PATTERN BASED SENTIMENT ANALYSIS ON SOCIAL NETWORKING SITES

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

ANU SHARMA

11503248

Supervisor

SAVLEEN KAUR

Assistant Professor

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

April 2017

LOVELY PROFESSIONAL UNIVERSITY

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

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

COURSE CODE : CSE546 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0004

Supervisor Name : Savleen Kaur **UID :** 18306 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Anu Sharma	11503248	2015	K1518	9780863681

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

PROPOSED TOPIC : Pattern-Based Sentiment Analysis on Social Networking Sites

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	7.40
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.00
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.00
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.40
5	Social Applicability: Project work intends to solve a practical problem.	7.20
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.20

PAC Committee Members		
PAC Member 1 Name: Prateek Agrawal	UID: 13714	Recommended (Y/N): Yes
PAC Member 2 Name: Pushpendra Kumar Pateriya	UID: 14623	Recommended (Y/N): Yes
PAC Member 3 Name: Deepak Prashar	UID: 13897	Recommended (Y/N): Yes
PAC Member 4 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 5 Name: Anupinder Singh	UID: 19385	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Pattern-Based Sentiment Analysis on Social Networking Sites**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11011::Dr. Rajeev Sobti **Approval Date:** 28 Oct 2016

4/26/2017 1:06:55 PM

ABSTRACT

To analyze or mining the microblog data is one of the hot topic from recent few years. Sentiment analysis is one of the techniques to analyze the microblog data. Sentiment analysis refers to identification of attitudes and opinion expressed by online to a specific topic or specific product. Sarcasm is one of the most ironies widely used in microblogs social network websites. Sarcasm is a different way to convey a message information one person to another. It is might be used in different manner like mockery to someone or lough on someone. Sarcasm detection is one of the important concepts to improve the data analysis, and improve the automatic sentiment analysis. We propose a pattern based approach for Sarcasm Detection on twitter (tweets). We proposed a technique how we do a sentiment analysis by using a pattern based approach. We also study the importance of each proposed set of features and its value to the classification.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "PATTERN-BASED SENTIMENT ANALYSIS ON SOCIAL NETWORKING SITES" 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 Ms. Savleen Kaur (18306). 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

Anu Sharma

11503248

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled **"PATTERN-BASED SENTIMENT ANALYSIS ON SOCAIL NETWORKING SITES",** submitted by **ANU SHARMA (11503248)** 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

Savleen Kaur (18306)

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

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant guidance crowned our efforts with success.

I sincerely express our deep gratitude to the management of our college for giving us liberty to choose and to work on the most relevant project i.e. "**A PATTERN BASED SENTIMENT ANALYSIS ON SOCIAL NETWORKING SITES**". I am thankful to **Dr. Dalwinder Singh** (HOD, CSE Department) for ensuring that we have a smooth environment in the university by providing us with the best suitable mentors according to our field. I would also like to thank the Research and Development department (R&D department).

I would like to thank my guide **Ms**. **Savleen Kaur (18306)**, **Assistant Professor, CSE Department,** who encouraged and insisted me in the formulation of problem definition & without her valuable guidance and constant inspiration it would have been difficult for me to prepare this project report.

TABLE OF CONTENTS

LIST OF ABBRIVATIONS

ABBRIVATION	MEANING OF ABBRIVATION
SA	Sentiment Analysis
Max. Ent.	Maximum Entropy
Rand. Forest	Random Forest
k-NN	K nearest Neighbour
SVM	Support Vector Machine
NLTK	Natural Language Toolkit
FED	Feature extraction and description
MHP	Multiple factor based Hybrid Pattern
SSAM	Sentiment and sarcasm analysis module
Ws	Words list
Sd	Sentic Dictionary
EvL	Emotion value obtained
EoF	End of file
Tdm	Tweet data matrix
Tw	Current tweet
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

LIST OF FIGURES

LIST OF TABLES

Checklist for Dissertation-III 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

Today is the world of information world, in this domain everything goes on internet and every single person use this services that provide on internet like E-Commerce network site (Flipkart.com, Snapdeal.com, Amazon.in etc.) And many community networking web sites like (Facebook.com, Twitter.com etc.). Sentiment analysis help to taking out the product and comment analysis that posted by a client on internet. Mining these remark help the production to improve their creation quality and give a new idea how they can show there products in web sites. There are many procedures that are used in sentiment analysis they can effort on dissimilar way that ways are like lexical-based sentiment analysis is not fine performance in some category like of comments take an example of bookmyshow.com this is an operational movie reservation site in this site users can post there remark about movie illustration posted a comment for movie like a movie is too good but the climax is not impressive these kind of review lexical-based classification is not work as much satisfactory level. Numerous sentiment analysis method.

Sentiment analysis is not one work in social network or manufactured goods review field but it can also very beneficial in other application area such as psychology, sociology perspective, political polls and business intelligence. Figure 1 shows how sentiment analysis can be done on data, following is the detail explanation for it

- **Data Collection:** The first step is data gathering data is collect from social networking site like Twitter, Facebook, and Blogs or may be gather from e-commerce site. This collected data is loud and cannot give respectable information so examination can be done on this data. Natural Language Processing or other script mining can be finished on this data to abstract an information
- **Data Preparation:** The another step is data grounding in this step loud data can be cleaned and prepare for sentiment analysis
- **Sentiment Detection:** In sentiment detection remark opinion in extract for data these opinion may be a facts, attitude or entities of review's that posted by user online about the product.

- **Sentiment Classification:** Text is classified according to negative, positive, better, amazing, bad, worst, good, and wonderful. The classification can be done on my point of views
- **Output:** The final step is represent the output of analysis data, the output is show in a graphical layout it may be in line chart, bar chat, pie chat or in any other graphical representation it include time average a point of classification that use in analysis on data for sentiment analysis.

The way of sentiment analysis can done show in figure 1.

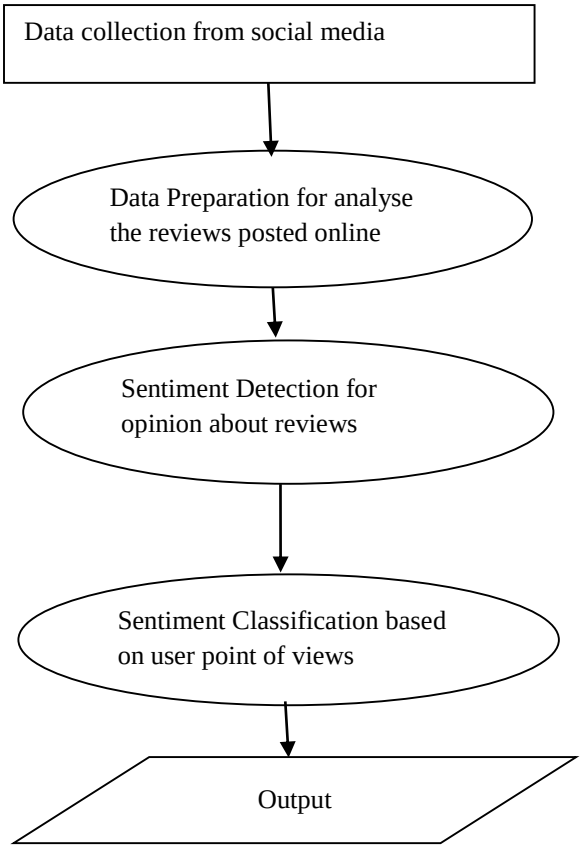

Figure 1.1 Sentiment Analysis for user review posted in online.

With the rise quality of social media, these micro-blog websites (like Twitter, Instagram) is a huge quantity of knowledge is generated. We all know that the internet is a huge collection of networks. The age of web has changed the approach of individuals specific

their thoughts and feelings. Each person is connected with other by the help of web social network sites and other microblog websites.

Twitter turn out to be one among the largest internet interesting point for those that give their thought, also they can share these opinion about that event etc. Preceding time and year, Twitter data is grow to extend, so constituting typical example of the supposed massive information. Nowadays, agreeing to its certified website, Twitter has quite 288 billion dynamic consumers, and quite five hundred billion twitter tweets are shown every day. Several corporations or civil service have remained curious about the information for the aim of learning the opinion of individuals on the way to politically aware events, movies current product. In twitter the limitation on characters that are written in tweet, so that's why the people prefer informal language to write their feelings and that is difficult to analyze. And use of sarcasm makes it more difficult to analyze. We can say that, Sarcasm is when a person say something but his/her intension is different from their words. In this paper author tell that in sentiment analysis sarcasm can change the polarity of messages.

We tend to suggest a method that depends on writing patterns, and we can consider dissimilar types of sarcasm and at that time identify their sarcastic tweets.
The main contributes of this paper are.

- We tend to suggest a collection of Pattern Based options, along with other choices to categorize tweets.
- We analysis the main way which irony is recycled in microblog websites.
- We can identify an effective way to spot ironic tweets and improve the efficiency of sentiment analysis
- Categorize the tweets into dissimilar modules (The proposed classification of tweets dataset is hypothetical)

Social networks in addition microblogging websites like twitter take the topic to several studies within the fresh a small number of years. Programmed sentimentality study and judgment mining gift a boiling issue of study. Community webs present a large supply of knowledge instead of the attitudes of a user's, however absolutely random classes of consumers and clients that are using these manufactured goods services. Though, nice one

to the similar language used, the existence of non-textual content and there for the usage of colloquial speech word sand shortenings, arrangement of information take out from such social networks social network sites is quite a difficult job. The hidden sentiment identification that identify the original feeling rather than the sentiment polarity, which means that, capacity have a dissimilar sentiment polarity for the same word. To proof of identity of these importance and polarity of the words among other as the greatest difficult and challenging task in front of a sentiment analysis of social networking websites. In Twitter such ironies tweets are text like that, example: - *"The all product is incredible!!!" "Well, it's true but these products are so costly no one can buy these products😢"* You can think it's a complement but see the second line, the buyer is explicitly tell what he want to say.

Proceeding a linked context, the state of the art planned methods are mostly specifying in the positive, negative and neutral sentimentality organization. In alternative words, they categorize texts either obsessed by "positive" also "negative", or into "positive", "negative" plus "neutral". Still, to review the belief of a user, it might be additional attention grabbing to travel shallower within the grouping, and find out the sentimentality unseen behind his post. We can identify the pattern that help to identify whether a tweet is sarcastic or not and build a framework that help us to experiment to identify sarcasm in tweets.

1.1 WHY DATASET TWITTER IS USED FOR SENTIMENT ANALYSIS?

Twitter is one of the most famous social networking website. In which people can openly share their opinion on every topic like Policies, Business and any other topic. Twitter is very different from other social networking web sites like Facebook etc. because they do not provide a privacy on posts. Twitter tweets are limited by the character in post it allow only 140 word in single tweet so that why to analyze the sentiment of tweet is difficult. Now these day sarcasm is one of the hot topic that mostly used in twitter by using a hash tag. So that's why am choosing a twitter dataset of my research.

1.2 SENTIMENT ANALYSIS

Sentiment Analysis is a way that convey that product information is satisfactory to buy a product or not? So that's why sentiment analysis is important to do. I twitter the textual information is tell about the objective information it not convey the attitude or opinion of the tweet and some time the person write something and its meaning is something else, so that we can need to do sentiment analysis.

Sentiment Analysis is a technique that detect the attitude, opinion and behaviour of the tweets and any other social networking dataset. In sentiment analysis the word like opinion, view, belief and sentiment have different meanings.

- **Opinion:** The opinion is interested in question like example "I do not like demonetization" in this example the person show there interest in demonetization. Every person have their different opinion about same statement
- **View:** A view is a subjective idea about statement
- **Belief:** A belief is a planned acknowledgement about scholar content
- **Sentiment:** Sentiment is represent the felling of person

How sentiment expressions work in this example "I do not like demonetization"

<SENTENCE>= I do not like demonetization

<OPINION HOLDER>=Author

<OBJECT>=demonetization

<FEATURE>= do not

<OPINION>=like

<SENTIMENT>=negative

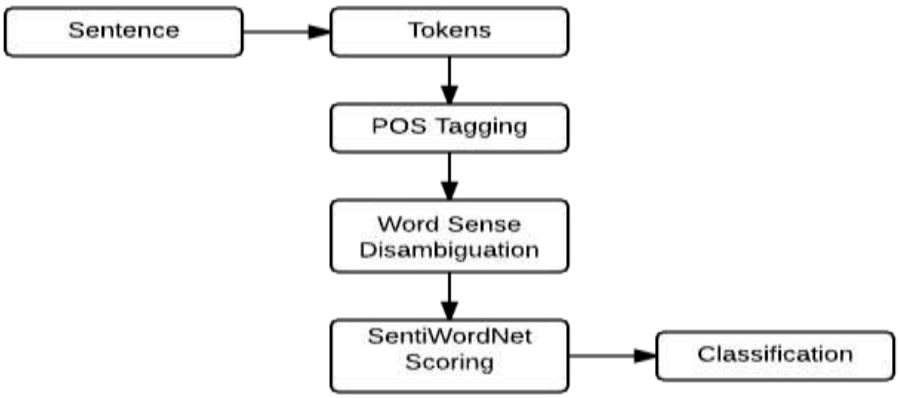

Figure 1.2 Sentiment Analysis Architecture

1.2.1 PHASES FOR SENTIMENT ANALYSIS

1. **Pre-processing of Dataset:** In present work twitter data is used that contain reviews for a particular topic or product these reviews are expressed by different peoples by their own ways. Twitter dataset is used to detect the sarcastic tweets after finding them labeled as sarcastic or non-sarcastic, if the tweet is sarcastic it labeled as positive and if it is not then labeled as negative. The twitter data is inconsistence and contain a repeated text to remove these inconsistence we can applied pre-processing. To perform pre-processing applied following steps:
    - Eliminate the URL form tweets and remove the handle @username
    - Replace the similes with words e.g. ☺ To happy
    - Eliminate the symbols and punctuations
    - Eliminate stop words
    - Eliminate the other language tweets like Hindi, Punjabi, and Spanish etc.
2. **Trained Dataset: -** In supervised learning approach the trained dataset is help to predict the result and provide an appropriate results. We can trained a data based on labels positive for sarcastic or negative for non-sarcastic.

3. Classification types that are used in current system

1. **Naive Bayes:** Naive Bayes is a classification method that is based on Bayesian Classification. The Bayesian Classification is a supervised machine learning approach or method. A Bayesian Classification is also a statistical classification method that assume the probabilistic model and help to getting a proper outcome by determining the probabilities.

   The Naive Bayes classification is planned when predictors are self-regulating from one of alternative within each session, but it will be work fine in practice even when that individuality statement is not correct. It classified data in two steps:

   1. **Training step:** Use the trained data, in this method evaluate the constraints of a possibility distribution, and guesses the predictors are provisionally self-regulating given the class.
   2. **Prediction step:** This step is done for any unseen data, in this the applied approach is computes the probability of the sample data that relate to the class. And after that the classifier is classifies the test data according to the large probability of next class.

$$C^* = argmac_c P_{NB}(c|d)$$

$$P_{NB}(c|d) = \frac{(P(c)) \sum_{i=1}^{m} p(f|c)^{n_{i(d)}}}{P(d)}$$

Equation 1 Naive Bayes

2. **SVM (Support Vector Machine):** SVM is a controlled machine learning approach that is used for classification and regression propose.

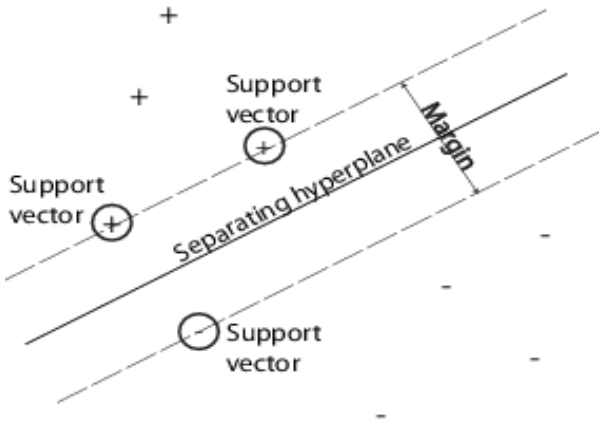

Figure 1.3 Support Vector Machine

The advantages of SVM are:

- Support vector machine be situated in most effective in high dimensional dataset
- SVM is also effective in such cases where number of input is grater then numeral of samples.
- This is a most memory effective because it use support vector (teaching points) in decision function

The disadvantages of support vector machines are:

- In any case the number of features that are extract by user is greater than the number or sample then in this case the performance is likely compromised.

Classification algorithm differential analysis shown in table 1.1

Table 1.1 Classification algorithm differential analysis

Features	Naive Bayes	Max Entropy	Boosted Tree	Random Forest
Based on	Bayes theorem	Feature Based Classifier	Decision Tree Learning	Decision Tree Aggregation
Simplicity	Very Simple	Hard	Moderated	Simple
Performance	Better	Good	Good	Excellent
Accuracy	Good	High	Poor	Excellent
Memory requirement	Low	High	Low	High
Other Applications	Spam Detection, Sentiment Analysis, Document classification	Diagnosis test in pathology labs	Classifying Cardiovascular outcomes	Bio medical application
Result Accuracy over a period of time	Variable	Consistent	Incremental	Incremental

Advantages and disadvantages of supervised algorithm is shown in table 1.2

Table 1.2 Advantages and disadvantages of algorithms

Algorithms	Benefits	Drawbacks
Naïve Bayes	• Self-determining Assumptions • Can be expert with small amount of facts • Executes well with self-determining features	• Restrictions in applicability • We could drop performance with the assumptions
Max Entropy	• execution is good with depended features	• Less Performance with self-determining features. • The feature selection could develop a complex

1.2.2 APPROACHES FOR SENTIMENT ANALYSIS

1. **Lexicon Based:** Lexicon based approach is an unsupervised technique, in this approach no need to maintain a large amount of training data set and rules. Which makes whole process is much faster. Lexicon approach is divided into dictionary-based and corpus-based to analyze the sentiment polarity. There are 5097 negative and 2533 positive word in lexicon linguistic dictionary all of words are define strong and weak polarity.
    - **Dictionary Based approach:** The main strategy of dictionary based it's working on manually create set of opinion that are repeated and then find their synonyms and antonyms by iterations and save these word in seed list these iterations repeat until no synonyms and antonyms are found. After that manually remove and correct errors [4].The limitation of dictionary approach its low applicability.
    - **Corpus Based approach:** Corpus approach improves the limitations and help to improve the finding opinion in particular area or orientation

Limitation of Lexicon: approach is that its cannot show high quality result in big amount of data, such that to analyse the movie review comments that posted

online this can't analyse well but this approach is good for small review data set like Facebook post comments or tweets.

2. **Machine Learning Approach:** Machine learning approach (ML) is use on many learning algorithms that are used for sentiment analysis in give dataset. ML is usually divided into supervised and unsupervised approach. Supervised ML approach have a pre-define or large amount of trained data set rules, But in unsupervised ML approach don't have any trained data set that's why it's difficult to find the level of trained dictionary rules in data set. In machine learning first trained the algorithm with some know data rules before apply it to a actual dataset. In machine learning the algorithm by supervised or unsupervised method.

   - **Supervised Machine Learning:** In supervised machine learning approach is made a model that is based on the prediction of evidence that are appear in uncertainty. The applied algorithm is responsible to find out the pattern in data that provide by user and the system is learn from the observation. When this observation is exposed the system performance is increased.

     Supervised machine learning a trained dataset is used to provide a correct outcomes, the outcomes of result is based on the classification and extraction of features from the data and labelled them as correct relation. In supervised machine learning have following algorithms linear classifier (it can use support vector machine and neural networks), Rule-based, Probabilistic classifier (it can use Naïve Bayes, Bayesian Networks)

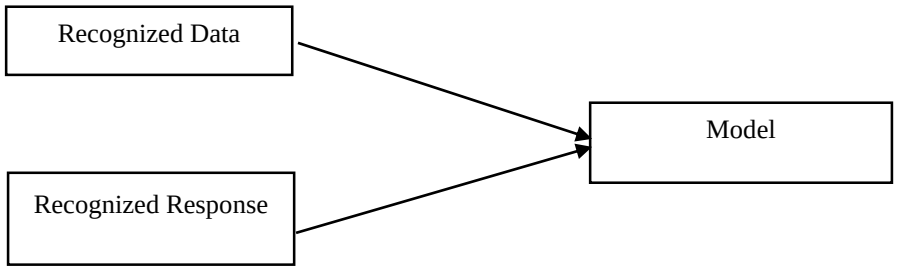

Figure 1.4 Supervised learning trained a data model

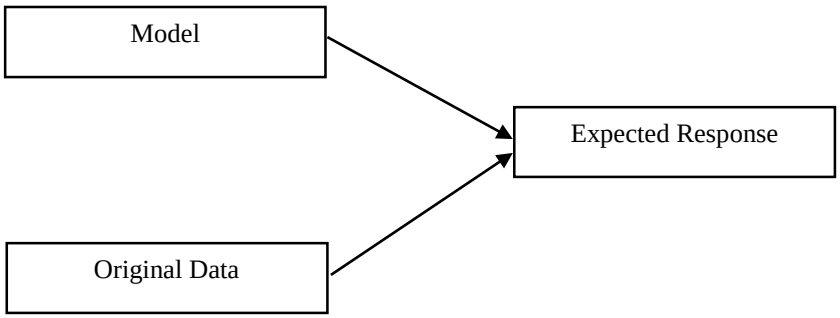

Figure 1.5 Use trained model to get a predicted result

The supervised learning is divided into two broader categories classification and regression

Some most common algorithms for supervised machine learning.

- Support Vector Machine
- Naive Bayes
- Decision Tree
- Neural Network
- Nearest Neighbor (k-NN)

- **Classification:** In classification the data is divided according to a target point like example data can be divided into positive or negative form or may be is positive, negative and natural form.
- **Regression:** In regression data can be divided into continuous form e.g. of decision tree and neutral network is work in regression

Some common Regression methods:

1. **Linear Regression:** Linear regression is a statistical demonstrating technique used to define an uninterrupted response variable as a function of one or more analyst variables. It can help you recognize and calculate the performance of complex structures or analyze investigational, financial, and biological data.
2. **Nonlinear Regression:** Nonlinear regression is a statistical method that helps describe nonlinear relationships in

investigational data. Nonlinear regression models commonly assumed to be parametric, where the model is labelled as a nonlinear equation. Typically machine learning approaches are used for non-parametric nonlinear regression.

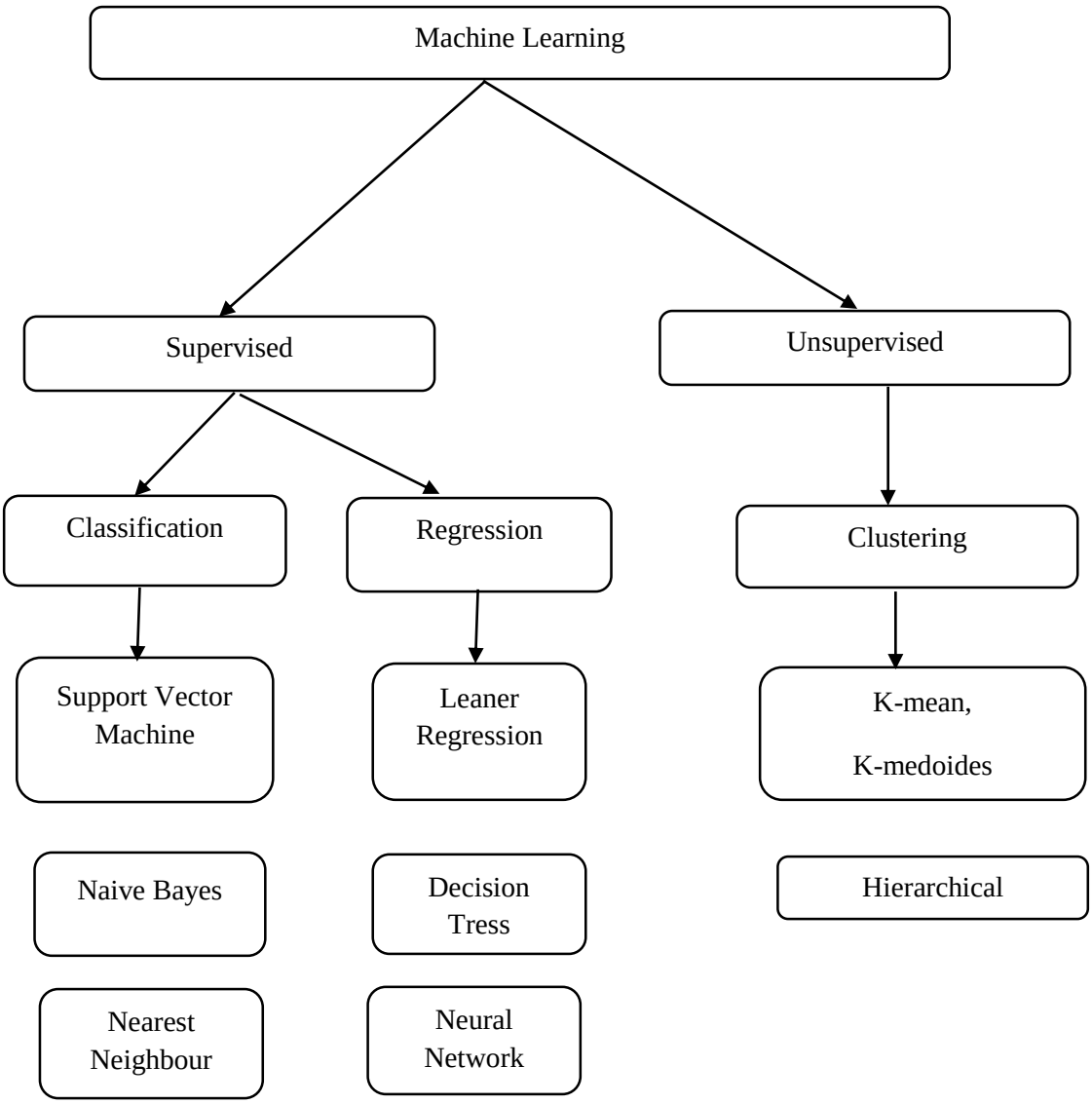

Figure 1.6 Classification of Machine learning

- **Un-supervised Machine Learning:** Unsupervised machine learning approach do not contain any trained dataset. In this approach do not have a correct target it used in clustering.

  Common clustering algorithms include:

  - **Hierarchical Clustering:** Use this for make a hierarchy of clusters by using a cluster tree
  - **K-Means clustering**: dividing data into k distinct clusters based on the detachment to the centroid of a cluster
  - **Gaussian mixture models**: this approach is a mixture of multivariate normal density mechanisms

Table 1.3 supervised Machine Learning Algorithms

Algorithms	**Approach**	**Accuracy**
Naive Bayes	Calculate the probability of element and then multiply with must likelihood to get final probability	Low
Support Vector machine	To calculate sentiment analysis SVM use discriminative classifier	High
Centroid Classifier	This algorithm assign the centroid vector (CV) to different training classes and use CV to get similar values.	High
K-Nearest Neighbour	In this algorithm classification can be done on the basis of similar score of neighbour	High

Table 1.4 Applications of Sentiment Analysis

Applications	Modules
Business Intelligence	User reviews, Product Reputation in market, Online product branding, Product disappointment factors, E-Commerce, Sell of product review.
Political Poll	Polling on news articles, Clarify politician status on voting application.
Psychological	Understand consumer buying habit, Improve selling of product.
Sociological	Help to improve manufacture to redesign product.

1.3 PROPOSED APPROACH

1.3.1 Pattern Based Approach to detect Sarcasm

1. **Features Extraction:** Actuality a complicated method of speech, irony is used for various proposes. Whereas expansion the information, the annotators resolved that these functions drop principally, however not all together, in three classifications: irony as wittiness, irony as cry and irony as shunning.
   - **Irony as wittiness:** when cast-off as a wittiness, irony is cast-off with the aim of being comic, somebody pays certain superior sorts of dialogues, have a habit of to embellish, or uses an attitude that's totally dissimilar from that when he dialogs classically to create it simple to identify. In community webs, voice tone are restore into different sorts of characters: use of upper case words, exclamation (!) and question marks (?), also as some irony-related emoticons.

- **Irony as cry:** Once used as shout, irony is working to point out however irritated or irritated person is. Therefore, it tempts to point out however unhealthy the condition is by exploitation overstatement or by using identical positive languages to explain a negative condition.
- **Irony as shunning:** It mentions to the case once the individual needs to escape giving a clear answer, thus, kind's use of irony. During this case, the individual hires sophisticated judgments, unusual words and some unusual terms.

Four families are used to extract the opinions:

a. Sentimentality related features: An actual in grace form of irony that's wide utilized in each regular conversations similarly as short messages like tweets, is once associate showing emotion confident appearance is employed in a damaging context. An identical method to specific irony is to use vocabularies having unreasonable sentiments. This kind of irony we have a tendency to qualify as "whimper" is incredibly mutual in community webs and social networking websites. Show that this sort of irony are often a detected once a confident statement, typically a verb or a linguistic verb, is gathering with a negative scenario (example, "I love being unnoticed all the time"). They planned a lexicon-based method that studies the potential positive expressions and negative things and used it to find such distinction in unknown tweets. But, knowledge all potential negative conditions needs a giant and supply and maybe not feasible as a result of negative things are unpredictable. We choose an additional straight forward, yet more overall method. We tend to think about any reasonably inconsistency between sentiments of words also as alternative elements inside the tweet. And so, to spot and count such variation's we tend to extract gushy elements of the tweet and total them. For this drive, we tend to keep dual lists of words fit as "positive words" and "negative words".

b. Punctuation-related options: Sentiment-related options aren't enough to find all types of irony that may be gift. Additionally, they are responsibility not make use of all the portions of the tweet. Therefore, additional options are to be taken out. As declared before, sarcasm could be a classy type of language: it is not only acting with words and meanings, however additionally it works activity aspects like low-slung tones facial motions or overstatement. These features are interpreted into an explicit use of punctuation or replication of vowels once the note is written. To discovery such aspects, we have a tendency to extract a group of options that we have a tendency to be suitable as punctuation-related options. For every tweet, we have a tendency to analyze the subsequent values:

- Variety of scream letters
- Variety of query letters
- Variety of spots
- Variety of all-capital words
- Variety of speech marks

c. Syntactical and linguistics features: Beside with the punctuation-related options, certain common languages are used occasionally during a satiric situation. It is likely to associate these languages with the punctuation to decide whether or not what's same is satiric or may be no. Also, in different suitcases, individuals have a habit of to create sophisticated judgments or usage unusual words to create it an unclear to the auditor to induce a transparent response. This is often public on irony is employed as "sarcasm", wherever the individual's drive is to skin his actual touch or opinion by exploitation satire. Later, we cutting the next option that replicates these features.

- Usage of rare arguments
- Numeral of unusual words
- Being of shared satiric expressions
- Varity of interjections
- Variety of happy words

In explicit, the feature “Existence of common satiric expression” is take out within the similar method we have a tendency to extract the features capable as “pattern-related”.

d. Pattern options: Common satiric expressions are quite shared even in spoken communication. Though, their variety is little they are not single and greatest of the tweets in each our training and take a look at sets don't contain them. That being the case, we tend to dig any and excerpt another set of options. The idea of our pattern-related options is impressed from the work of in his method, the writer confidential words into 2 categories: great occurrence words and content words supported their frequency of look in his knowledge set associate degreed outlined a pattern as an “ordered sequence of high frequency words and slots for content words”. That approach, have it is some potential to observe irony. Therefore, we tend to propose additional well-organized and reliable patterns.

We divide words into 2 classes: a primary one referred to as “CI” containing words of that the content is very important and another mentioned as “GFI” containing the arguments of that the linguistic perform is more necessary. If a word fits to the primary class, its variant forms of the same word otherwise, it's replaced it by a definite expression. The classification into categories is completed supported the part of language tag of the word within the tweet.

Chapter 2

LITERATURE SURVEY

Mondher Bouazizi and Tomoaki Ohtsuki, 'A Pattern-Based Approach for Sarcasm Detection on Twitter', 3536.c (2016).[1] In this paper author detect a sarcasm on twitter based on pattern based approach. Sarcasm is states as say something which meaning is opposite to of the saying word. The detection of sarcasm is very difficult in microblogging web sites. In twitter is difficult to detect. The author can use a pattern based approach in this approach he can divide a words into HFWs and CWs and use a hashtag #sarcasm use a 6000 tweets dataset which contain hash tags and use a OpenNLP and for classification use a toolkit weka for classification and libsvm.

Diana Maynard and Mark A Greenwood, 'Who Cares about Sarcastic Tweets ? Investigating the Impact of Sarcasm on Sentiment Analysis' [2] In this paper author consider that sarcasm is also an important in sentiment analysis for social media like micro blogging websites, Twitter, Instagram etc. The author can consider that in some way hashtag (#) contain the sarcasm and sentiments of human beings, and to find that sarcasm developed a hashtag tokeniser for GATE. This tokeniser is built for easily find the sarcasm with in the tweets or massage. To detected the sarcastic tweets created a manually list of all sarcastic hashtag in a corpus of a random tweets. After collected a list extended automatically groups of hashtags where at least one hashtag contain #sarcasm hashtag using GAZETTER LIST COLLECTOR GATE plugin.

Komalpreet Kaur Bindra, Asst Prof and Ankita Gupta, 'Tweet Sarcasm : Mechanism of Sarcasm Detection in Twitter', 7.1 (2016), 215–17.[3] In this paper author consider that sarcasm is a linguistic phenomenon in which people state the opposite of what they actually mean. To detect the sarcasm in twitter the data is collected in to ways 1.) Build an online corpus of sarcastic (S) Negative (N), Positive (P). (2.) Way to collect tweets by using twitter API and collect the tweets that contain hashtag (#sarcasm, #sarcastic). The main advantage of using Twitter API that we can have enough sample to fulfill our requirement. Use different steps for feature extraction. (A) n-grams (B) Sentiments by using SentiWordNet (C) Pattern extraction: words are divided into two categories Content words

(CWs) and high-frequency words (HFWs). And algorithm for classification is Machine learning Support vector machine and Logistic Regression. Logistic Regression in binary response variable that related to set of art variables.

Juan M Soler, Fernando Cuartero and Manuel Roblizo, 'Twitter as a Tool for Predicting Elections Results', 2012.[4] In his paper author use a twitter as tool to predict the elections results. We all know that twitter is one of the most popular social networking website to share their opinion without any fear so in the meantime of elections every person given their opinions about the candidates and predict who is win. In this paper author us a data of Spanish election in 2012 and use a tool known as TARA TWEET tool this tool work on two main objectives. 1) Tara tweet allows the monitoring of social conversation in twitter through some hashtags define by the user. 2) It counts keywords which users have introduced in the certain of specific experiment definition.

Umesh Rao Hodeghatta, 'Sentiment Analysis of Hollywood Movies on Twitter', 2013, 1401–4.[5] In this paper author perform sentimentality analysis for Hollywood pictures on twitter, online reviews are create a buzz about movie in market so analyze this buzz perform a sentiment analysis . In this paper the author divided the task into two categories firstly it divided the tweet data into positive, negative and cognitive parts and secondly it collect data according to regions and countries so that easily find out the opinion of people from different regions. They use n-grams and machine learning algorithm naïve-Bayes and Max. Entropy.

Oren Tsur and Ari Rappoport, 'ICWSM – A Great Catchy Name : Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews', 162–69 [6] This paper focused on sentiment analysis on online product reviews and find the sarcastic reviews. The author can used a semi-supervised approach to find the sarcastic reviews he can used an amazon book sell data and perform an analysis. The novel semi-supervised algorithm is work into two categories first it can define the semi-supervised pattern acquisition and second detection of sarcasm. The pattern acquisition technique is work in two ways first find the most frequent word and second content words and categories all data according to that pattern and find the sarcastic reviews.

Clint Burfoot and Timothy Baldwin, 'Automatic Satire Detection : Are You Having a Laugh ? University of Melbourne University of Melbourne', 2009, 161–64.[7] In this paper author focus to determine that the newswire articles (Newswire is a service that provide a latest news stories via satellite, the internet etc.) are true or sarcastic. Author can user a support virtual machine for feature scaling) and number of lexical and semantic feature types.

The main contribution of this paper is. 1) It introduce a novel way to computational linguistic and machine learning and provide a slandered dataset for research on newswire sarcastic detection. 2) Also developed a method which adept identification of sarcasm based on simple beg-of-words features. It build a corpus consists of 4000 newswire document and 233 sarcastic news articles and use binary feature weights and Bi-normal separation feature scaling.

Tony Veale and Yanfen Hao, 'Detecting Ironic Intent in Creative Comparisons'.[8] In this paper author frequently used inclosing device for linguistic irony the symbol to show how irony is regularly marked in ways that it make computationally achievable to detect irony symbol is web data. They create a decent large corpus for examination web scraping symbols to recognize the most exciting characteristics of ironic judgments and provide a new algorithm for extrication ironic from non-ironic symbols. In this paper focus on one common form of ironic explanation the humorous symbol and develop a multi-pronged method to different ironic to non-ironic. Use Google API as interface to text of web and WordNet.

Roberto González-ibáñez and Nina Wacholder, 'Identifying Sarcasm in Twitter : A Closer Look', 2011, 581–86.[9] In this paper author report a technique for building a corpus of ironic twitter messages in which resolve of the sarcasm of each note has been prepared by its author. Corpus compare sarcastic remark in twitter remark express positive or negative and sarcastic boldness. Influence of lexical and practical factors on machine learning efficiency for identify sarcastic comments. And associate the performance of machine learning techniques and human judges. Using a hashtag #bicycling, #happy, #sarcasm. Lexical factors by unigrams and dictionary based approach and for classification

use Support vector machine with sequential minimal optimization (SMO) and logistic regression (LOgR).

Marina Boia and others, 'A :) Is Worth a Thousand Words : How People Attach Sentiment to Emoticons and Words in Tweets', 2013 [10] Emoticons are usually used to ready positive or negative sentiment on Twitter. In this paper learning with living consumers to regulate whether emoticons are used to only highlight the sentimentality of tweets, or whether they are the main elements ringing the sentiment. Author invent that the sentimentality of an emoticon is in important agreement with the sentiment of the entire tweet. Thus, emoticons are appreciated as predictors of tweet sentiment and necessity not be ignored in sentiment association. In the first part, current a live user culture on the use of emoticons on Twitter. Inspect the link between the sentimentality of emoticons, the sentiment of associated words, and the sentiment of tweets. Second part, we present two approaches that generate Twitter subjectivity lexicons from sentiment seeds. We obtain lexicons from emoticons and emotion words and calculate them to strengthen the findings of the user study.

Sitaram Asur and Bernardo A Huberman, 'Predicting the Future With Social Media', 2010. This paper is fouce on predicting the future by social media, Social networks is the most popular web sites in world like twitter, facebook, Linked in or e-commerce sites like amozn, sanpdeal . In this paper author predict the real world outcomes like tweets and moviews revenues from Box-Office. In social media movies name is one of the most interesting thing and when movie trailer is out user can predict the future of movie and movie producers and co-sopnser try to permote their movie it create a buzz in box-office. The main goal of this paper is assume how buzz oan attention in created before movie relase and how assume its revenue. Dynamic LMC classifier is use by author to predict the futur of movies LMC is a language classifier that help in sentiment analysis and that is based on tranined data set rules or dictionary.

Chetan Kaushik and Atul Mishra, 'A S CALABLE , L EXICON B ASED T ECHNIQUE FOR', 4.5 (2014). In this paper describe the efficiently perform sentiment analysis on large dataset. Sentiment analysis or opinion mining is one of the best way to

analyze the social media or product revenue in market. Sentiment analysis is classified according to positive negative or neutral reviews according to this classification analysis on review can be done that posted by consumer online. In this paper big data can be analysis frequently or give faster and accurate result, in this paper using a lexicon sentiment approach and lexicon dictionary that is supervised learning and result is compare on the bases of speed and accuracy.

Antonio Moreno-ortiz and Chantal Pérez Hernández, 'Lexicon - Based Sentiment A Nalysis of Twitter Messages in Spanish', 2013, 93–100. [13] This paper based on lexicon sentiment analysis of twitter in Spanish. Lexicon is a semi-supervised learning methods in first step manually dictionary is created for sentiment word and these words are store in seed list. Lexicon method is faster than machine learning method. Author use a lexicon dictionary for tweet analysis in Spanish language by classifying it into positive negative and neutral and compare the result on the basis of accuracy in English tweet analysis results.

Cataldo Musto, Giovanni Semeraro and Marco Polignano, 'A Comparison of Lexicon-Based Approaches for Sentiment Analysis of Microblog Posts'. [14] This paper is based on a judgment of lexicon based method of sentimentality analysis for microblog. These microblog contain like twitter tweets and other social network web sites. Sentiment analysis is based on aspect like positive negative and neutral in data set. Sentiment analysis can be done in two ways first one is supervised and second is unsupervised. In this paper author does a lexicon based semi-supervised approach and compare result in four lexical resources, SentiWordNet, WordNet Affect, MPQA and SentiNet. In the experiment the efficiency of method was estimated against two-state-of-art dataset.

Haseena Rahmath P and Tanvir Ahmad, 'Sentiment Analysis Techniques - A Comparative Study', 17.4 (2014), 25–29. [15] In this paper author survey on sentiment analysis and classification that are used in opinion mining or sentiment analysis. Today's world is an information world in this big amount of data is generated on internet from any social media web site, e-commerce site etc. There are some classification can be done on data to perform a sentiment analysis the analysis can be done by supervised or unsupervised

learning methods. Supervised approach has a trained data set and unsupervised approach has not any trained data set. Author can be describe the why how text can be extract from big dataset and different classification can be apply. Data can be classified on document level, Text level and sentence level. There are some open challenge in sentiment analysis to find the complexity at polarity level and speedup the analysis speed and accuracy according to different dataset.

Francesco Barbieri and Horacio Saggion, 'Modelling Irony in Twitter', 2014, 56–64. [16] In this paper author investigate the automatic detection of irony casting it is a classification problem and propose a model capable of detecting irony in the social network such as Twitter. Model based on lexical features out performs a word-based baseline previously used in opinion mining and achieve state-of-art performance. Reduce the complexity of the problem by studying irony detection in the microblogging services twitter. Author can use a hashtag to make a corpus #irony #education #humor and #Politics. They use supervised machine learning method random forest and decision tree and implementation by weka tool kit.

Francesco Barbieri, Horacio Saggion and Francesco Ronzano, 'Modelling Sarcasm in Twitter , a Novel Approach', 2014, 50–58. [17] In this paper author represent a novel computational model capable to detect sarcasm in social network twitter. Unlike previous system, it does not include pattern of word as features. Our novel approach has seven set of lexical features aim to detect sarcasm by its inner structure, abstracting from the use of specific terms. They adopted 60,000 tweets and divided into equal topics Sarcasm, Education, Irony, Politics, Newspaper, comedy. They can use twitter API to get tweet data. And use supervised machine learning with decision tree classifier and implemented with Weka toolkit.

Santosh Kumar Bharti, 'Parsing-Based Sarcasm Sentiment Recognition in Twitter Data'.[18] Sentiment Analysis is a method to identify people's opinion, attitude, sentiment, and emotion towards any exact target such as persons, events, topics, product, civil service, services etc. Sarcasm is a special kind of sentimentality that include of words which mean the conflicting of what you actually want to say (particularly in order to abuse or wit someone, to show irritation, or to be funny). People often stated it verbally through

the use of weighty tonal stress and certain gestural clues like rolling of the eyes. In this paper two approaches use to detect sarcasm in the test of twitter data first one is Analyzing based lexicon generation algorithm, second is detect sarcasm based on the incidence of exclamation word. The grouping of two approaches is also shown and associate with existing state-of-art approach to notice sarcasm. Interjection word like Wow, Oh, Aha etc. that are used in tweets.

Ellen Riloff and others, 'Sarcasm as Contrast between a Positive Sentiment and Negative Situation'. [19] In this paper author current a novel bootstrapping procedure that mechanically learns lists of positive sentiment idioms and negative situation idioms from ironic tweets. We show that classifying contrasting contexts using the idioms learned through bootstrapping profits improved recall for ironic recognition. The goal of our investigation is to identify ironic that arises from the difference between a positive sentiments mentioning to a negative condition. A key challenge is to mechanically identify the negative "conditions", negative conditions are unenjoyable activities include going to the giving an exam, and doing work on holidays. Objectionable states include being ignored. Author can use hybrid approach with combination of Bootstrapped Lexicons union SVM Classifier.

Ashwin Rajadesingan, Reza Zafarani and Huan Liu, 'Sarcasm Detection on Twitter : A Behavioral Modeling Approach', 2015, 97–106 [20] Sarcasm is a difficult task to detect in text even it is too difficult to detect sarcasm by human begin also. Sarcasm is a linguistic phenomenon that meant opposite about their meaning. In this paper author approach is automatically detect a sarcasm by using linguistic cues. Author used a behavioral approach to find a sarcasm in content it go throw the user old post and analysis the behavior of user and then decide the content is sarcastic or not. In this approach author can analysis the past tweets of user that define T and user is define U, This is different from old approach because in this author can analysis the old tweets T of user and then decide the tweet is sarcastic or not sarcastic. Author can use a SCUBA (Sarcasm Classification using Behavioral Modeling) and to detect the sarcasm in tweets use a following point that help to detect the sarcasm 1) It's complexity to detect expression 2)

It's way to conveying a emotions 3)It is way to find out the familiarity 4) It's way to written an expression. Its model accuracy is 79.83% and its use a random classifier an n-gram (bi-gram and tri-gram).

Manju Venugopalan, 'Exploring Sentiment Analysis on Twitter Data', 2015, 0–6. [21] The growth and popularity of microblogging web sites is increase day by day, for this reason mining about sentiment and opinion is become a boots topic for research. In this paper author exploring a sentiment analysis on twitter and mining the data about latest mobiles in market and mining the reviews about these product it analyze the consumer reviews about particular product. Author can proposed an approach that is Hybrid model that use for classification of sentiment about a particular domain and tweets about that specific domain and classify the tweets according to that particular product domain. The author can analyze and extract the reviews about popular smart phone and compare it with its past few years. For classification use a support vector machine and NLTK for language pre-processing. The result is improved over uni-gram (bi-gram and tri-gram)

Anurag P Jain, 'Sentiments Analysis Of Twitter Data Using Data Mining', 2015, 807–10.[22] The extreme growth of internet and social media websites the researcher get attracted for do research on particular product, event and any hot topic that posted or trending on social media. Everyone in interested for knowing but is going on in other's life and how they can get know about them for this propose a social media is one of the platform that provide this kind of facility. In this paper author can propose an approach in this they can analysis and classification sentiment of social post by using a classifiers. Author can use K nearest neighbour, Random Forest, Naive Bayes and hybrid approach and also compare the performance of each algorithm with other, collect data from twitter by using twitter API and perform pre-processing by using NLTK and clean the dataset and perform a classifier the highest accuracy is k-NN.

Nehal Mamgain, Ekta Mehta and Ankush Mittal, 'Sentiment Analysis of Top Colleges in India Using Twitter Data', 2016, 1–6.[23] In this paper author can perform Sentiment Analysis on top of the colleges of India like NIT, IIT etc. the motive behind this

analysis, the world of today is totally based on internet and the social media websites the opinions of these websites and make importance so to analysis these reviews and opinion in good way is very important. These reviews makes a good or bad image in person mind about that particular product, In many cases every person go on internet to check that particular college website to check their reviews and then decide what he want to do. In this paper author proposed an approach which they can do sentiment analysis on top college reviews and provide a result they can use a Twitter tweets to perform experiments and remove a duplicate tweets. The model proposed is based on Bayes'. Perform a comparison between Naive Bayes and SVM and also perform artificial neural network. The accuracy of Naive Bayes is higher.

Chapter 3

PRESENT WORK

3.1 PROBLEM FORMULATION

The sentiment analysis is the technique to analyze the behaviours, attitude and opinions of the users. The sentiment analysis is applied on the social networking sites to analyze the behaviour of users. In present world analyze the behaviours of users or customers is very important for business and also for mining a social networking data. Because the growth of internet is fast as compare to previous years in present everyone is accessing the social networking sites to update views or opinion about every events. Youth is very active on these social networking sites like Twitter and Facebook.

Twitter is one of the interesting social networking website that used by youth and wide range of people. The region behind choosing a twitter data as a research work is; there are no privacy on tweets and user can mention every one whom he/she want to tag a tweet. This facility is not provide by a Facebook and any other social networking sites like if you write a tweet on twitter you can tag this to anyone by mention @username example @narendramodi this tweet is shown by everyone twitter account who are following the Mr. Narender Modi ji and notification is send to Mr. Narender Modi Ji. Another region to pick Twitter for research work is that it allows only 140 words in a tweet that is very less to express a complete felling of user so that's why take a twitter data for my research work. As we know there are a lot of research is done on twitter in last few year's but we have done Sarcasm Detection in twitter tweets that is different from normal sentiment analysis.

Sarcasm detection is little bit different from sentiment analysis because sarcasm is a linguistic phenomenon in which a user say opposite about their felling like example "Today is wonderful day working on Sunday!!!" In this example user use a sarcastic words which sense positive but not really due to punctuation marks. In this work Pattern Based approach is used to detect sarcasm in twitter dataset.

Base paper [1] proposed a different kind of feature that cover the sarcasm. They used Support vector Machine, k- Nearest neighbours, Maximum entropy and Random forest for their work and accuracy is 83.1%. Tools used are apache OpenNLP and WEKA for classification.

In base paper the future scope is to perform a sentiment analysis on rich dataset and opinion mining. The limitation in base paper is that they perform their work on twitter dataset, they didnot check their model on other kind of dataset, also they clean their dataset tweets manually and after that perform a classification on it but in our model we cannot clean data manually we can fetch data from twitter according to our need, we are not fetch duplicate tweets, location or retweets.

In our thesis work three dataset are used, two dataset belongs to twitter and one is movies reviews dataset. For classification used two classifier that belong to supervised machine learning:

- Naive Bayes probabilistic classifier
- Support Vector Machine linear classifier

For programming environment used a Python 2.7 script to perform a work, we used NLTK and textblob for pre-processing. Also import other library that are important in work.

3.2 OBJECTIVES

1. To study and analyze various sarcasm detection techniques for sentiment analysis.
2. Implement a Naïve Bayes and Support Vector Machine classifier to detect a sarcasm.
3. Improved texture extract features as compare to existing model.
4. Improved the overall accuracy for sarcasm detection.
5. To design the new feature extraction and description (FED) using multiple factor based hybrid pattern (MHP) module to overcome the shortcomings of the existing techniques.
6. To implement the newly designed feature descriptor module (FED-MHP) with all essential input and output parameters according to the workflow.
7. To implement the sentiment and sarcasm analysis module (SSAM) for the social datasets using the pre-classified training data for supervised classification
8. To integrate the SSAM module with the FED-MHP in order to build the complete sarcasm analysis model.

3.3 RESEARCH METHODOLOGY

Sarcasm detection on social media websites is one of the kind of sentiment analysis that is used for analyse the opinion and sentiment of user about particular topic or event. The meaning of sarcasm is define by Sanford university is that the meaning of sentence is opposite from their original meaning so, detection of sarcasm in text is little bit difficult as compare to normal sentiment analysis. The sarcasm detection is very helpful to improve the sentiment analysis is social websites add some extra effect in sentiment analysis.

The base paper [1] is based on sentiment analysis (sarcasm detection) from the social network sites (twitter). They proposed four set of features that cover the different kind of sarcasm and define a pattern to detect the sarcasm in tweets and classified sarcastic tweets into sarcasm or non-sarcasm, they used Random Forest, SVM, k-NN, and Max. Entropy algorithms for classification and the highest accuracy of sarcasm detection in Random Forest the accuracy is 83.1% and precision is 91.1% but other algorithms used are not giving good results.

In our research work Naive Bayes and Support Vector Machine are used as supervised classifier. The performance of Naive Bayes and Support Vector Machine is higher as compare to classifiers implemented in base paper. The accuracy of Naive Bayes is 88.59% and precision is 86.57%. The accuracy of SVM is 93.97% and precision is 90.87% and generate a new pattern that extract a better features as compare to base paper.

The methodology that is used in our research work is divided into these steps:

a. **Twitter Streaming API v 1.1 use for scraping tweets**

   Twitter API are able in two type first is REST API and second one is STREAMING API, in our experiment used a streaming API the difference between streaming and rest API is that in REST API we can modify the query and modify the account without permission of user but we can used OAuth authentication to perform a query. The STREAMING API is little bit different from REST it provide a tweets based on user query example he/she want tweets regarding #happy, #demonetization so he can query only that particular hashtag to get the same result

based on real time and user name. In our experiment used a STREAMING API version for getting a tweets from my own account.

To scrap a tweets from account we can use an apps.twitter.com web site and create a new application to get a four creational that is important to for working on Twitter streaming API. These creational are:

1) Consumer key (API key)
2) Consumer secret key (API Secret)
3) Access token
4) Access token secret

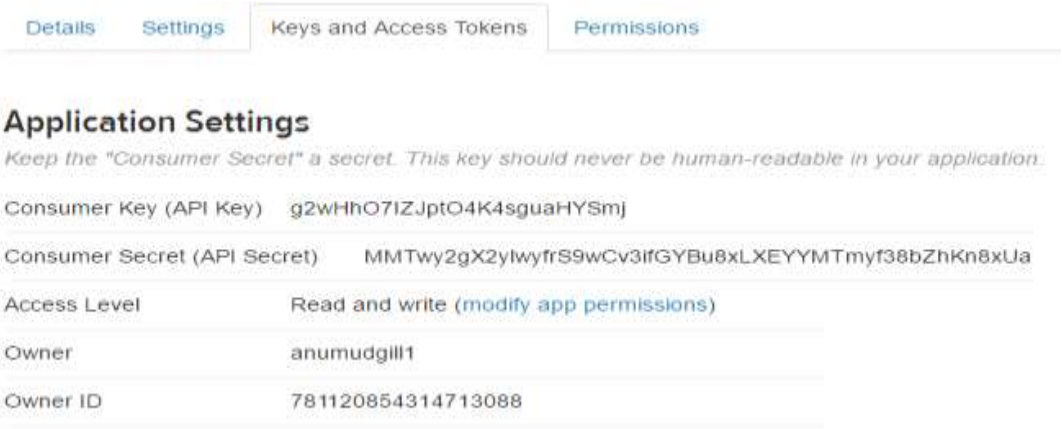

Figure 3.1 Twitter application settings

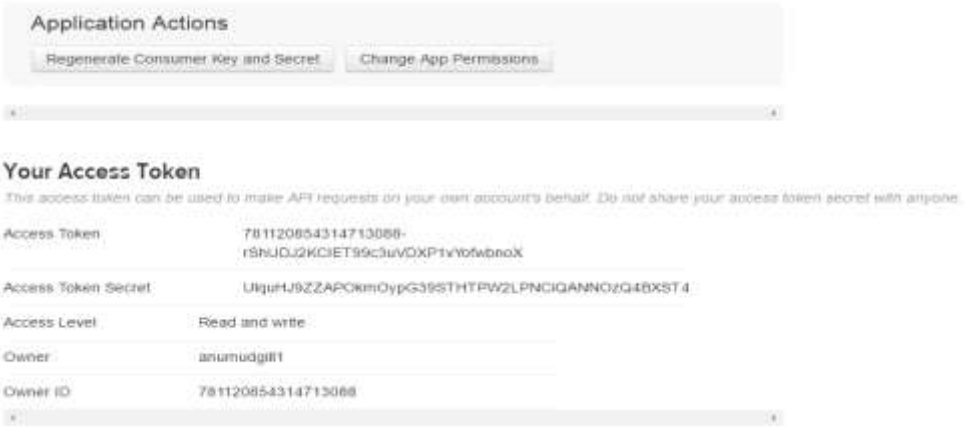

Figure 3.2 Twitter application access token

b. Tokenization Model

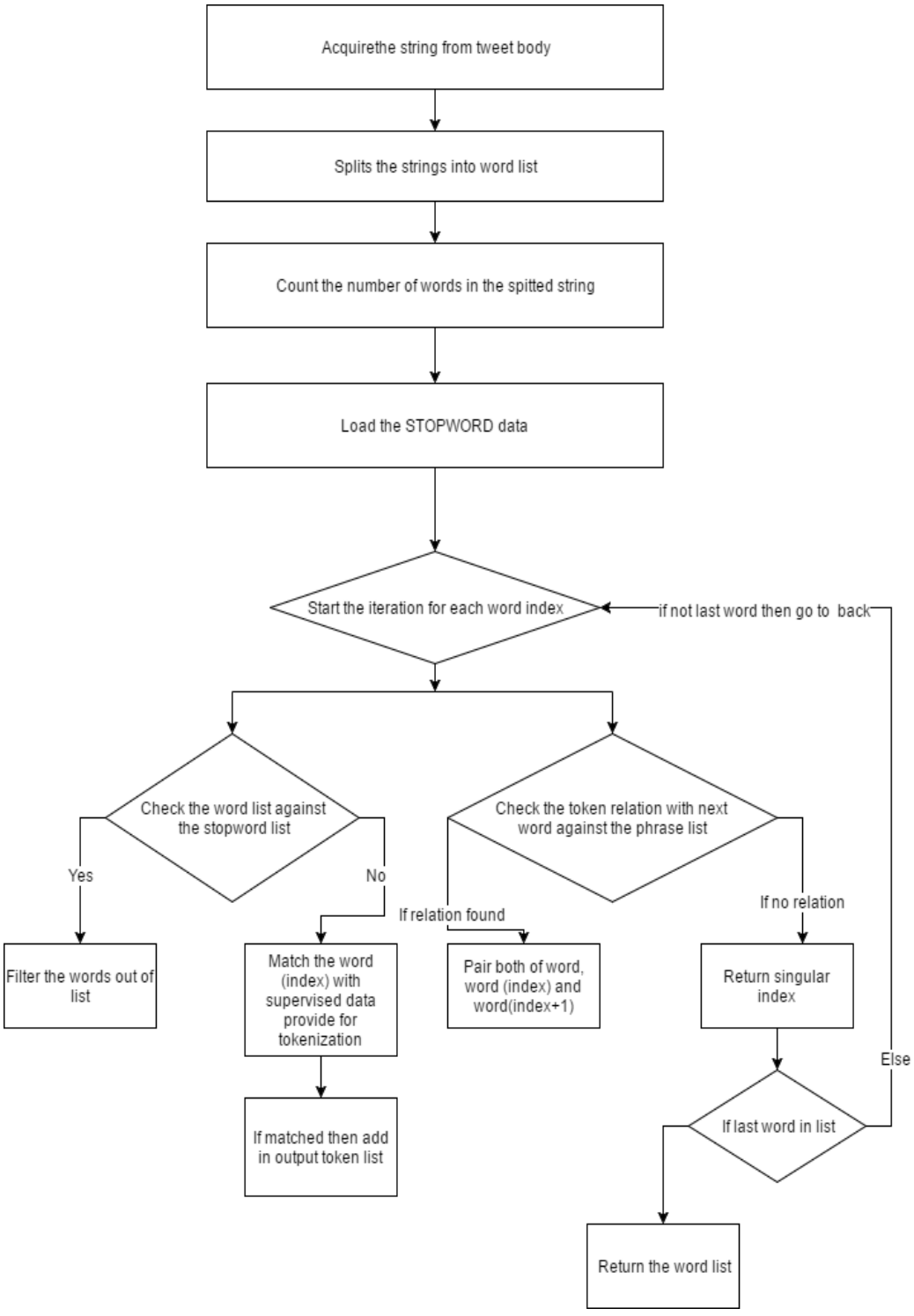

Figure 3.3 Tokenization Process

The working of flow diagram is

1. Acquire the string from the message body
2. Split the string into the word list
3. Count the number of words in the spitted string
4. Load the STOPWORD data
5. Start the iteration for each word (index)
    a. Check the word (index) against the STOPWORD list
    b. If the word (index) match return true
        i. Filter the word out of the list
    c. Otherwise Match the word (index) with the supervised data provided for the tokenization
    d. If the token matches the data in the supervised lists
        i. Add to the output token list
    e. Check the token relation with the next word against the phrase data
    f. If relation found
        i. Pair both of the words word (index) and word (index + 1)
    g. Otherwise return the singular word (index)
    h. If it's the last word
        i. Return the word list
    i. Otherwise GOTO step 5(a)

c. Sentiment Analyzer

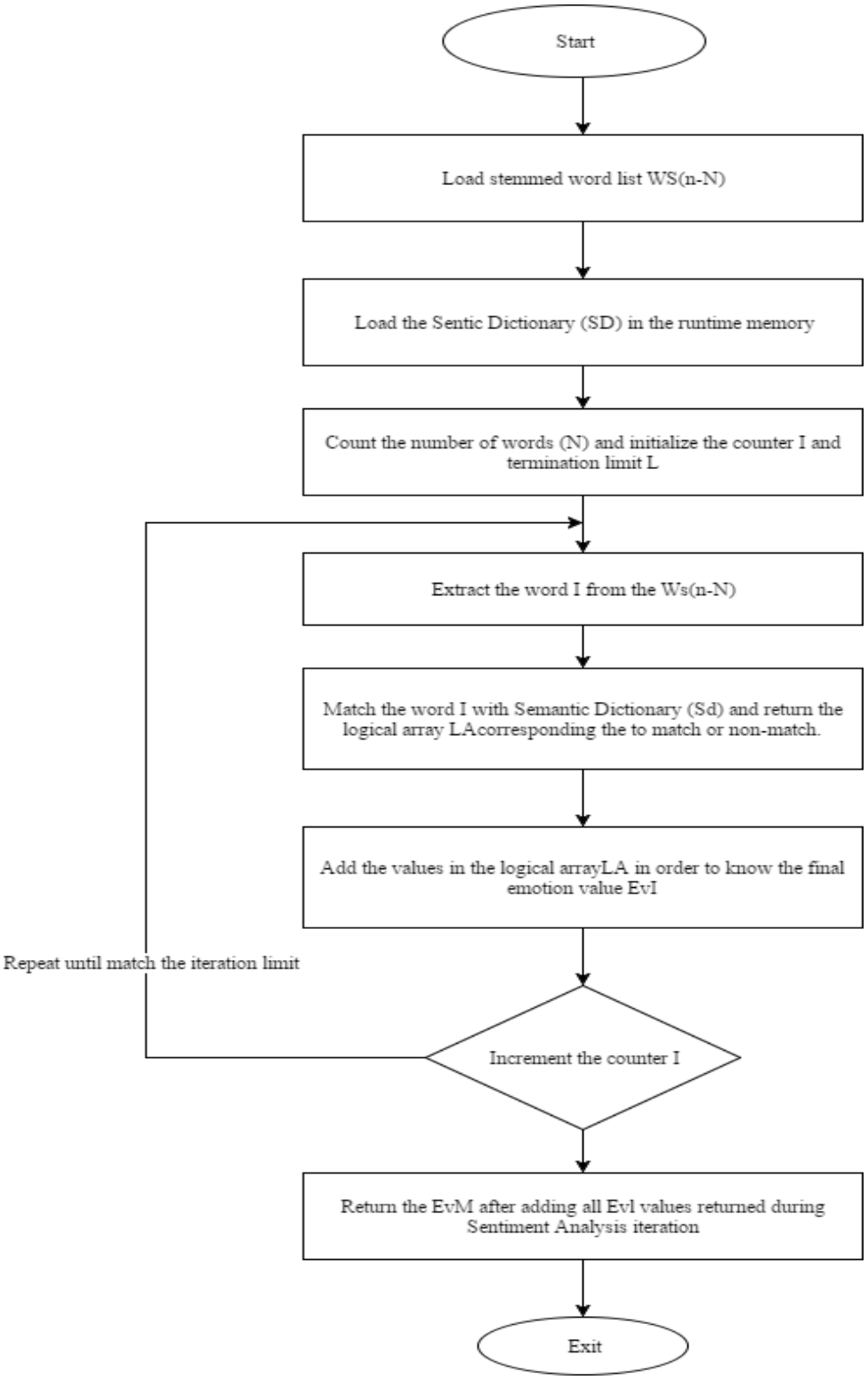

Figure 3.4 Sentiment Analyzer

The explanation of Sentiment analyzer flow diagram is:

1. Load the stemmed word list Ws(n-N)
2. Load the Sentic dictionary (Sd) in the runtime memory.
3. Count the number of words (N) and initialize the counter I and termination limit L
4. Extract the word I from the Ws(n-N)
5. Match the word I with Sentic Dictionary (Sd) and return the logical array LA corresponding the to match or non-match.
6. Add the values in the logical array LA in order to know the final emotion value EvL(Emotion Value obtained)
7. Increment the counter I
8. Repeat the steps 4 to 7.
9. Exit the loop if I matches the termination limit
10. Return the EvM ((Emoticon value Matrix)after adding all EvL values and returned during sentiment analysis iteration
11. Exit

d. Sarcasm Detection

Input 1: X (Training Matrix)

Input 2: T (Testing Vector) ← EvM (Emoticon value Matrix)

Input 3: Number of Neighbors

1. Assign the activation function to the Classification Model and assign → phi
2. Initialize the offshoot value for the desired classifier function over {phi} and return →phiA
3. Initiate the classifier {Naive Bayes or Support Vector Machine} model
    a. Acquire X and rearrange this to the matrix of input nodes (denoted with i)
        i. Initiate the iterative function over every input object (i)
        ii. Prepare the output vector for each input object as derivative of X
        iii. Return the computational cost
    b. Perform the processing over the current input node matrix
        i. Perform the computation over each object
        ii. Compute the probability matching between test data and training data rows
    c. Find the most matching row in the training data
    d. Find the category according to the matching component's class
    e. Return the decision logic

Workflow Diagram Pattern for Sarcasm Detection

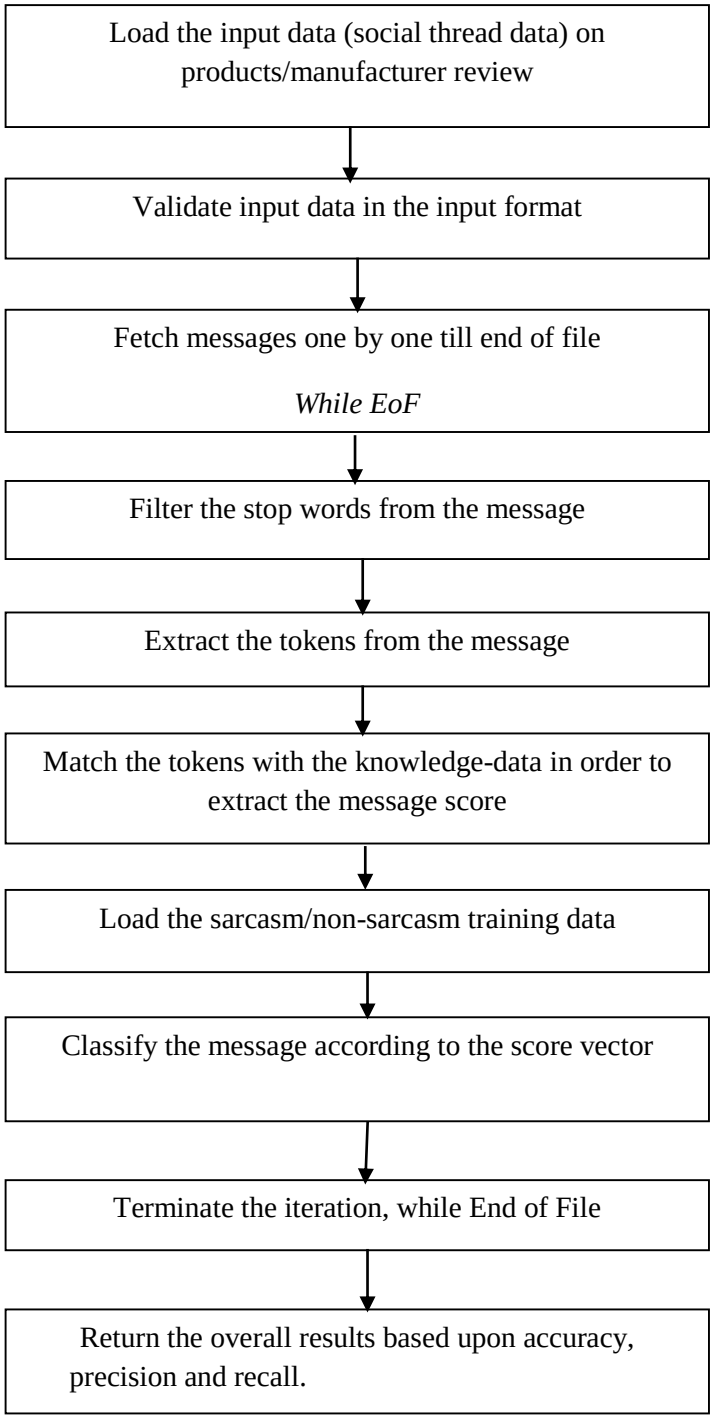

Figure 3.5 Pattern for Sarcasm Detection

Hybrid pattern Extraction with classification model

1. Acquire the tweet data obtained from the API in json format
2. Count the rows in the tweet data matrix
3. Iterate for every row in tweet data matrix (Tdm)
    a. Read the current tweet (Tw) from the Tdm
    b. Convert the tweet string to lowercase
    c. Normalize the string to make it process able through NLP processors
    d. Replace the URL with the word “url”
    e. Replace the string “@username” with the word “at_user”
    f. Remove the hashtags from the string
    g. Remove the number values from the input string
    h. Remove the special characters from the input string
    i. Convert the string to Unicode string
    j. Apply the tokenization on the string
        i. Split the string in space separated words
        ii. Discover the punctuations in the input string
        iii. Re-split the string based upon the discovered punctuations
        iv. Return the tokens array
    k. Replace the internet slangs with the original syntactic replacements in the tokens array
    l. Convert the tokens to the string
    m. Re apply the tokenization on the re-prepared string in strep 3(l)
        i. Apply N-gram analysis
        ii. Extract the multi-word keywords according to the NLTK dictionary
        iii. Return the new tokens array
    n. Remove the stopwords from the extracted keywords under N-gram analysis
    o. Extract the subjective words
        i. Keep the verbs in the tokens array
        ii. Keep the adverbs in the tokens array

iii. Keep the adjectives in the tokens array
iv. Keep the nouns in the tokens array
v. Remove the pronouns from the tokens array
vi. Remove the propositions from the tokens array
vii. Remove remaining attributes from the tokens array

p. Add the output to the processed array

4. Acquire the training data
5. Process the training data
6. Apply the classification & Return the classification results
7. Compute the classification performance on the basis of Precision, Recall, F-measure and Accuracy
8. Return the performance parameters

Chapter 4

RESULTS AND DISCUSSION

4.1 EXPERIMENTAL RESULTS

In our thesis work I can generate a hybrid pattern that is combination of syntax related features and punctuations related features to extract features from dataset. Naive Bayes and Support Vector Machine are supervised classifier that are used for classification in research work. The performance of Naive Bayes and Support Vector Machine is higher as compare to base paper [1]. The accuracy of Naive Bayes is 88.59%, precision is 86.57% and the accuracy of SVM is 93.97%, precision is 90.87%.

In this research work we used three dataset for experiment. Two datasets are belong to twitter and one for movies review dataset. For working environment used Python 2.7 script and NLTK (Natural Language Toolkit) is used for pre-processing and tokenization.

The results are calculated from parameters listed below:

a. Accuracy: Accuracy is used to find and evaluating matrix to see the effectiveness of the classifier algorithm to calculate the accuracy the formula is use is:

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Equation 2 calculate accuracy

b. Precision: A precision is used to calculate the exactness (Quality) of the classifier in research methodology precision is known as Type I error which linked as false positive value (FP) means the Null hypothesis is true but it is rejected in result in other way we can say that the condition or value is present in result but actually it is not present. To calculate the precision following equation is used.

$$\frac{TP}{TP + FP}$$

Equation 3 calculate precision

c. **Recall:** A recall is used to calculate the completeness (Quantity) of the classifier in research methodology a precision is known as Type II error which linked as false negative value (FN) that occurs when the null hypothesis is rejected but by mistake it should not be rejected.

$$\frac{TP}{TP + FN}$$

Equation 4 calculate recall

d. **F-measures:** F-measures is a combination of precision or recall it will produce a single matrix of both precision and recall. To calculate the F-measures use the following equation.

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

Equation 5 calculate f-measures

In our research work we have used Spyder python IDE for programming platform. After successfully installation of sypder python IDE we import "Tweepy" and pass the twitter application generated credentials (Consumer key, Consumer secret key, Access token, Access secret key), and scrap the tweets from current account of user.

In our research, tweets are classify into two labels sarcasm and non-sarcasm. Then categorized into two dataset test and train. We have taken 1000 tweets in sarcasm test dataset and 2850 tweets in non-sarcasm test dataset. For train dataset 7000 tweets are taken for both sarcasm and non- sarcasm. After getting training dataset Naive Bayes probabilistic classifier and Support Vector machine linear classifier are applied on the test dataset, the result is shown in tables 4.1, 4.2 and 4.3.

Naive Bayes and SVM classifier result for dataset 1 (twitter) shown in table 5

Table 4.1 result for classifier algorithms on dataset 1

	Overall Accuracy	**Precision**	**Recall**	**F-measures**
Naive Bayes	88.6	86.8	82.5	84.3
SVM	93.7	90.9	94.7	92.7

Naive Bayes and SVM classifier result for dataset 2(twitter) shown in table 6

Table 4.2 result for classifier algorithms on dataset 2

	Overall Accuracy	**Precision**	**Recall**	**F-measures**
Naive Bayes	82.9	81.7	75.8	77.8
SVM	81.5	77.9	79.3	78.5

Naive Bayes and SVM classifier result for dataset 3(Movies reviews) shown in table 7

Table 4.3 result for classifier algorithms on dataset 3

	Overall Accuracy	**Precision**	**Recall**	**F-measures**
Naive Bayes	79.2	72.1	87.6	73.5
SVM	93.6	86.2	96.2	89.9

In figure 4.1, 4.2 and 4.3 the current system classifier result is shown in bar chart, the performance of support vector machine classifier is better as compare to Naive Bayes the

overall accuracy of Naive Bayes is 88.59% and precision is 86.76% and the overall accuracy for SVM is 93.97% and precision is 90.89 %.

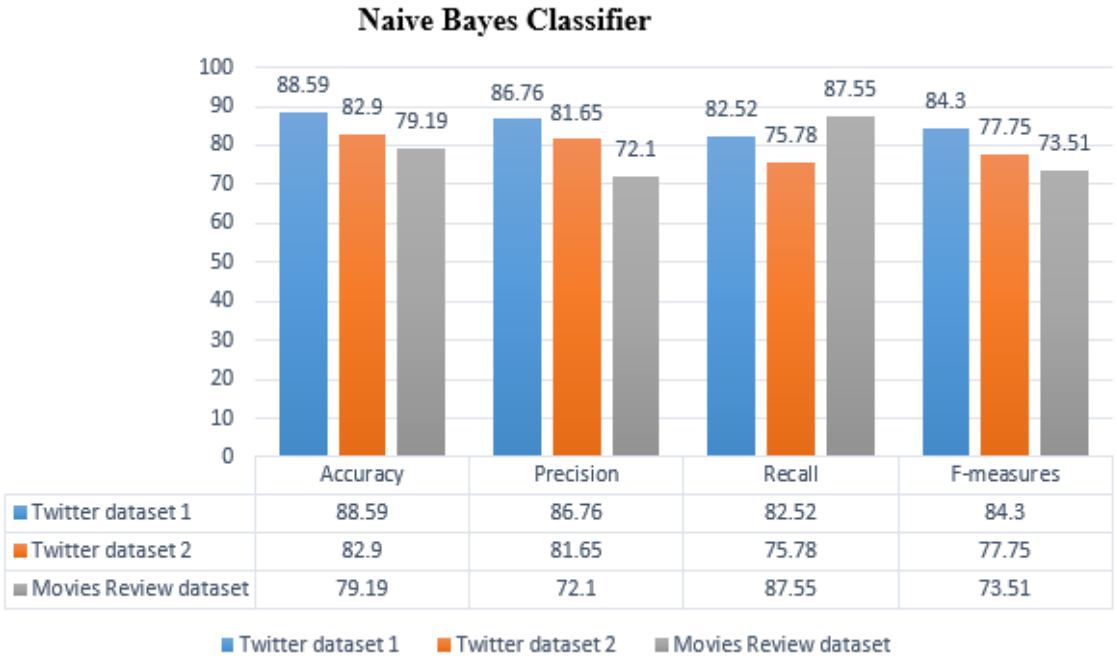

Figure 4.1 Naive Bayes classifier

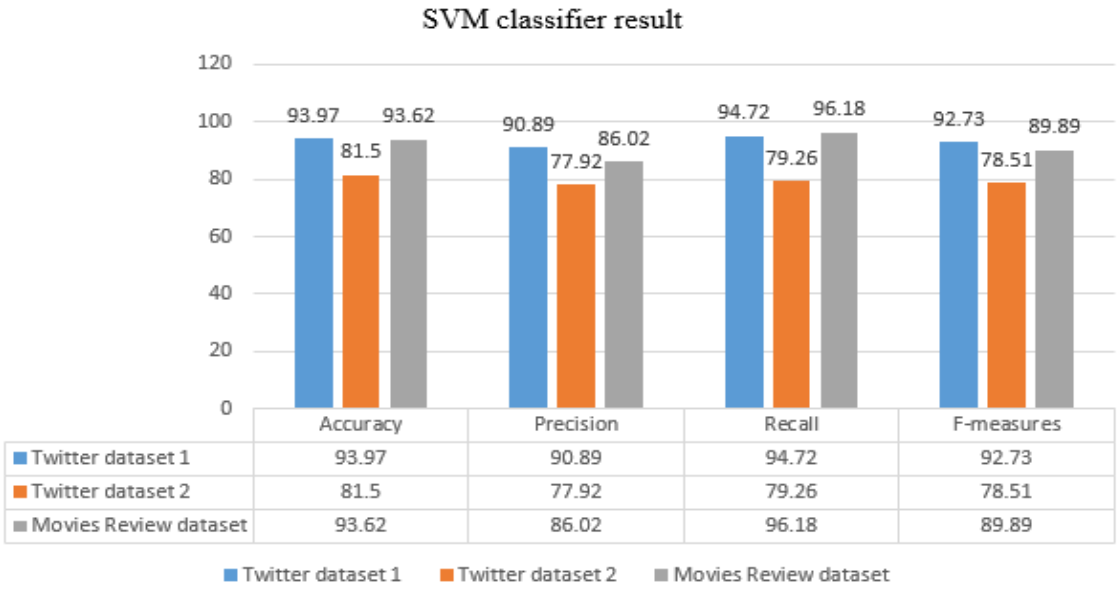

Figure 4.2 SVM classifier

4.2 COMPARISION WITH EXISTING TECHNIQUE

The base paper [1] the accuracy of applied classifier shown in fig 4.2 they used four algorithms. Random Forest is performed best out of three algorithms the overall accuracy of existing model is 83.1% and precision is 91.1% but in current model the overall accuracy in SVM is 93.97% and precision is 90.89 % which is higher than the existing model accuracy of SVM classifier. In current model Naive Bayes gives overall accuracy is 88.59% and precision is 86.76%. The results of existing model is shown in table 4.4 and table 4.5 is belong to current model which perform in proposed classifiers and results are shown in bar chart.

Table 4.4 For existing model Classifier algorithms performance

	Overall Accuracy	**Precision**	**Recall**	**F-measures**
Random Forest	83.1%	91.1%	73.4%	81.3%
SVM	60.0%	98.1%	20.4%	33.8%
K-NN	81.5%	88.9%	72.0%	79.6%
Max. Entropy	77.4%	84.6%	67.0%	74.8%

Table 4.5 For current model classifier algorithms performance

	Overall Accuracy	**Precision**	**Recall**	**F-measures**
Dataset 1				
Naive Bayes	88.6	86.8	82.5	84.3
SVM	93.7	90.9	94.7	92.7
Dataset 2				
Naive Bayes	82.9	81.7	75.8	77.8
SVM	81.5	77.9	79.3	78.5
Dataset 3				

Naive Bayes	79.2	72.1	87.6	73.5
SVM	93.6	86.2	96.2	89.9

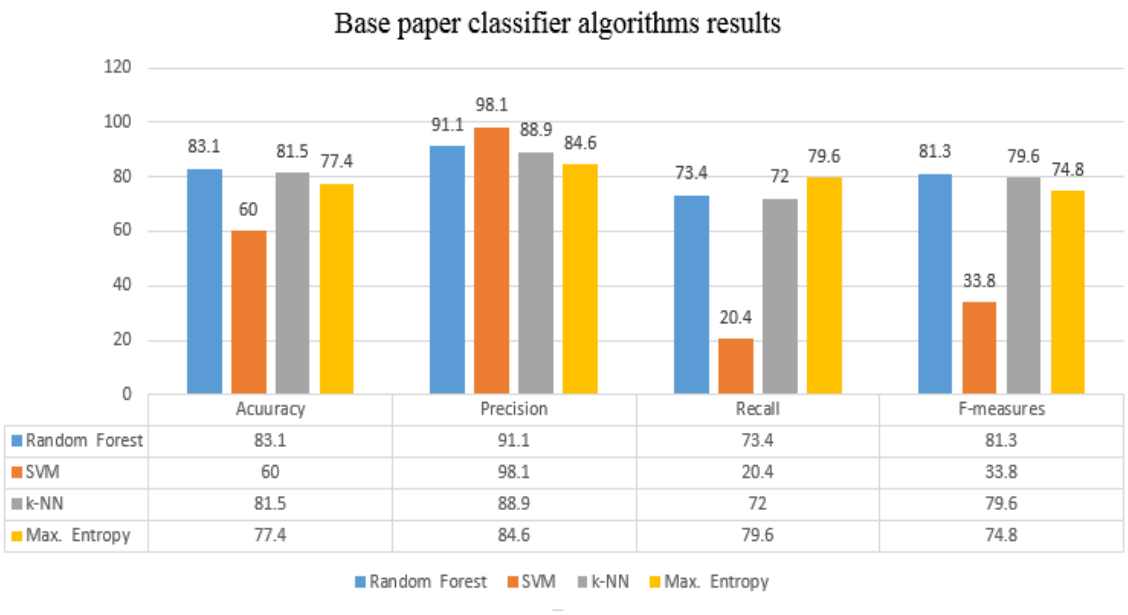

Figure 4.3 Base paper classifier algorithms

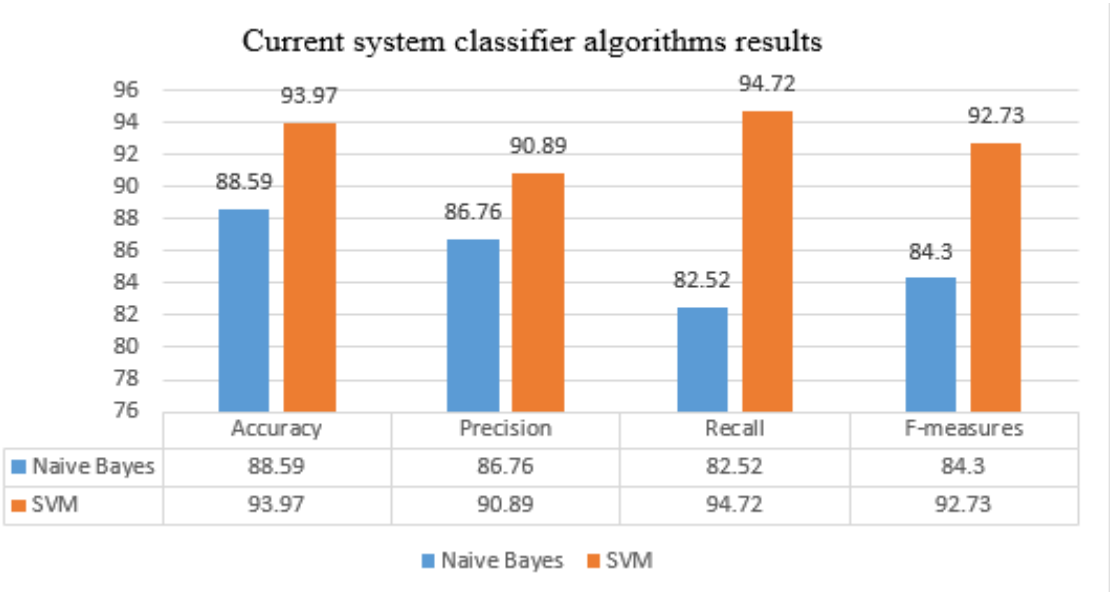

Figure 4.4 Current system classifier result

In base paper [1] the accuracy of classification of the test for each family of feature is shown in figure 4.5 and the current model accuracy is shown in figure 4.6, in this the existing model is perform on four parameters to classify the features these parameters are sentiment related features, punctuation related features, syntax related features and pattern related features. The overall accuracy for all four parameters are shown graphically in figure 4.5. The result of current model is based on these four parameters but its overall accuracy is higher as compare to existing model that is shown in figure 4.6

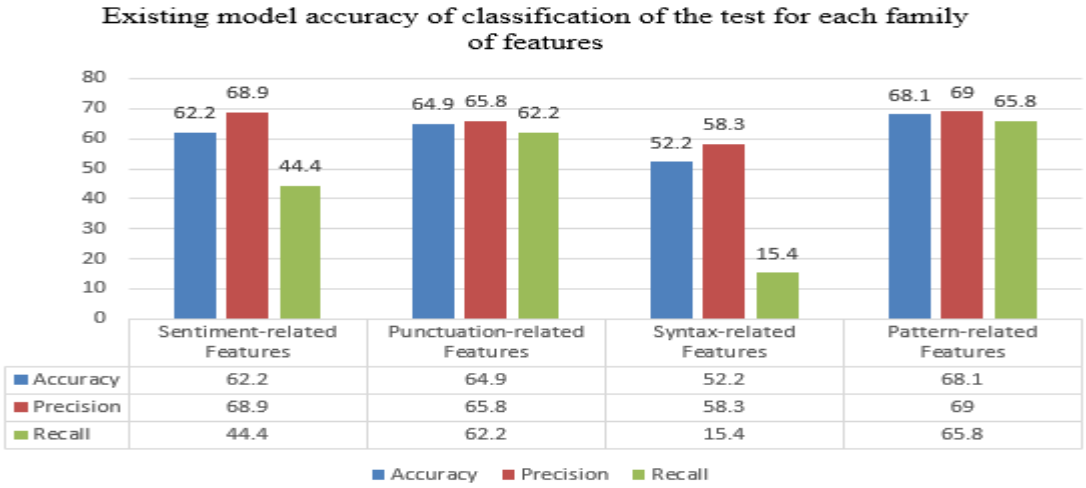

Figure 4.5 Existing model accuracy of classification of the test for each family of features

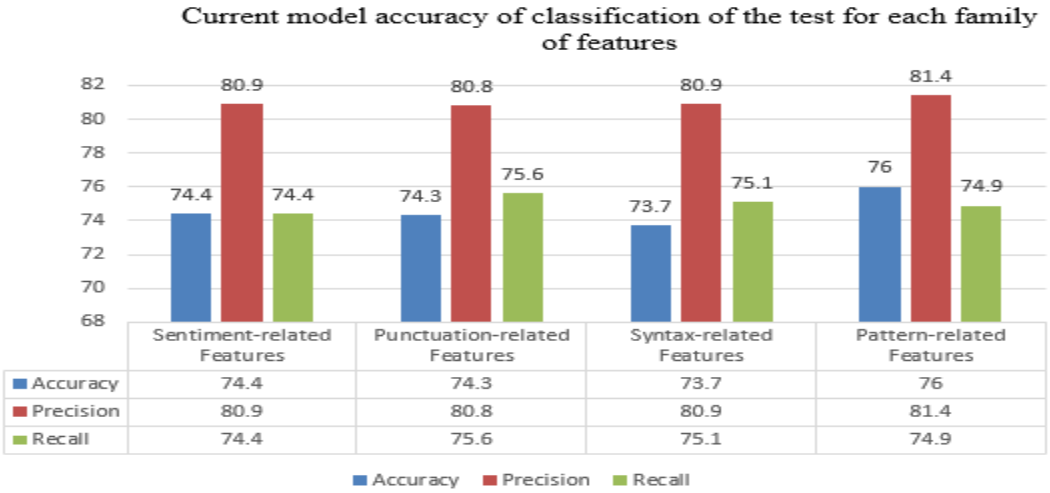

Figure 4.6 Current Model Accuracy of classification of test for each family of features

In figure 4.7 and 4.8 shown the accuracy of classification during cross-validation of each feature. The feature parameters are sentiment related features, punctuation related features, syntax and pattern related.

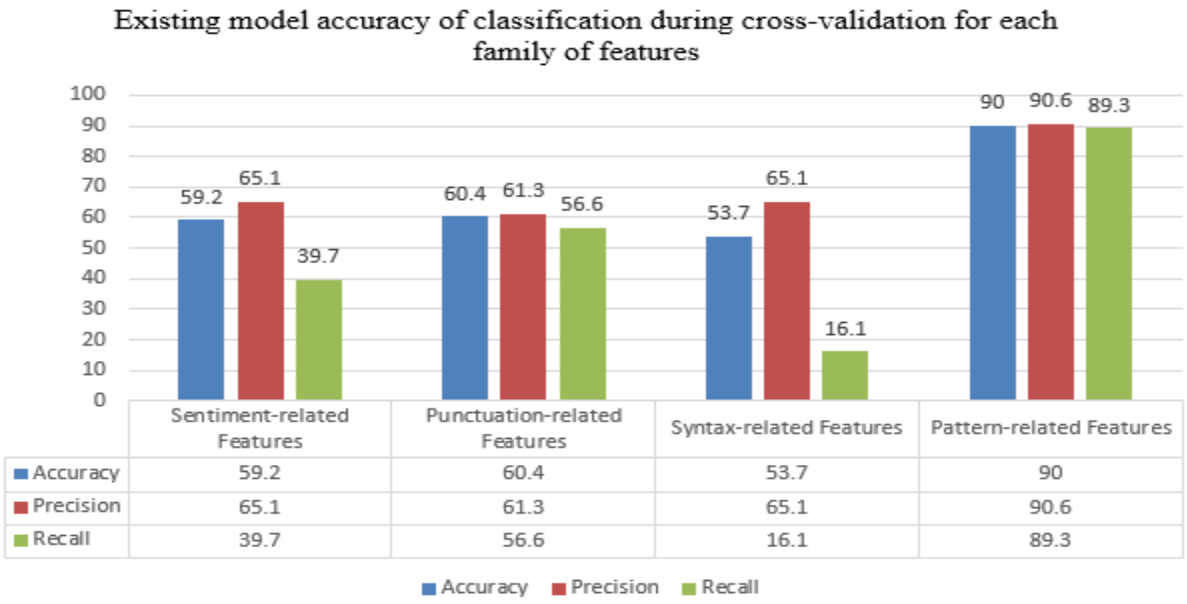

Figure 4.7 Existing model accuracy of classification during cross-validation

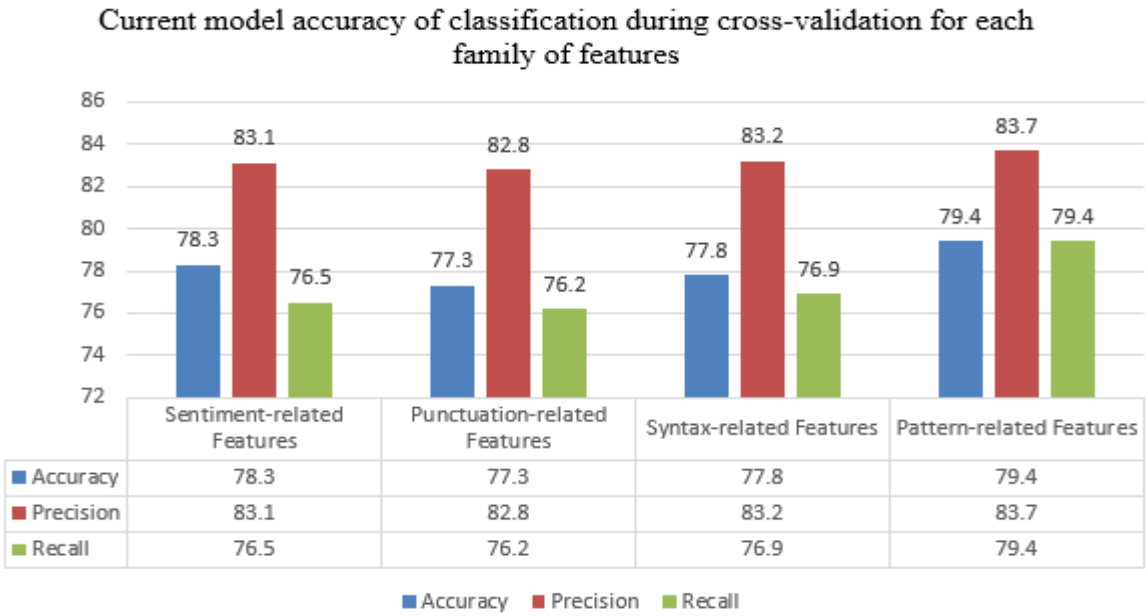

Figure 4.8 Current model accuracy of classification during cross-validation

In figure 4.9 and 4.10 the accuracy of classification using all features during training set-cross-validation and on the test set. The parameters of testing is train set cross-validation, test set before enrichment, test set after enrichment and the overall accuracy of existing model is shown is figure 4.9 and current model accuracy shown is figure 4.10

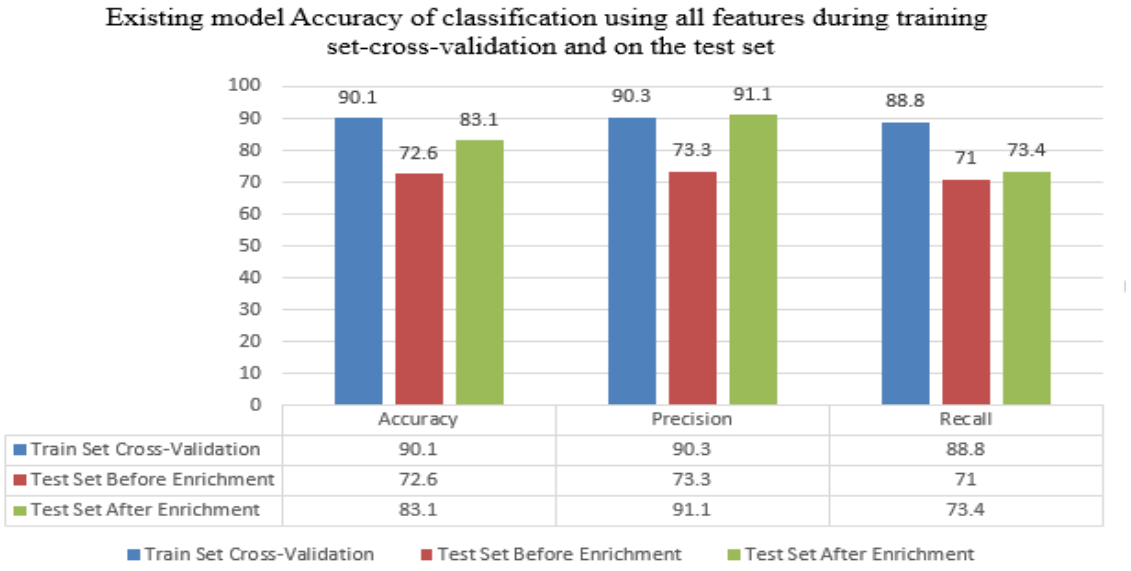

Figure 4.9 Existing model accuracy on training and test data set

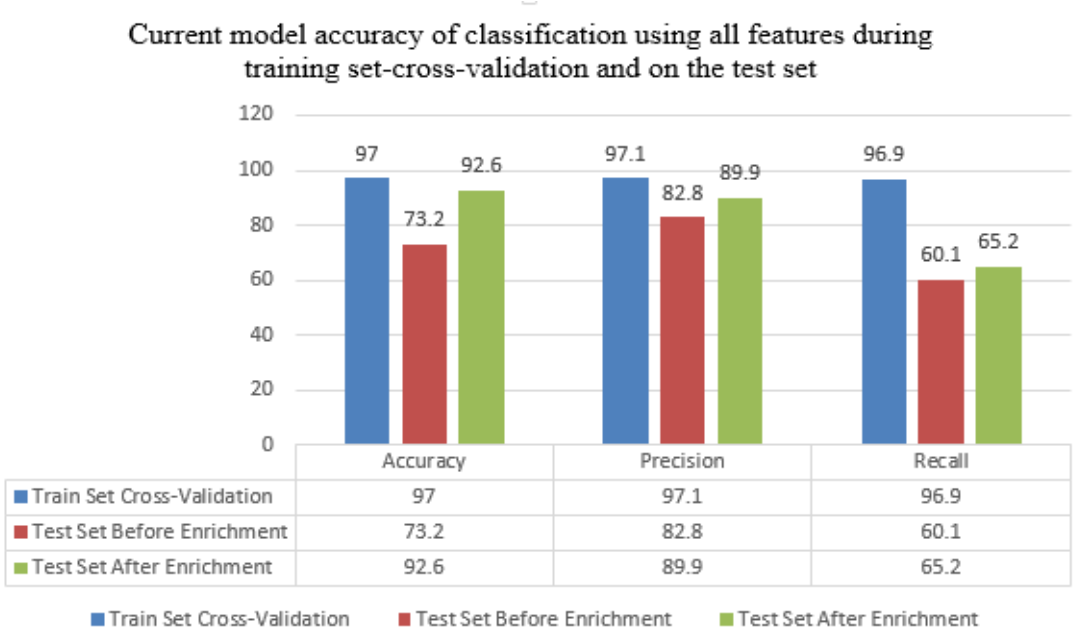

Figure 4.10 Current model accuracy of classification on training data and test dataset

Chapter 5
CONCLUSION

5.1 CONCLUSION

Sentiment analysis is one of wide area of research and improvement in there techniques and classification approaches. Sentiment analysis help in opinion mining and text summarization that help in various way not in social network sites it's also help in politic, Business intelligence, Public activates. Sentiment analysis or opinion mining is help in many ways like help to improve the sale of product and also help to find the consumer point of view about product, Sentiment analysis can be done on approaches like machine learning that use trained set of data or we can also say this is a supervised approach. Sarcasm detection is add on additional benefits in sentiment analysis on social networking sites the meaning of sarcasm is that way to say something that opposite about their meaning. Sarcasm detection is help to analyse the social networking websites data better that add more impact on sentiment analysis. In current model used a pattern based approach to detect a sarcasm in twitter dataset and movies review dataset. Applied a two classifier to perform classification one is probabilistic classifier Naive Bayes the overall accuracy of is 88.59% and precision is 86.76% and other one is linear classifier support vector machine overall accuracy is 93.97% and precision is 90.89%.

5.2 FUTURE SCOPE

In the future, the proposed work can be enhanced for the automatic dataset classification for eligible and non-eligible tweets, specifically written in other languages using the English alphabetic sequence for the purpose of earlier cleaning. The swarm intelligent algorithms, such as ant colony optimization (ACO), bee swarm optimization (BSO), etc. can be incorporated for the flexible and robust learning of the emotions amongst the given dataset. The deep emotion analysis can been utilized, which can classify the anger, disgust, joy, satisfactions, un-satisfaction, etc. for the purpose of deep emotion learning for sarcasm detection.

The proposed work can be enhance by finding the areas on geo location where mostly sarcastic tweets are generated based on timings like morning or evening. To get that information, generate a pattern that is based on behaviour of user tweets and analyse the past posted tweets of that user account after analyse generate a result.

REFERENCES

[1] M. Bouazizi and T. Ohtsuki, “A Pattern-Based Approach for Sarcasm Detection on Twitter,” vol. 3536, no. c, pp. 1–11, 2016.

[2] D. Maynard and M. A. Greenwood, “Who cares about sarcastic tweets ? Investigating the impact of sarcasm on sentiment analysis.”

[3] K. K. Bindra, A. Prof, and A. Gupta, “Tweet Sarcasm : Mechanism of Sarcasm Detection in Twitter,” vol. 7, no. 1, pp. 215–217, 2016.

[4] J. M. Soler, F. Cuartero, and M. Roblizo, “Twitter as a Tool for Predicting Elections Results,” 2012.

[5] U. R. Hodeghatta, “Sentiment Analysis of Hollywood Movies on Twitter,” pp. 1401–1404, 2013.

[6] O. Tsur and A. Rappoport, “ICWSM – A Great Catchy Name : Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews,” no. 9, pp. 162–169.

[7] C. Burfoot and T. Baldwin, “Automatic Satire Detection : Are You Having a Laugh ? University of Melbourne University of Melbourne,” no. August, pp. 161–164, 2009.

[8] T. Veale and Y. Hao, “Detecting Ironic Intent in Creative Comparisons.”

[9] R. González-ibáñez and N. Wacholder, “Identifying Sarcasm in Twitter : A Closer Look,” no. 2010, pp. 581–586, 2011.

[10] M. Boia, B. Faltings, C. Musat, and P. Pu, “A :) Is Worth a Thousand Words : How People Attach Sentiment to Emoticons and Words in Tweets,” 2013.

[11] S. Asur and B. A. Huberman, “Predicting the Future With Social Media,” 2010.

[12] C. Kaushik and A. Mishra, “A S CALABLE , L EXICON B ASED T ECHNIQUE FOR,” vol. 4, no. 5, pp. 35–43, 2014.

[13] A. Moreno-ortiz and C. P. Hernández, “Lexicon - Based Sentiment A nalysis of Twitter Messages in Spanish,” pp. 93–100, 2013.

[14] C. Musto, G. Semeraro, and M. Polignano, “A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts.”

[15] H. R. P and T. Ahmad, “Sentiment Analysis Techniques - A Comparative Study,” vol. 17, no. 4, pp. 25–29, 2014.

[16] F. Barbieri and H. Saggion, “Modelling Irony in Twitter,” pp. 56–64, 2014.

[17] F. Barbieri, H. Saggion, and F. Ronzano, “Modelling Sarcasm in Twitter , a Novel Approach,” pp. 50–58, 2014.

[18] S. K. Bharti, “Parsing-based Sarcasm Sentiment Recognition in Twitter Data.”

[19] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, “Sarcasm as Contrast between a Positive Sentiment and Negative Situation.”

[20] A. Rajadesingan, R. Zafarani, and H. Liu, “Sarcasm Detection on Twitter : A Behavioral Modeling Approach,” pp. 97–106, 2015.

[21] M. Venugopalan, “Exploring Sentiment Analysis on Twitter Data,” pp. 0–6, 2015.

[22] A. P. Jain, “Sentiments Analysis Of Twitter Data Using Data Mining,” pp. 807–810, 2015.

[23] N. Mamgain, E. Mehta, and A. Mittal, “Sentiment Analysis of Top Colleges in India Using Twitter Data,” pp. 1–6, 2016.

[24] Hernández A, Sanchez Victor, Sánchez G, Pérez, Toscano, Mariko Nakano and Victor Martinez,” Security Attack Prediction Based on User Sentiment Analysis of Twitter Data”, 2016

[25] A. P. Jain, “Sentiments Analysis Of Twitter Data Using Data Mining,” pp. 807–810, 2015.

[26] M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L. Haug, and M. Hsu, “Visual Sentiment Analysis on Twitter Data Streams,” pp. 277–278, 2011.

[27] M. Venugopalan, “Exploring Sentiment Analysis on Twitter Data,” pp. 0–6, 2015.

[28] G. D. Rajurkar, “2015 International Conference on Computing Communication Control and Automation A speedy data uploading approach for Twitter Trend And Sentiment Analysis using HADOOP,” 2015.

[29] Pagolu V S, Challa K. R., Panda G., “Sentiment Analysis of Twitter Data for Predicting of Stock Market Movements” 2016

[30] C. Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang, “SentiView : Sentiment Analysis and Visualization for,” vol. 43, no. 6, pp. 620–630, 2013.

[31] A. S. S. Analysis, R. M. Eshleman, and H. Yang, “‘Hey #311, come clean my street!,’” 2014.

[32] Gautam G, Yadav D, “Sentiment Analysis of Twitter Data Using Machine Learning”, 2014

[33] H. T. Gemilang, A. Erwin, and K. I. Eng, “Indonesian President Candidates 2014 Sentiment Analysis by Using Twitter Data,” pp. 4–7, 2014.

[34] Gokulakrishnan B, Priyanthan A, Ragavan T, Prasath N, “Opinion Mining and Sentiment Analysis on a Twitter Data Stream”, 2012

[35] P.Grandin and J.M Adan, “Piegas: A System for Sentiment Analysis of Tweets in Portuguese”, vol. 14, no.7, pp. 3467-3473, 201

PLAGIARISM REPORT