, social media data mining become very popular in field of EDM (Educational data mining). As social media data is very helpful in extracting the student problem and addressing them.

Nowadays most hot research area in social media data mining is developing a system which helpful in explain current social and political trend. In recent research social media data mining results are better the traditional polls. Moreover, social media data analysis is capable enough to portrait the social character of any public figure (celebrity, politician). Taking this into account our research aim is to develop a system which can assist the people in understanding the current social and political trends. It is planned to produce a final product which is user friendly and close to real time, so that a large number of people can enjoy service of developed product

Chapter 4
OBJECTIVES OF THE STUDY

Today's world is full of competition and fast paced, people are interested in abstract of a thing rather than complete knowledge. So, our motivation behind this research is to produce an efficient system that can assist people in understanding social and political dynamic through the summary of actors (celebrity, leader, other social representative) of the society. Our research goal will be to produce an efficient system in terms of: accuracy of results and real time working.

Given below are definitions of terms used in objectives-

i. **Actor-** Term actor represent a person (celebrity, leader and social representatives).
ii. **Confidence-**Term confidence is used to define people sentiment related to an actor.
iii. **Topics-**Topics define the issues and events with whom an actor is more concerned.
iv. **Summary presentation interface-** It is group of graphs and plots.

Thus, the main objectives of our research can be defined as:

- To develop a methodology that is capable enough to find out confidence of a social actor considering specific time and event.
- To develop a methodology that can extract the topics in which target social actor is interested, further comparison among actors/entities can be done on the basis of topic distribution.
- To propose a summary presentation interface of the target individual.

Chapter 5
RESEARCH METHODOLOY

In proposed methodology R studio (IDE) will be used for the purpose of data analysis.

Proposed methodology for analysis include four major phases-

A. Collection of data of target figure from his/her social media account.

B. Preprocessing of data.

C. Mining the data to find hidden patterns and trends

D. Presenting the different Knowledge Pattern

A. Collection of data of target figure from his/her social media account

Basically, this methodology is focusing on extracting data from two platforms-

1. Facebook

The simple way to connect to Facebook API and extracting the data is by using the Facebook APP. Following are the steps involved in this-

a) Install and load require package in R studio.

b) Create app on Facebook developer platform

c) Establish Connection between R studio and Facebook through app authentication key.

d) After successful establishment of connection, extract the required data.

There are many other methods available for data extraction such as Facebook Graph API and third-party tools.

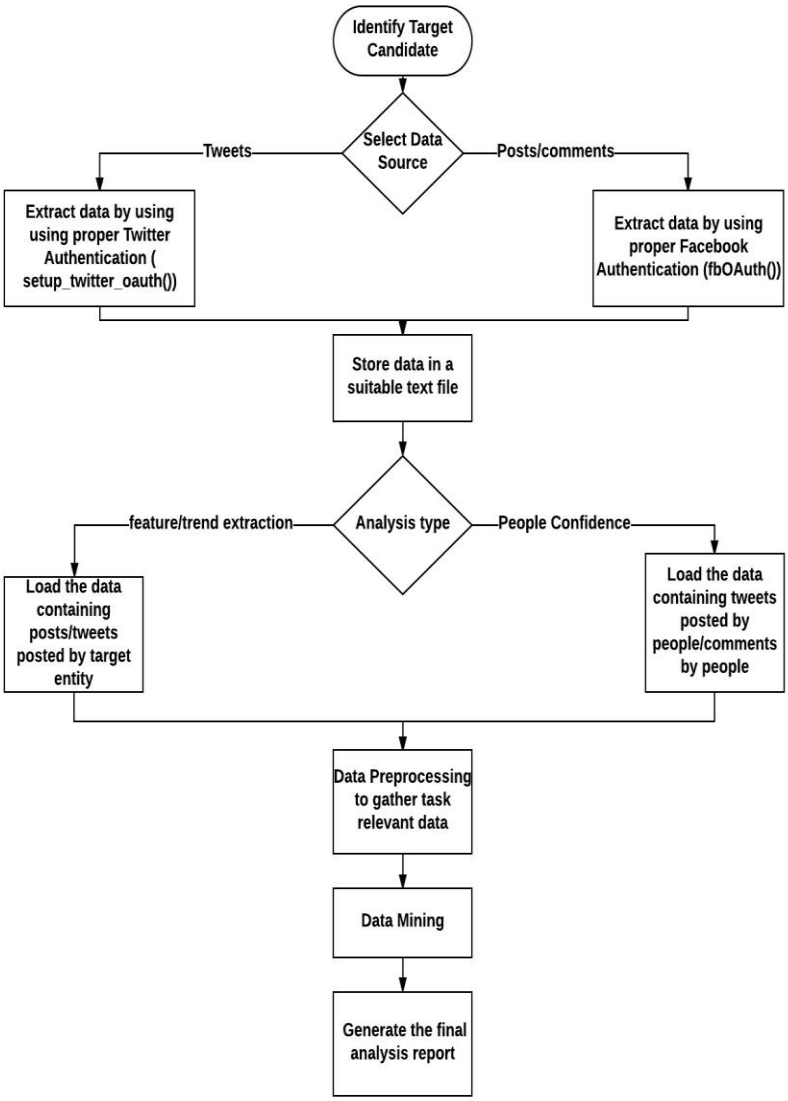

Figure 5.1: Representation of methodology.

2. Twitter

In order to extract the data from twitter we can use Twitter App. Following are steps involved in this-

a) Install and load required packages in R studio to support twitter connection.

b) Create an app on twitter and get the Consumer key (API key) and Secret key (API Secret).

c) Establish the connection between twitter app and R studio with the help of authentication key mentioned in previous step.

d) After successful authentication extract the required data.

B. Preprocessing of Data

Initial phase helps in removal of all other data present in all other languages except the English language. For the next level of preprocessing first we need to load the target file in a user defined object. Further the target data columns are loaded into the data corpus (collection of documents containing (natural language) text. In next step the operations are carried out on the data corpus to clean the data. Figure 5.2 shown below presents the complete procedure of preprocessing.

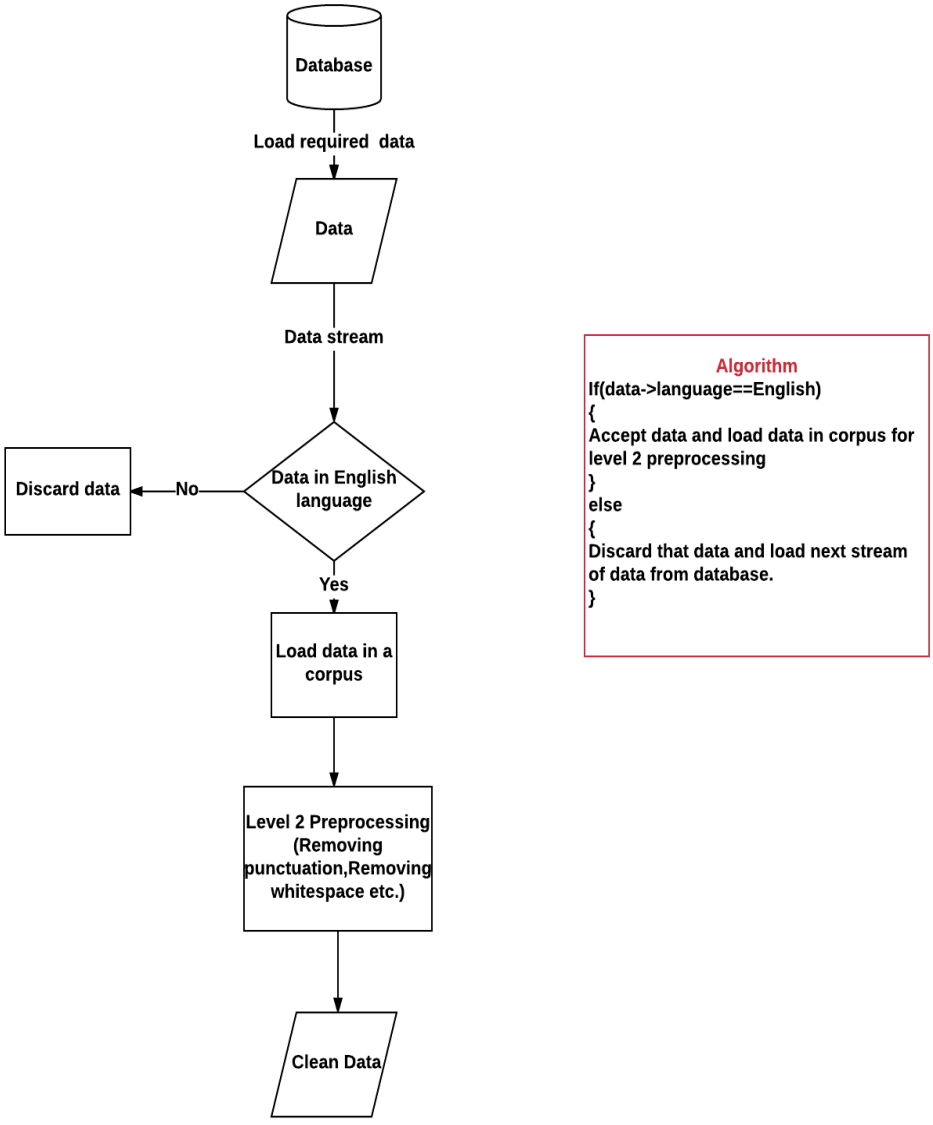

Figure 5.2: Data preprocessing steps

Level 2 of preprocessing of data basically involves:

a) Removing punctuations-English language is supported by different punctuation marks such as dot (.), coma (,) etc. However, punctuations are meaningless whenever we have to perform analysis, so it is become important to remove the punctuation marks.

b) Removing white spaces-It might be possible that extracted text contain unwanted whitespaces, which may act as noisy data during analysis. For better results, it is advisable to remove the white spaces.

c)Converting all the text into lower case-Most of the analysis/mining code treat are case sensitive, so to reduce errors it is advisable to have our all our text in same case (lower case or upper case.

d)Removing the stop words of English-Stop words are comprised of general words which is to support our sentence such as I, me, my, do, should etc. However, these are not important from the data analysis point of view, so it is advisable to remove such words.

e) Stemming of the words- For the analysis purpose it is important to convert all the words to base words such as played is converted to play.

f) Removing other unwanted words or symbols- It might be the case that we want to remove certain targeted word from the text file for more specific results.

C. Mining the data to find hidden pattern and trends

In our proposed methodology we are basically focusing on analysis for the purpose of feature extraction of target candidate and confidence of people related to the candidate. So for the general analysis we can use the Data Corpus and Document term matrix for presenting the generating the knowledge based on the frequency of words such as word cloud and 'n grams'.

For more depth analysis one need to perform natural language analysis. There are certain open sources libraries are available that we can integrate with R studio such as OpenNLP (provided by Apache). The basic aim of semantic analysis is to generate a plot that represents the score of a candidate on the basis of kind of word used in his/her posts/tweets. Further we can identify confidence of a candidate among people with the help of semantic analysis of people comment and tweets data.

To find out the confidence score the people opinion data can be broken down into single words (tokens). For the purpose of converting the text document into stream of tokens some natural language processing tool is required such as OpenNLP by Apache. Further with the assistance of

NLP tool word can be categorized into a scale (-5 [very negative word] to +5 [very positive word]) as shown in **Figure 5.3**.

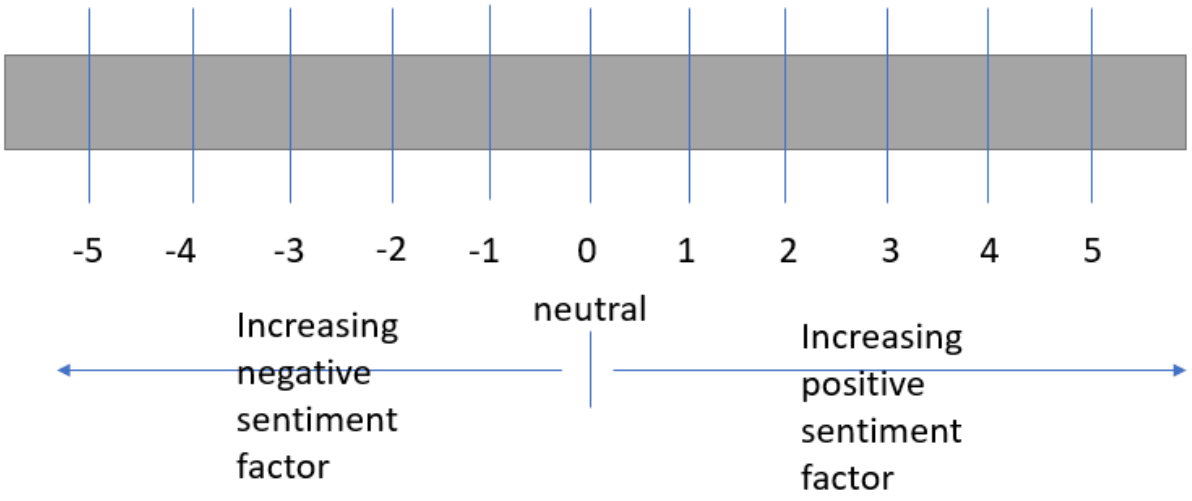

Figure 5.3: Proposed scale for calculating sentiment score for each word/token.

After finding the sentiment factor of each token in the text document, the next step involves summation of sentiment factors which gives overall sentiment score of the document. Given below is the mathematical expression defining the method to calculate overall sentiment score of group of tokens in text document-

$$f(text\_tokens) = \sum_{i=1}^{n} (a_i) - [sarcasm\ factor]$$

n=total no of tokens in the text

a_i=associated impact factor with ith token

Sarcasm factor is based on the degree of sarcasm present with in text document

We can extend this analysis and can present the different topics/issues upon which a candidate is more focused. Basically, the distribution of topics related to target candidate in extension of the previous work done in the base paper presented by Saud Alashri et al [1].

D. Presenting the Knowledge

To present the knowledge different graphical structures can be used-

- Word Cloud is an image comprised of words; in which size of word depend on the frequency of word in the document.

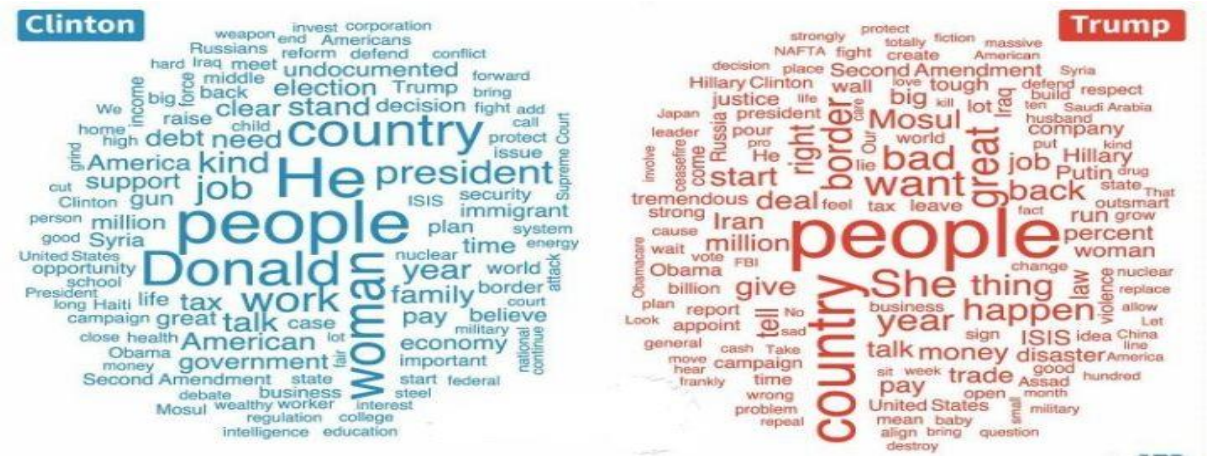

Figure 5.4: Word Cloud

- N- gram in Corpus is to present the group of words which are used together.
- Bar graphs representing the sentiment score related to different emotions as shown in **Figure 5.5**.

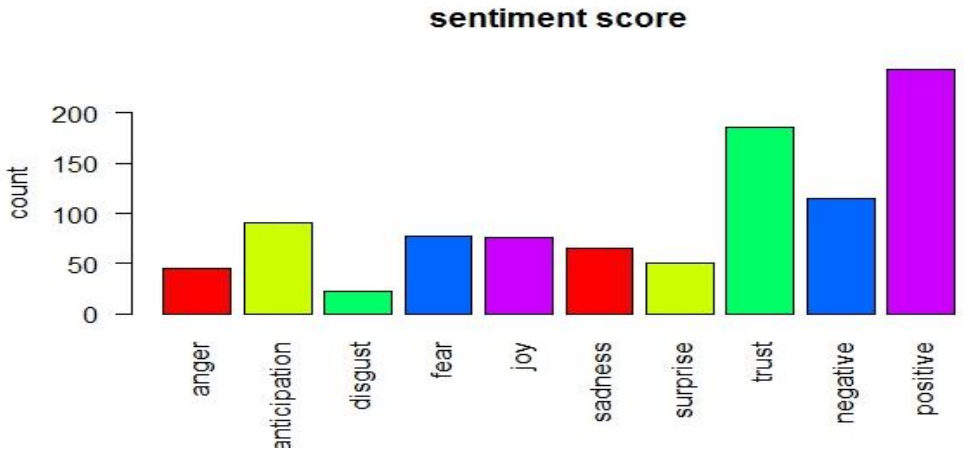

Figure 5.5: Bar graph representing sentiment score.

Further we can import the Google visualization in R studio to present other trends with the help of more versatile plots.

Chapter 6
SUMMARY AND CONCLUSION

The main motivation behind this research is to define the methodology by using which people can perform the data analysis to find out the issues on which actors (celebrity, leader, other social representative) of society are focusing. Moreover, methodology can be used by the actors to check their reputation and confidence among people via semantic analysis of comments on their posts. In future, our target is to build a functional model on the basis of the proposed methodology. The research work will also be carried out to extend the previous work of Saud Alashri et al [1] to find out more efficient way to find out topic distribution curve.

LIST OF REFERENCES

I. Books

Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts And Technique (third edition)", Morgan Kaufmann,USA.

II. Research papers

[1] S. Alashri, S. S. Kandala, V. Bajaj, R. Ravi, K. L. Smith, and K. C. Deusouza, "An Analysis of Sentiments on Facebook during An Analysis of Sentiments on Facebook during the," no. August, pp. 795–802, 2016.

[2] A. P. Jain and V. D. Katkar, "Sentiments analysis of Twitter data using data mining," *2015 Int. Conf. Inf. Process.*, pp. 807–810, 2015.

[3] A. Teixeira, "Data extraction and preparation to perform a The example of a Facebook fashion brand page."

[4] J. Treboux, F. Cretton, F. Evéquoz, A. Le Calvé, and D. Genoud, "Mining and visualizing social data to inform marketing decisions," *Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA*, vol. 2016–May, pp. 66–73, 2016.

[5] M. Korakakis, E. Spyrou, and P. Mylonas, "A survey on political event analysis in Twitter," *Proc. - 12th Int. Work. Semant. Soc. Media Adapt. Pers. SMAP 2017*, no. March 2006, pp. 14–19, 2017.

[6] D. Antonakaki, D. Spiliotopoulos, C. V. Samaras, S. Ioannidis, and P. Fragopoulou, "Investigating the complete corpus of referendum and elections tweets," *Proc. 2016 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2016*, pp. 100–105, 2016.

[7] R. Jose, "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation," no. November, pp. 638–641, 2015.

[8] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, "140 characters to victory?: Using Twitter to predict the UK 2015 General Election," *Elect. Stud.*, vol. 41, pp. 230–233, 2016.

[9] A. Giachanou, I. Mele, and F. Crestani, "Explaining Sentiment Spikes in Twitter," *Proc. 25th ACM Int. Conf. Inf. Knowl. Manag. - CIKM '16*, pp. 2263–2268, 2016.

[10] V. Kagan, A. Stevens, and V. S. Subrahmanian, "Using twitter sentiment to forecast the 2013 Pakistani election and the 2014 Indian election," *IEEE Intell. Syst.*, vol. 30, no. 1, pp. 2–5, 2015.

[11] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, "PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis," *Knowledge-Based Syst.*, vol. 69, no. 1, pp. 24–33, 2014.

[12] Z. Wang, V. Joo, C. Tong, X. Xin, and H. C. Chin, “Anomaly detection through enhanced sentiment analysis on social media data,” *Proc. Int. Conf. Cloud Comput. Technol. Sci. CloudCom*, vol. 2015–Febru, no. February, pp. 917–922, 2015.

[13] D. M. E.-D. M. Hussein, “A survey on sentiment analysis challenges,” *J. King Saud Univ. - Eng. Sci.*, no. April, 2016.

[14] M. Kanakaraj, R. Mohana, and R. Guddeti, “NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers,” *Signal Process. Commun. Netw. (ICSCN), 2015 3rd Int. Conf.*, pp. 1–5, 2015.

III. Websites

https://www.omnicoreagency.com/facebook-statistics/

https://www.omnicoreagency.com/twitter-statistics/

https://www.lucidchart.com/

http://ieeexplore.ieee.org/Xplore/home.jsp