TWITTER BASED SENTIMENTAL ANALYSIS FOR PREDICTING ELECTION

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

Shanu Sahni

Registration number

11501404

Supervisor

Mrs. Maneet Kaur

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

Month April, Year2017

ABSTRACT

Data mining is a task which is used to find out the hidden pattern or information to analysis any subject. Now a day's alot of research is going on web mining i.e.to mine the web content for analysis. Web mining can be further classified into following categories i.e., static web mining, dynamic web mining, dynamic web mining is also known as Data Stream (DS).

In our research the main aim is to perform the text mining over the real time data to predict the result of election that which party will win the state election held on India. In our work we get the data from social media (tweeter) where the citizen gives their opinion towards different political party and analysis the sentiments to conclude the result.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled **"TWITTER BASED SENTIMENTAL ANALYSIS FOR PREDICTING ELECTION"** in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mrs. Maneet Kaur I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Shanu Sahni

Reg. No. 11501404

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled **"TWITTER BASED SENTIMENTAL ANALYSIS FOR PREDICTING ELECTION"**, submitted by **Shanu Sahni** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Mrs. Maneet Kaur

Date:

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: _____
   HoD Name: _____
   Date: _____

2) **Neutral Examiners:**

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   Internal Examiner

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

I have taken efforts in this dissertation-II. However it, would not have been possible without the kind support and help of my mentor **Mrs. Maneet Kaur.** I would like to extend my sincere thanks to him. I am highly indebted for his guidance and constant supervision as well as for providing necessary information regarding the dissertation and also for their support in completing the Dissertation-II.

I would like to express my gratitude towards **Er. Dalwinder Singh HOD (CSE),** members of Lovely Professional University for their kind co-operation and encouragement which help me in completion of this dissertation.

My thanks and appreciation also go to my colleagues in the doing the Dissertation –II and people who have willingly helped me out with their abilities.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

Checklist for Dissertation-II Supervisor

Name: _____ UID: _____ Domain: _____
Registration No: _____Name of student:_____

Title of Dissertation:

_____

- ☐ Front pages are as per the format.
- ☐ Topic on the PAC form and title page are same.
- ☐ Front page numbers are in roman and for report, it is like 1, 2, 3…….
- ☐ TOC, List of Figures, etc. are matching with the actual page numbers in the report.
- ☐ Font, Font Size, Margins, line Spacing, Alignment, etc. are as per the guidelines.
- ☐ Color prints are used for images and implementation snapshots.
- ☐ Captions and citations are provided for all the figures, tables etc. and are numbered and center aligned.
- ☐ All the equations used in the report are numbered.
- ☐ Citations are provided for all the references.
- ☐ **Objectives are clearly defined.**
- ☐ Minimum total number of pages of report is 50.
- ☐ Minimum references in report are 30.

Here by, I declare that I had verified the above mentioned points in the final dissertation report.

Signature of Supervisor with UID

CHAPTER 1

INTRODUCTION

Data mining is a task which is used to find out the hidden pattern/information to analysis any subject. Now a day's a lot of research is going on web mining i.e.to mine the web content for analysis [1]. Web mining can be further classified into following categories i.e., static web mining, dynamic web mining, dynamic web mining is also known as Data Stream (DS).

Stream mining is an area that gaining lots of practical significance and finding various application areas related to medicine, computer science, stock market prediction, online data generation etc. Since in web technology (stream data) has a challenging task because they are real time data which changes rapidly over the time [5]. In-stream mining, a huge amount of online data is generated from several things like sensors, internet relay chat, twitter, Facebook, online transactions etc.

Information Stream Mining is the way toward separating learning structures from persistent, fast information records. Figure 1.1 shows how mining will perform and if the memory is full then its reverse back otherwise knowledge integrated.

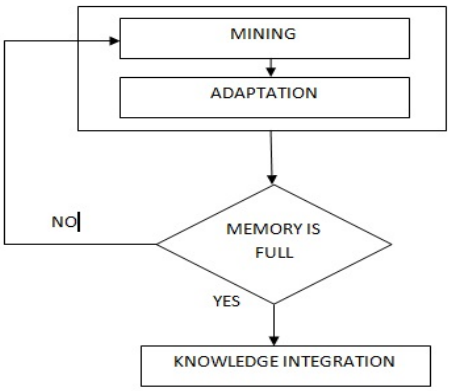

Figure 1.1: Flow Diagram of Stream Mining

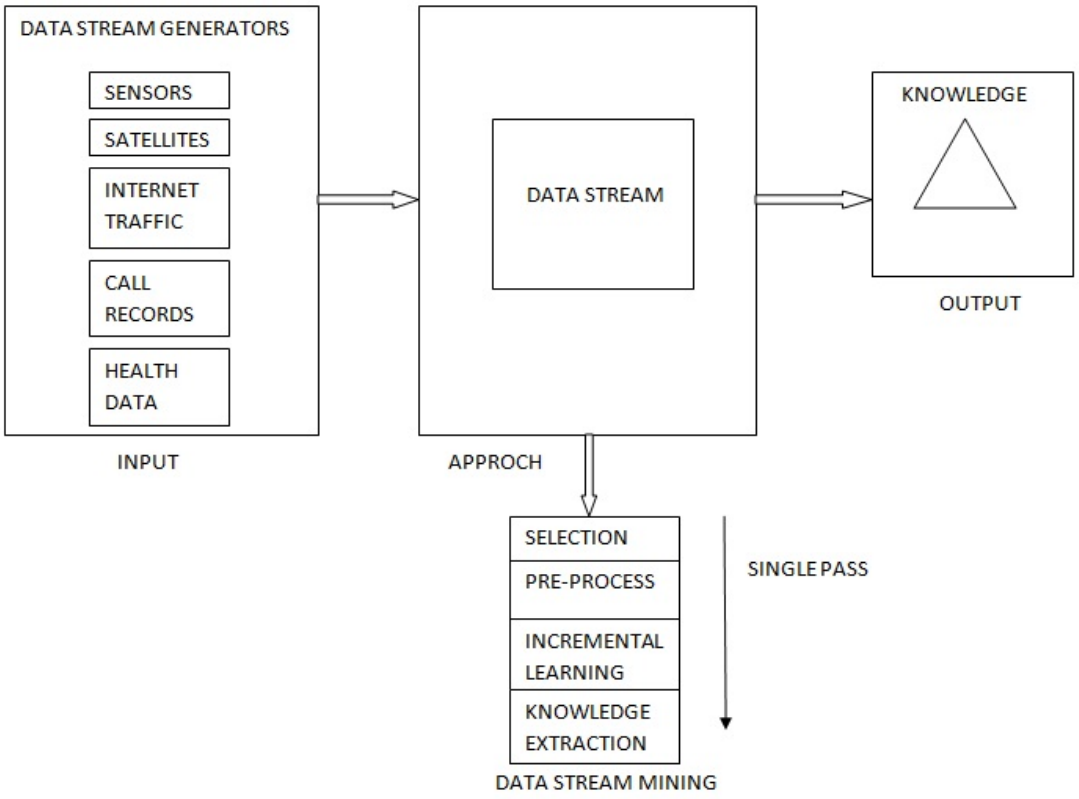

Figure 1.2: Process of Data Streaming

Stream mining over twitter data is an area where lots of research is going on because twitter is miniaturized scale blogging administration that checks with a huge number of clients from everywhere throughout the world [23]. Sentimental analysis has improved in the last few year as well as its applications. The figure 1.2 shows that how data were taken and perform data stream to extract knowledge. This is used for product marketing for recognition of anti-social behavior. The advances in Facebook, twitter, YouTube and other smaller scale blogging and long range informal communication destinations have contributed change to the social locales as well as have in a general sense changed the way we utilize these locales and how we share our emotions, our perspectives with the more extensive gathering of people [22]. A huge number of messages seen day by day in prominent sites that provide access to smaller scale blogging [18]. Users write about their life and share opinions on a variety of topics and discuss current issues happened in the world. In light of a free arrangement of messages and a simple availability of miniaturized scale blogging stages, Internet clients tend to move from customary

specialized instruments, (for example, conventional web journals or mailing records) to small scale blogging administrations. The same number of clients post about items and administrations they utilize or express their

Political and religious perspectives, smaller scale blogging sites get to be distinctly profitable wellsprings of individuals' assessments and suppositions. Such information can be productively utilized for versatile UI [16]. Data we get from these sources can be used in opinion mining and sentiment analysis tasks. For example:

i. What do people think about these persons who post their status (comments)?
ii. How positive (or negative) are people about anything?

Previous work incorporates by the author Turney and Pang who connected the diverse techniques to identifying the extremity of an item survey of the product and film audits individually. There work is at the record level [23]. One can likewise arrange an archive's extremity on a multi-way scale, which was endeavoured by Pang and Snyder among others: Pang and Lee extended an essential assignment of grouping a motion picture survey either positive or negative to foresee star evaluations on either a 3 or a 4 star scale, while Snyder played out a top to bottom investigation of eatery audits, anticipating appraisals for different parts of the given eatery, for example, the sustenance and air (on a five-star scale) [24]. Despite the fact that in most actual arrangement techniques, the impartial class is overlooked under the presumption that unbiased writings that lie too close the limit of the paired classifier in which a few analysts recommend that, in each extremity issue the three most important classifications must be recognized. Additionally, it can be demonstrated on particular classifiers, for example, the Max Entropy and the SVM can profit by the presentation of a nonpartisan class and enhance the general exactness of the characterization [25]. An alternate strategy for deciding slant is the utilization by a scaling framework where the words ordinarily connected with a negative, nonpartisan or positive notion with them were given a related number as a - 10 to +10 scale (most negative to best This makes it conceivable to changed slant by a given term with respect with its surroundings (for the most part on that level of the sentence) [22]. At the point when a bit of no structured content is examined utilizing common dialect preparing, every idea in the predetermined environment is given the score to view the

way of feeling words that identify with an idea and its related score. This permits development to a more modern comprehension of opinion since it is currently conceivable to confirm the assumption estimation of an idea with respect to changes that may encompass it. Words, for instance, that escalate, unwind or refute the estimation communicated by the idea can influence its score.

1.1 TEXT MINING

Content mining additionally alluded to as content information mining, generally comparable to content examination, is the way toward getting top notch data from content. Amazing data is ordinarily determined through the contriving of examples and patterns through means, for example, factual example learning.

Content mining is the examination of information contained in regular dialect content. The use of content mining systems to take care of business issues is called content investigation [27].

Content mining can help an association infer conceivably significant business bits of knowledge from content based substance, for example, word records, email and postings via web-based networking media streams like Facebook, Twitter, and LinkedIn. Mining unstructured information with regular dialect preparing (NLP), measurable displaying and machine learning systems can challenge, notwithstanding, on the grounds that characteristic dialect content is frequently conflicting [32]. It contains ambiguities created by conflicting linguistic structure and semantics, including slang, dialect particular to vertical businesses and age gatherings, two-sided sayings and ridicule.

1.2 DATA FILTRATION

The data filtration is also known as data pre-process which means information sifting in IT can allude to an extensive variety of techniques or answers for refining informational collections. This implies the informational indexes are refined into essentially what a client (or set of clients) needs, without including other information that can be dreary, unimportant or even delicate [31]. Diverse sorts of information channels can be utilized to correct reports, inquiry comes about, or different sorts of data results.

1.3 WHY PRE-PROCESSING REQUIRED

i. Inadequate: lacking characteristic qualities, without specific traits of intrigue, or containing just total information [26].
ii. Loud: containing mistakes or exceptions
iii. Conflicting: containing errors in codes or names, Assignments in information pre-preparing
iv. Information cleaning: fill in missing qualities, smooth load information, distinguish or expel exceptions, and resolve irregularities.
v. Information joining: utilizing various databases, information 3D shapes, or records.
vi. Information change: standardization and collection.
vii. Information lessening: decreasing the volume yet delivering the same or comparable investigative outcomes.
viii. Information discretization: some portion of information diminishment, supplanting numerical qualities with ostensible ones.

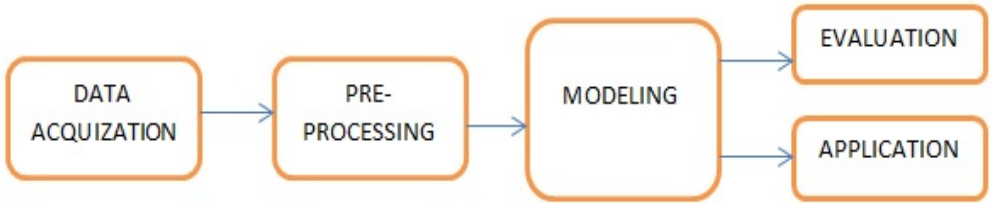

Figure1.3: Process of Text mining

Information pre-processing is an information mining procedure that includes changing crude information into a reasonable organization [30]. Genuine information is regularly deficient, conflicting, as well as failing in specific practices or inclines, and is probably going to contain numerous blunders. Information pre-processing is a demonstrated strategy for settling such issues. Information pre-processing gets ready crude information for further handling. Figure 1.3 shows that first data acquisition is done then pre-process and in finally, we evaluate the knowledge by using application software.

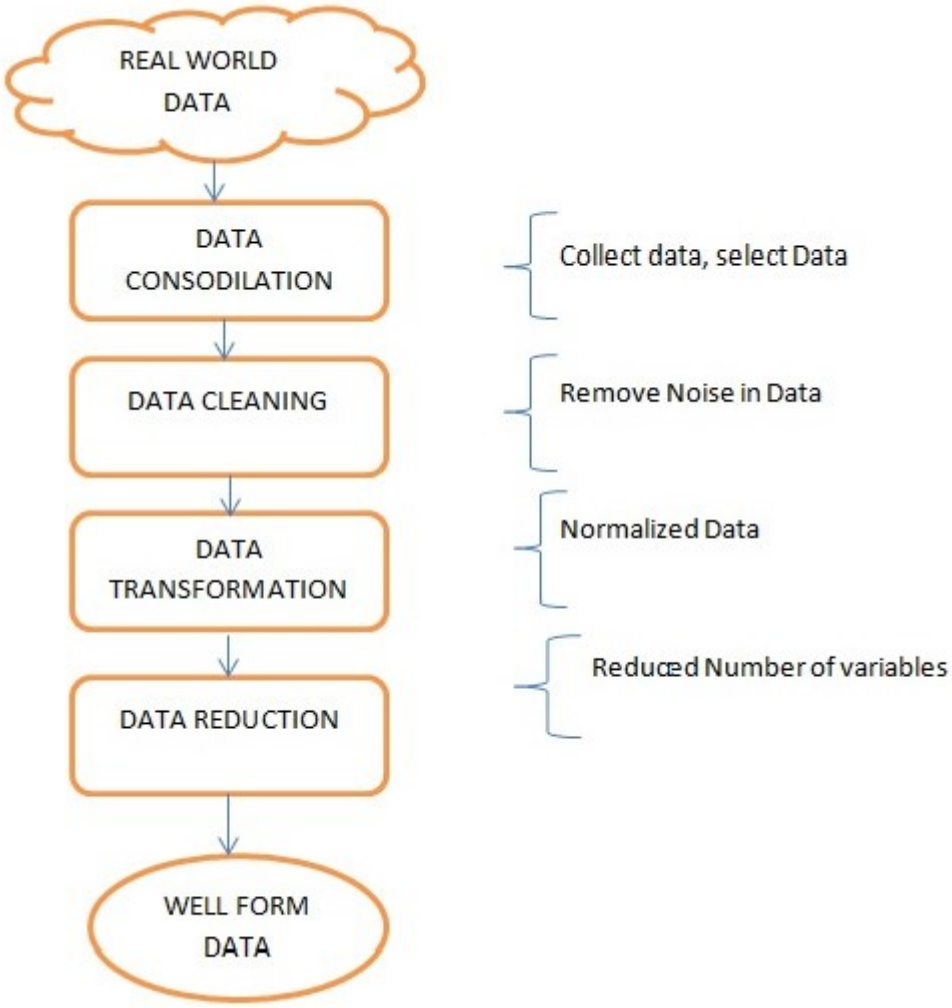

Figure 1.4: Pre-Processing

Figure 1.4 explain the mechanism of the pre-processing of data i.e. data collection, removing noise, normalization of data to remove duplicity and at last find the fine or clean data which is suitable for mining.

CHAPTER 2

LITERATURE REVIEW

M.S.B. Phridvi Raj, "Data Mining- The Past, Present, and Future", 2015, [11] is generally a review paper in which the author tries to catch attention to the importance of the data mining especially the stream mining. This paper provides detail information that how the concept of data mining is use in past and present to fetch the information from the bulk of storage or data as well as try to provide the future of research work which can be done. The paper tells about the data stream which can be classified into two types first one is Offline data stream and the second one is online data stream and the different types of an algorithm which is used in the data mining as well as their disadvantages such as Frequent pattern, Naive Bayes, Decision Tree etc. The author tries to provide the information that the algorithm which is used for Offline data mining is not suitable for the Online data mining because in the online world the data change rapidly according to time so the Online stream data mining is a challenging task in nowadays. The paper also discusses some important challenge occur in the stream data mining such as handling stream clustering, classification and topic detection which is the major research topic for the future work in the field.

Eduard Hromada, **"Mapping Of Real State Price Using Data Mining Techniques", 2015,** [6] author provide that how, Data Mining techniques is used in the mapping of real estate price i.e. the paper describe the software which is used for actual state assessment, mapping and analysing of actual estate ad distributed on the web before developing system the author make a proper Literature Survey find how the prize and advertisement differs from actual fact .The basic function of the software can be explained below

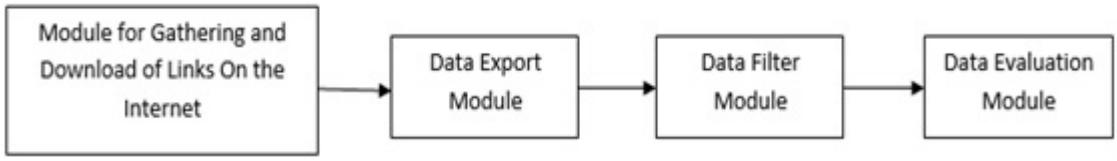

Figure 2.1: The basic structure of software

In figure 2.1 explain the basic structure of software which explain the step by step process of data evaluation which includes data export, data filtrations and then creates the data model.

The information obtaining is sent in the accompanying Categories:-

i. Apartment for the deal.
ii. Apartment for rent.
iii. Apartment for the deal.
iv. Allotment for the deal.
v. Corporate plots for the deal.

So by the help of output which is generated by the software contain the above module is used as a ground for allocating investment or holder decision for both familiar people as well as companies. The paper also provides the challenges which were faced by the developer and still some problem which can be used for the next research work for the same.

Dr. Rajeev Tripathi and Dr. Santosh, "A Quick Review Of Data Stream Mining Algorithms", 2016, [13] author proposed a review about the limitation of traditional data in supporting streaming application have been understood and why we developed the new system for manage the stream data. The paper provided the information that how data mining affects the data streaming, the author review the whole process for data stream for data, proper algorithm and model are analyzed. This paper provided the basis algorithm used in the data streams which can be listed as Segmentation, Clustering, Sampling, Sketching, Aggregation Sliding Window, and Damped Window Algorithm Output Granularity. The paper also provide the challenges such as visualization of data, Efficient querying Mechanism, real data are irregular and unpredictable hence an algorithm should be able to manage the traffic, the technique should be intelligence to differentiate between noise and concept etc. The above challenge given by the author in this paper can be our future work.

V. Punna Rao and Sagar Galanin, "Visualization Of Streaming Data Using Social Media", 2016, [12] the author proposed a new system/method to discuss the existing

analysis of twitter and Facebook dataset with text mining approaches such as Natural Language Processing Algorithm and R- Programming. They give a new approach that automatically classified the sentiments of tweets, posts, comments taken from the twitter as well as a Facebook dataset. These tweets and comments are classified into positive, negative and neutral with respect to a query term of people using graph. The tool is divided into four parts first one data which is taken from Twitter/Facebook, the second one is Natural Language Processing which is used to derived meaning by the computer such as done by a human. The third one is R programming which used for statistical computing and graphics as well as analysis the data and the last one is Lucene which is simple a java based search library which can add in any tool to make the capability of search. The main proposed methodology given by the author in this paper was explained with the help of flow chart i.e.in figure 2.2 which explain the system architecture of the used by the author in which they take data from the web which is in graph format and apply business intelligence and then apply data mining.

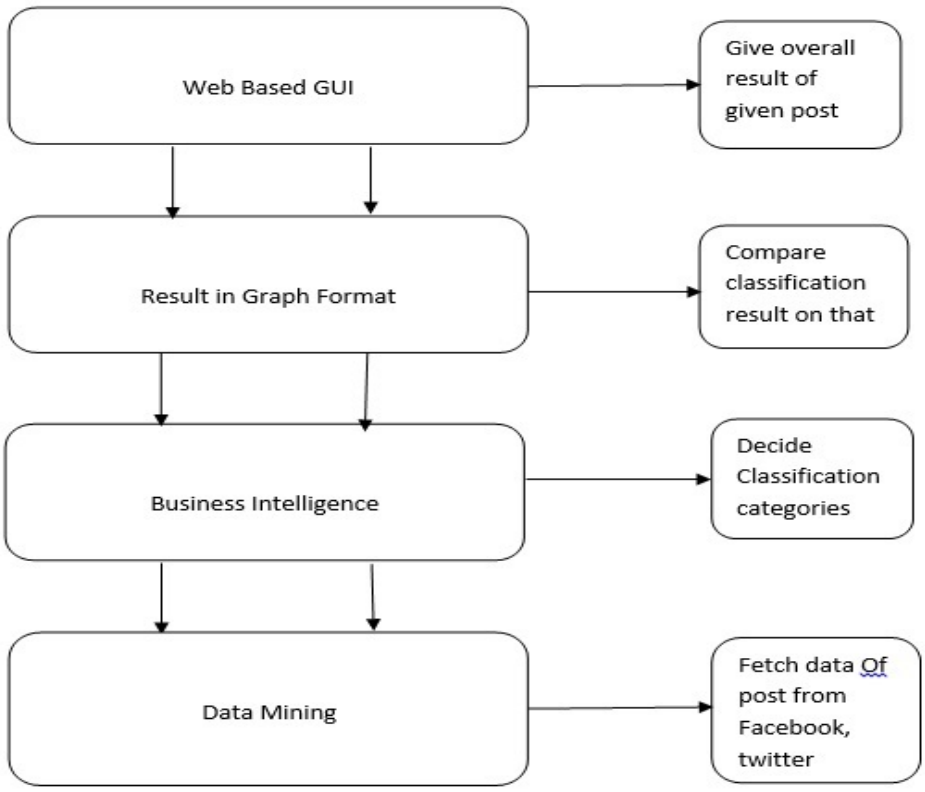

Figure 2.2: System Architecture

Tongwei Yuan and Peng Chen, "Data Mining Application In E-Government Information System", 2012, [2] the author proposed a present more develop Association Analysis Model after it gives an outline of information mining technique and behaviors a formal depiction. This paper told the significance of the security of the information as in nowadays the entire world relies on upon web framework so security of information is the real worry to keep the framework from the infection assault, the spillage of privileged insights, framework disappointment. This paper also provided the security survey done by the FBI and the total loss occurs by the threaten the security of the system. The paper also provides the overview of the different method of data mining such as K-mean algorithms, Decision tree, artificial neural Network etc.

Jerome Treboux, Fabian Cretton, Florian Evequoz, Anne, "Mining and Visualizing Social Data to Inform Marketing Decisions", 2016, [15] the creator indicates how the mining and representation of online networking information can be useful really taking shape of the market choice, they introduce a strategy that plans to comprehend the advertising need of an organization and build up the framework/apparatus that can bolster it, the philosophy was partitioned into two utilize case with Swiss organization, the entire technique utilized and characterized as a part of this paper depended on CRISP-DM (Cross-industry standard procedure of information mining). The entire procedure contain seven stage and the work begin with the distinguishing proof of customer needs through meeting the rest of the means are recognizable proof of objective, characterize of precondition, characterize of post condition, depict of principle stream and the last stride is portrayal of exemption and the assistance of this they made a device which gives the investigation result based on the client criticisms and remark which settle on choice/showcase wanting to enhance the business procedure

K. Santhisree and Dr. A. Damodaramin, **"Web Usage Data Clustering using Dbscan algorithm and Set Similarities", 2003,** [30] the creator attempt to exhibit another harsh set Dbscan grouping calculation which recognizes the conduct of client page guest and request of an event of the visit. They exhibit the exploratory outcomes on MSNBC web route dataset as demonstrate that the roughly set of Dbscan grouping calculation is good productivity for execution bunching on web mining. This paper built up another Rough

arrangement of Dbscan Clustering calculation and introduced an exploratory outcome on msnbc.com which was helpful for finding the client to get designs and the request of visits of the hyperlinks of the every client and the bury bunch likeness among the groups. The roughly set Dbscan clustering algorithm is efficient when compared to the rough set agglomerative clustering As in rough set clustering the elements can be present in more than one cluster (soft clustering), whereas in our proposed algorithm the elements will not occur in other clusters.

Eduard Hromada, **"Mapping of real estate prices using data mining techniques", 2015,** [6] in which the authors portray the imaginative programming that is utilized for land assessment and mapping and investigating of land promotion. The author of this paper offers an objective for fair-minded assessment for value improvement on land advertises. The author presents data in light of broad research and a lot of factual information which has been gathered persistently from the year 2007 until today. The principle idea of this exploration paper was the measurements correlations which can be made by the product empower experts and looks into in the field of land to pick up knowledge on the genuine change in the market costs of land in the Czech Republic. This yield might be utilized a reason that suitable ventures and lodging choices for both regular people and organizations. We have seen a relentless long haul decline of land market costs since on second quarter of 2008. The negative pattern does not appear to an altogether evolving. Albeit land media endeavor to present positive data, there is no sign that costs ought to begin ascending in all districts of the Czech Republic (the special case is just the locale Prague and area Middle Bohemia).

Emanuele Vimeo, Luca One, and Davide Annuitant, "Condition Based Maintenance in Railway Transportation Systems Based on Big Data Streaming Analysis", 2015, [4] the creator say in regards to Streaming Data Analysis (SDA) of Big Data Streams (BDS) for Condition Based Maintenance (CBM) with regards to Rail Transportation Systems. SDA of BDS is the issue of breaking down, displaying and extricating data from gigantic measures of information that ceaselessly originate from a few sources continuously through computational mindful arrangements. Among others, CBM for Rail Transportation is a standout amongst the most difficult SDA issues, comprising of the

usage of a prescient upkeep framework for assessing the future status of the observed resources keeping in mind the end goal to decrease dangers identified with disappointments and to maintain a strategic distance from administration disturbances. The author utilizes the Online Support Vector Regression computational mindful Models for the heuristic approach. For this reason, we propose to abuse the Online Support Vector Regression for overhauling the model when new information gets to be distinctly accessible. Our Proposal likewise comprises a model choice system that can improve the exchange off between Accuracy of the last model and assets required with a specific end goal to play out the model choice stage itself, an extra alluring component, for the most part, ignored because of its computational necessities.

Saranya Balaguru, Rachel Nallathamby, C.R. Rene Robin, "Novel Approach for Analyzing the Social Network", 2015, [1] in this paper the author propose a calculation to prepare an extensive symmetric framework of billion scale chart keeping in mind the end goal to concentrate information from diagram dataset. These fascinating examples are found by calculation of a few Eigen qualities and Eigenvectors. The principle challenge in breaking down the chart information are streamlining the diagram, tallying the triangles, discovering trusses. These difficulties are tended to in the proposed calculation by utilizing orthogonalization, parallelization and blocking methods. The proposed calculation can keep running on exceptionally versatile Map Reduce environment. We utilize an interpersonal organization dataset (Facebook around 2 to 7 TB of information) to assess the calculation. The primary point is to outline an Eigensolver that finds the top-k Eigen estimations of vast symmetric lattice framed from Billion scale chart in a parallel situation. The effective parallel environment for preparing the web scale chart is HADOOP. The plan of Eigensolver needs watchful consideration in picking the calculation. We pick the successive strategy and plan it in a manner that it will keep running in parallel. The principal commitment is to enhance the versatility, adaptability, and proficiency contrasted with other Eigensolver.

Girija Chetty, Matthew White, Farnaz Aether, "Smart Phone Based Data Mining For Human Activity Recognition", 2015, [3] in this paper the author say in regards to

the Automatic action acknowledgment frameworks to catch the condition of the client and its surroundings by abusing heterogeneous sensors, and allow constant observing of various physiological signs, where these sensors are appended to the subject's body In this paper they introduce novel information scientific plan for shrewd Human Activity Recognition (AR) utilizing PDA inertial sensors in view of data hypothesis based component positioning calculation and classifiers in light of arbitrary backwoods, gathering learning and apathetic learning. We inspected a few learning methodologies and discovered sluggish learning, irregular timberlands and outfit learning based ways to deal with being promising as far as movement characterization exactness, show building time for program order, and Confusion network, with trial approval on openly accessible action acknowledgment dataset.

Siti Khadijah Mohamad, Zaidatun Tasir, "Educational data mining: A review", 2013, [7] in this paper the author says in regards to the Data Mining is exceptionally helpful in the field of training particularly when looking at understudy's learning conduct in the web-based learning environment. This is because of the capability of information mining in examining and revealing the concealed data of the information itself which is hard and exceptionally tedious if to be done physically. The motivation behind this survey is to investigate how the information mining was handled by past researchers and the most recent patterns on information mining in instructive research. With respect to future research, maybe we can move our concentration from the e-learning, towards the utilization of long range interpersonal communication devices like Blog and Facebook since these applications as of now increased high prominence among understudies and appropriate to be utilized to connect with the understudies with synergistic learning [2021]. We might, of course, encounter some problems, like difficulties in gathering the log data since these applications are not able to provide us with the logs of learner activity as compared to other e-learning applications, but then again, this can be encountered by integrating the Google Analytics tool into the blog environment and the log data can be exported later for further analysis using the data mining techniques.

Xinzhi Wang and Xiangfeng Luo, "Sentimental space based analysis of user personalized sentiments", 2015, [2] in this paper the author focus on the establishment of user sentimental space obtain from online documents to analysis the user behavior. The author takes the three parameters which are used in their work, these parameters are Affection, sentiments behavior, and attributes of the user. In this paper, the author used the support vector machine to train the system for the analysis of the result. The author also used some basics law in support of their work to analysis the sentiments. These basics laws are as follows:

i. Law of sentimental inertia.
ii. Law of origin asymptotically stable
iii. The law of sentimental confliction
iv. The law of sentimental diffusion

In short, this paper used some basic law and attributes to analysis the user behavior and makes the classification of the user based on their sentimental model alone.

Shoiab Ahmed and Ajit Danti, "Novel Approach for Sentimental Analysis and Opinion Mining based on Sentimental WordNet using Web Data", 2016, [27] in this paper the author novel approach is proposed in view of Sentiment WordNet, which produces tally of score words into seven classifications, for example, solid positive, positive, feeble positive, impartial, frail negative, negative and solid negative words for the supposition mining undertaking and assessed utilizing machine learning calculations like Naïve Bayes, SVM and Multilayer Perception (MLP). The proposed approach is probed motion picture and item web areas and got higher achievement rate regarding precision measured by different devices like Kappa insights with an exactness of 77.7% and has brought down mistake rates. Weighted normal of various exactness measures like Precision, Recall, TP Rate, F-Measure rate portrays higher proficiency rate and lower FP Rate for Naïve Bayes and MLP models. The creator first gathers the information from the web source then evacuate the stop word i.e. pre-handle the information and apply the pre-characterized calculation to get the outcome.

Prashant Kumar and Dhruv Mahajan, "An Approach Towards Features Specific Opinion Mining and Sentimental Analysis Across E-Commerce Websites", 2014, [31] author wants to analysis the user sentimental or opinion about particular products from the different websites where the users are generally active to purchase the goods. Their Research helps for the manufacture as well as the customer before purchasing the product. In their study, they take the different product example like phone, Dell laptops, and their reviews to analyze the behavior of the users for that particular product, and at last, they provided the label to the product as per the analyzed behavior.

D.A. Adeniyi, Z. Wei, Y.Dongguan, "Automated Web Usage Data Mining And Recommendation System Using K-Nearest Neighbour (KNN) Classification Method", 2014, [29] The related work is from K-NN Algorithm, which works on click-stream mining. In this work, they take the data from the random click wise event of the user, and for their field of interest, they make the clusters. The whole information into different categories so that when the user again visited the website they don't need to search their interest of area, in fact, the website automatically refer the interest of the user. Before making the clusters the data should be extracting from the click stream, prepares and remove noise. They use K-NN and Excluding distance for their work.

Saiyan Dai, Ling Chen, "An Algorithm of Mining Frequent Closed Itemsets In The Data Stream", 2016, [26] This paper introduced a new algorithm AFPCFI-DS, for searching the data in FP-tree. It is the enhancement to FP-tree algorithm which provides very fast execution of the system as compared to the previous algorithm and solves the problem of using too much space to search in a moment. According to this paper, AFPCFI-DS is performed very much better as compared with the moment. In this work first, they take the data stream and begin the first tree and repeat it until the condition is not false. Then repeat the new transaction from stream data and add transaction after all these steps delete the oldest transaction from the window until the last condition is not satisfied. By the help of algorithm, the author improved the performance of FP-tree.

K. Kannika Paremeswari, Dr. Antony Selvadoss Thanamani, **"Frequent Item Mining Using Damped Window Model"**, **2014,** [10] author provides the concept of data mining using the Damped Window Algorithm. In Damped Window the input dataset is always less or equal to the user pre-defined window size so that at the particular time of the system can handle a particular number of the transaction which reduced the overloading of the memory as well as increased the performance and accuracy of the system. Another task done by the Author to reduced the cost provides the Algorithm.

R. Mythily, Aish Banu, Shriram Raghunathan, "Clustering Model For Data Stream Mining", 2015, [8] the author try to aggregate the different news content from the data stream. The data stream have different research problems i.e. the user wants a particular topic from the news but he or she cannot find out quickly so author try to develop a cluster based system in which for a particular topic we created a cluster that means the whole dataset is divided into different clusters and each cluster contain specific information so that when the user try to find out any news as per their interest they just go for their cluster and he or she will easily get the results.
Here in this paper, the author used RSS feeds and histogram window concept for making the clusters and implementing the hypothesis.

Shamila Naseer, Muhammad Awais, Khurram Shehzad, Usman Naeem, Mustansar Ali Ghazanfar, "Frequent Pattern Mining Algorithms For Finding Associated Frequent Patterns For Data Stream: A Survey", 2014, [9] author try to say that the purpose of analyzing of huge transaction in database many of algorithm can be used to find frequent pattern. In this paper, the Author used a different algorithm to find the more frequent item set so that they compare different algorithm like Apriori algorithm, FP-Growth algorithm, Rapid Association Rule Mining (RARM), Association Sensor Pattern Mining of Data Stream (ASPS). By comparative analysis, they easily draw the conclusion, by scan and less execution time of many frequent patterns mining algorithm. They produced results on the basis of storage of structure and format of data. On the basis of comparison, it is clear that the ASPS is more flexible and take very less time to produce the results or saves time to drawn the results also very flexible in deletion and

addition of transaction moreover this algorithm does not use a double scan of the database. ASPS used less memory as compared to another algorithm.

Jing Guo, Peng Zhang, Jianlong Tan, Li Guo, "Mining Hot Topics From Twitter Streams", 2012, [15] the author wants to say that in earlier we mining hot topics from the Internet. The Internet is a huge source of code used Clustering algorithm where we get a huge amount of data and we face difficult to get accurate output because as data is huge so processing time is also more. The procedure is first we take the data from Internet i.e. data acquisition then use web text pre-processors then use Text vectorization after that make text clusters finally evaluate the hot topics and get the output of hot topics.

But now these days mining the hot topics are quite easy because of Twitter analysis. The Twitter text used sparse attributes and rapidly spreading to detect the hot topics with the help of frequent pattern stream mining. The technique of Twitter is Ist take data from Twitter then use data acquisition then used twitter text pre-processors then used frequent twitter topics mining and at the end get the output. The implementation steps are less and get more accurate results also more flexible explain in figure 2.3 which shows the process of extraction of hot topics from twitter which includes twitter for data source then apply pre-processing then mine the frequent item and on that basis find the hot issue.

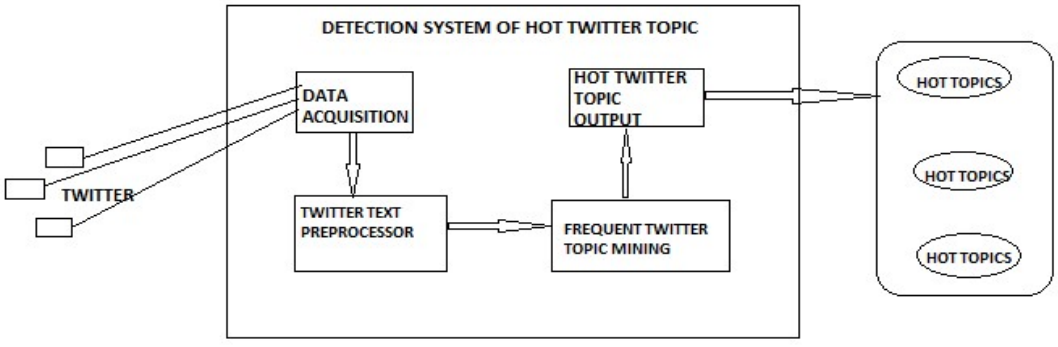

Figure 2.3: Detection System of Hot Topic Twitter

Zhijuan Xu, Lizhen Liu, Wei Song, Chao Du, "**Conflicting Views Analysis Algorithm"2015,** [20] author try to light upon the day by day increasing the popularity of social networking websites. In today's life, most of the people express their feelings on the Internet i.e. social networking websites; these are blogs (Personnel), reviews, and editorials. The sentimental analysis is one of the mind readers of human by posting their reviews on the Internet or tries to know the views of people. In this paper, the Author used SVM, Naïve Bayes, and Decision Tree Classifier. The main motive of this paper is to know the views or feelings of people by their emotions that they share on the internet and what should the effect on society. At last, by comparison of SVM and Naïve Bayes, the Naïve Bayes is better than SVM.

Adhi Dharma Wibawa, Pramana Yoga Chandra and Kusuma Surya Sumpeno, "Social Media Analysis of BPS Data Availability in Economics Using Decision Tree Method", 2016, [21] The main aim of Badan Pusat Statistik (BPS) is to provide the reliable and accurate data for the benefit of society and Government. For analysis the advantage of BPS, they take data from society and door to door and social media (Twitter).And try to compare those results with the existing data for better enhancement of BPS. Mostly they used Twitter data because data (text) of Twitter is short in length or sparse attributes and rapidly spreading and mine that twitter data and make the different classes by Decision tree method and Analysis the data through social media (Twitter). They used methodology very first they collect the tweets from Twitter then pre-process it and use classification by using Decision Tree then analysis BPS data availability in most used keywords ten identify that keywords and at last after evaluation they evaluate results. The structure is explained in figure 2.4 which explain the procedure of BPS which includes tweets from twitter pre- process then classified into a decision tree, then validate it and analysis the data.

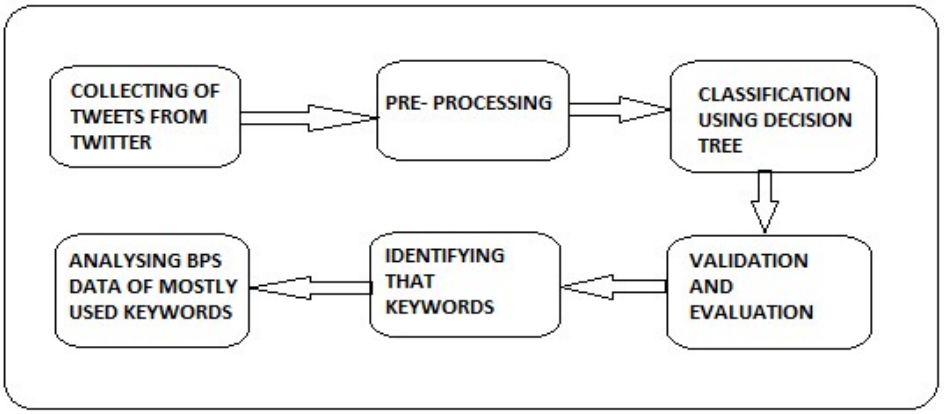

Figure 2.4: Procedure of BPS Data using Decision Tree

Marc Cheong, Vincent Lee, **"A Study On Detecting Patterns in Tweets Intra-Topics Users And Message clustering", 2010,** [14] author wants to detect the hidden pattern of Twitter messages (Tweets) and try to analysis that data for the better decision. For extracting Tweets they used unsupervised learning features as machine learning clustering tools. The main motive is to detect the hidden data is that the Author wants to know the popular topics or hot topics that cannot easily find by normal web data. The Methodology is first they take twitter data from twitter API and collect the messages (Tweets) after that they used User Demographics and message slats and then at last Demographics and user habits data. By this procedure, they detect the popular topics of the current era.

Simon Fong, Richard Khoury, "Sentimental Analysis of Online News Using MALLET", 2013, [17] author selects the news related data. The main motive to take the news data is data related to news is unemotional i.e. reviews, editorials and blogs. The emotional data is difficult to manage i.e. to divide into the positive and negative part. They divide their respected data also into positive and negative parts but the Author also phase problem for divide and manage the data because every data has a different meaning. But as compare to emotional data it is easy. The Author use MALLET (Machine Learning for Language Toolkit) and for implementation, they used the comparison of Six algorithms. The algorithms are Naive Bayes, Maximum Entropy,

Decision Tree, C4.5 Decision Tree, Winnow, and Balanced Window. After Comparison the Author considers Naïve Bayes is best performed out of another five algorithms.

R. Mohammad, P. Effat, "Improved Algorithm for Leader Election In Distributed System", 2010, [19] Now in today's life, Social Media is the main platform to voice their sentimental and opinions or it is the main platform to analysis the main popular/ hot topics. Out of them one of the major topics is Political Entities. In this paper, they used classification algorithm with supervised learning algorithm like Naïve Bayes algorithm, SVM. And the accuracy of these algorithms depends upon the training set of data. In this paper, the author takes trained data from Twitter because Twitter text used rapidly spreading and sparse attributes to detect the popular topics and predict the output through scalable machine learning model. By the text classification, they compared the results of Donaid Trump and Hillary Clintor using their own sentimental classifier and predict the election results.

The given figure 2.5 explain the entity classifier which is subdivided into two types, first one is sentimental classifier1 and sentimental classifier 2 and compare the result obtain by this two classifier.

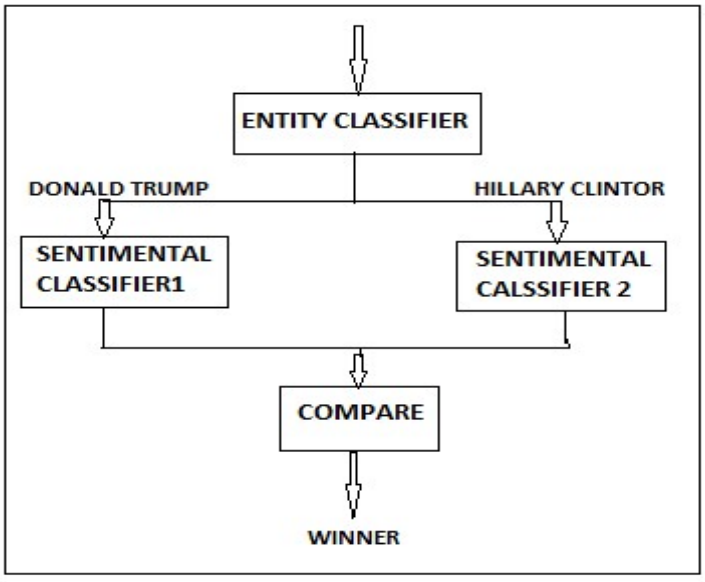

Figure 2.5: Model of Election Result Predication

CHAPTER 3

PRESENT WORK

In our research work, we divide the whole work into three parts which are gave as problem formulation, the objective of the study and the research methodology.

3.1 PROBLEM FORMULATION

Stream mining over twitter data is an area where lots of research is going on because twitter is miniaturized scale blogging administration that checks with a huge number of clients from everywhere throughout the world. Sentimental analysis has improved in the last few year as well as its applications. This is used for product marketing for recognition of anti-social behavior. The advances in Facebook, twitter, YouTube and other smaller scale blogging and long range informal communication destinations have contributed change to the social locales as well as have in a general sense changed the way we utilize these locales and how we share our emotions, our perspectives with the more extensive gathering of people.

The Sentimental examination on Social Media is an important field of study nowadays and will be in future. It is a procedure of discovering/deciding a concealed passionate tone behind the arrangement of words which is utilized to pick up a comprehension of the conclusion, feelings, conduct of the general population communicated in the online mode. The utilization of Sentimental examination is extremely expensive nowadays it is utilized wherever, for example, as a part of business, science, governmental issues social and so forth. Sentimental Analysis has been more than just a social sensible instrument. It is a most challenging field of research because the opinion of the people can randomly change with respect to time. In any case, it is a field that is up 'til now being examined, despite the way that not at magnificent lengths as a result of the multifaceted plan of this analysis. This field has limits that are exorbitantly jumbled for machines, making it difficult to get it. The ability to appreciate the joke, misrepresentation, positive feelings or negative suppositions has been troublesome, for machines that need feelings. The present

system cannot have more than 70% exactness the opinions portrayed by people. So the examination of online networking information to foresee the future or any condition is an awesome territory of research in nowadays.

3.2 OBJECTIVES

The objective of the study is mentioned below:

i. To analysis the sentiments of the user who updated their views/thoughts on social media.
ii. The events which occur/happen in the world has its distinct effect on different regions of the world. The focus will be to analyze the recent news and what is people opinion about it whether people are happy or sad about it, whether people agree or disagree with it and majorly which part of the globe is happy or sad, agree or disagree.
iii. The work will help to understand the effect of the event and reasons behind people opinion.

3.3 RESEARCH METHODOLOGY

In our work the methodology explain in three points, the first point explain the algorithm of our work which include the whole process and framework of our work, in the second part we describe the implementation tools that we used for the implementation of our work and at last i.e. in the third part we introduced the flow chart diagram which represents a graphical view of our research work.

3.3.1 RESEARCH METHODOLOGY

The main motive of the new proposed algorithm is to analysis the sentiments of the user by the mining of social media in an optimistic way, the algorithm which is used in my research work to analyze the sentiments of users motivated by the concept of probability and the Euclidean distance. “The probability is the uncertainty i.e. the chance of occurrence from a given set of data.The main reason that I selected this phenomenon is because of its simplicity and most suitable method of creating the classifier”.

“Decision tree is one of the simplest supervised learning algorithms that solve the well-known classification problem. The procedure follows a simple and easy way to classify a given data set through a certain number of classes (assume k clusters).”A decision tree is a flowchart-like structure in which each interior hub speaks to a "test" on a trait (e.g. regardless of whether a coin flip comes up heads or tails), each branch speaks to the result of the test, and each leaf hub speaks to a class mark (choice taken in the wake of registering all traits). The ways from root to leaf speak to arrangement rules. For example

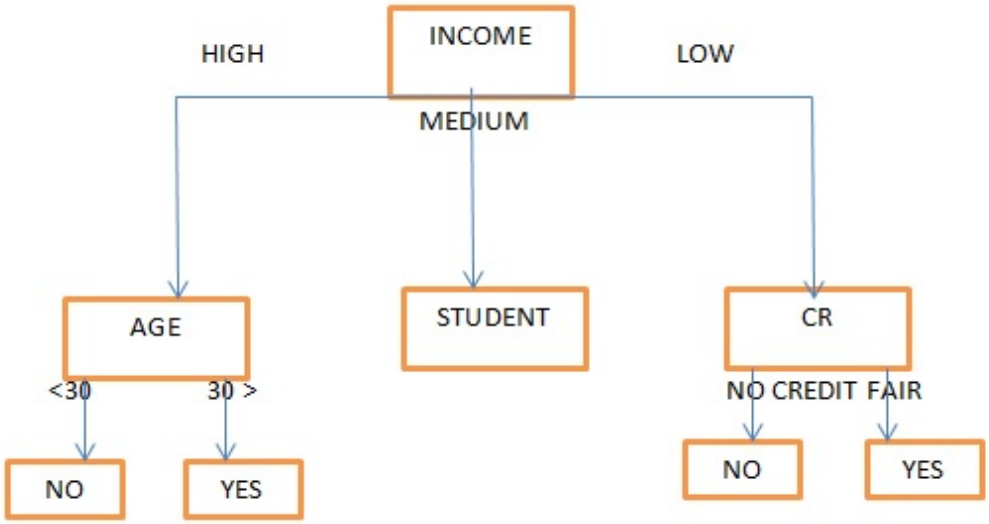

Figure 3.1: Example of Decision Tree

In figure 3.1 shows the decision tree which explains who get the loan on the basis of age, income, and profession.

“In machine learning, naive Bayes classifiers are a family of a simple probabilistic classifier based on applying in Bayer’s theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.”

$$P(C_K|X) = \frac{P(C_K).P(X|C_K}{P(X)} \quad \text{.....(i)}$$

Equation (i) represents the Naive Bayes equation.

i.e. posterior = (prior * likelihood)/ evidence.

The whole methodology of my work is explained in the following points:-

i. To fetch the data from Social media site.
ii. Pre-Process the coming data to remove the stop words and repetition of the words.
iii. Divide the data set into different labels i.e. provide the positive or negative label as per the tweets we get from the twitter.
iv. Now calculate the region wise the opinion of people which is done by using the Naïve Bayes algorithm.
v. After that apply the algorithm of the decision tree to classify the given review by the user of different parties.
vi. Now the system gives the overall opinion of people on the basis of above steps.
vii. At last, we predict the result.

In this way future about the opinion mining and sentimental analysis can be predicted.“Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of independence between every pair of features.

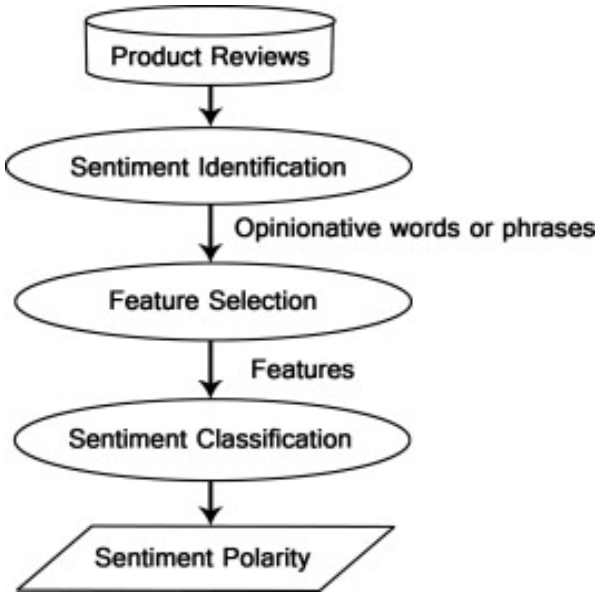

Figure: 3.2 (Structure Diagram of Sentimental Analysis)

The figure 3.2 shows the selection of product review and check the polarity of the sentiments i.e. Collect the review and identified according to product apply features selection, classified it and then check the polarity.

3.3.2 IMPLEMENTATION TOOLS

In our work for the implementation of our hypothesis, we chose the R studio for the experimentation because R is a framework which provides the statistical tools and different packages for the analysis of the data. It is a statistical computing and graphics language which is very similar to S language developed in Bell laboratory. It provides some strong features like

i. Huge data handling capacity and storage.
ii. Different library packages for analysis as well as graphical visualization.
iii. Inbuilt different classification algorithm for supervised learning
iv. Provide a good environment for the data mining analysis.

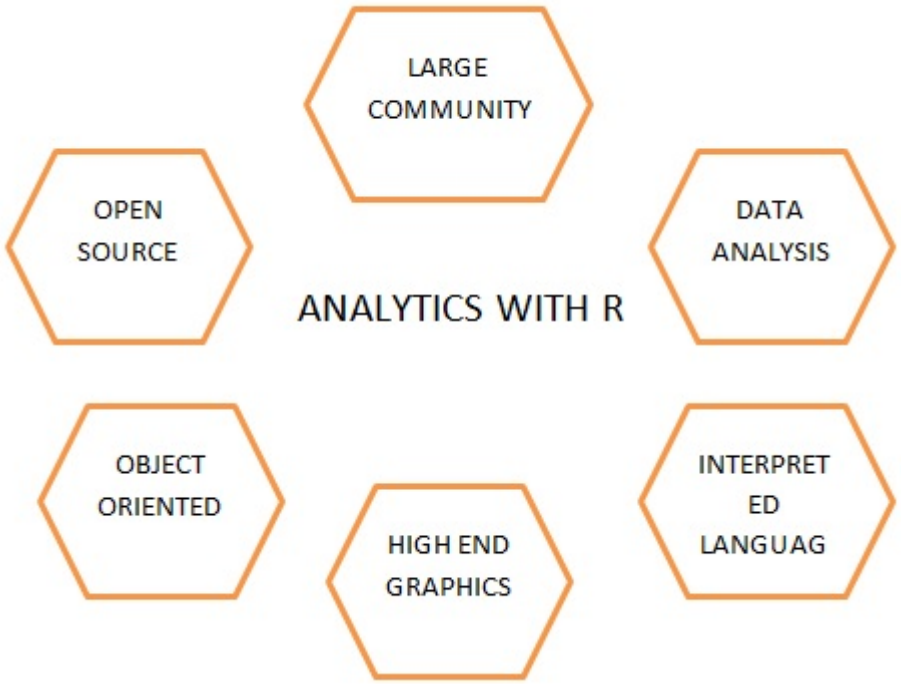

Figure 3.3: Reason for choosing R

In the figure 3.3 explain the characterized of the R language i.e. it is open source, object oriented, high-end graphics etc.

3.3.3 FLOW OF IMPLEMENTATION

In this section we deal with the graphical representation of our work, this provides a view of the flow of our work i.e. the overall structure of the research methodology. In figure 3.4 we take data from twitter, pre- process it applies C- tree classifier (Decision tree) and tries to predict the election results.

The research methodology is concluded by the flowchart described below:

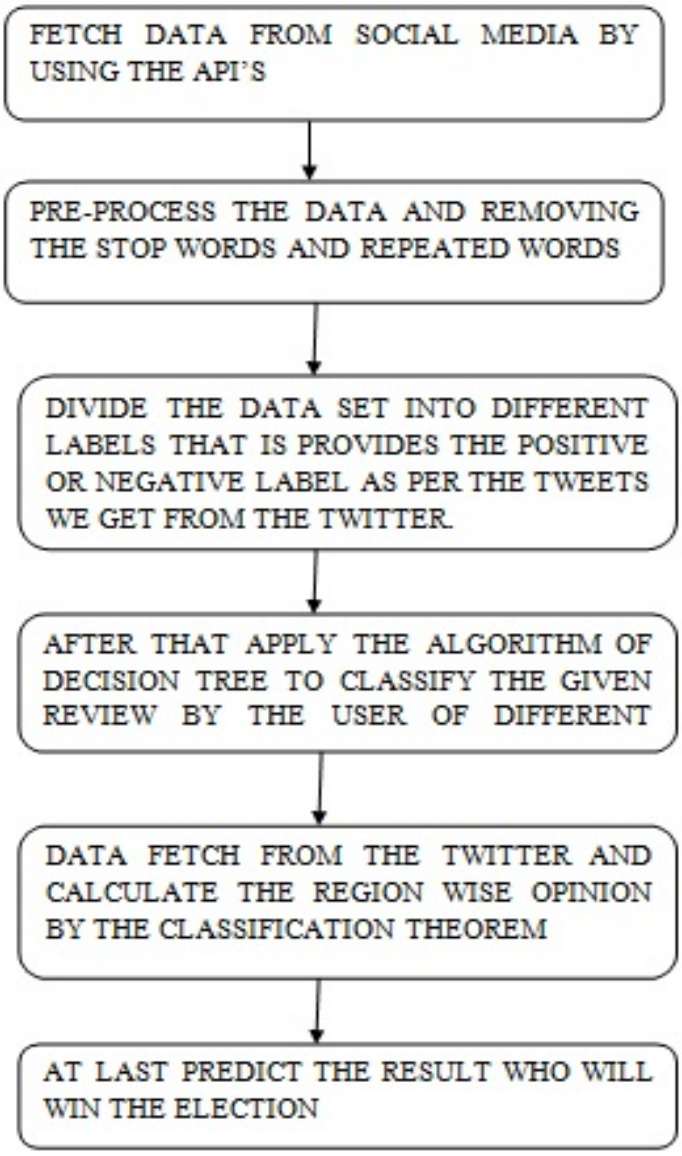

Figure 3.4: Flow Diagram of Research methodology

CHAPTER 4

RESULT AND DISCUSSION

4.1 EXPERIMENTAL RESULT

Every task is performing to get an output; the main aim of our work is also to get a result. The output of the task is defined whether the work is done accurately or not. The main aim of our research is to predict the election. As India is a democratic country where we chose our representative through election the result should come after some days and tells which party win the election, keeping this thing in the mind we try to develop the system which predict the result that which party win the election on the basis of the comments given by the citizen of the country by the way of twitter.

The success of the system depends on the rate of accuracy and the precision value. If the system gives high accuracy that means our work goes in the right direction.

After using the Classification based algorithm for Sentimental Analysis or Opinion mining, the expected outcomes are:

i. Reducing completion time.
ii. Minimizing the error rate.
iii. Finding the behavior of the users.
iv. The increase in the performance of Analysis.
v. Self-trained algorithm for predicting future.

Consequently, the overall efficiency of Analysis is expected to improve by minimizing speculative executing.

In our research work, we take the data from the social media site like twitter so in the given figure 4.1 and 4.2 we show that how we take the data for our work i.e. data taken from twitter as per the tweets.

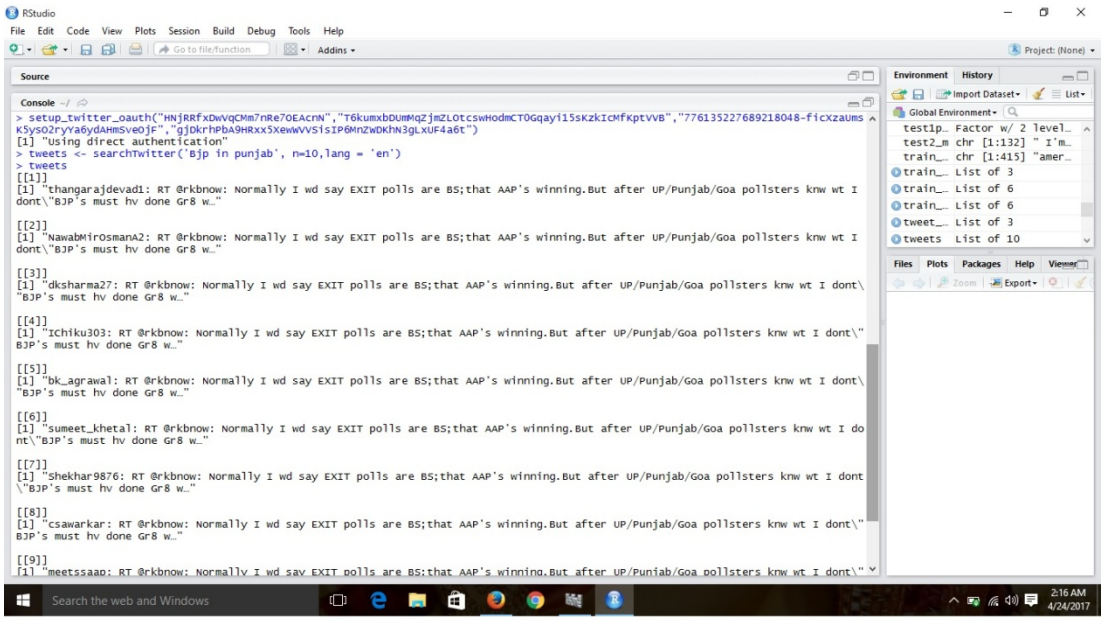

Figure 4.1: Tweets fetch by R

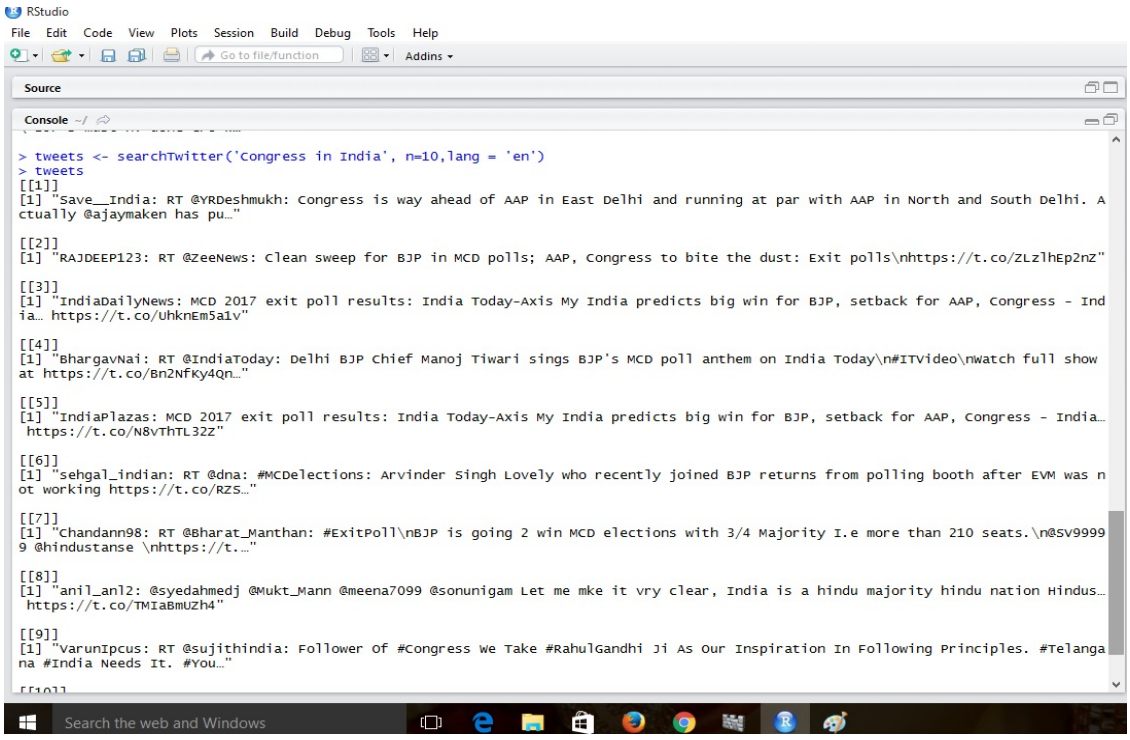

Figure 4.2: Tweets related to Congress

Frequency tells us the intensity of the things which is regularly happened. In the text mining, the frequency of words has very much importance which tells that in the database which words is regularly used and what is its frequency of occurrence, in the given diagram below we represent the frequency of top 10 words by the help of bar chart.

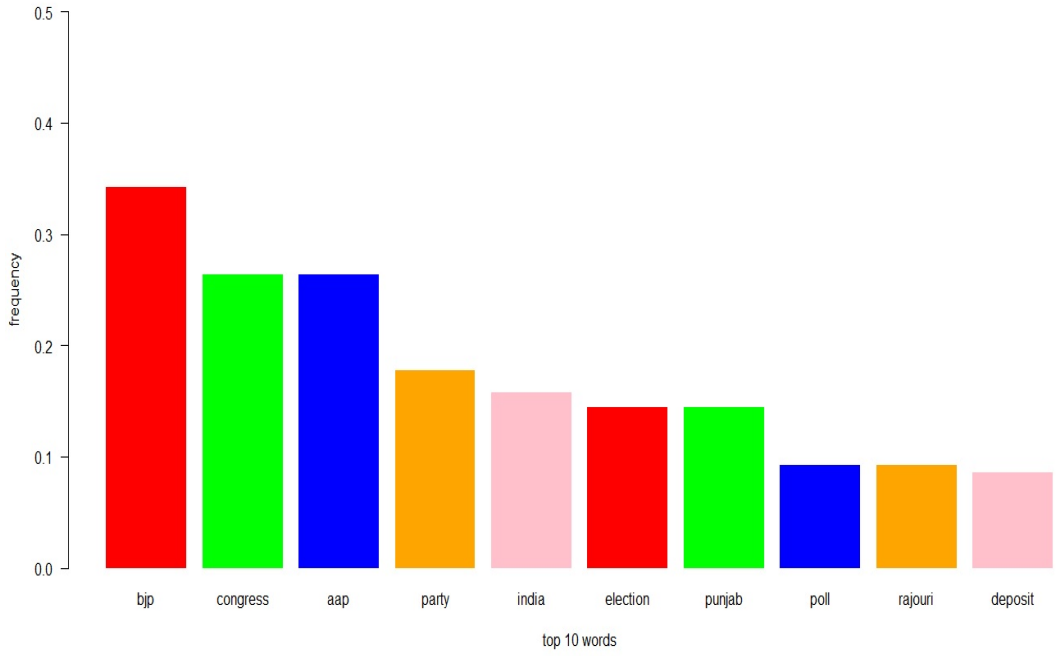

Figure 4.3: Word Frequency Corresponding to Election.

In figure 4.3, the same color represents the relation between each other i.e. red color graph shows that bjp wins the election whereas green color shows that in Punjab the chances of congress party to win the election.

The concept of using decision tree is to show the output with the help of tree and nodes because it gives the best representation of the result as well as it is the simplest classifier which is good for the text mining. The tree structure of our system is given in the below diagram.

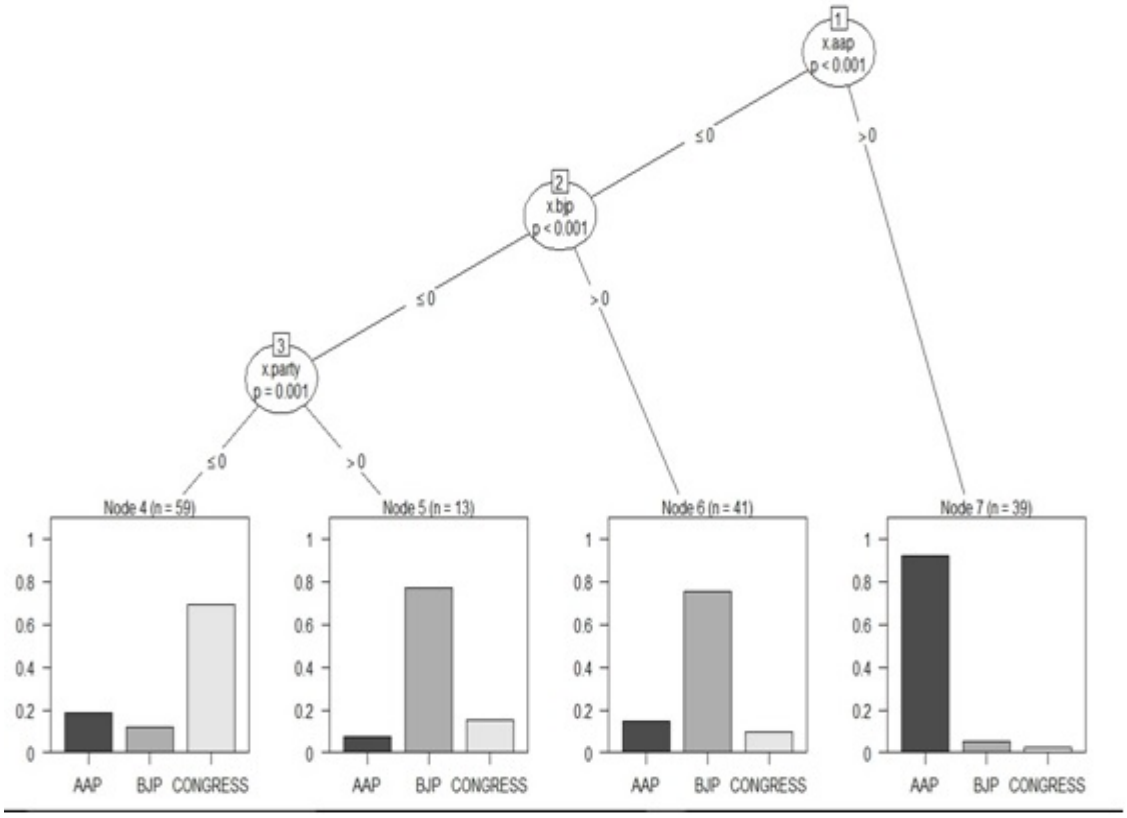

Figure 4.4: Decision Tree For Election Prediction

The figure 4.4 we represent the outcome of our prediction in the form of a tree, in which we see that there are three nodes should be generated and as per the analysis, they show which party has maximum chance to win the election with the help of bar graph. The visualization predicts that the chances of BJP to win the election are more in another state as compared to other parties but in one place the congress has a chance to beat the BJP. Thus the motive of our research work is to predict the correct outcome of the decision before the result. The prediction should be visualized by the help of tree and nodes so can anyone can able to find out the result by seeing the graph i.e. its provide the feature of simplicity.

The nature of the precise and auspicious outcomes must be surveyed preceding discharge. In the event that blunders in the outcomes happen, they ought to be specifically adjusted and people, in general, to be educated as quickly as time permits i.e. the success or failure totally depend on the accuracy, that how much our research gives the accurate result so that we are able to rely on that result. As the importance of the accuracy is more in any research

so our aim is also to make a system which predicts the result in more accurate way and we did it by gaining the accuracy of our system to approximately 77%

The given figure 4.5 shows the accuracy of our system generated by applying a decision tree algorithm over a set of data. The figure is given below.

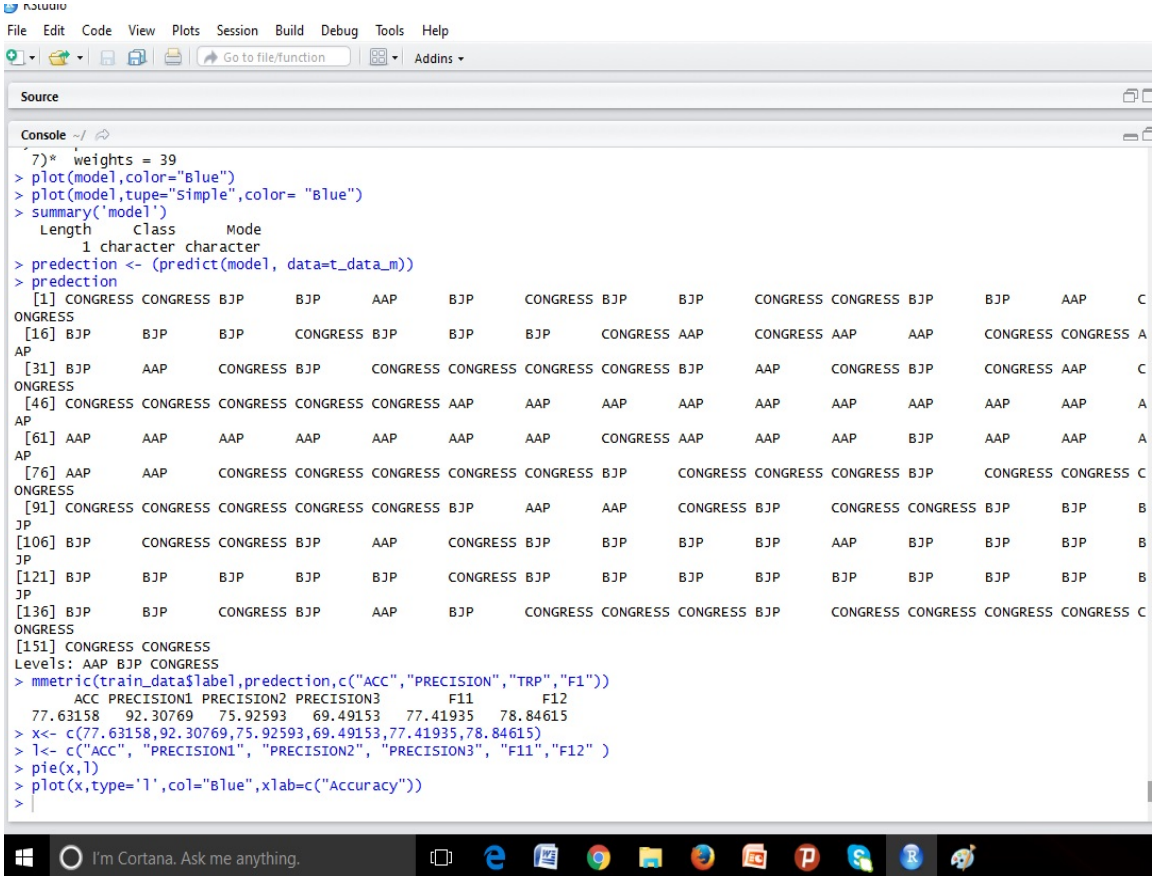

Figure 4.5: Accuracy of our system

The accuracy of the Naive Bayes algorithm generated by the system over the same data is given in figure 4.6.

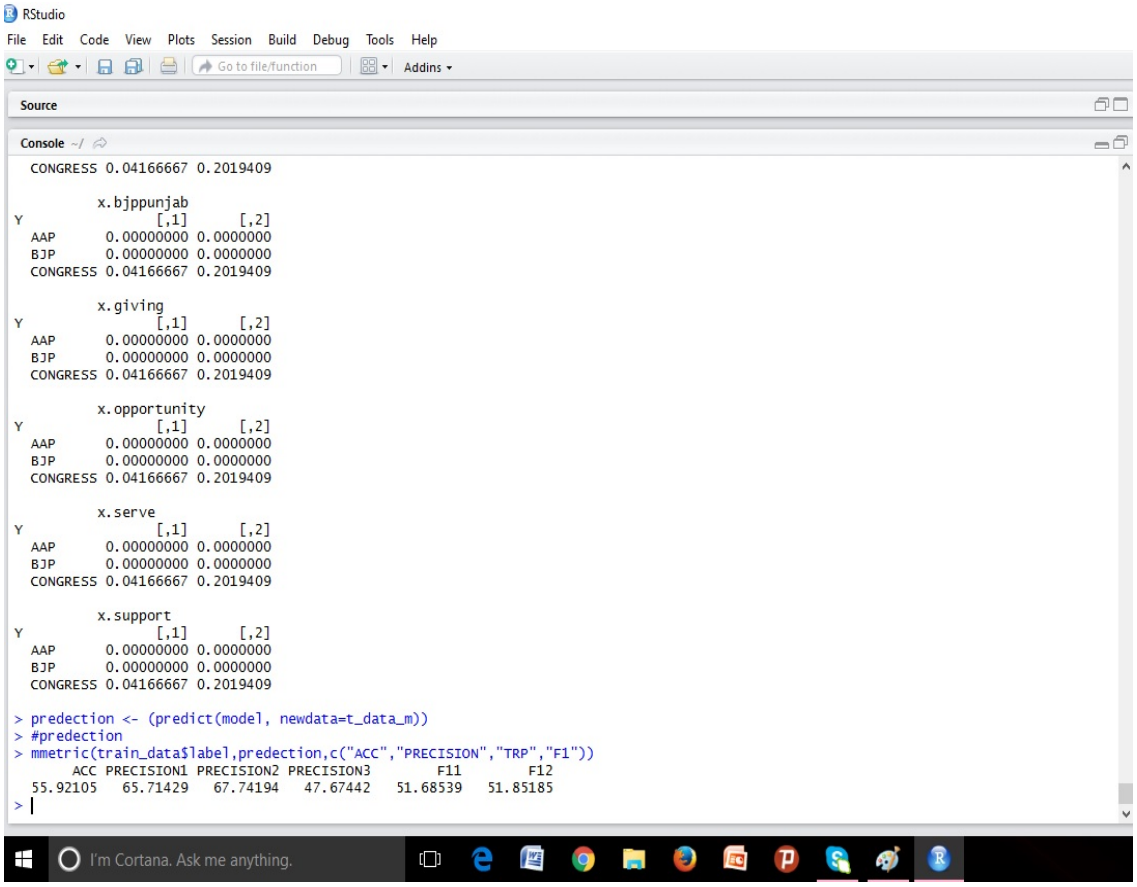

Figure 4.6: Accuracy of Naive-Bayes

Our system also calculate the most frequent words which is used in the data set and on that basis we are able to find out the result, so the given figure 4.7 tells the input words taken by the system from the data set and also shows that how many conditional nodes should be generated by the system using the predefined data.

The figure of output is given below:

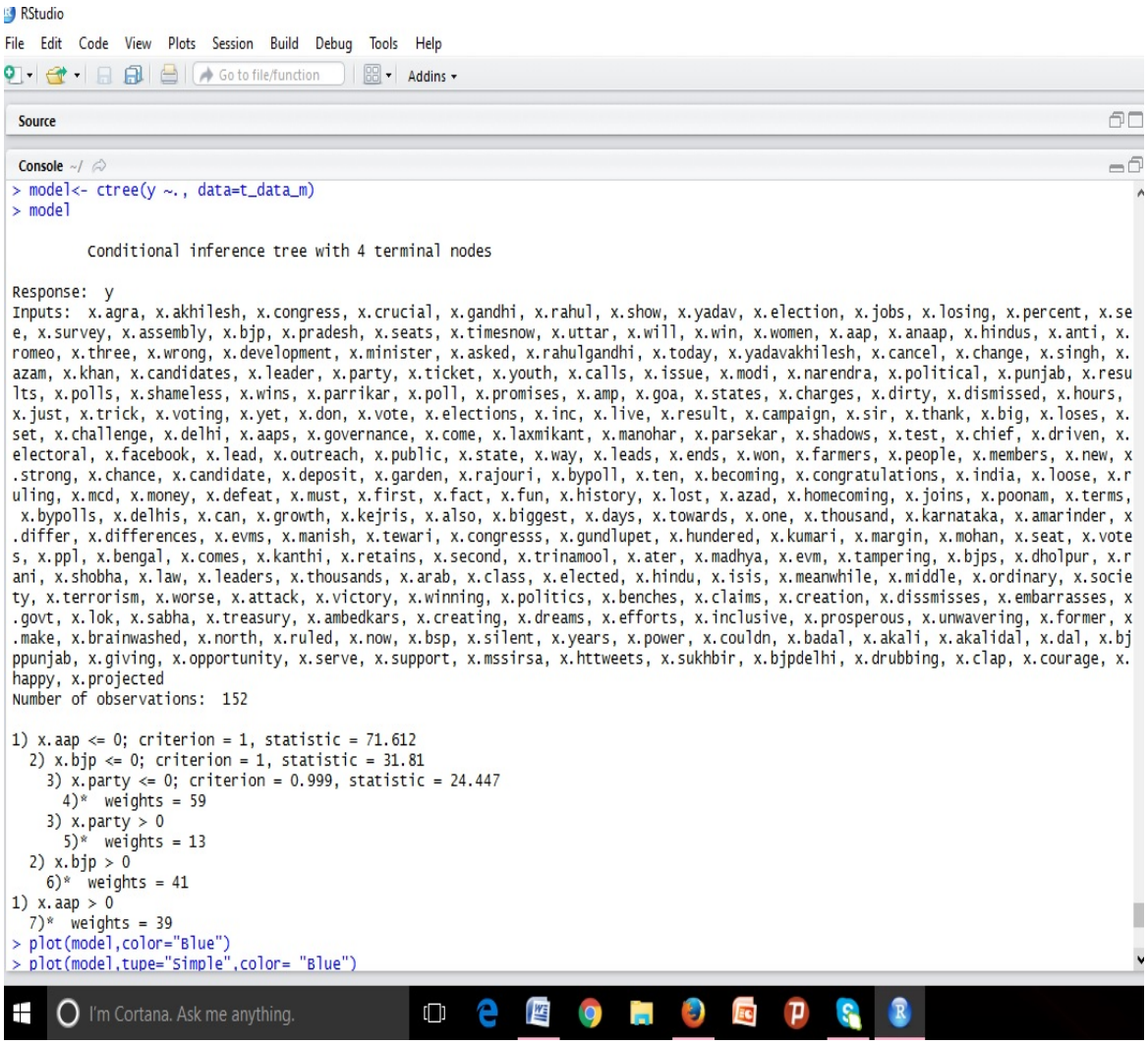

Figure 4.7: Output of our system

The overall result of our research work is visualized in the form of pie chart where we show the accuracy, precision values, F11 and F12 values and find that the accuracy, as well as the precision, is much greater than the other algorithm or system generated by apply same data. Which shows that our system is more optimized than others. The representation of the pie chart is given in figure 4.8 below.

In the figure the blue area represents the accuracy rate of the system, the orange color area represents the precision1, gray color area represents the precision 2 whereas the yellow color area represents the value of precision 3, the dark blue color area represents the F11.

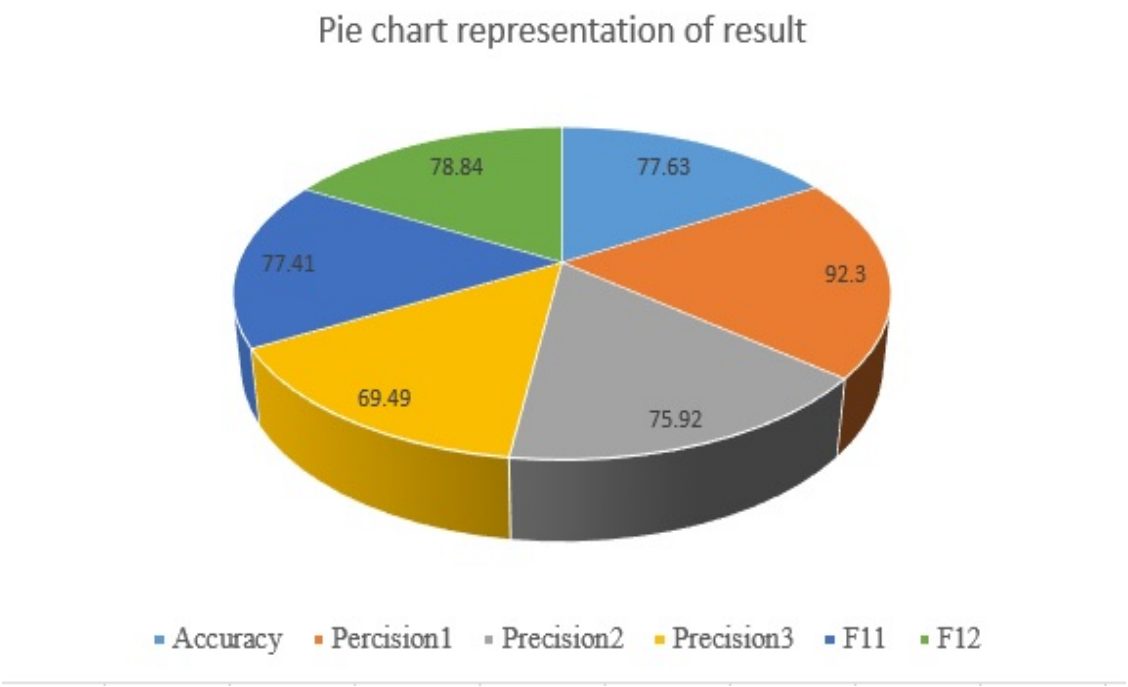

Figure 4.8: Pie chart representation of result

4.2 COMPARISON WITH EXISTING TECHNOLOGY

In the given figure 4.9 below we compare the result of our system with the naive Bayes algorithm and find that the result and accuracy generated by our system is much better than the naive Bayes, the comparison of the result generated by the system using two different classifiers:

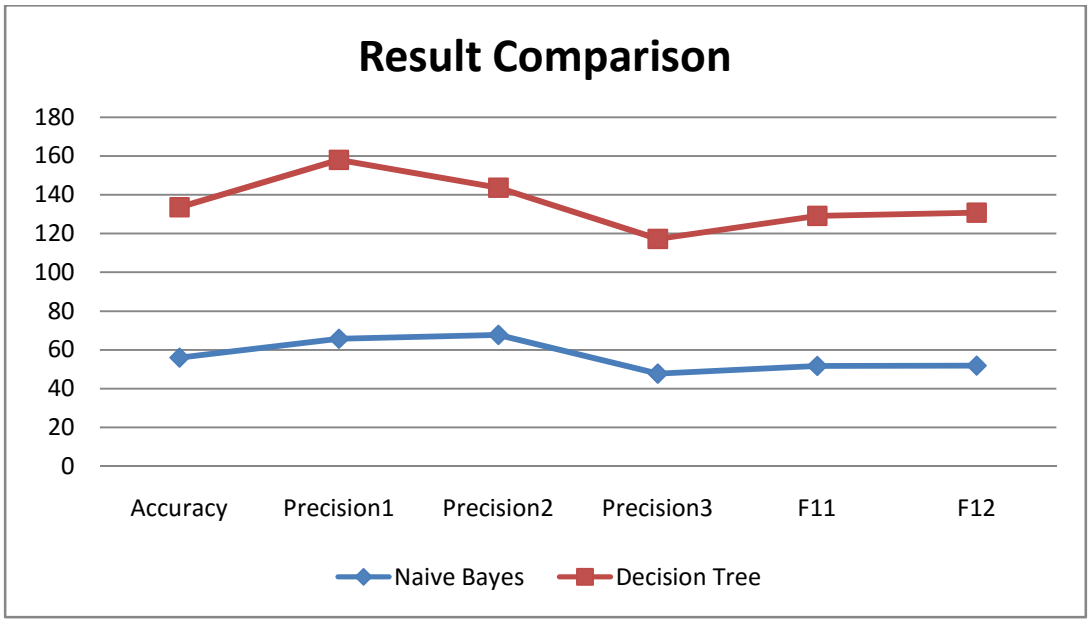

Figure 4.9: Result Analysis

In the given figure the red line shows output rate of a decision tree which is generated by our system whereas the blue line represents the Naive Bayes result over the same data, this comparison clearly represents that our system is more accurate and efficient than others.

The comparison of the accuracy, precision1, precision2 and precision3 of the two different classifiers is explain given below by the help of figure. In the given bar graph the dark pink color bar represents the accuracy, precision1, precision2 and precision3 of our system i.e. decision tree whereas the blue color bar graph represents the accuracy and precision values of the naive Bayes. The given figure 4.10 clearly shows that our system is much better than the other system.

The bar graph of the result comparison between naive Bayes and the decision tree is given below.

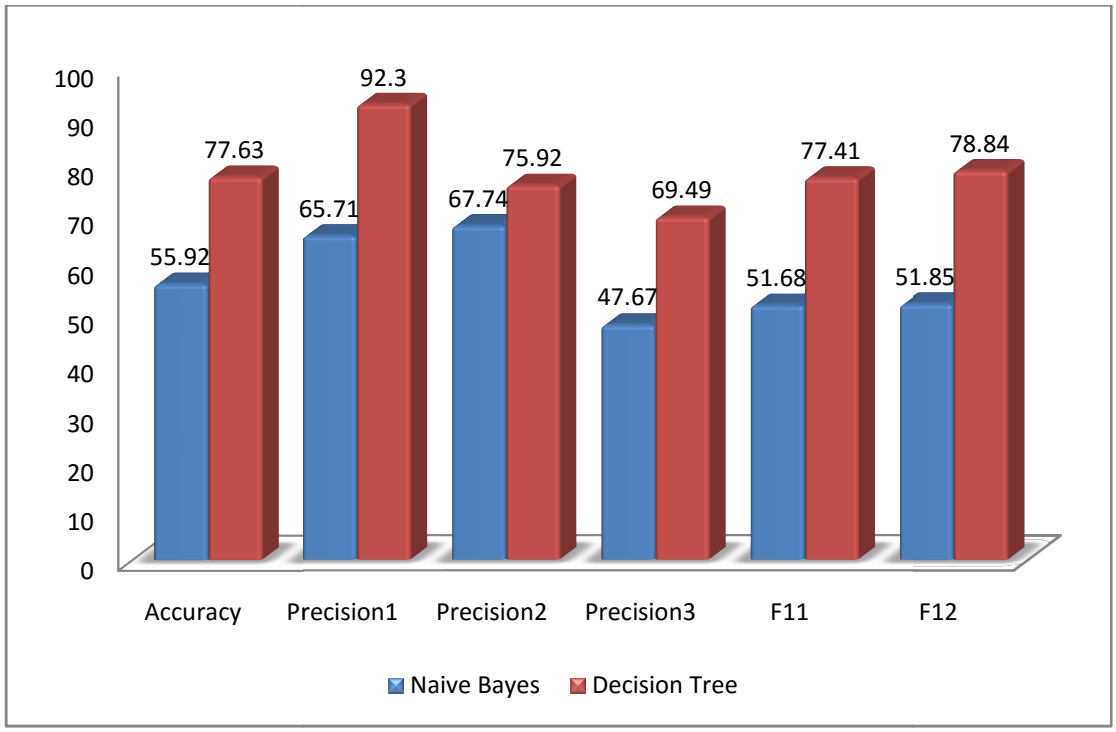

Figure 4.10: Result Comparison of Different Classifier

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

The work being done on the point is boundlessly slender and as it addressed the issue of USER nostalgic investigation and not SNA. Fusing this will be the following stride in accomplishing better results. Moreover better fuse with long range informal communication destinations and different Facilities also, gathered android gadgets can help our program to accomplish a more sweeping knowledge.

The issue is that most supposition examination calculations utilize straightforward terms to express estimation about an item or administration. Not with standing, social elements, semantic subtleties and varying settings make it greatly hard to transform a string of composed content into a straightforward expert or con sentiment. The way that people regularly differ on the opinion of content represents how enormous an undertaking it is for PCs to get this privilege. The shorter the string of content, the harder it gets to be. Opinion investigation scrapping vast information sets have additionally enhanced slant mining.

5.2 FUTURE WORK

Our research likes to think, there is a developing interest for philosophical assumption assessment of content, in various ranges of application. It is insufficient to state that content is generally positive or general negative. Clients might want to know which isolate points are discussed in the content, which of them are certain and which are negative, so in future our work may extend for predicting other concepts which are helpful for the citizen, as well as there, are lots of work can be done to enhance the algorithm and increase the accuracy as well as the reliability of the system.

REFERENCES

[1] S. Balaguru, R. Nallathamby, and C. R. R. Robin, "A NOVEL APPROACH FOR ANALYZING THE SOCIAL NETWORK," *Procedia - Procedia Comput. Sci.*, vol. 48, no. Iccc, pp. 686–691, 2015.

[2] P. Chen, "Procedia Engineering Data Mining Applications in E-Government Information Security," vol. 29, pp. 235–240, 2012.

[3] G. Chetty, M. White, and F. Akther, "Smart Phone Based Data Mining For Human Activity Recognition," *Procedia - Procedia Comput. Sci.*, vol. 46, no. Icict 2014, pp. 1181–1187, 2015.

[4] E. Fumeo, L. Oneto, and D. Anguita, "Condition Based Maintenance in Railway Transportation Systems Based on Big Data Streaming Analysis," *Procedia - Procedia Comput. Sci.*, vol. 53, pp. 437–446, 2015.

[5] M. M. Gaber, "Granularity-Based Approach," vol. 6, pp. 47–66.

[6] E. Hromada, "Mapping of real estate prices using data mining techniques," *Procedia Eng.*, vol. 123, pp. 233–240, 2015.

[7] S. Khadijah and Z. Tasir, "Educational data mining : A review," *Procedia - Soc. Behav. Sci.*, vol. 97, pp. 320–324, 2013.

[8] R. Mythily, A. Banu, and S. Raghunathan, "Clustering Models for Data Stream Mining," *Procedia - Procedia Comput. Sci.*, vol. 46, no. Icict 2014, pp. 619–626, 2015.

[9] S. Nasreen, M. Awais, K. Shehzad, and U. Naeem, "Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams : A Survey," *Procedia - Procedia Comput. Sci.*, vol. 37, pp. 109–116, 2014.

[10] K. K. Parameswari and A. S. Thanamani, "Frequent Item Mining Using Damped

Window Model," vol. 3, no. 9, pp. 7902–7905, 2014.

[11] M. S. B. Phridviraj and C. V Gururao, "Data mining – past , present and future – a typical survey on data streams," *Procedia Technol.*, vol. 12, pp. 255–263, 2014.

[12] V. P. Rao, S. Galande, A. Nalla, S. Devghare, and S. Kadam, "Visualization of Streaming Data Using Social Media," no. 6, pp. 602–605, 2016.

[13] R. Tripathi and S. K. Dwivedi, "A Quick Review of Data Stream Mining Algorithms," no. 7, 2016.

[14] M. Cheong and V. Lee, "A study on detecting patterns in Twitter intra-topic user and message clustering," *Proc. - Int. Conf. Pattern Recognit.*, pp. 3125–3128, 2010.

[15] Jing Guo, Peng Zhang, Jianlong Tan, Li Guo, "Mining Hot Topics From Twitter Streams", pp. 211–214, 2012.

[16] D. Fajardo-Delgado, J. A. Fernández-Zepeda, and A. G. Bourgeois, "Randomized self-stabilizing leader election in preference-based anonymous trees," *Proc. 2010 IEEE Int. Symp. Parallel Distrib. Process. Work. Phd Forum, IPDPSW 2010*, 2010.

[17] S. Fong, Y. Zhuang, J. Li, and R. Khoury, "Sentiment Analysis of Online News Using MALLET," *2013 Int. Symp. Comput. Bus. Intell.*, pp. 301–304, 2013.

[18] T. B. Mirani and S. Sasi, "Sentiment analysis of ISIS related Tweets using Absolute location," pp. 1140–1145, 2016.

[19] R. Mohammad, P. Effat, N. Yazdani, M. E. P, A. Dadlani, and A. Khonsari, "Improved Algorithms for Leader Election in Distributed Systems," *2010 2nd Int. Conf. Comput. Eng. Technol.*, pp. 6–10, 2010.

[20] Z. Xu, L. Liu, W. Song, H. Wang, C. Du, and J. Lu, "Conflicting views analysis algorithms," *Proc. 2015 4th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2015*, no. Iccsnt, pp. 526–529, 2016.

[21] P. Yhoga and C. Kusuma, “Social Media Analysis of BPS Data availability in Economics using Decision Tree Method,” no. February, pp. 148–153, 2016.

[22] D. Donato, A. Gionis, G. Mishne, R. Batool, A. M. Khattak, J. Maqbool, “A Survey on Sentiment Analysis Algorithms for Opinion Mining,” *Proceeding 2015 1st Int. Conf. Wirel. Telemat. ICWT 2015*, vol. 6, no. 6, p. 12, 2013.

[23] S. Nasreen, M. A. Azam, K. Shehzad, U. Naeem, and M. A. Ghazanfar, “Frequent pattern mining algorithms for finding associated frequent patterns for data streams: A survey,” *Procedia Comput. Sci.*, vol. 37, pp. 109–116, 2014.

[24] U. Franke and M. Rosell, “Prospects for detecting deception on twitter,” *Proc. - 2014 Int. Conf. Futur. Internet Things Cloud, FiCloud 2014*, pp. 528–533, 2014.

[25] S. Jamoussi and H. Ameur, “Dynamic Construction of Dictionaries for Sentiment Classification,” *2013 Int. Conf. Cloud Green Comput.*, pp. 418–425, 2013.

[26] Saiyan Dai, Ling Chen, “An Algorithm of Mining Frequent Closed Itemsets In The Data Stream”, *2016 Int. Conf. Inf. Commun. Technol. Converg.*, pp. 141–146, 2016.

[27] Shoiab Ahmed and Ajit Danti, “Novel Approach for Sentimental Analysis and Opinion Mining based on Sentimental WordNet using Web Data," *2016 Int. Conf. Electr. Electron. Optim. Tech.*, pp. 3318–3323, 2016.

[28] N. B. Model, “Naive Bayes Algorithms,” pp. 1–7.

[29] D.A. Adeniyi, Z. Wei, Y.Dongguan, “Automated Web Usage Data Mining And Recommendation System Using K-Nearest Neighbour (KNN) Classification Method”, pp. 97–104, 2014.

[30] K. Santhisree and Dr. A. Damodaramin, “Web Usage Data Clustering using Dbscan algorithm and Set Similarities”, *Proc. Twent. Int. Conf. Mach. Learn.*, vol. 20, no. 1973, pp. 616–623, 2003.

[31] Prashant Kumar and Dhruv Mahajan, “An approach towards features specific Opinion Mining and Sentimental Analysis Across E-Commerce Websites”
ICISA 2014 - 2014 5th Int. Conf. Inf. Sci. Appl., pp. 2–5, 2014.

[32] H. Zhang, “The Optimality of Naive Bayes Naive Bayes and Augmented Naive Bayes,” *Am. Assoc. Artif. Intell.*, 2004.

APPENDIX

ABBREVIATIONS USED

ACL: Access Control List

CFO: Central Force Optimization

CRISP-DM: Cross industry standard procedure of information mining

D- Tree: Decision Tree

D-Window: Damped window

FP: Frequent pattern

i.e.: That is to say

NV: Naïve Bayes

SNA: Social Networking Analysis

SVM: Support Vector Machine

viz.: As follows