A COMPARATIVE STUDY OF EMOTION MINING ON SOCIAL MEDIA DATA

BY

IFTEKHAR UDDIN YOUSUF
ID: 151-15-235

AND

ABIR AHAMED RAJU
ID: 152-15-519

AND

MOHAMMAD ABDUR RAKIB
ID: 152-15-540

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Farzana Akter
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Bulbul Ahmed
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
APRIL 2019

APPROVAL

This Project titled "**A Comparative Study of Emotion Mining on Social Media Data**", submitted by Iftekhar Uddin Yousuf ID: 151-15-235, Abir Ahamed Raju ID: 152-15-519 and Mohammad Abdur Rakib ID: 151-15-540 to the Department of Computer Science and Engineering, Daffodil International University, has been acknowledged as tasteful for the halfway satisfaction of the necessities for the level of B.Sc. in Computer Science and Engineering and affirmed as to its style and substance. The presentation has been held on 6th April 2019.

BOARD OF EXAMINERS

_____

Dr. Syed Akhter Hossain **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. S M Aminul Haque **Internal Examiner**
Assistant Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Saif Mahmud Parvez **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Mohammad Shorif Uddin **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University
Daffodil International University

DECLARATION

We therefore announce that, this research work has been finished by us under the supervision of Farzana Akter, Lecturer, Department of CSE, Daffodil International University. We also declare that neither this endeavor nor any bit of this endeavor has been submitted somewhere else for honor of any degree or diploma.

Supervised by:

_____

Farzana Akter
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

_____

Bulbul Ahmed
Lecturer
Department of CSE
Daffodil International University

Submitted by:

_____

Iftekhar Uddin Yousuf
ID: 151-15-235
Department of CSE
Daffodil International University

_____

Abir Ahamed Raju
ID: 152-15-519
Department of CSE
Daffodil International University

_____

Mohammad Abdur Rakib
ID: 152-15-540
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

In the first place, we express our heartiest thanks and thankfulness to god-like God for His wonderful blessing makes us possible to complete the most recent year adventure/transitory occupation adequately.

We extremely appreciative and wish our significant our obligation to Farzana Akter, Lecturer, Department of CSE Daffodil International University, Dhaka. Profound Knowledge and distinct fascination of our manager in the field of "Emotion Mining" to do this undertaking. His perpetual tolerance, academic direction, consistent support, steady and vigorous supervision, helpful feedback, important guidance, perusing numerous second rate drafts and amending them at all stage have made it thinkable to finish this task.

We might want to offer our heartiest thanks to Bulbul Ahmed, Lecturer and Dr. S.M. Aminul Haque, Associate Head, Department of CSE, for his benevolent aid to complete our venture and furthermore to other employee and the staff of CSE division of Daffodil International University.

We might want to thank our whole course mate in Daffodil International University, who took part in this talk about while finishing the course work.

At last, we should recognize with due regard the consistent aid and patients of our folks.

ABSTRACT

Sentiment analysis or emotion mining is the method of computational and natural language processing based techniques which are used to extract, identify or characterize subjective information, such as opinions, comment that expressed in a given piece of text. The main contributions of this paper involve the sophisticated categorizations of the trend of research on sentiment analysis and its allied areas as well as to identify the most and least commonly used feature selection techniques to find out research gaps for future research. This paperwork gives a detailed analysis of the recent sentiment analysis schemes and directed towards new avenues for future research work in this region.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 History of Sentiment Analysis

Emotions are a part of parcel of human nature that can be considered as inherited. Also it has been found that different human being expression of a particular emotion is uniform. Emotion mining is the method of detecting, analyzing human's felling towards different proceedings, services, or any other particular interest. The most interesting and trending topic for the research field is emotion mining was started since the 20th century. It is also known as the detection or finding the opinions from the positive text and negative text. Besides, this era of web 2.0 end user Produced data over the Internet has prolonged more and more rapidly. Social Media platform and commercial website. For instance, Facebook, Instagram, Twitter, LinkedIn, Amazon, IMDB offer a platform to share their experiences, knowledge and views on the recent trend of politics, economics and other global- critical issue. Emotion mining accumulates online documents ranging from twitters, Facebook; product reviews blogs and other social media platform. As we know, Human decision making is always influenced by others thinking, ideas and opinions. While making any purchase online consumer usually checks opinions of others about the product. Emotion mining is the automated mining of attitudes, opinions, and emotions from text, speech, and database sources through Natural Language Processing (NLP). Emotional mining implicates classifying opinions in a text into categories like "positive" or 'negative' or "neutral". It's often mentioned as subjectivity analysis, Sentiment analysis and appraisal extraction. Sentiments or Opinions contain public generated content about products, services, policies and politics. Customers have a habit of trust the opinion of other consumers, particularly those with past experience of a product or service, rather than company marketing. Social Media are influencing customer preferences by understanding their approaches and behaviors.

A lot of research work is being done in the field of emotional mining due to its importance in the marketing level competition and the changing needs of the people. It requires the usage of a training set for its performance. We present a table summarizing all the studied work. In this comparative study, we have taken a systematic literature review process to identify areas well focused by researchers, least addressed areas are also emphasized giving an opportunity to researchers for

further work, weakness, strengths, threats and opportunities, whose factors represent, in a sense, future work to be carried out. We have also tried to identify most and least commonly used feature selection techniques to find research gaps for future work. Finally, the emotion mining schemes have been compared in terms of accuracy with respect to some algorithms. Thus this paperwork provides a detailed analysis of the recent emotion mining schemes and throws light on new avenues for future research work in this area.

CHAPTER 2

ALGORITHM

2.1 Some Algorithm Used in Sentiment Analysis

Support Vector Machine (SVM): It is a discriminative classifier formally defined by a separating hyper plane basically used for text categorization. It also can provide a good performance in high-dimensional feature space. A SVM algorithm expresses the instances as points in space, charted so that the examples of the altered categories are separated by a clear margin as extensive as possible. It provide the best results than Naive Bayes algorithm and sentiment classification tools. The basic idea is to find the hyper plane which is showed as the vector w which separates document vector in one class from the vectors in other class.

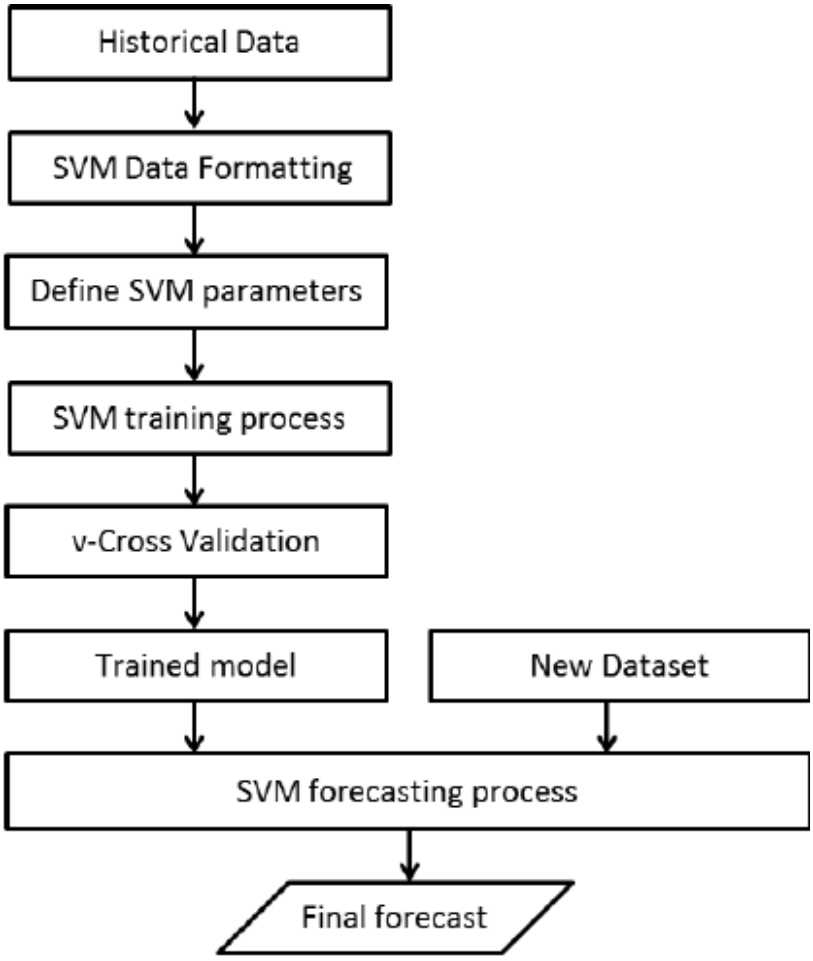

Figure 2.1: Support Vector Machine

Naive Bayes Classifier: The Naive Bayes Classifier algorithm is based on the alleged Bayesian theorem and is mainly suited when the dimensionality of the inputs is high. Even though its easiness, Naive Bayes can frequently outperform more sophisticated classification technique. To determine the concept of Naïve Bayes Classification; consider the example showed in the illustration below. As specified, the objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently existing objects.

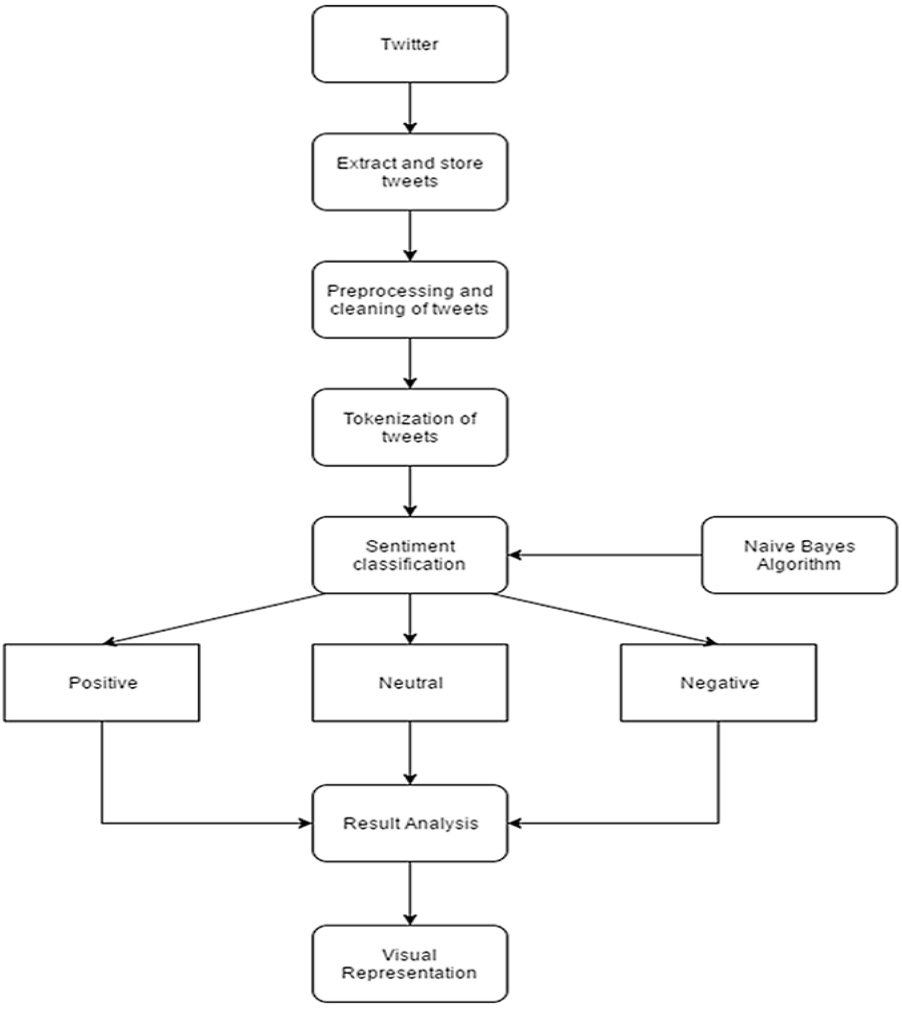

Figure 2.2: Naïve Bayes Algorithm

K-Means: *K*-means clustering is a type of unsupervised learning that is used when you have unlabeled data for instance, data without defined categories or groups. The aim of this algorithm is to discover groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data

2. Labels for the training data (each data point is assigned to a single cluster)

Rather than describing groups before seeing at the data, clustering permits you to find and study the groups that have designed organically. The "Choosing K" section below describes how the number of groups can be determined. Every centroid of a cluster is a gathering of feature values which describe the resulting groups. Scrutinizing the centroid feature weights can be used to qualitatively understand what kind of group each cluster represents.

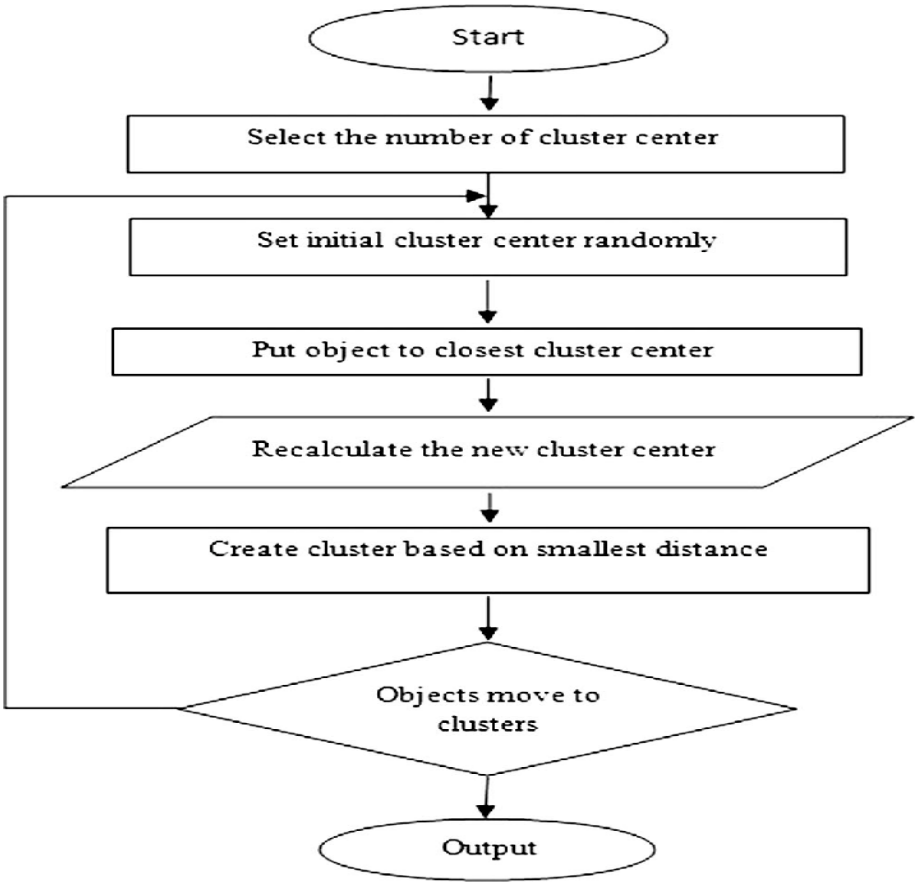

Figure 2.3: K-Means

Natural Language Processing (NLP): The core of Natural Language Processing lies in influencing PCs to comprehend the regular linguistic. That is not easy undertaking, though. PCs can comprehend the systematized type of information like spreadsheets and the tables in the database, however, human messages, languages, and voices structure an unstructured classification of information, and it gets worrying for the PC to acquire it, and there emerges the condition for Natural Language Processing. There's a great contract of characteristic language information out there in different structures and it would get remarkably simple, if the PCs can comprehend and process that information. We can make the models as per anticipated yield in various ways. People have been composing for a great many years, there are tons of writing pieces accessible, and we should influence the PCs to get that. In any case, the assignment is never going to be simple. There are different difficulties drifting out there like understanding the right significance of the sentence, right Named-Entity Recognition (NER), right forecast of different grammatical forms, co-reference goals.

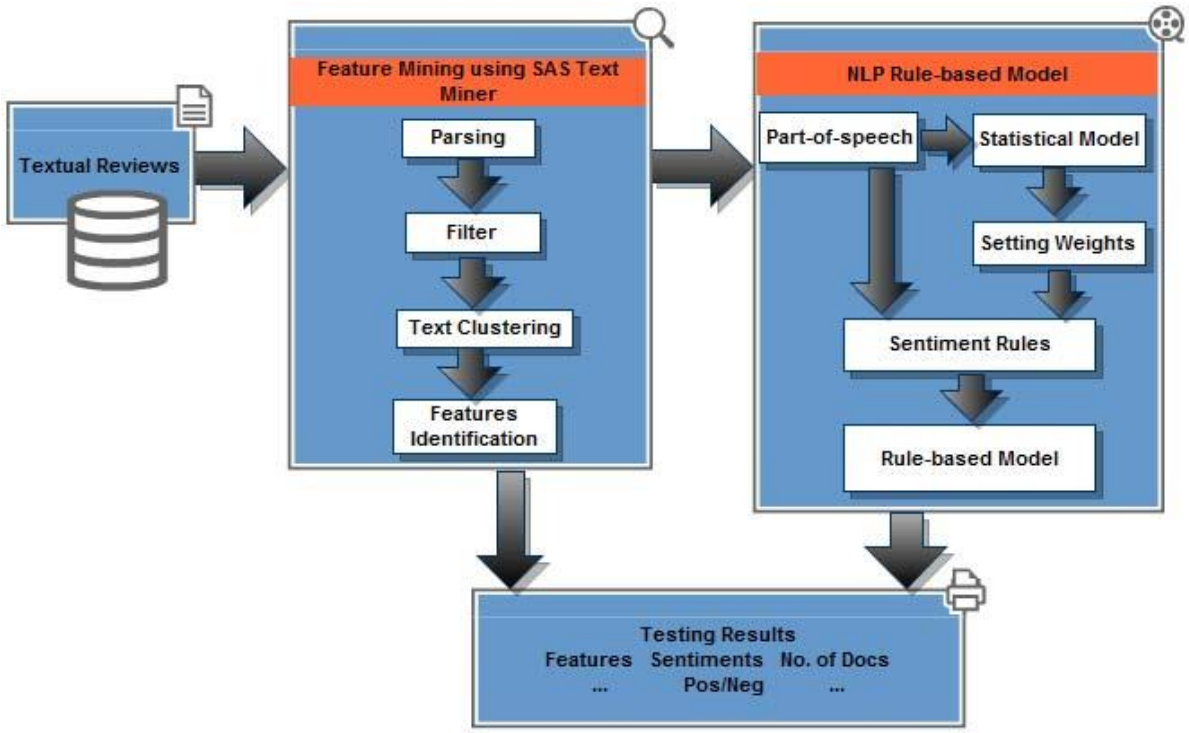

Figure 2.4: Natural Language Processing

2.1.1: Classification Tools

The paper gives a diagram of concentrating on the slant grouping venture; specifically beginning of this review the paper an arrangement of (I) assumption characterization approaches us for highlights/procedures and focal points/confinements and (ii) instruments regarding the distinctive methods utilized for assessment examination. Diverse application fields of supposition examination, for example, business, politics, open activities and fund are additionally talked about in the paper. The slant order methodologies can be grouped in: (I) AI (ii) vocabulary based (iii) hybrid system. Calculations of the accompanying strategies: (1) Decision Tree, (2) Rules Induction, (3) Clustering, (5) Artificial Neural Network (ANN), (6) Bayesian classifier and (6) a Support Vector Machine. The order is normally acquired by administering adapting yet can likewise be performed by unsupervised learning, for example where the class isn't utilized or obscure as in the Clustering procedure.

2.1.2: Sentiment Analysis Tools for Twitter

Numerous instruments have been created as of late for investigating notions in short casual web-based social networking writings. We analyzed a genuinely differing set of 20 Twitter opinion examination apparatuses. These devices included uninhibitedly accessible frameworks created in scholastic settings, business API-based devices requiring a month to month membership, and even a couple of calculations distributed in the NLP writing.

2.1.3: Feature Extractions Tools

The choice of the instruments to test was done as needs be to the 5 best Data Mining devices. The decision was at the caution of ease of use; all have a Graphical User Interface (GUI) yet just four are conceivable to use without scripting. To utilize JHepWork is fundamental to have capabilities to Jython programming language. We chose to contemplate just those that an examiner can utilize: KNIME 2.6.0, Orange 2.6, Rapid Miner 4.6 and Weka 3.6

2.1.4: Feature Extraction

The relative information is extracted is a given dataset which was managed in the previous stage. After that, this information is occupied into consideration to analyze the negative and positive and polarity in a section that is significant to decisive the opinion of the characters .The main features like record or documents are measured as feature vectors that are used for the classification. There are some steps of a feature extraction. Different types of features, identified from literature review on sentiment analysis are categorized as under. Morphological Types, Frequent Features, Implicit Features are the types of feature categorization. Appropriate feature selection methods in sentiment analysis have got important part for recognizing relevant features and increasing classification accuracy. It is the methods that are gathered into four main categories NLP or heuristic based, Statistical, Clustering based and Hybrid [26].

CHAPTER 3
EMOTION MINING APPROACHES

3.1 Literature Review

Sentiment analysis is the mechanized procedure of understanding an assessment about a given subject from composing or spoken language. In reality as we know it where we create 2.5 quintillion bytes of information consistently, assumption investigation has turned into a key apparatus for understanding that information. This has enabled organizations to get key bits of knowledge and mechanize all sorts of procedures. Yassine & Hajj work on their paper about Facebook data. They proposed another structure for portraying enthusiastic collaborations in informal organizations, and afterward utilizing these qualities to recognize companions from colleagues. The objective is to extricate the passionate substance of writings in online interpersonal organizations. The intrigue is whether the content is an outflow of the essayist's feelings or not. For this reason, content mining methods are performed on remarks recovered from an informal organization. The system incorporates a model for information gathering, database diagrams, information handling and information mining steps. The casual language of online informal communities is a primary concern to consider previously playing out any content mining strategies. This is the reason the system incorporates the advancement of uncommon dictionaries. In general, the paper displays another viewpoint for contemplating fellowship relations and feelings' appearance in online social systems where it manages the idea of these locales and the nature of the language utilized. It considers Lebanese Facebook clients as a contextual investigation. The method received is unsupervised; it essentially utilizes the k-implies grouping calculation. Investigations appear high exactness for the model in both deciding subjectivity of messages and foreseeing fellowship [1]. Zhang et. al. build up an assessment ID framework called SES which executes three diverse feeling ID calculations. In the 2nd calculation, they figure supposition ought not be essentially delegated negative, positive, and target, however a persistent score to replicate assessment degree. All word scores are determined dependent on an extensive volume of client audits. Because of the unique qualities of internet based life, writings, they propose a third calculation which takes images, invalidation word position, and area explicit words into the record. Besides, an AI demonstrates is utilized on highlights got from yields of three calculations. They direct their examinations on client remarks from Facebook and tweets from twitter. The outcomes demonstrate that using Random Forest will obtain a superior exactness than choice

tree, neural system, and calculated relapse. They additionally propose an adaptable method to speak to record slant dependent on assessments of each sentence contained. SES is accessible on the web [2]. Walt, R. apply data mining and AI systems to anticipate clients' identifying characteristics (explicitly, the attributes of the Big Five identity display) utilizing just statistic and content based characteristics separated from their profiles. We at that point utilize these forecasts to rank people as far as the five qualities, anticipating which clients will show up in the top or base 5% or 10% of these qualities. Our outcomes demonstrate that when utilizing certain models, we can locate the top 10% most Open people with almost 75% exactness, furthermore, over all attributes and bearings; we can foresee the top 10% with at any rate 34.5% exactness (surpassing 21.8%, which is the best precision when utilizing only the best-performing profile quality). These consequences have security suggestions regarding enabling publicists and different gatherings to focus on a specific subset of people reliant on their identity characteristics [3]. Akaichi, J. centralized the use of content digging for opinion order. The outline is performed on Tunisian clients' statuses on Facebook posts amid the "Arabic Spring" time. Their point is to remove valuable data, about clients' assessments and practices amid this delicate and noteworthy period. For that reason, they propose a strategy dependent on Naive Bayes and Support Vector Machine (SVM). They furthermore develop a notion vocabulary, in light of the emojis, contributions and abbreviations', from extricated statuses refreshes. Additionally, we play out some relative investigations between two AI calculations SVM and Naive Bayes through a preparation show for slant order. This paper investigates the utilization of content mining procedures for supposition characterization. This is performed on statuses refreshes so as to dissect the Tunisians' practices amid the unrest (December-January, 2011). For this reason, they pick an irregular populace having Facebook accounts. It incorporates guys and females, understudies, specialists, housewives, and so forth. The time of focused populace is changing somewhere in the range of 21 and 54 years of age [4]. Troussas et. al. said that in their paper work the essential and basic thought is that the reality of realizing how individuals feel about specific themes can be considered as a grouping task. Individuals' sentiments can be sure, negative or impartial. An assessment is regularly spoken to in inconspicuous or complex courses in content. An online client can utilize an assorted scope of different strategies to express his or her feelings. Aside from that, s/he may blend objective and abstract data about a specific subject. In addition,

information accumulated from the World Wide Web frequently contains a great deal of commotion. Undoubtedly, the errand of programs estimation acknowledgment in online content turns out to be progressively troublesome of all the previously mentioned reasons. Consequently, they present how estimation examination can help language learning, by invigorating the instructional process and exploratory outcomes on the Naive Bayes Classifier [5]. Gil, G.B. adopt a regulated strategy to the issue, yet influence existing hashtags in Twitter for structure our preparation information. At long last, we tried the Spanish passionate corpus applying two distinctive AI calculations for feeling distinguishing proof coming to about 65% exactness. They manage the subject of perceiving individuals' feeling setting by investigating information from Twitter (Microblogging stage) [6]. Isah, H. work in advancement with commitments, including: the improvement of a system for social occasions and examining the perspectives and encounters of clients of medication and corrective items utilizing AI, content mining and opinion mining; the utilization of the proposed system on Facebook remarks furthermore, information from Twitter for brand investigation, and the depiction of step by step instructions to build up an item security vocabulary and preparing information for demonstrating an AI classifier for medication and restorative item conclusion expectation. The underlying brand and item correlation results mean the value of content mining and assessment investigation via web-based networking media information while the utilization of a machine learning classifier for anticipating the notion introduction gives a helpful device to clients, item producers, administrative and implementation offices to screen brand or item supposition drifts so as to act in case of unexpected or noteworthy ascent in negative assumption [7]. Bhuta et. al. said that their methods have been utilized for breaking down the assessment of content, be it an archive or a tweet, are Investigated. These methods run from straightforward dictionary based ways to deal with direct learning techniques. The learning based techniques incorporate Naive-Bayes classifiers, Maximum Entropy technique and Support Vector Machines. A half breeds system, mark proliferation, which utilizes a blend of the above techniques and furthermore joins a Twitter Follower diagram for mark dispersion is likewise talked about. Notwithstanding the procedures for notion investigation, the paper likewise features a number of issues and moves that should be defeated for notion investigation of Twitter information. These issues for the most part incorporate the impediments and focal points of various classifiers, flexibility to Twitter traditions and language

use and furthermore the connection between auxiliary properties of systems and inferring the slant of a populace [8]. Ortigosa, A. presents a new method for sentiment analysis in Facebook that starting from messages written by users, supports to extract information about the users' sentiment polarity (positive, neutral or negative), as transmitted in the messages they write. We have implemented this method in SentBuk, a Facebook application also presented in this paper. SentBuk retrieves messages written by users in Facebook and classifies them according to their polarity, showing the results to the users through an interactive interface. It also supports emotional change detection, friend's emotion finding, user classification according to their messages, and statistics, among others. On the other hand, the students' sentiments towards a course can serve as feedback for teachers, especially in the case of online learning, where face-to-face contact is less frequent. They discovered that, for users with higher activity is detected and positive sentiment changes,. Conversely, the negative sentiment changes for every user, the action is either rather higher or rather lower, but not similar to the regular activity [9]. Ngoc, P.T. proposed a content-based positioning technique in which the client commitment and the remark extremity are altogether considered. The client remark breaks down utilizing a dictionary based approach. They apply the proposed strategy for the genuine Facebook dataset gathered utilizing the Social Packets crawler. The outcome demonstrates that the positions of pages evaluated by our strategy are near the positions assessed by commitment based strategy. All the more critically, by concerning the remark extremity, our page positioning is increasingly exact with respect to client assessment [10]. Duwairi et. al. managed feeling investigation in Arabic surveys from an AI point of view. Three classifiers were connected with an in-house created data set of tweets/remarks. Specifically, the Naïve Bayes, SVM and K-Nearest Neighbor classifiers were kept running on this dataset. The outcomes demonstrate that SVM gives the most elevated exactness while KNN gives the most astounding review [11]. Kandasamy, K. focus extra on spammers on Twitter. And Twitter is a micro blogging web-page that permits just a more extreme of 140 characters in each message (tweet). The four noteworthy sorts of spammers on Twitter that we have deliberated in this paper are: Marketers, Phishers, Malware Propagators, and Adult Content Propagators. They are suggesting an application which can characterize a Twitter client into spam or genuine. To achieve this, a coordinated methodology that contains URL examination, Normal Language Processing and Machine Learning systems are

utilized [12]. Bing et. al. proposed a technique to dig Twitter information for answers to these inquiries. In particular, they propose to utilize an information mining calculation to decide whether the cost of a choice of 30 organizations recorded in the NASDAQ and the New York Stock Exchange can really be anticipated by the given 15 million records of tweets such as- Twitter messages. They do as such by extricating vague printed tweet information through NLP procedures to characterize open assumption, at that point utilize an information mining strategy to find designs between open feeling and genuine stock value developments. With the proposed calculation, they figure out how to find that it is feasible at the stock shutting cost of certain organizations to be anticipated with a normal precision as high as 76.12%. In this paper, they depict the information mining calculation that we use and talk about the key discoveries in connection with the inquiries presented [13]. Chang, W.Y. present their primer trials on tweets estimation examination. This trial is intended to remove conclusion dependent on subjects that exist in tweets. It identifies the assumption that alludes to the particular subject utilizing Natural Language Processing systems. To group estimation, our trial comprises of three principle steps, which are subjectivity characterization, semantic affiliation, and extremity order. The trial uses supposition vocabularies by characterizing the linguistic connection between conclusion dictionaries and subject. Exploratory outcomes demonstrate that the proposed framework is working superior to current content feeling examination instruments, as the structure of tweets isn't same as custom content [14]. Xie, Y. said on their work a multilingual assumption ID framework (MuSES) actualizes three distinctive notions distinguishing proof calculations. The main calculation expands past compositional semantic principles by adding rules explicit to internet based life. The second calculation characterizes a scoring capacity that estimates the level of a supposition, rather than just ordering a slant into parallel polarities. Every such score is determined depending on an expansive volume of client audits. Because of the unique qualities of internet based life messages, a third calculation takes images, nullification word position, and area explicit words into the record. Likewise, a proposed mark frees procedure exchanges multilingual assessment, learning between various dialects. The writers direct their tests on client remarks from Facebook, tweets from Twitter, and multilingual item audits from Amazon [15]. Kanakaraj, M. includes investigating the inclination of the general public on a specific news from Twitter posts. The key thought of the paper is to build the exactness of order by including

Natural Language Processing Techniques (NLP) particularly semantics and Word Sense Disambiguation. The mined content data is exposed to Ensemble order to dissect the estimation. Gathering arrangement includes consolidating the impact of different autonomous classifiers on a specific arrangement issue. Trials led show that troupe classifier beats conventional AI classifiers by 3-5%. They have broken down the tweets are brought utilizing the Twitter API vl.l. The execution of Decision Tree, Random Forest, extremely API vl.l gives a progressively advanced programming interface Randomized Trees and Decision Tree relapse with Ada through which the list items can be gotten in Object Lift Classifiers on Twitter estimation examination [16]. In [17], the author has proposed the fundamental goal of this investigation is to think about the general exactness of two AI procedures (calculated relapse and neural organize) as for giving a positive, negative and unbiased supposition for stock-related tweets. The two classifiers are looking at utilizing Bigram term recurrence (TF) and Unigram term recurrence - reverse record term recurrence (TF-IDF) weighting plans. Classifiers are prepared utilizing a data set that contains 42,000 naturally commented on tweets. The preparation dataset frames positive, negative and unbiased tweets covering for innovation related stocks (Twitter, Google, Facebook, and Tesla) gathered utilizing the Twitter Search API. Classifiers give the equivalent results as far as in the general precision (58%). Be that as it may, experimental tests demonstrate that utilizing Unigram TF-IDF outflanks TF. The author in [18] has analyzed the propound of this study is to inquire influencers on Twitter to find the characteristics of their tweets through PIAR, an exclusive data mining research tool developed by the University of Salamanca that combines graph theory and social influence theory. Additionally, it delivers insights for practitioners and marketers on how to discover influencers talking about their brands by observing tweets' content. Thus, the main purpose of this study is to investigate users' Twitter behavior and its impact on their social influence. Sharma et. al. utilized Twitter Archived device to get tweets in Hindi language. They performed information (content) mining on 42,235 tweets gathered over a time of a month that referenced five national ideological groups in India, amid the crusading time frame for general state decisions in 2016. They utilized both managed and unsupervised methodologies. They used Dictionary Based, Naive Bayes and SVM calculation manufacture our classifier and grouped the test information as positive, negative and unbiased. They distinguished the feeling of Twitter clients towards every one of the thought about Indian ideological groups. The aftereffects of

the examination for Naive Bayes were the BJP (Bhartiya Janta Party), for SVM it was the BJP (Bhartiya Janta Party) and for the Dictionary Approach it was the Indian National Congress. SVM anticipated a 78.4% shot that the BJP would win more races in the general race because of the positive supposition they got in tweets. As it turned out, the BJP won 60 out of 126 voting public in the 2016 general race, definitely more than some other ideological group as the following party (the Indian National Congress) just won 26 out of 126 voting demographics [19]. As proposed in [20], the researchers acquaint an answer with concentrated furthermore, break down remarks of bosses` understudies from the Facebook scholarly gathering. The proposed strategy is actualized utilizing Facebook diagram API to remove the remarks and afterward those are characterized into three gatherings (for example positive, negative and impartial) utilizing Bayesian Network probabilistic models. The said framework may assist organizations with improving the learning condition by giving a distinctive performance and dissecting diverse suppositions from the understudies. Pandey, A.C. proposed a novel meta-heuristic technique (CSK) which depends on K-means and cuckoo seek. The proposed strategy has been utilized to locate the ideal group heads from the nostalgic substance of Twitter dataset. The adequacy of proposed strategy has been tried on various Twitter datasets and contrasted and standard tile swarm enhancement, differential development, quick look, improved cuckoo seek, Gauss-based cuckoo hunt, and two n-grams strategies. Test results and statistical investigation approve that the proposed technique beats the current strategies. The proposed technique has hypothetical ramifications for the future research to dissect the information produced through interpersonal organizations/media. This technique has additionally summed up practical suggestions for structuring a framework that can give decisive surveys on any social issues [21]. In [22], the researchers said that the internet based life has tremendous and prominence among every one of the administrations today. Information from SNS (Social Network Administration) can be utilized for a great deal of targets. For example, expects or notion investigation. Twitter is an SNS that has colossal information with client posting, with this critical measure of information; it has the capability of look into identified with content mining and could be exposed to estimation investigation. In any case, taking care of such an enormous measure of unstructured information is a troublesome undertaking; AI is required for taking care of such enormity of information. Profound learning is on the AI technique that utilizes the profound feed forward neural system with many

concealed layers in the term of neural system with the consequence of the trial about 75%. Bouazizi et. al. proposed an example based methodology that goes more profound in the characterization of writings gathered from Twitter. They characterize the tweets into 7 unique modules; anyway the method can be hurried to arrange into more modules. Tests demonstrate that our methodology achieves a precision of characterization equivalent to 56.9% and an accurate dimension of nostalgic tweets equivalent to 72.58%. In any case, the approach turns out to be precise in twofold order such as- order into "positive" and "negative" and ternary characterization order into "positive", "negative" and "impartial": in the previous case, they achieve an exactness of 87.5% of the equivalent data set utilized in the wake of expelling impartial tweets, and in the last case, they achieved an exactness of the order of 83.0% [23]. Karan & R.S. said that their framework is the primary goal is to examine tweets and result to positive, negative and impartial structure. The extra element is an area acknowledgment of tweets. In pre-preparing the proposed framework defeats the slang with online just as disconnected slang word reference and gets ongoing information. The ongoing tweets will assist a client with knowing about current actualities. The spell remedy is a vital pre-preparing part defeat in this framework. The name and substance extraction will tell the client that tweets have a place with any individuals or association. The area of tweets could be utilized for any item audit, political issues, see, and so on the extra component of the area make framework extraordinary and slang substitution increment the precision of framework and make framework increasingly dependable. The framework connected method lays Naive Bayes and semi administered. The yield is shown in type of pie diagrams or visual charts [24]. Arora et. al. proposed a short review of work done on sentiment analysis over social media applications beside with various stages and levels of sentiment analysis has been argued. This paper deliberates sentiment analysis, its phases, and levels along with procedures. It will only work for English language but in future it can be used for Multilingual. Also, various opinion summarization algorithms can be applied to generate a summary of reviews are given by users [25].

The following table shows the algorithms used on various platform.

Table 3.1: Classification and Feature extraction techniques Table

Serial No.	Classification and Feature extraction techniques	Author Name	Platform
01	K-means clustering	Mohamed Yassine, Hazem Hajj	Facebook
02	Random Forest, Neural Network	Kunpeng Zhang, Yu Cheng, Yusheng Xie, Daniel Honbo Ankit Agrawal, Diana Palsetia, Kathy Lee, Wei-keng Liao, and Alok Choudhary	Facebook, Twitter
03	Machine Learning Technique	Randall Wald and Taghi Khoshgoftaar, Chris Sumner	Facebook
04	Support Vector Machine(SVM) & Naive Bayes	Jalel Akaichi, Zeineb Dhouioui, Maria José López-Huertas Pérez	Facebook
05	Naive Bayes	Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, Jaime Caro	Facebook
06	Natural Language Processing(NLP)	Gonzalo Bl´azquez Gil, Antonio Berlanga de Jes´us, and Jos´e M. Molina Lop´ez	Twitter
07	Naive Bayes	Haruna Isah, Paul Trundle, Daniel Neagu	Facebook, Twitter
08	Naive Bayes, SVM	Sagar Bhuta, Avit Doshi, Uehit Doshi, Meera Narvekar	Twitter

09	Lexicon based	Alvaro Ortigosa, José M. Martín, Rosa M. Carro	Facebook
10	Lexicon based	Phan Trong Ngoc, Myungsik Yoo	Facebook
11	Naive Bayes, SVM & KNN	Rehab M. Duwairi, Islam Qarqaz	Twitter
12	NLP, Naïve Bayes, SVM	Kamalanathan Kandasamy, Preethi Koroth	Twitter
13	Natural Language Processing(NLP)	LI Bing, Keith C.C. Chan, Carol OU	Twitter
14	NLP	Wei Yen Chong, Bhawani Selvaretnam, Lay-Ki Soon	Twitter
15	Compositional Semantic Rule, Numeric Sentiment Identification, Bag-of-Words and Rule-Based	Yusheng Xie, Zhengzhang Chen, Daniel K. Honbo, Kunpeng Zhang, Yu Cheng, Ankit Agrawal, and Alok N. Choudhary	Facebook, Twitter
16	Natural Language Processing(NLP), SVM, Naive Bayes, Random Forest, MaxEntropy	Monisha Kanakaraj and Ram Mohana Reddy Guddeti	Twitter
17	Neural Network, SVM	Mohammed Qasem, Ruppa Thulasiram, Parimala Thulasiram	Twitter
18	Graph theory	Eva Lahuerta-Otero, Rebeca Cordero-Gutierrez	Twitter
19	Naive Bayes & SVM	Parul Sharma, Teng-Sheng Moh	Twitter
20	Bayesian Network	Nisha Tanwani, Sandesh Kumar, Akhtar Hussain Jalbani, Saima	Facebook

		Soomro, Muhammad Ibrahim Channa, Zeeshan Nizamani	
21	K-means	Avinash Chandra Pandey, Dharmveer Singh Rajpoot, Mukesh Saraswat	Twitter
22	Neural Network	Adyan Marendra Ramadhani, Hong Soon Goo	Twitter
23	Pattern-Based	Mondher Bouazizi, Tomoaki Ohtsuki	Twitter
24	Naive Bayes	Ritu S.Karan, Pooja L. Kasar, Kavita K. Shirsat, Reshma Chaudhary	Twitter
25	Naive Bayes, SVM	Akankasha and Bhavna Arora	Twitter, Facebook

We have prepared another table to represents the accuracy rate of different sentiment analysis techniques on social media data.

Table 3.2: Accuracy Rate Table

Serial No.	**Classification and Feature extraction techniques**	**Accuracy Rate %**	**Platform**
01	Machine Learning Technique	75%	Facebook
02	Natural Language Processing(NLP)	65%	Twitter
03	Lexicon based	83.27%	Facebook
04	Natural Language Processing(NLP)	76.12%	Twitter
05	Neural Network, SVM	58%	Twitter
06	Graph theory	36%	Twitter
07	Naive Bayes & SVM	78.4%	Twitter
08	Neural Network	75%	Twitter
09	Pattern-Based	83%	Twitter

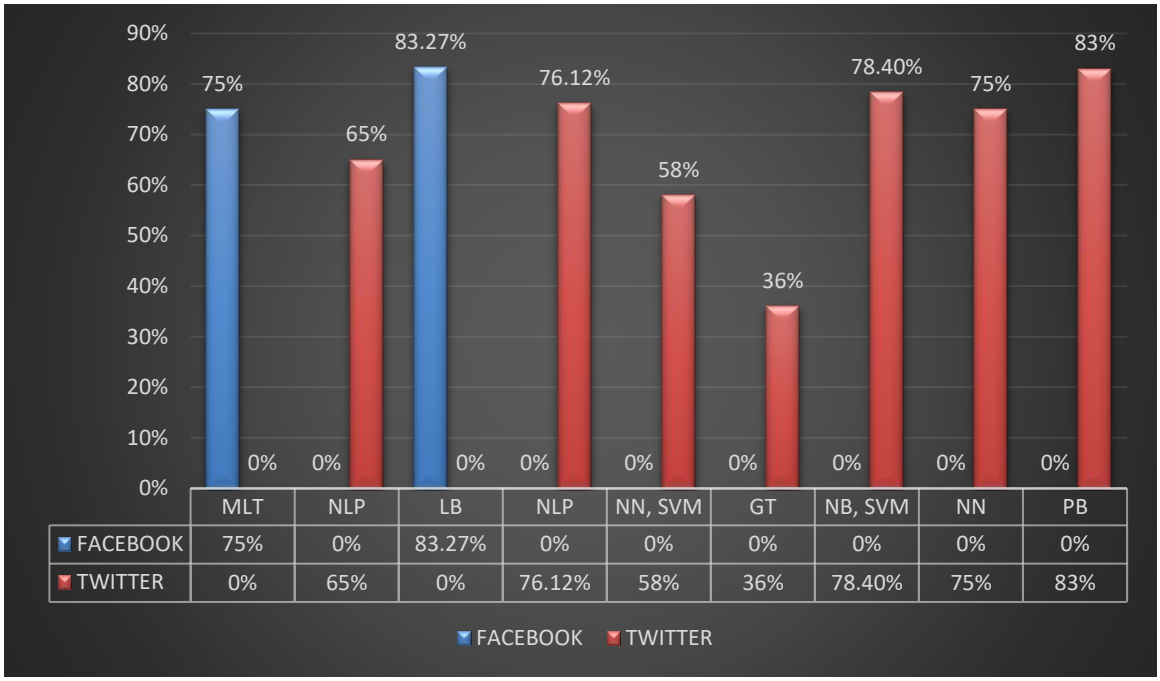

Figure 3.2: Accuracy rate for Algorithms on Social Media Data

Table 3.3: Uses of SVM on social media data

Serial No.	Publishing Year	Platform
1.	2013	Facebook
2.	2014	Twitter
3.	2014	Twitter
4.	2014	Twitter
5.	2015	Twitter
6.	2015	Twitter
7.	2016	Twitter
8.	2019	Facebook, Twitter

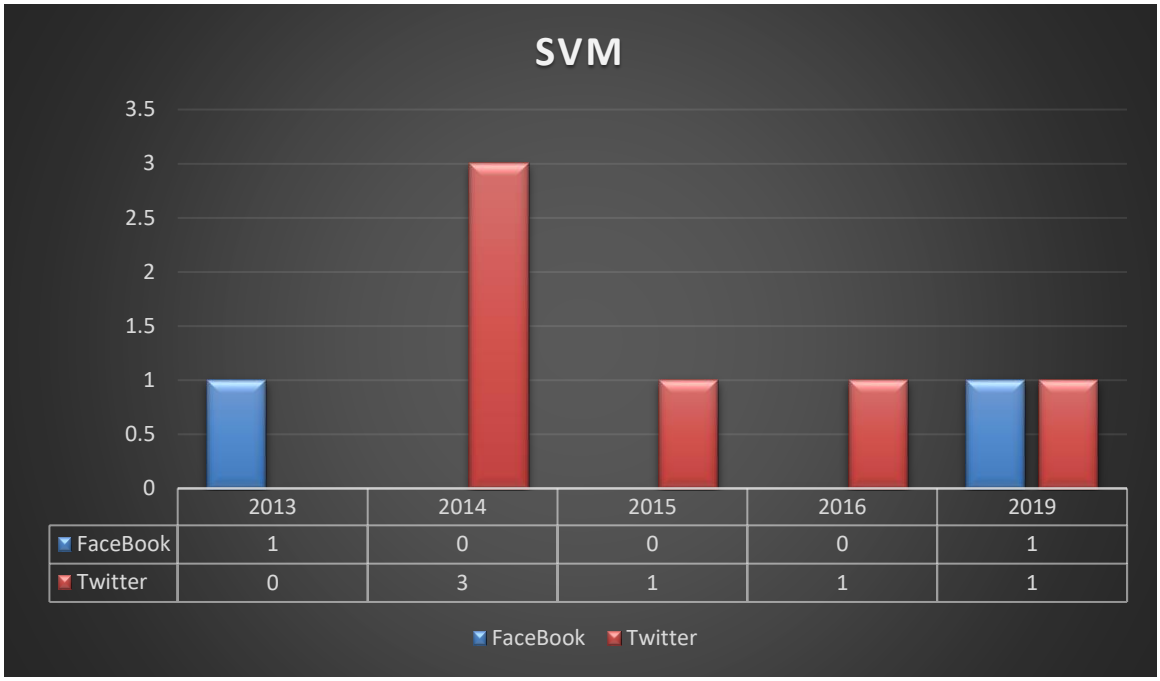

Figure 3.3: Bar Diagram of SVM

Table 3.4: Uses of Naïve Bayes on social media data

Serial No.	**Publishing Year**	**Platform**
1.	2013	Facebook
2.	2013	Facebook
3.	2014	Facebook, Twitter
4.	2014	Twitter
5.	2014	Twitter
6.	2014	Twitter
7.	2015	Twitter
8.	2016	Twitter
9.	2018	Twitter
10.	2019	Facebook, Twitter

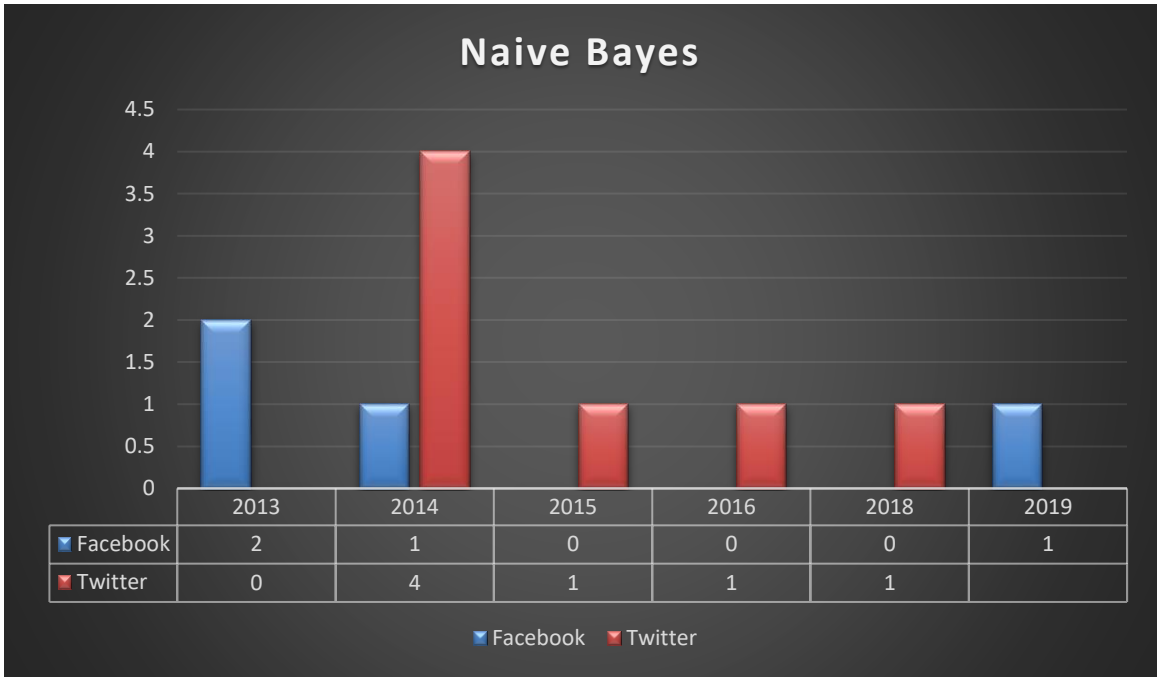

Figure 3.4: Bar Diagram of Naïve Bayes

Table 3.5: Uses of Natural Language Processing on social media data

Serial No.	**Publishing Year**	**Platform**
1.	2013	Twitter
2.	2014	Twitter
3.	2014	Twitter
4.	2014	Twitter
5.	2015	Twitter

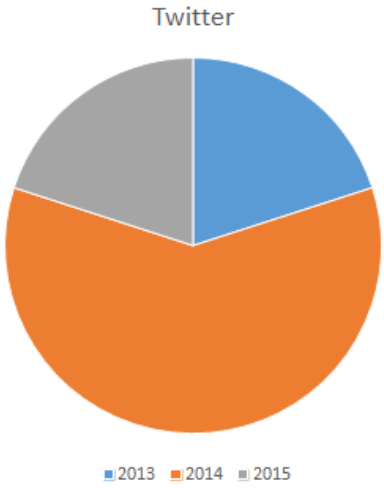

Figure 3.5: Pie chart of NLP

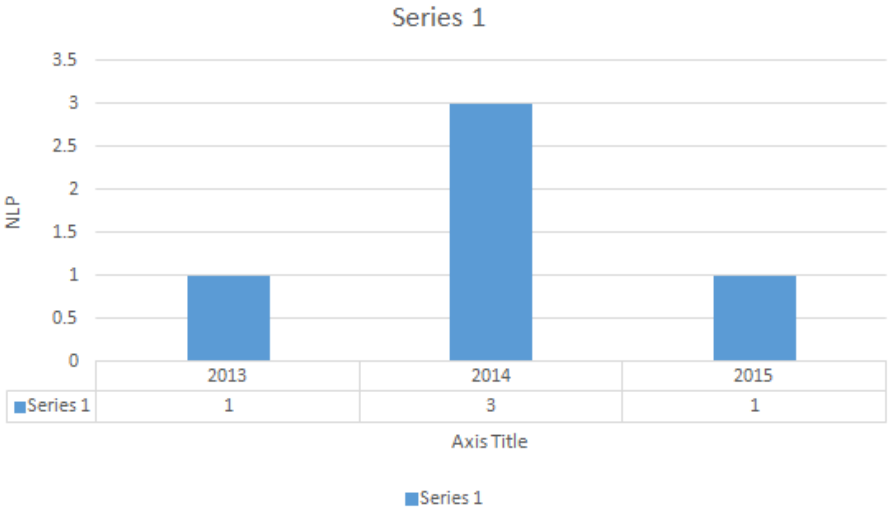

Figure 3.5.1: Bar Diagram of NLP

CHAPTER 4

Future Scope Further Development

4.1 Future Scope Further Development

One of the key factors that impact the social media is the way the users express their emotion online. In this comparative study we considered only Facebook and twitter platform. Improving Accuracy rate of each algorithm can be the future direction of research work for all researchers in the field of Emotion mining or sentiment analysis. In accordance with this effort, we plan to start an elaborate study of the emotion mining problem with a various number of platforms like Instagram, IMDB and other Microblogging site.

Conclusion

In this paper we concentrate only on opinion mining and show the position of this field in social media mining. Emotion mining is a rising research field so, in this paper we have concentrated on related work performed in this area between 2010 to 2019 and identify directions for future work. As described we found the most used algorithms in this area of research are K-Means, lexicon based, SVM, Naïve Bayes, NLP, Random forest. Improving Accuracy rate of each algorithm can be the future direction of research work for all researchers in the field of Emotion mining. Current studies are intended to pave the way for further researches and development activities by identifying weaknesses and deriving guidelines. In accordance with this effort, we plan to start an elaborate study of the emotion mining problem with a number of platforms.

Reference

[1] Yassine, M., & Hajj, H. (2010, December). A framework for emotion mining from text in online social networks. In *2010 IEEE International Conference on Data Mining Workshops*(pp. 1136-1142). IEEE.

[2] Zhang, K., Cheng, Y., Xie, Y., Honbo, D., Agrawal, A., Palsetia, D., ... & Choudhary, A. (2011, December). SES: Sentiment elicitation system for social media data. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 129-136). IEEE.

[3] Wald, R., Khoshgoftaar, T., & Sumner, C. (2012, August). Machine prediction of personality from Facebook profiles. In *2012 IEEE 13th International Conference on Information Reuse & Integration (Iri)* (pp. 109-115). IEEE..

[4] Akaichi, J., Dhouioui, Z., & Pérez, M. J. L. H. (2013, October). Text mining facebook status updates for sentiment classification. In *2013 17th International conference on system theory, control and computing (ICSTCC)* (pp. 640-645). IEEE.

[5] Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J. (2013, July). Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. In *IISA 2013* (pp. 1-6). IEEE.

[6] Gil, G. B., de Jesús, A. B., & Lopéz, J. M. M. (2013, May). Combining machine learning techniques and natural language processing to infer emotions using Spanish Twitter corpus. In *International Conference on Practical Applications of Agents and Multi-Agent Systems* (pp. 149-157). Springer, Berlin, Heidelberg.

[7] Isah, H., Trundle, P., & Neagu, D. (2014, September). Social media analysis for product safety using text mining and sentiment analysis. In *2014 14th UK Workshop on Computational Intelligence (UKCI)* (pp. 1-7). IEEE.

[8] Bhuta, S., Doshi, A., Doshi, U., & Narvekar, M. (2014, February). A review of techniques for sentiment analysis Of Twitter data. In *2014 International conference on issues and challenges in intelligent computing techniques (ICICT)* (pp. 583-591). IEEE.

[9] Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior*, *31*, 527-541.

[10] Ngoc, P. T., & Yoo, M. (2014, February). The lexicon-based sentiment analysis for fan page ranking in Facebook. In *The International Conference on Information Networking 2014 (ICOIN2014)* (pp. 444-448). IEEE.

[11] Duwairi, R. M., & Qarqaz, I. (2014, August). Arabic sentiment analysis using supervised classification. In *2014 International Conference on Future Internet of Things and Cloud* (pp. 579-583). IEEE.

[12] Kandasamy, K., & Koroth, P. (2014, March). An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques. In 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science (pp. 1-5). IEEE.

[13] Bing, L., Chan, K. C., & Ou, C. (2014, November). Public sentiment analysis in Twitter data for prediction of a company's stock price movements. In *2014 IEEE 11th International Conference on e-Business Engineering* (pp. 232-239). IEEE.

[14] Chong, W. Y., Selvaretnam, B., & Soon, L. K. (2014, December). Natural language processing for sentiment analysis: an exploratory analysis on tweets. In *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology* (pp. 212-217). IEEE.

[15] Xie, Y., Chen, Z., Zhang, K., Cheng, Y., Honbo, D. K., Agrawal, A., & Choudhary, A. N. (2014). MuSES: multilingual sentiment elicitation system for social media data. *IEEE Intelligent Systems*, *29*(4), 34-42.

[16] Kanakaraj, M., & Guddeti, R. M. R. (2015, February). Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)* (pp. 169-170). IEEE.

[17] Qasem, M., Thulasiram, R., & Thulasiram, P. (2015, August). Twitter sentiment classification using machine learning techniques for stock markets. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 834-840). IEEE.

[18] Lahuerta-Otero, E., & Cordero-Gutiérrez, R. (2016). Looking for the perfect tweet. The use of data mining techniques to find influencers on Twitter. *Computers in Human Behavior*, *64*, 575-583.
[19] Sharma, P., & Moh, T. S. (2016, December). Prediction of indian election using sentiment analysis on hindi twitter. In *2016 IEEE International Conference on Big Data (Big Data)*(pp. 1966-1971). IEEE.
[20] Tanwani, N., Kumar, S., Jalbani, A. H., Soomro, S., Channa, M. I., & Nizamani, Z. (2017, November). Student opinion mining regarding educational system using facebook group. In *2017 First International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT)* (pp. 1-5). IEEE.
[21] Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, *53*(4), 764-779.
[22] Ramadhani, A. M., & Goo, H. S. (2017, August). Twitter sentiment analysis using deep learning methods. In *2017 7th International Annual Engineering Seminar (InAES)* (pp. 1-4). IEEE.
[23] Bouazizi, M., & Ohtsuki, T. (2017). A pattern-based approach for multi-class sentiment analysis in Twitter. *IEEE Access*, *5*, 20617-20639.
[24] Karan, R. S., Shirsat, K. K., Kasar, P. L., & Chaudhary, R. (2018, March). Sentiment Analysis on Twitter Data: A New Aproach. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-4). IEEE.
[25] Arora, B. (2019). A Review of Sentimental Analysis on Social Media Application. In *Emerging Trends in Expert Applications and Security* (pp. 477-484). Springer, Singapore.
[26] Koncz, P., & Paralic, J. (2011, June). An approach to feature selection for sentiment analysis. In *2011 15th IEEE International Conference on Intelligent Engineering Systems*(pp. 357-362). IEEE.