

Sentiment Analysis On Twitter data Using Machine Learning Techniques

THESIS SUBMITTED IN PARTIAL FULFILMENT OF REQUIREMENT
FOR THE AWARD OF THE DEGREE OF

**Master of Technology
in
Software Engineering**

Under the guidance of
Mr. Sanjay Patidar
(Assistant Professor, Department of Computer Science and
Engineering)
Delhi Technological University

Submitted By
ABHILASH MITTAL
(Roll No. 2K17/SWE/01)



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)
Shahabad Daulatpur, Main Bawana Road, Delhi-110042

June 2019

DECLARATION

I hereby declare that the thesis work entitled “*Sentiment Analysis on Twitter Data Using Machine Learning Techniques*” which is being submitted to Delhi Technological University, in partial fulfilment of requirements for the award of degree of Master of Technology (Software Engineering) is a bonafide report of Major Project-II carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

Place: Delhi

Name- Abhilash Mittal

Date:

Roll No. 2K17/SWE/01

CERTIFICATE

This is to certify that Project Report entitled “*Sentiment Analysis on Twitter Data Using Machine Learning Techniques*” submitted by Abhilash Mittal (roll no. 2K17/SWE/01) in partial fulfilment of the requirement for the award of degree Master of Technology (Software Engineering) is a record of the original work carried out by him under my supervision.

Place: Delhi

SUPERVISOR

Date:

Mr. SANJAY PATIDAR

(Assistant professor)

Department of Computer Science and Engineering

Delhi Technological University

Bawana Road, Delhi -110042

ACKNOWLEDGEMENT

I am very thankful to **Mr. Sanjay Patidar** (Assistant professor, Computer Science and Engineering Department) and all the faculty members of the Computer Science and Engineering Department of Delhi Technological University. They all provided us with immense support and guidance for the project.

I would also like to express my gratitude to the university for providing us with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions.

I would also like to appreciate the support provided to us by our lab assistants, seniors and our peer group who aided us with all the knowledge they had regarding various topics.

ABHILASH MITTAL

Roll No. 2K17/SWE/01

M. Tech. (Software Engineering)

Delhi Technological University

ABSTRACT

Twitter is the popular micro blogging site where thousands of people exchange their thoughts daily in the form of tweets. The characteristics of tweet is to be short and simple way of expressions. though this thesis will focus on sentiment analysis of twitter data. The research area of sentiment analysis are text data mining and NLP. By using different supervised machine learning techniques we will perform the sentiment analysis on twitter data. However we will focus on techniques and types of sentiment analysis where we will perform how to extract tweets from twitter. Further we will compare different machine learning techniques on the same dataset and also find some standard measures.

TABLE OF CONTENTS

a) Candidate's Declaration.....	i
b) Certificate.....	ii
c) Acknowledgement.....	iii
d) Abstract.....	iv
e) Table of Contents.....	v
f) List of Tables.....	viii
g) List of Figures.....	ix
h) List of Symbols, Abbreviations and Nomenclature.....	x
1. Introduction.....	1
1.1. Introduction to python.....	1
1.2. Introduction of Anaconda and Jupyter.....	2
1.3. What is Sentiment Analysis.....	2
1.4. Sentiment Analysis Classification.....	3
1.4.1. Machine learning techniques.....	4
1.4.1.1. Supervised learning.....	4
1.4.1.1.1. Naïve Bayes.....	5
1.4.1.1.2. Support Vector machine.....	6
1.4.1.1.3. K-Nearest Neighbour.....	7
1.4.1.1.4. Decision Tree.....	9
1.4.1.2. Unsupervised learning	9
1.4.1.3. Reinforcement learning.....	10
1.4.2. Lexicon Based Approach.....	10
1.4.2.1. Dictionary based approach.....	10
1.4.2.2. Corpus based approach.....	11
1.5. Natural Language Processing.....	12
1.5.1. Morphological Processing.....	12
1.5.2. Syntax and semantic analysis.....	12
1.5.3. Pragmatic Analysis.....	13

1.6. Need of Sentiment Analysis.....	14
1.6.1. Industry evolution.....	14
1.6.2. Research demand.....	14
1.6.3. Decision making.....	14
1.6.4. Understanding contextual.....	14
1.6.5. Internet marketing.....	15
1.7. Applications of sentiment analysis.....	15
1.7.1. Word of mouth.....	15
1.7.2. Voice of voters.....	15
1.7.3. Online commerce.....	16
1.7.4. Voice of market.....	16
1.7.5. Brand reputation management.....	16
1.7.6. Government.....	16
2. Literature Survey.....	18
3. Problem statement.....	21
3.1. Motivation.....	21
3.1.1. The consumer perspective.....	22
3.1.2. The producer perspective.....	22
3.2. Objective.....	23
3.3. Level of analysis.....	23
3.3.1. Document level analysis.....	23
3.3.2. Sentence level analysis.....	24
3.3.3. Aspect level analysis.....	24
3.3.4. Comparative sentiment analysis.....	24
3.3.5. Sentiment lexicon acquisition.....	24
3.4. PoS Tagging.....	25
3.5. Twitter dataset.....	26
3.6. Sentence weightage.....	27
3.7. Hashtag.....	27
3.8. Abbreviations and Redundant/Repeated letters.....	28
4. Implementation and methodology.....	29

4.1. Proposed methodology.....	29
4.2. Algorithm.....	30
4.3. Data extraction through Twitter API.....	30
4.4. Data collection.....	30
4.5. Pre-processing of Twitter data.....	31
4.6. Classification of machine learning classifier.....	33
4.7. Code.....	34
5. Results and analysis.....	36
5.1. Tweets collected from twitter.....	36
5.2. Weekly wise tweet collected.....	37
5.3. Performance metrics of sentiment classification.....	37
5.3.1. Precision.....	38
5.3.2. Recall.....	38
5.3.3. Accuracy.....	38
5.4. Results of classifier for twitter data.....	39
5.5. Results of four weeks of twitter data.....	41
6. Conclusion and future scope.....	47
7. References.....	48

LIST OF TABLES

Table 4.1 Removed and modified contents.....	32
Table 4.2 Sample and clean data.....	33
Table 5.1 Weekly wise report of twitter data.....	37
Table 5.2 Results of classifier on gaganyaan data.....	40
Table 5.3 Results report of first week.....	42
Table 5.4 Results report of second week.....	43
Table 5.5 Results report of third week.....	44
Table 5.6 Results report of fourth week.....	45

LIST OF FIGURES

Figure 1.1 Sentiment analysis Techniques.....	3
Figure 1.2 SVM classifier uses hyper-plane for classification.....	7
Figure 1.3 Recommender system using KNN algorithm.....	7
Figure 1.4 Concept search using KNN algorithm.....	8
Figure 1.5 Learning through decision tree.....	9
Figure 1.6 flow diagram of lexicon based approach.....	11
Figure 1.7 Steps of Natural Language Processing.....	13
Figure 3.1 PoS Tagging.....	25
Figure 4.1 flow diagram of twitter sentiment analysis.....	29
Figure 5.1 Tweets collected from twitter using Twitter API.....	36
Figure 5.2 performance of classifiers.....	41
Figure 5.3 Weekly wise report of classified data.....	46

LIST OF ABBREVIATIONS

NLTK :	Natural Language Toolkit
NLP :	Natural Language Processing
SA :	Sentiment Analysis
NB :	Naïve bayes
SVM :	Support Vector Machine
DT :	Decision Tree
KNN :	K - Nearest Neighbor
WOM :	Word of Mouth
VOM :	Voice of the Market
BRM :	Brand Reputation Management
API :	Application Programming Interface
URL :	Uniform Resource Locator
RT :	Re-Tweet
LOL :	Laughing Out Loudly
REST :	REpresentational State Transfer

Chapter 1

Introduction

Now a days twitter, facebook, whatsapp are getting so much attention from people and also they are getting very much popular among people. Sentiment analysis provides many opportunities to develop a new application. in the industrial field, sentiment analysis has big effect, like government organization and big companies, their desire is to know about what people think about their product, their market value. the aim of sentiment analysis is to find out the mood, behavior and opinion of person from texts. for the sentiment analysis purpose, social networking used the various sentiment analysis techniques to take the public data. Sentiment analysis widely used in various domain such as finance, economics, defense, politics. The data available on the social networking sites can be unstructured and structured. almost 80% data on the internet is unstructured. Sentiment analysis techniques are used to find out the people opinion on social media. Twitter is also a huge platform in that different idea, thought, opinion are presented and exchanged. It does not matter where people came from, what religious opinions they hold, rich or poor, educated or uneducated, they comment, compliment, discuss, argue, insist.

1.1 Introduction of python

in this thesis we are going to use python. Python is high level programming language. It is robust and versatile programming language. In python there is no need to compile the code because an interpreter is used in this which makes the testing and debugging with very high speed. an open source libraries is available for python.

It is very popular programming language . therefore it can be used in such as web development, software development, System scripting. It can works on various platform such as R, Raspberry, windows etc. syntax of python is similar to English language that help the programmer to write the less lines of code compare to other different programming language. The most updated version of python is python3. It is the updated version of python2 which is quite popular.

1.2 Introduction of Anaconda and jupyter

anaconda is nothing but a bunch of popular python packages. And a packet manager called conda(similar to pip). this python package are very popular in data science communities. Some of the popular packages are numpy, scipy, jupyter, nltk, scikit-learn etc. anaconda consist several python libraries. A light weight version of anaconda is also available called mini conda.in addition anaconda supplant their own package called conda. It is very efficient than PIP.

Jupyter is a interpreter that is based on browser that help you to work on python and R. anaconda consist jupyter libraries. You can consider jupyter as a notebook which is digital that provides you an chance to execute commands draw charts and takes notes. Data scientist used this as on prior basis. This is very helpful tool if you are learning python and R. jupyter is much better than shell.

Jupyter is amazing tool for the analytical work where you could show your code in “modules” adding common formatting option between modules and include of formatted output of modules and generate the graph in well suited manner in other modules code.

Jupyter assure reproducibility in other’s work. Therefore if someone come back after few months then by seeing the code he/she will easily get what someone has tried to do. And can exactly tell which code run which conclusion and visualization.

1.3 What is sentiment analysis

Sentiment analysis is a process of computationally identifying and categorizing opinions from piece of text, and determine whether the writer’s attitude towards a particular topic or the product is positive, negative or neutral. For instance suppose you want to buy a product. so before purchasing a product. You look for the feedback like what the other customer have to say about that particular product whether it is good or bad and you analyze it manually by looking at their feedback. now consider at the company level how did the company analyze what their customer is thinking about their product. Generally they do not have one or more customer. they do have millions of customers. So what they will do. So here company needs to do sentiment analysis. To know whether their product is actually doing good in the market or not.[19]

1.4 Sentiment Analysis Classification

Based on the different perspective. Sentiment analysis has different variety of class. In which only one is used in sentiment classification techniques. This is classified into two other approaches i.e. machine learning approach and lexicon based approach. We can add one more techniques i.e. hybrid approach.

There are three main classification level i.e. sentence level, document level, and last one is aspect level. Based on sentiment analysis, polarities can be classified into three classes such as positive neutral or negative.

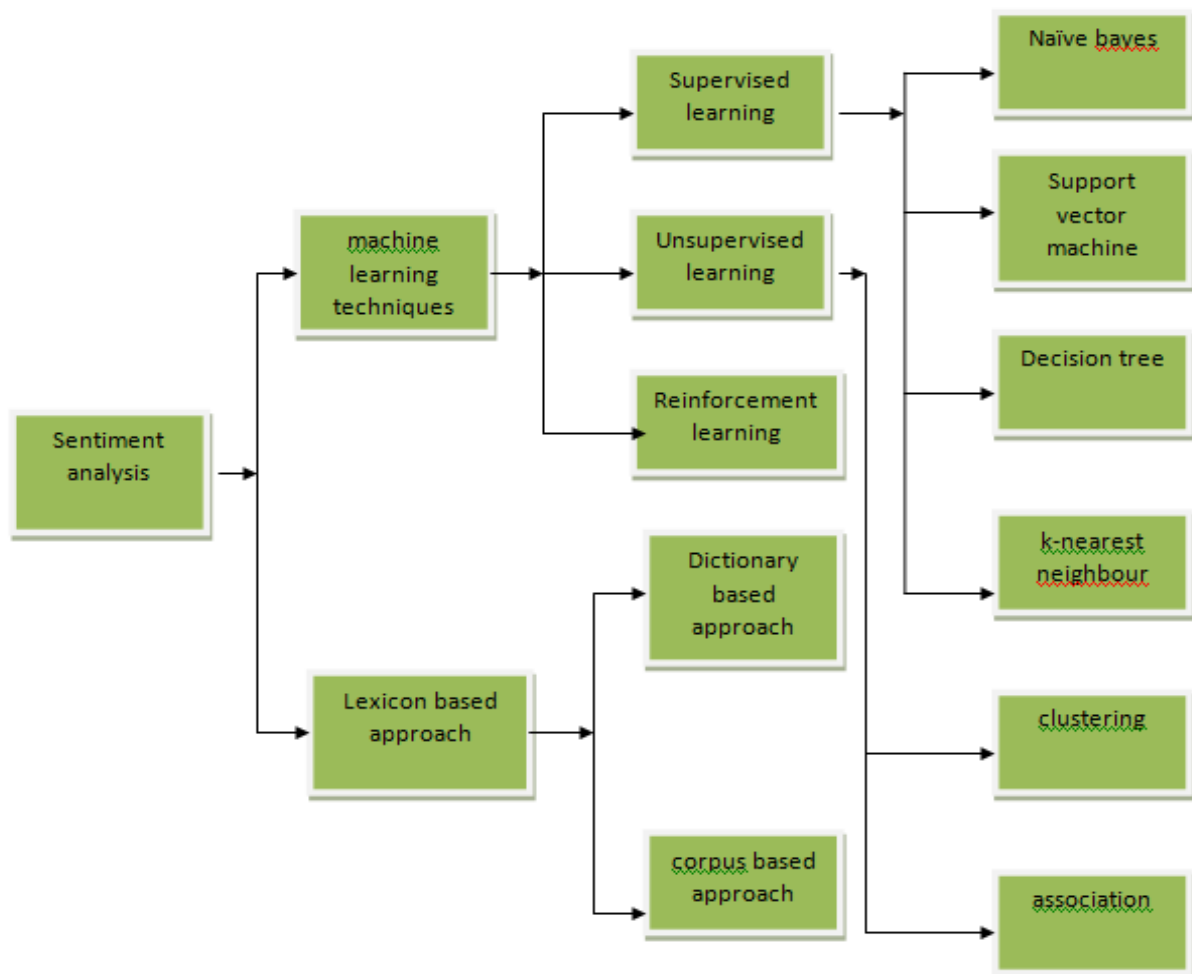


Figure 1.1. Sentiment Classification Techniques

Sentiment analysis is an area where we can classify the various techniques. It is the most popular area of research. It is notably classified into two types such as machine learning based approach and lexicon based approach. Lexicon based approach basically focuses on negative and positive terms and it is further classified into two types i.e. dictionary based and corpus based. Moreover, the machine learning approach is focused on two techniques namely supervised and unsupervised approach. Furthermore, supervised techniques are also classified that we will discuss in the next section.

1.4.1 Machine Learning techniques

To classify the text classification problem in sentiment analysis, machine learning is used. In this, to train a model, training data records are used which are later used to identify the predicted model without level. Each and every record is labeled into different classes. When we give a new unlabeled record to the model, then the model will label that dataset into different classes. There are three types of different classes such as positive, negative, and neutral. Generally, the neutral class is a mixed opinion. Rarely we consider the neutral class. Eventually, machine learning techniques are of three types i.e. supervised learning techniques, unsupervised learning techniques, and reinforcement techniques.

1.4.1.1 Supervised learning

In the field of machine learning, different classification techniques are used to classify the unlabeled data. These techniques use different classifiers for training the dataset. Examples of machine learning classifiers include naïve Bayes, support vector machine, KNN, and Decision Tree. These can be classified as supervised machine learning classifiers which require a training data set as prior. In supervised machine learning, we do have several data points that describe features/variables and target variables. The aim in supervised learning is to predict the final outcome variable given the predictor variable. The goal of supervised learning is to automate time-consuming or expensive manual tasks. For instance, “doctor’s diagnosis”. And we can make predictions about the future for instance “will a customer click on an ad or not?”. In supervised learning, there are many ways to get labeled data such as you can perform an experiment to label the data. Or you can do crowd-sourcing labeled data. In any of the above cases, the output is known so we can make predictions on new data for which we do not know the output. Therefore, there are many ways to perform

supervised learning in python such as scikit-learn it includes scipy libraries. It include other libraries as well such as tensorflow and keras. There are some supervised learning technique which we are going to explain below.

1.4.1.1.1 Naïve bayes(NB)

Naïve bayes theorem is a classification method with the independent assumption between the predictors. In other words the approach of particular predictor of one class is not connected to closeness of some other class. Naïve bayes is a “probabilistic classifier”. Lets take an instance ,an apple may be considered a fruit if it is red in color, and if it is round in shape and if its diameter is consider to be three inches approximately.[20] Despite of these feature are dependent on one another or in the presence of another feature. All these independent properties contribute to find the probability of naïve classifier that this is an apple. Naïve bayes is beneficial for big data sets and can be build easily. Let us consider a class variable ‘y’ and a dependent vector from x_1 to x_n . So according to naïve bayes:

$$P(y | x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n | y)}{P(x_1, \dots, x_n)}$$

so according to mutually independent assumption:

$$P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y),$$

for each value of i this function behaves:

$$P(y | x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i | y)}{P(x_1, \dots, x_n)}$$

Joint model could be expressed as:

$$P(y | x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i | y)$$

$$\Downarrow$$

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i | y),$$

1.4.1.1.2 Support vector machine(SVM)

It is a supervised machine learning technique. SVM is well known perform in sentiment analysis. The important information is represented in two vector where every vector is of size k. there is a classifier that separate the data in such a manner that margin should be maximum. SVM is used in sentiment classification and it perform better than naïve bayes in term of classification problem. By using the hyper-plane. We do implement SVM classification , regression and other works. These are the groups of hyper-plane and works in high dimensional space. therefore we can conclude that to achieve a good separation between any class The distance should be maximum for the nearest data point which is also called as functional margin. thus if the margin would be less then the generalization error between the classifier would be less. A hyper-plane instance is shown below.[25]

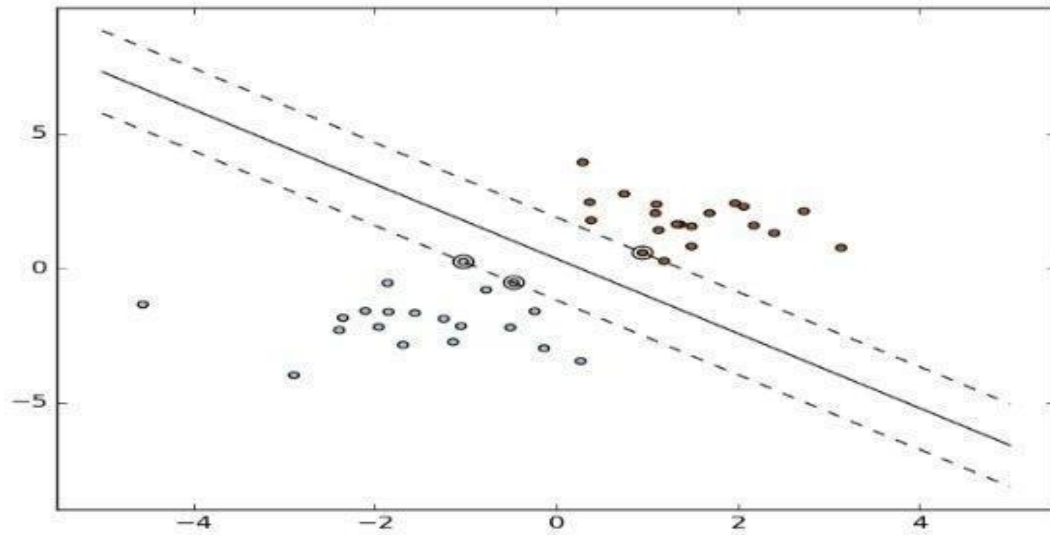



Figure 1.2 SVM classifier uses hyper-plane for classification

1.4.1.1.3 K-Nearest Neighbour(KNN)

k-NN is very simple algorithm that stores all the available cases and classifies the new data or case based on similarity measure. it uses the entire dataset in its training phase.



amazon
Recommender System
Industrial
Application of
KNN
Algorithm



Customers who bought this item also bought

Apple
Apple iPhone 8 (Space Grey, 64GB)
★★★★☆ 281 customer reviews | 195 answered questions

M.R.P. ₹ 7,400.00
Price: ₹ 62,450.00 FREE Delivery.
You Save: ₹ 7,437.47 (99%)
Inclusive of all taxes

EMI starts at ₹2,959 per month. Options +
Want it faster? Available with **A. Fulfiller** FREE delivery from another seller at ₹ 63,925.

Only 1 left in stock.
Delivery to pincode 603203 - Kanchipuram between Jul 11 - 13. Details
Sold and fulfilled by Nationalshope (4.5 out of 5 | 190 ratings)
8 offers from ₹ 63,925.00

Colour: Space Grey


Size name: 64GB
64GB 256GB

- 11.93 centimeters (4.7-inch) capacitive touchscreen with 1334 x 750 pixels resolution
- iOS v11 operating system with 1.2GHz Apple A11 Bionic hexa core processor, 2GB RAM, 64GB internal memory and single SIM
- 1821mAh lithium-ion battery
- 1 year manufacturer warranty for device and in-box accessories including batteries from the date of purchase
- See more product details

Report incorrect product information.

LIMITED QUANTITY The order quantity for this product is limited to 1 unit per customer. Please note that orders which exceed the quantity limit will be auto-cancelled. This is applicable across sellers.

Page 1 of 10

Figure 1.3 recommender system using KNN algorithm

For instance if apple look most similar to banana, orange ,melon rather than a monkey, dog or cat most likely apple belong to the group of fruits. In general k-nn is used in search application where you are looking for the similar item.In KNN, k denotes the number of nearest neighbor which are holding class of the new data or the testing data. KNN is used at the industries level.

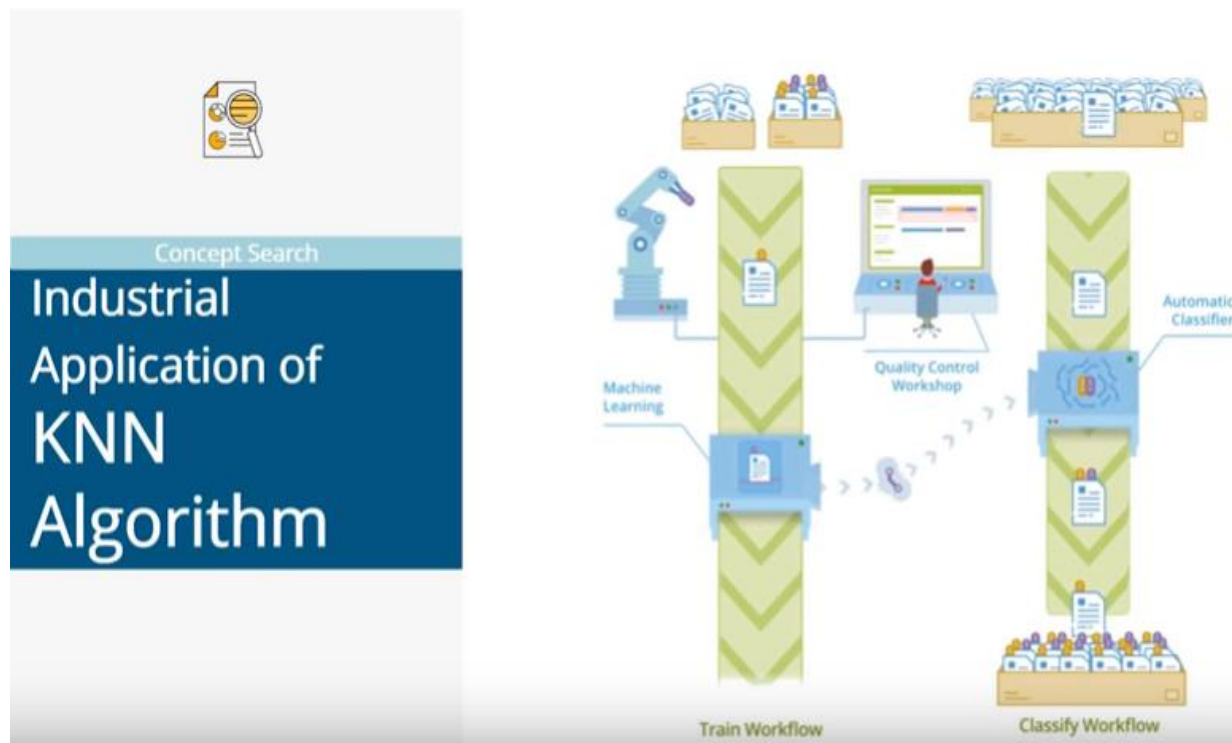


Figure 1.4 concept search using KNN algorithm

The biggest use case of KNN search is recommended system. recommended system is a automated form of shop counter guy. When you will ask for the product it not only show to you the relevant product but also suggest you or recommend you the product related to your relevant product that you want to buy. KNN algorithm applying to recommending product like in amazon and for recommending media in Netflix. In the above diagram indicate the concept search or searching semantically similar documents and classifying documents containing similar topics.

1.4.1.1.4 Decision Tree

Decision tree is a classification algorithm which comes under the supervised machine learning techniques. It is a graphical representation of all the possible solutions to a decision based on certain condition. Such as for every node there is two path “should we go through it is not” and it will come until we will get our aim. In the field of prediction and classification, decision tree play major role. There are various applications of decision tree. Decision tree is used in product design when there are series of decision and the next outcome depends on the previous outcomes. Another application of decision tree is when the user has some objective he wants to get max. profit or he may want to optimize the cost

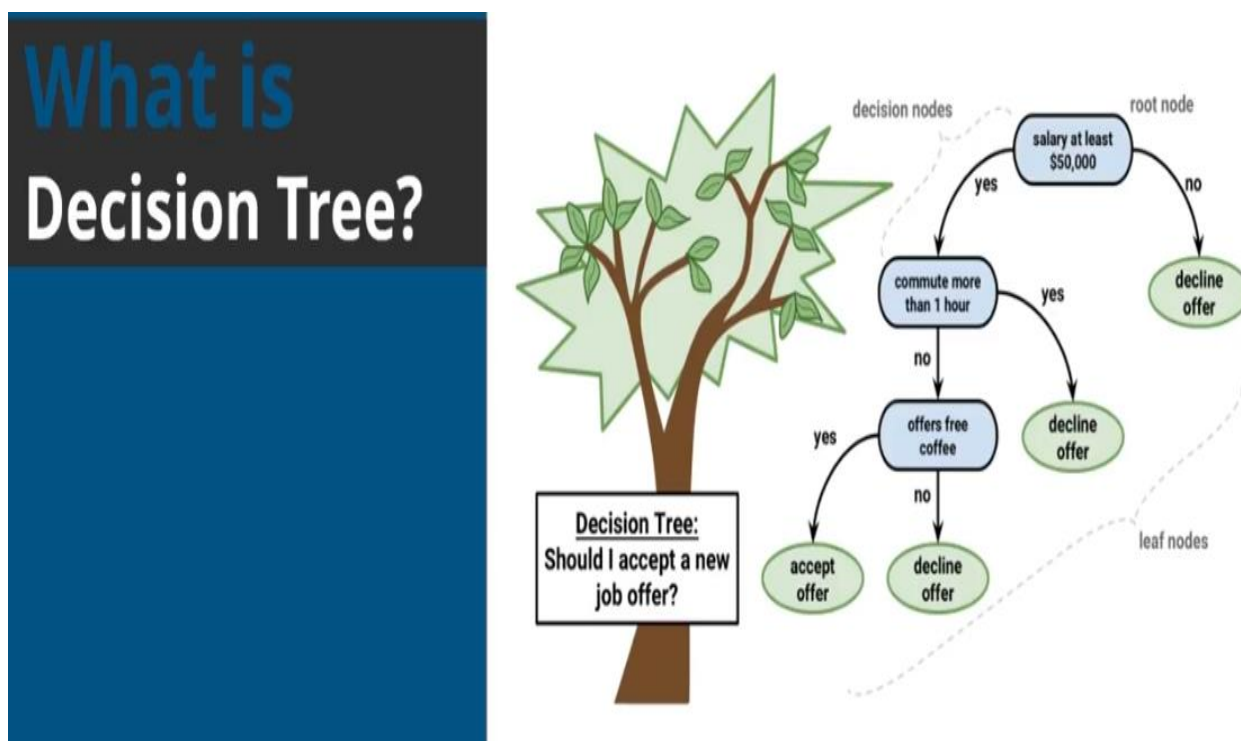


Figure 1.5 learning through decision tree

1.4.1.2 Unsupervised learning

It is an approach of machine learning. It does not use the information that is neither classified nor labeled. it allows the algorithm to perform on that information without human guidance. The task in the unsupervised learning to grouping of unsorted data according to patterns, differences, similarities without any previous information of data. In unsupervised learning there is no human

guidance provided that means there is no training of data given to the machine. Therefore in unlabeled data, machine is unable to find the hidden structure by ourselves. Unsupervised learning is further classified into two categories i.e. clustering and association.

1.4.1.3 Reinforcement learning

It is a well approach of machine learning . It performs suitable action in a generic situation to get the maximum awards. When the agents performs correct then it get rewards otherwise penalties if perform incorrectly. The agents started learning without taking interruption from human to maximize the awards and to minimize the penalty. Reinforcement learning is different from supervised machine learning in such a way that in supervised machine learning the training dataset knows the solution to the problem already. so this model is already trained with the known solution. but in case of reinforcement learning there is no already known solution to the problem but the agents has to decide what should be perform on the given action. there is no training dataset. It learns from the experience.

1.4.2 Lexicon-Based Approach

Lexicon based approach comes under the unsupervised learning approach. But it uses a dictionary for antonyms and synonyms of sentiments phrases and words with their corresponding opinionated guidelines. Furthermore lexicon based approach is further classified into two approaches i.e. dictionary and corpus based approach which is generally used by lexicon based to find sentiment.

1.4.2.1 Dictionary based approach

It is used to compile the opinionated words. Generally dictionaries contains list of positive and negative sentiments. The procedure for the dictionary based approach is very easy. In the first stage we manually collect the list of known positive and negative words. Then the algorithm search for the synonyms and antonyms in the wordNet or online dictionaries for growing this dataset. Later they updated the wordlist if found. Followed by next iterations. Then this process continues till we got no words to update the dictionaries. Finally When this process is finished then we clean the list manually.

1.4.2.2 Corpus based approach

It is mainly used to find the new opinionated words from a corpus by using a list of known sentiment words. And the other main use of corpus based approach is to build a sentiment dictionary with the help of other words. Generally this approach is not dominant compare to dictionary based because it needs dictionary of all English words. Corpus based approach is further divided into two techniques i.e. semantic and statically.

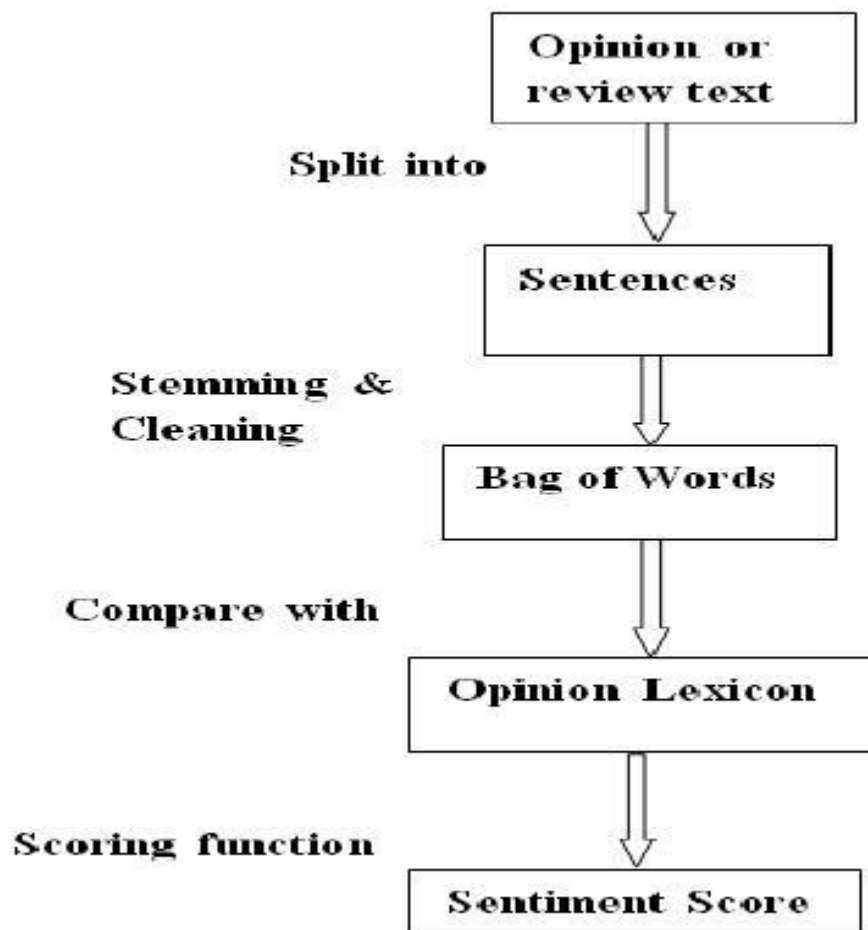


Figure 1.6 flow diagram of lexicon based approach

1.5 Natural language processing

Natural language processing deals with techniques that can analyze, compute, and represent the data at various level analysis of languages for the purpose to make the machine process like human language for different disciplines and applications. NLP algorithm highly depends on machine learning techniques with the major being numerical. there are various types of natural language processing namely[26].

1.5.1 Morphological processing

morphological processing deals with meaningful components of words. In morphological, strings are broken down into the sets of token corresponding to the punctuation, words, sub-words etc. however in morphological, transformation will not be happen by adding prefix or suffix but could be by other major changes

1.5.2 Syntax and semantic analysis

syntax analysis deals with the study of structural relationship among words. It can be used by two ways. The first use is to identify that the sentence is well structured or not and the second use is to break down the sentence into a structure that give correct syntactic relation. The syntax analysis can also be gained by a set of syntactic rules and by the help of parser by taking the word from dictionary.

1.5.3 Pragmatic analysis

pragmatic analysis deals with the study of use of language to achieve goals. It is a part of process to extract the particular information from the text.

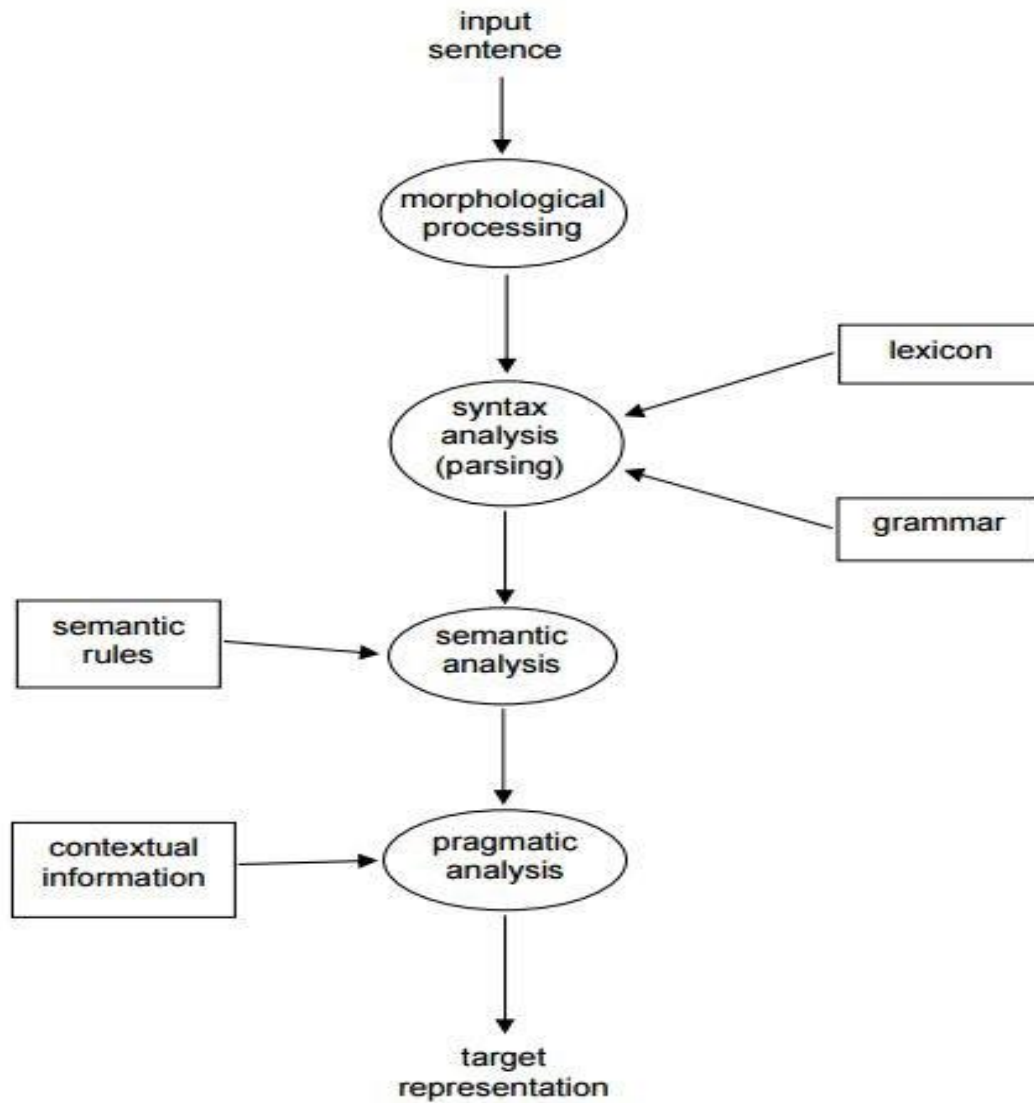


Figure 1.7 Steps of Natural Language Processing

1.6 Need of sentiment analysis

1.6.1 Industry evolution

from the industry point of view there is no need of fully unstructured data. Only the beneficial information is required. Sentiment analysis is useful for the industry purpose because they extract the useful information from the data. It can yield an excellent opportunity to the industries to value their earnings and audience. Many industries that carry out commercial activities with consumers will benefit in the case of restaurants, hotels, entertainment, mobile customers, retailers or travel, hospitality

1.6.2 Research demand

the other key reason behind the successively growth of SA is that it deals with the research in evaluation, classification, views, and opinion and . The current solutions for the analysis of sentiment and the analysis of opinion are evolving rapidly, including the reduction of human effort that is needed to labeled the comments. The research topic will also be based on well-established computer disciplines, such as text extraction, artificial intelligence, machine learning, NLP and, voting applications and automated content analysis. etc.

1.6.3 Decision Making

Any person who kept their information on different web applications, blogs and social media websites. to obtain the relevant information from the social web you required a particular method for analyzing the data and, therefore, return some useful results. It would be very difficult for the company to organize the survey on the daily basis and took the user opinion advices or views based on their best product. People reviews and opinion help the other people while taking the best decisions and also help in the field of research and business areas.

1.6.4 Understanding contextual

it is very difficult for the machine to understand the human language because it is going to be very complicated gradually. And that could be shows nuances, slangs, spelling error, cultural

variations. therefore a better system is needed that can help the human to understanding the machine language.

1.6.5 Internet marketing

the other big reason for the successively growth of sentiment analysis is online marketing which is being used by the big companies and industries. on the daily basis these companies took the data i.e. user opinion from the internet about their brands, products, social post , blogs, events. therefore sentiment analysis works as a tool for marketing.

1.7 Applications of sentiment analysis

Sentiment analysis has many applications in the field of NLP. Due to growth in the area of sentiment analysis of feelings, data from social media are in great demand. for the improvement process so many big companies got adopt the sentiment analysis. Some of the main applications are mentioned below.

1.7.1 Word Of Mouth(WOM)

Word of mouth (WOM) is deals with the process by which information is passed from one person to another. It helps people while making decisions. The major task of WOM is to classify documents based on WOM into the positive and negative. WOM has great effect on marketing strategies and customers behaviour. Consumers influence affects the informative quality of WOM. If the quality of WOM is high then it more influence the customer behavior the low quality review product. Therefore most WOM focused model works on the quantity WOM . they do not consider the WOM quality. However, A document with good quality information is more reliable than a document with less quality information.

1.7.2 Voice of voters(VOV)

now a days, debate on politics is going on every corner of the country. It is going popular day by day. Every political parties spent a lot of money for the good advertising of their parties and they also spent money on voters before some days back from voting. Instead of doing a lot of

these things if they took the people opinion and reviews, that would be very effective. Therefore sentiment analysis not only helps the political parties but also helps in doing the analysis of news. People from other countries have already taken the sentiment analysis.

1.7.3 Online commerce

Internet has the huge collection of e-commerce websites. These websites take the reviews of customers after buying the product. After obtaining such information from different areas such as quality details of user and services, the user of the company experiences the suggestion, product review and features. These companies collect all the information such as reviews and opinion from the internet and convert these data into geographic form.

1.7.4 Voice of market(VOM)

Whenever any company does launch their product, the customer is very curious to know about the reviews of product, rating of product, description of product, price of product and other details of product. So sentiment analysis helps these companies while making new marketing strategies and advertising to promote their product. It helps the customer while selecting the best product among all.

1.7.5 Brand reputation management(BRM)

Sentiment analysis also helps to dictate the brand position. How is the brand position. Are the services provided by the brand good or not. Do the employees of the brand behave good with customer or not. They generally focus on product and brand. They do not focus on customer. Therefore these opportunities are created to manage and strengthen the reputation of brand organization.

1.7.6 Government

Sentiment analysis is doing help of organization to offer the different types of services to public. To analyze the negative and positive views of government, best result must be generated. Therefore sentiment analysis is pretty much helpful in different fields such as taxation, decision making policies, recruitments, and to evaluate the social strategies. There are some almost equal

techniques which provide the priorities and services as citizen. There is one major problem which could be applied on this in the country such as India where we speak different different languages. In India, sometimes we people speak the hybrid language i.e. the mixture of two or more language i.e. (English Punjabi).and it is very common.

Chapter 2

Literature Survey

In this chapter, we will discuss different literature survey of different author. There are many researchers who has done great contribute in this area. In sentiment analysis, researchers has done various research on sentiment analysis by using different techniques. Here we are going to discuss the some researches that will help us to know about the sentiment analysis in depth.

Dhiraj gurkhe, Niraj pal and Rishit Bhatia discussed how twitter data is processed firstly they collected data from various sources and eliminate those feature which does not contribute to find any polarity and then this data send into the sentiment classification engine i.e. naïve bayes classification algorithm which will calculate the probabilities i.e. how much data is corrected and predict the sentiment for the given query.[1]

M.bouazizi, T. ohtsuki have discussed the tweets which contain more than one sentiment called as multi class sentiment analysis. Where they have identify the exact sentiment conveyed by the user rather than the whole sentiment of the tweet. To identify this thing they have also used SENTA tool. They proposed an approach, with the help of this approach they have calculated the sentiment score whoever sentiment is having highest score that will be considered this process is called as “Quantification”. [2]

Geetika gautam, Divakar Yadav have discussed about customer review classification for which they have used twitter dataset which is already labeled. In this task they have used machine learning based algorithm i.e. naïve bayes, SVM, maximum entropy. They have worked on Python and NLTK for training the SVM, naïve bayes, maximum entropy. Naïve bayes is better techniques in term of accuracy and gives the better result compare to Maximum entropy. We can get the better result with compare to SVM by using the SVM with unigram model. And then further accuracy can be improved by semantic analytic followed by wordNet.[3].

Akshay Amolik, Niketan jivane, Mahavir Bhandari, Dr. m.venkatesan, a highly suitable model have discussed in his paper which will take the twitter data of upcoming Hollywood and bollywood movies. They are able to this task with the help of classifier and features like SVM and naïve bayes. Both of them are used for high accuracy but in terms of precision naïve bayes is better than SVM and if we talk about recall then SVM is better than naïve bayes. By increasing the dataset we can increase the classification accuracy.[4]

Subhabrata Mukherjee, Akshat Malu, Balamurali A.R, Pushpak Bhattacharya have discussed a hybrid system named as TwiSent which will resolve problem like spam tweet ,pragmatics, noisy text. Twisent consist of spell checker and pragmatics handler. spell checker finds the noisy text whereas pragmatics handler handles the pragmatics in tweets. Twisent gives better result compare to C-feel-IT system. The accuracy of finding the negative sentiment of TwiSent system is high the C-Feel-IT.[5].

Dmitry Davidov, Oren Tsur, Ari Rappoport in this paper they have proposed a supervised sentiment classification framework which is based on twitter data. They have used K nearest neighbor and feature vector. the basic purpose of this framework is to identify and distinguish between sentiment types defined by smiley and tags.[6]

Neethu M S, Rajasree R, the author have used the machine learning techniques in this survey paper to explore the twitter data related to electronic product. They have used feature vector for the tweets classification . they have used three types of classifier i.e. SVM, naïve bayes, maximum entropy, and these classifier were tested using Matlab simulator. SVM and naïve bayes classifier are implemented using built in function. Whereas MaxEnt classifier is used by MaxEnt software. So basically the all classifier have nearly the same performance.[7]

Pulkit et al. built and proposed a model which extract tweet from twitter based on the post terror activities. they made their study on terrorist attack which was occurred in uri on 18 september 2016. They considered 59,988 tweet which had taken after the attack. They consider only those tweets which has #UriAttack, #uriattack. #uriattacks. They have used the naïve bayes and SVM to extract the last re-tweet time and number of re-tweet.[8]

Sudarshan Sirsat et al. proposed a technique in sentiment analysis on twitter data where they have collected reviews of the product. They have used naïve Bayes algorithm which perform better in term of accuracy and efficiency. They have extracted 200 tweets where the average length of tweet was 70.105. the aim of this research is to identify the characteristic of tweet like how many times the tweet was liked and how many times they have re-tweet the tweet.[9]

Hetu et al. proposed a model in sentiment analysis on twitter data based on anaconda python. They extract the dataset from kaggle in which they classify the people emotions based on positive and negative reviews. This model gives high accuracy on large dataset.[10]

Ali hasan et al. proposed a model using the hybrid approach that comprise sentiment analyzer machine learning. They took only those tweet that is followed by the hashtag(#) and contain the current political trends. Basically this model converts the urdu tweet into English tweet. They took 1690 tweet for training data and 400 for testing the data. They have used the naïve bayes and SVM classifier for training the dataset in weka and building a model. They have used three different libraries to calculate the subjectivity and polarity.[11]

Feddah AlhumaidiAl Otaibi et al. proposed a model by using the supervised and unsupervised algorithm. They wanted to know that which restaurant has more popularity between mcdonald and kfc by using the sentiment analysis. Moreover , they extracted 7000 tweets of both the restaurant by twitter API. The tweet was in English and they used R programming language. Because R programming language can perform big computational task. They have used several machine learning techniques but they found MaxEnt has performed better result compare to other technique. Moreover they have also found KFC have many neutral tweet. And McDonald have more positive and negative tweet.[12]

Chapter 3

Problem Statement

we have given a large collection of tweet that contain multiple types of features and opinions. Our task is to extract the opinion from the dataset that describes the target feature and distribute it as positive or negative, neutral. Sentiment analysis deals with process of extracting the features from people views, opinion, thoughts and feelings, which they used to post on social websites. The outcome of SA is the classification of human language into the classes such as positive negative and neutral. Huge amount of data is generated from the social networking sites. these data can be in any form whether unstructured or structured, generally these type of data are unstructured. unstructured data do not convey any meaning until unless it is not analyzed. Therefore to utilize these unstructured data, there is a need of performing the sentiment analysis on this data i.e. take the valuable feature from these data and classify them into classes. Now a days , performing sentiment analysis on data is very important because the data on the internet is growing with very high rate and people are pretty much affecting with the opinion of other people and they are distracting by this unstructured data. In today's world if someone wants to buy anything or he wants to watch movie or he wants to sell anything on the internet then before doing any of the above activity, that person will go through the thoughts of other people that means what other people thinks about that activity so it is very important. therefore we can conclude that we should generate that type of system which could automatically perform sentiment analysis on this big amount of data.

3.1 MOTIVATION

It is an open research field area and it is used in various real world application. People are using Forums, facebook, twitter, blogs and many other application on internet to express their views or opinions. With the help of these applications people are coming closer: communication in one click away[21]. Twenty years ago there was no social media at that time , people used short messaging services(SMS) to talk with people with high national and international charges. Now a days SMS services has been evolved by sending messages from one person to many person with the cheapest price. Many telecom operators provided this services but the twitter is the one who lead it. In today's era twitter is growing very fastly. Millions of people do billions of tweets

daily approximately hundreds tweets per seconds. Tweets posted by people is not necessary in English language. It could be in various local language. For business intelligence purpose these data are very confidential because some major companies wants to know "why isn't consumer buying our laptops? ", "why the competitors products are outselling our products".

3.1.1 The consumer perspective

With the help of feedback in sentiment analysis consumer can make choices. In older times people were asked to friends, relatives, for the reviews of any product. But now a days with the help of internet or social media they can take the people opinion from the internet. This information is very useful for the people who are planning to buy a particular product. With help of this they can make decision "should we buy this product or not?". This kind of planning is generally binary Either he will purchase it or he will not purchase it. User could not take the decision as he go through the large amount of data. Therefore this process must be automated because it involve great use of technology which will generate bad or good opinion, finally helping the users by taking the decision.[22]

3.1.2 The producer perspective

Consumer decisions go hand in hand with producer decisions. While consumers are busy sharing their opinions online, producers monitor their behavior to make business decisions. This scenario creates a multi-user consumption and producer system in which the nature of consumer spending encourages other consumers to choose to buy products and encourages the producer to sell the product. The producer learns the evolution of sales of products and services through opinions and reschedules his future business plan. The products and services express their impression and usefulness through a series of decisions of consumers and producers.

3.2 Objective

Our main objective of this thesis is to do the sentiment analysis on “GAGANYAAN” like people thoughts about “will be able to launch our space station or not”. Which we will get from twitter. Therefore to get this aim, we will build a classifier using machine learning technique which will collect from huge data from the internet and will perform sentiment analysis on this

3.3 Level of analysis

Sentiment analysis is the area of study where we deliberate the people’s behavior related to any topic, about any chronicle. It is highly produces the big problem zone. It is having different tasks and multifarious names e.g. sentiment mining, subjectivity analysis, affect analysis, review mining etc. There are three different levels of sentiment analysis which are as follows

3.3.1 Document level analysis

Opinions are feelings, notions or expressions about a event or comment. Numerous data in the network or in meetings allow people to express their evaluation in the form of surveys and comments. At a time when opinions are provided in the form of research, rather than merely positive or negative, the recognition of true opinions would require a subjective examination of the words used in the investigation. At this document level the task is to determine the the overall opinion of the document. It assumes that sentiment analysis at document level expresses opinions at a single entity.

Opinions about a product will not be the same as those of a person to another person. There are only two types of courses, either positive or negative. A legitimate case: Review of an object: "I reported on a new iPhone two days before, it's a good phone, the touch screen is fast, the clarity of the voice is better, like this phone." The words that were used to be used. Objective feelings are measured using the star chart or review, where 4 or 5 stars are safe and 1 or 2 stars are negative.

3.3.2 Sentence level analysis

This method is used to provide useful data during the search because the polarity of the sentence will be perfect. At this level of feeling analysis, look at the sentences that contain opinions and give opinions as negative or positive. Our task is to determine whether each and every sentence expressed a positive or negative or neutral opinion. Neutral generally means no opinion

3.3.3 Aspect level analysis

Document level and sentence level analysis works admirably when it refers to a single element. Again, many times people talk about components with different points of view or qualities. Also, they will have undeniable feelings about the distinctive elements. This happens frequently in the review of things and in social dialogues. A valid example: "I am a great Nokia phone, I like the look of the phone, the screen is huge and clear, the camera is extraordinary, but there are also some disadvantages: Burning and accessing WhatsApp are annoying." Ask the points The positive and negative aspects of this review mask vital information about the thing, so that the analysis of feelings based on aspects focuses on the recognition of all the reports in a given register and in the points. in which the feelings are expressed. It is also called feature level analysis. At this level we can find what people likes or dislikes but on the sentence level and document level we can't find what people likes or dislikes. This level perform finer-grained analysis. it is closely related to tasks like feature based opinion mining and opinion summarization.[23]

3.3.4 Comparative sentiment analysis

Users use to express various emotions in articles or brands. It can be the same item or the same brand. The purpose of comparative feeling is to discover the assumption of a relative sentence.

3.3.5 Sentiment lexicon acquisition

Sentimental analysis is a process that uses data to search for opinions and expressions of that data. In the analysis of feeling, we will see two types of positive and negative classes. Given a statement: "Auto X is superior to anything that is automatic." This statement does not show to

which class this statement belongs. Similarly, these types of sentences / documents are analyzed using three systems: manual methodology, dictionary approach, and corpus approach.[24]

3.4 POS Tagging

POS marking is extremely valuable in Opinion Mining. When we need to examine a document or sentence first, we need to focus the subjective data of the particular file or phrase. POS tagging helps us find parts of the word's discourse. After extracting these words, we can perform different activities and reach a conclusion. POS Tagging is done using the HMM model that used tokenize and tag the words further to name the elements.

The word in the content (or phrase) is tagged by a POS tagger for the purpose of naming each name, thus allowing the machine to do something with it. It looks something like this:

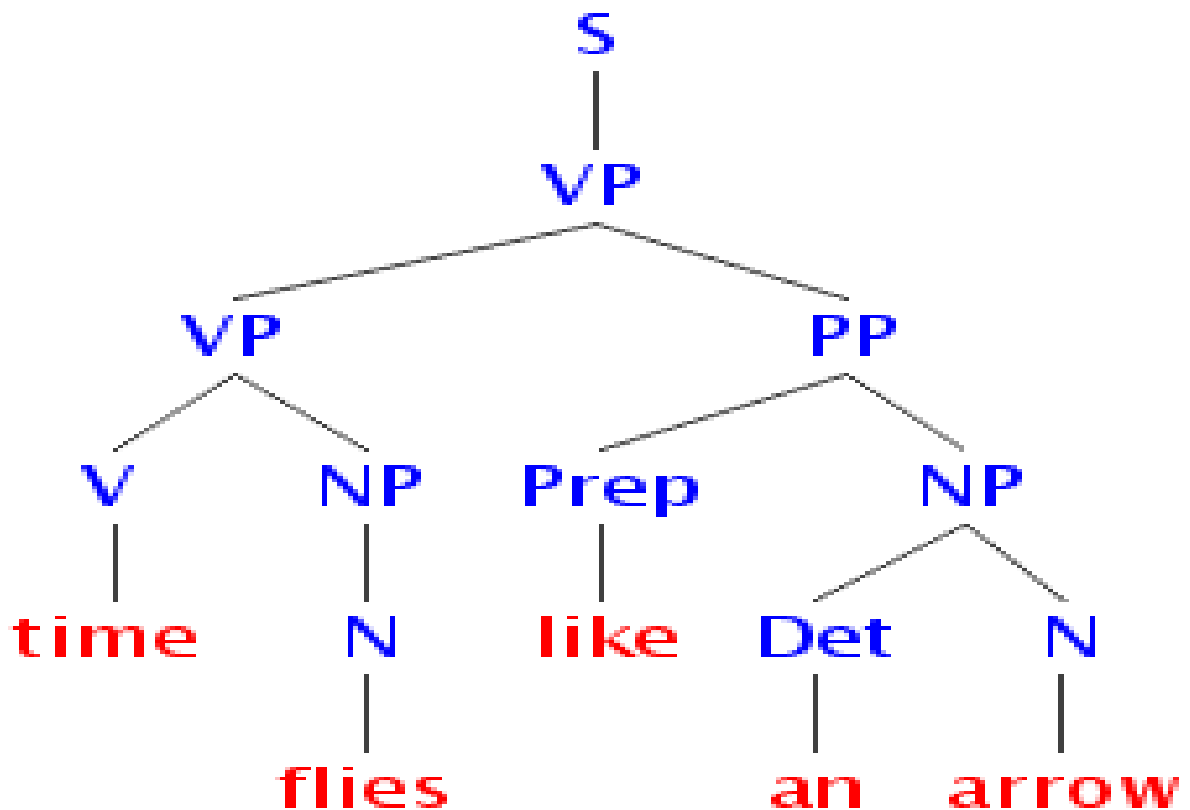


Figure 3.1 POS Tagging

We are extracting sentiment orientation(SO) from the example. For instance we might have taken Amazing + phone which is:

[JJ] + [NN] (or descriptive word took after by thing in human)

The reverse of this may be “Repulsive” for instance. On this stage the machine will try to arrange the words in an emotive scale.

It deals with sufficient number of views from the normal sentimental knowledge, we have gathered. This will allow the machine to say something like: "The big ones like the new iPhone" and they tell us like "Most people hate the new iPhone". They do not

3.5 Twitter dataset

it is our aim to classify the tweets in various sentiment classes to perform sentiment analysis on twitter data. In this research field, to train a model several approaches have developed which is also used for testing to check its efficiency. It is challenging to perform sentiment analysis on twitter data. There are some reasons defined for this

- **Bound tweet Size**

size of the tweets is bounded i.e. it is having only 280 characters which generate intensive statement.

- **By Using slang words**

these slang words are bit different from English words. If we used slang words in our sentence then that slang words make an approach outmoded.

- **Twitter features**

twitter is permissive to use of URL's,hashtag(#) and user reference. In the twitter feature different processes are used to compare with other different words.

- **User diversity**

there are various ways to convey their opinion, people's used different language between the tweets while others used encore words or symbols to convey their emotions.

All the above problems are required to be faced in the pre-processing section. The microblogging steps are used by different people to express their opinions on various topics. Therefore, it is an important source of people's opinions.

- Twitter contains a large amount of content posts and is constantly growing. The assembled corpus can be huge.
- The Twitter crowd ranges from normal client to superstars, delegates from organizations, legislators and even national presidents. In this perspective, it is possible to collect messages from clients of various social groups and intrigues bunch
- Customers in many countries are talking to Twitter. Even though American customers are winning, it is possible to collect information in several languages.

Writings that contain positive opinions, for example, satisfaction, allegation or pleasure. Targeted messages that simply express a certainty or do not express feelings, carry out an analysis of our body and show that it is possible to create an exercise book that uses the body collected as learning information. Given a message, indicate if it is a positive, negative or impartial sentiment. For messages that convey both a positive sentiment and a negative sentiment, the most valid opinion should be chosen.[27]

3.6 Sentence weightage

- If a tweet contain more than one sentences then the upcoming sentence will be having more weightage.
- This is due to the tendency of most tweets to be of a conclusive nature.
- It can improved the accuracy by approximately 3% if we test it on small group of tweets

3.7 Hashtag

- To get the information from the tweet we will use the hashtag(#).
- hashtags for instance: #indiawonworldcup ,#goodmorning and etc.
- these hashtags are not the complete sentences however they are used to express the views of people.

- Before going to treat them it is compulsory to analyze it.
- these hashtag #happy, # sad, # cool, etc. do convey enough knowledge to find the polarity of tweets.

3.8 Abbreviations and Redundant/Repeated letters

- Due to the simplicity of the Twitter dialect, some words (usually, concluding words) are often misspelled or underlined, so the classifier may not respect the polarity of this word (for example, loooooooooove). the authentic word. (for example, love) during training.
- if in the words, if a letter is coming more than thrice continuously then these occurrences are replaced twice in the letter. for example. haaaaaaaaappy would change like haappy, goooooooooood would also change to goood.
- Create a general description of the most common and frequent shortcuts among the most used shortcuts.

Chapter 4

Implementation and Methodology

Yet Collecting the data is not very simple task. We do have to consider so many points while collecting the data. So in our thesis we will collect the dataset for training, testing and for sentiment analysis. This chapter study consist how data will collect, how data will processed , stored and mainly how to classify those data. Before moving on to this thing let's do discussion about proposed architecture.

4.1 Proposed Methodology

As mention earlier, our aim is to do sentiment analysis for twitter data. By using various kind of machine learning classifier, we will build a classifier. once it get trained then we will follow different step to sentiment analysis as mention in below diagram:

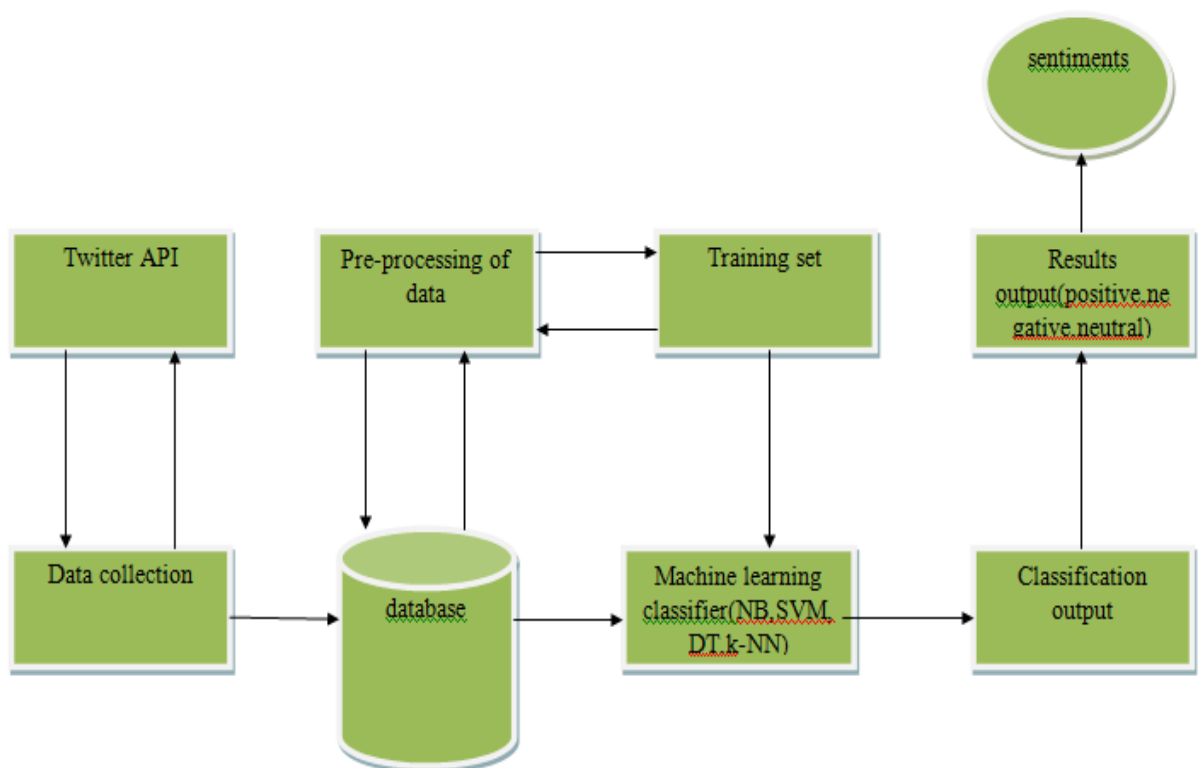


figure 4.1 flow diagram of twitter sentiment analysis

4.2 Algorithm

- in this step we will extract tweets for our new classifier by using the Tweepy API library in python.
- in this step after streaming the tweets, we will do preprocessing of those tweets so that they could work well in feature extraction and mining.
- now after completing the pre-processing, we will send our data to our new build classifier Which will classify our data into classes such as positive ,negative and neutral and our

classifier will also tell the accuracy.

Since, for analyzing the tweets , we have took the data from twitter that will stream into our database. Therefore we are going to use twitter application.

4.3 Data extraction through Twitter API:

With help of Tweepy, we can download the tweet from the twitter. Twitter API is of two type i.e, REST API and Streaming API. Rest API stands for REpresentational state transfer. So there is bit difference between these two. Rest API works on response and reply basis. In which client sends request to server and then server sends back request to client whereas in the case of streaming API it is different in respect to response. Whenever there is any update available on server. Then it started continuously to client.

4.4 Data collection

Twitter data- in order to extract data through tweeter API ,we should have account on twitter.com. it could easily access by filling sign up form on twitter website. After getting successful registration on twitter they will provide a valid username and a valid password , which you can use for further login. Once you done with this process, then you are allowed to do tweets ,retweet,likes etc. on any topic you want.[18]

Twitter is a platform which provide you to access data and also allow you to use it for self purpose. Indeed before using the twitter site firstly we must have to login on twitter then only we can access the data. This website ask you to provide the necessary detail for creating an application which later you could use for streaming purpose. When our API is generated . then

we will get access of some keys such as customer keys, access token key, customer secret key, access secret key. These key play important role when any user wants to get the data.

The major part of this thesis is to collect the tweet and to analyze the those tweet which are related to gaganyaan. In order to fetch the tweet from twitter, we have design a python script which will works on “tweepy” which is based on python library. So basically first we need to install the tweepy in our system.

Python is powerful programming language. In order to access the python services, python has very well designed libraries. Python has open source library called “tweepy” which empower python to connect with twitter and it also used their API to extract data which we are going to use in our program.

4.5 Pre-processing of twitter data

Twitter data may be in unstructured format that is not good for extracting feature. Tweets may consist of empty spaces, stop words, slangs, special characters, hashtag, emoticons, time stamps, abbreviations, URL's etc. for mining these data we should have to pre-process the data first by the using the functions of NLTK. While doing pre-processing our first aim is to extract message then we will remove all hashtags(#), empty spaces, repeating words, stop words(such as he, she, them, the etc.). emoticons and abbreviation will be replaced by their corresponding meaning such as :-), =D, LOL. They will be replaced by happy, laugh and laughing out loudly respectively. After done this thing we are ready to give this pre-process data to our new classifier for further process so we could get our required result.

We did code in python where we define function which would be used to get processed data.

- Remove quotes: give the access to user to eliminate the quotes from the tweet.
- Remove @- give the option to remove the @ symbol, delete @ together with the username or replace @ and the username with a word 'AT_USER' and append to the stop words.
- Remove URL's- URL stands for uniform resource locator. offers options to remove URLs or replace them with the word 'URL' and append to stop words.
- Removal of RT(Re-Tweet)- it deleted the RT word from the text.

- Removal of emoticons- replace emoticons with their correct meaning
- Removal of duplicates- delete all the duplicate word from the tweet.
- Removal of hashtag(#)- remove hashtags from the tweet.
- Removal of stopwords- delete all the stopwords from the tweet such as he, she, them because they do not convey meaning in classification.
- Removal of slang- remove all slangs with their specific meaning such as lol i.e laughing out loudly Etc.

Content	Action
Punctuation(!,?,.,,")	Removed
#word	Removed #word
ULR's and web links	Remove URLs or replaced with "URL" and then added in stop words
Number	Removed
Word not starting with alphabets	Removed
All Word	Stemmed all word (Converted into simple form)
Stop words	Removed
Emoticons	Replaced with respective meaning
White spaces	Removed

Table 4.1 Removed and modified content

Raw data	Clean data
# ISRO Not To Fly Living Being Before Actual Manned Space Mission	['isro', 'not', 'to', 'fly', 'living', 'being', 'before', 'actual', 'manned', 'space', 'mission']
ISRO not to have test flight with any living being before actual manned space mission	['isro', 'test', 'flight', 'living', 'actual', 'manned', 'space', 'mission']

Table 4.2 sample and clean data

After cleaning of data our next step is to classify the data into different classes by using some classifier. so we will used some supervised learning classifier.

4.6 Classification of machine learning classifier

We made a classifier in such that it consists of various kind supervised learning classifier which then classify the tweets into binary classes that is positive and negative. Python library did use in building a classifier, scikit-learn. Scikit-learn is a library which is used to perform machine learning in python. It is an open source library and also based on popular library such as numpy, scipy, matplotlib. Scikit has many tuning parameter along with wonderful documentation and support. Therefore it has many tools for classification, visualization, regression, clustering etc. through a simple command in python we can install scikit-learn in our system such as 'pip install scikit-learn'.

There are several classifier comes under the scikit-learn. Some of them we are going to explain below:

- Naïve bayes(NB)
- Support vetor Machine(SVM)
- Decision Tree(DT)
- K-nearest neighbor(k-NN)

4.7 Code

```

from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(trainmat, trainlab)
testNB = classifier.predict(testmat)
from sklearn.metrics import confusion_matrix
cmNB = confusion_matrix(testlab, testNB)
from sklearn.metrics import accuracy_score
print("NB accuracy: ",accuracy_score(testlab, testNB))
NB_Accuracy = accuracy_score(testlab, testNB)
print("NB Precision: ",cmNB[0][0]/(cmNB[0][0] + cmNB[1][0] + cmNB[2][0]))
NB_Precision = cmNB[0][0]/(cmNB[0][0] + cmNB[1][0] + cmNB[2][0])
print("NB Recall: ",cmNB[0][0]/(cmNB[0][0] + cmNB[0][1] + cmNB[0][2]))
NB_Recall = cmNB[0][0]/(cmNB[0][0] + cmNB[0][1] + cmNB[0][2])

```

####SVM

```

from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', gamma = 0.1, C=1, random_state = 0)
classifier.fit(trainmat, trainlab)
testSVM = classifier.predict(testmat)
from sklearn.metrics import confusion_matrix
cmSVM = confusion_matrix(testlab, testSVM)
from sklearn.metrics import accuracy_score
print("SVM accuracy: ",accuracy_score(testlab, testSVM))
SVM_Accuracy = accuracy_score(testlab, testSVM)
print("SVM Precision: ",cmSVM[0][0]/(cmSVM[0][0] + cmSVM[1][0] + cmSVM[2][0]))
SVM_Precision = cmSVM[0][0]/(cmSVM[0][0] + cmSVM[1][0] + cmSVM[2][0])
print("SVM Recall: ",cmSVM[0][0]/(cmSVM[0][0] + cmSVM[0][1] + cmSVM[0][2]))
SVM_Recall = cmSVM[0][0]/(cmSVM[0][0] + cmSVM[0][1] + cmSVM[0][2])

```

```

## KNN

from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
classifier.fit(trainmat, trainlab)
testKNN = classifier.predict(testmat)
from sklearn.metrics import confusion_matrix
cmKNN = confusion_matrix(testlab, testKNN)
from sklearn.metrics import accuracy_score
print("KNN accuracy: ",accuracy_score(testlab, testKNN))
KNN_Accuracy = accuracy_score(testlab, testKNN)
print("KNN Precision: ",cmKNN[0][0]/(cmKNN[0][0] + cmKNN[1][0] + cmKNN[2][0]))
KNN_Precision = cmKNN[0][0]/(cmKNN[0][0] + cmKNN[1][0] + cmKNN[2][0])
print("KNN Recall: ",cmKNN[0][0]/(cmKNN[0][0] + cmKNN[0][1] + cmKNN[0][2]))
KNN_Recall = cmKNN[0][0]/(cmKNN[0][0] + cmKNN[0][1] + cmKNN[0][2])

```

```

## Decision Tree

from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(trainmat, trainlab)
testDT = classifier.predict(testmat)
from sklearn.metrics import confusion_matrix
cmDT = confusion_matrix(testlab, testDT)
from sklearn.metrics import accuracy_score
print("DT accuracy: ",accuracy_score(testlab, testDT))
DT_Accuracy = accuracy_score(testlab, testDT)
print("DT Precision: ",cmDT[0][0]/(cmDT[0][0] + cmDT[1][0] + cmDT[2][0]))
DT_Precision = cmDT[0][0]/(cmDT[0][0] + cmDT[1][0] + cmDT[2][0])
print("DT Recall: ",cmDT[0][0]/(cmDT[0][0] + cmDT[0][1] + cmDT[0][2]))
DT_Recall = cmDT[0][0]/(cmDT[0][0] + cmDT[0][1] + cmDT[0][2])

```

Chapter 5

Results and Analysis

This chapter consist of various opinion mining result that we have achieved in the implementation.

5.1 Tweets collected from twitter

ISRO Not To Fly Living Being Before Actual Manned Space Mission: Official Link: <http://defencebuzz.org/2018/09/isro-not-to-fly-living-being-before-actual-manned-space-mission-official-826ttaab-2tt4>

ISRO Not To Fly Any Living Being Before Actual Manned Space Mission Read here : <http://dailyaddaa.com/Read/isro-not-to-fly-any-living-being-before-actual-manned-space-mission-54300.html> ...pic.twi

ISRO not to have test flight with any living being before actual manned space mission - The News Minute <http://dlvr.it/QkD2jZ> pic.twitter.com/SaT85DU6sv

ISRO not to fly any living being before actual manned space mission <https://newsinfo.in/news/technology/isro-not-to-fly-any-living-being-before-actual-manned-space-mission/> ...pic.twitter.com, Interesting timing. Feeling pressured about a second cis lunar manned mission announcement?

I hadn't realised that India is using the term "cosmonauts" in relation to its manned space mission. <https://www.thehindu.com/news/national/iaf-ready-for-space-challenge-says-air-chief/article2494909>

ISRO Not To Fly Any Living Being Before Actual Manned Space Mission <http://www.thehawk.in/science/isro-not-to-fly-any-living-being-before-actual-manned-space-mission-102712> ...

ISRO not to fly any living being before actual manned space mission <https://www.siasat.com/news/isro-not-fly-any-living-being-actual-manned-space-mission-1406867/> ...pic.twitter.com/xtz23gZpNE

India—Prime Minister Modi says "India will unfurl the tricolor in space" in first manned space mission by 2022—India will be 4th country to send humans into space—joining Russia, US, and China—In 2014

ISRO Not To Fly Living Being Before Actual Manned Space Mission: Official <https://glit.al/7TGuxf3k>

ISRO not to fly any living being before actual manned space mission | india news - <https://southasiansnews.com/2018/09/14/isro-not-to-fly-any-living-being-before-actual-manned-space-mission-india/>

I've just posted a new blog: ISRO not to fly any living being before actual manned space mission <https://ift.tt/2Myl9hf>

ISRO not to fly living being before actual manned space mission: Official <https://www.ndtv.com/india-news/isro-not-to-fly-living-being-before-actual-manned-space-mission-official-1916654> ...pic.twitte

Alot depends on how their first manned mission goes. Not looking long to wait for that.

ISRO not to fly any living being before actual manned space mission - <https://goo.gl/BZQN4G>

ISRO Not To Fly Living Being Before Actual Manned Space Mission: Official <https://goo.gl/fb/BKKnrc>

ISRO Not To Fly Living Being Before Actual Manned Space Mission: Official <https://dharanmind.wordpress.com/2018/09/14/isro-not-to-fly-living-being-before-actual-manned-space-mission-official/> ...

ISRO Not To Fly Living Being Before Actual Manned Space Mission: Official <https://www.news1inda.com/isro-not-to-fly-living-being-before-actual-manned-space-mission-official/> ...

ISRO Not To Fly Living Being Before Actual Manned Space Mission: Official <https://navkiratsinghmand.wordpress.com/2018/09/14/isro-not-to-fly-living-being-before-actual-manned-space-mission-offici>

ISRO Not To Fly Living Being Before Actual Manned Space Mission: Official <https://ift.tt/2Ndg2YT>

ISRO not to fly any living being before actual manned space mission <https://goo.gl/fb/YhAQXg> # samachar # news

ISRO not to fly any living being before actual manned space mission <https://m.economictimes.com/news/science/isro-not-to-fly-any-living-being-before-actual-manned-space-mission/articleshow/6588>

ISRO Not To Fly Any Living Being Before Actual Manned Space Mission <http://www.dailyaddaa.com/Read/isro-not-to-fly-any-living-being-before-actual-manned-space-mission-54300.html> ...

ISRO not to have test flight with any living being before actual manned space mission <https://www.thenewsminute.com/article/isro-not-have-test-flight-any-living-being-actual-manned-space-mission-1>

Figure 5.1 tweets collected from twitter using Twitter API

So above are the tweets dataset which we have fetch from the twitter with the help twitter API. These twitter dataset contain tweet which are related to “Gaganyaan”. Gaganyaan is a mission of space satellite which is run by Indian government. In which we will able to set our own station in

space. This space craft is designed in such manner that it will carry three people to space.[14][15][16]

5.2 Weekly wise tweet collected

Week 1		Week 2		Week 3		Week 4	
Date	No. of tweets	Date	No. of tweets	Date	No. of tweets	Date	No. of tweet
15/08/18	369	22/08/18	04	29/08/18	40	05/09/18	13
16/08/18	272	23/08/18	08	30/08/18	33	06/09/18	40
17/08/18	150	24/08/18	20	31/08/18	21	07/09/18	77
18/08/18	16	25/08/18	10	01/09/18	18	08/09/18	29
19/08/18	41	26/08/18	23	02/09/18	09	09/09/18	07
20/08/18	16	27/08/18	53	03/09/18	51	10/09/18	09
21/08/18	06	28/08/18	44	04/09/18	09	11/09/18	33
						12/09/18	18
						13/09/18	15
						14/09/18	40

Table 5.1 weekly wise report of twitter data

This table consist of four weeks of twitter data which is done by user. This data are fetched by twitter API. These data is in unstructured form so we will perform the sentiment analysis on this twitter data. After this we will perform the classification on this data in which we classify these data into different classes such as positive, negative, neutral. This table contain 1431 twitter dataset.

5.3 Performance metrics of sentiment classification:

Generally, to measure the performance of sentiment classification we use some predefined standards such as accuracy, precision, recall. Accuracy is dependent on two measure.

5.3.1 Precision

To get the right value of precision, we divide the total number of rightly classified positive observation by the total number of predicted positive observation. High precision denotes that the observation classified positive is indeed positive.

$$\text{Precision} = \frac{tp}{tp + fp}$$

5.3.2 Recall

It is the ratio between the right classified positive observation to the total number of positive observation. High recall denotes that the class is rightly classified.

$$\text{Recall} = \frac{tp}{tp + fn}$$

5.3.3 Accuracy

In order to find the which model gives better result, then it is necessary to find the accuracy. Accuracy for any model can be given as:

$$\text{Accuracy} = \frac{tp + tn}{tp + fn + tn + fp}$$

Where,

Tp=true positive, case was positive and it predicted positive

Tn=true negative, case was negative and it predicted negative

Fn=false negative, case was positive and it predicted negative

Fp= false positive, case was negative and it predicted positive

5.4 Results of classifier for twitter data:

In this thesis, different classifier have tested on same dataset in which some give best performance in terms of precision, recall and accuracy. These are data which we have fetched from the twitter dataset. This data contain the 'gaganyaan' tweet. Here below are performance of some classifier.

- Naïve bayes(NB):

```
NB accuracy: 0.6341463414634146
NB Precision: 0.6271929824561403
NB Recall: 0.89375
```

Naïve bayes classifier is tested on our dataset. Generally it works on large dataset and it is fast. It the accuracy 63% and precision is 62% and it gives better performance in term of recall i.e, 89%.

- Support vector machine(SVM)-

```
SVM accuracy: 0.7896341463414634
SVM Precision: 0.8531468531468531
SVM Recall: 0.7625
```

Support vector machine classifier generally it works better in small dataset. It give accuracy of 78% and moving on to side of precision and recall 85% and 76% respectively.

- Decision tree(DT):

```
DT accuracy: 0.7957317073170732
DT Precision: 0.8301886792452831
DT Recall: 0.825
```

this classifier is used to tested our dataset. It gives accuracy of 79% and it terms of precision it gives result 83% and recall 82%.

- K-nearest neighbor(k-NN):

```

KNN accuracy: 0.7073170731707317
KNN Precision: 0.8557692307692307
KNN Recall: 0.55625

```

It works well in accuracy it gives result 70% . precision and recall 85% and 55% respectively.

With the help of Tweepy API we collected total of 1431 tweets from twitter and did sentiment analysis on those tweets by using some supervised learning classifier such as support vector machine, k-nearest neighbor, naïve bayes, decision tree. with the help of these classifier we are able to find standard measure such as accuracy, precision and recall. Performance results of these classifier are mention in table below.

Supervised learning techniques	Accuracy	Precision	Recall
SVM	78.9	85.3	76.25
Naïve Bayes(NB)	63.4	62.7	89.3
Decision tree (DT)	79.5	83.0	82.5
k-NN	70.7	85.5	55.6

Table 5.2 results of classifier on “Gaganyaan” dataset

To find the superior one techniques among the all techniques, we have done a comparative analysis. It is found that DT has 79.5% accuracy which is the highest among all whereas naïve bayes has 63.4% accuracy that is least accuracy among all. So we can conclude that DT is best to find the accurate result. Moving on to the precision K-NN got 85.5 precision that is the highest among all. That means k-nn gives substantially relevant results than the irrelevant results. And on the other hand Naïve bayes got 89.3% recall which is the highest among all classifier that means our classifier returned most of the relevant results.

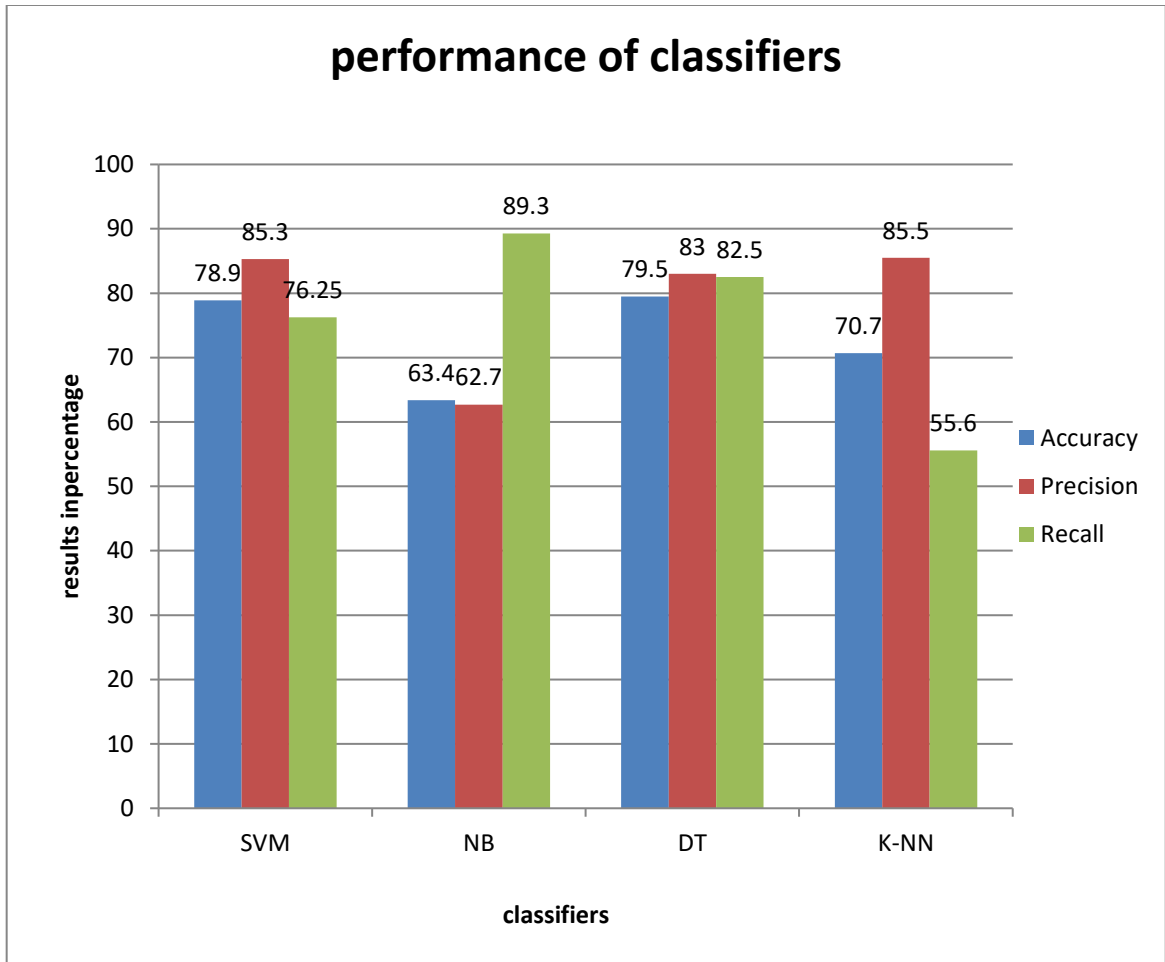


Figure 5.2 performance of classifiers

5.5 Results of four weeks of twitter data:

Here we going to re-present the result of every week. our classifier has classify the data into different classes such as positive, negative, neutral. And also count the total number of tweets.

Where positive tweets represents that people has given their opinion in favor of subject on the other hand negative tweet represents people has given their opinion against of subject. Where as neutral tweet denotes the mixed opinion.

- **First week results of classified data**

Week 1				
Date	Negative tweet	Neutral tweet	Positive tweet	Total tweet
15/08/18	15	42	312	369
16/08/18	10	104	168	272
17/08/18	02	63	85	150
18/08/18	03	09	04	16
19/08/18	02	21	18	41
20/08/18	02	13	01	16
21/08/18	02	03	01	06
Total	36	255	589	870

Table 5.3 results report of first week

The statistics of first week indicates the total around 870 tweets collected and out of which nearly 36 are negative and around 255 are neutral and nearly 589 are positive. In which on first date our classifier extract 15 negative tweet , 42 neutral tweet and 312 positive tweet. This things indicate that on first date people were more curious. And they give their reaction in favor of subject. The overall opinion of this week is positive which indicates that people are in favor of this mission.

- **Second week results of classified data**

Week 2				
Date	Negative tweet	Neutral tweet	Positive tweet	Total tweet
22/08/18	02	01	01	04
23/08/18	03	04	01	08
24/08/18	05	11	04	20
25/08/18	02	03	05	10
26/08/18	05	08	10	23
27/08/18	10	22	21	53
Total	27	49	42	118

Table 5.4 results report of second week

Statistics of second week represents mixed type of opinion. It is seen that percentage of positive tweets are more in first week compare to second week. Second week data includes 49 and 27 neutral and negative tweet respectively. Many number of people has given mixed opinion . they things depends on that mission of second week. Eventually the ratio of negative tweet of both week i.e first week and second week is almost equal. Which denotes that less number are people are against of this mission. Further more total number of tweet we got at the end of this week is 118 which is very less compare to first week.

- **Third week results of classified data**

Week 3				
Date	Negative tweet	Neutral tweet	Positive tweet	Total tweet
29/08/18	09	11	20	40
30/08/18	05	12	16	33
31/08/18	05	13	03	21
01/09/18	03	03	12	18
02/09/18	04	03	02	09
03/09/18	08	23	20	51
04/09/18	03	04	02	09
Total	37	69	76	181

Table 5.5 results report of third week

Records of third week represents more positive tweet. In this week people has given more positive opinion towards ‘gaganyaan’. But there is hike in number of negative tweet compare to second week. this week we got more positive tweet compare to second week but we got very less positive tweet compare to first week. We are continuously seeing that the negative tweet of all third week is almost same. But as of now we got drastic change between the positive tweet. Coming on to neutral tweet 69 people have given their mixed opinion. Finally at the end we got total 181 tweet .which is more than the second week but very less compare to first week.

- **Fourth week results of classified data**

Week 4				
Date	Negative tweet	Neutral tweet	Positive tweet	Total tweet
05/09/18	04	04	05	13
06/09/18	12	15	13	40
07/09/18	03	23	51	77
08/09/18	06	13	10	29
09/09/18	03	01	03	07
10/09/18	03	01	05	09
11/09/18	07	03	23	33
12/09/18	04	02	12	18
13/09/18	05	03	07	15
14/09/18	10	10	20	40
Total	57	75	149	281

Table 5.5 results report of classified data

Statistics of this week is very clear. We got 281 total number of tweets. Out of which 149 were positive tweets. And also we got 57 and 75 negative and neutral tweets respectively. In this week there is hike in positive number of tweet compare to second week and third week. And people has given more negative tweet compare to second week and third week. 75 people have give their mixed opinion.

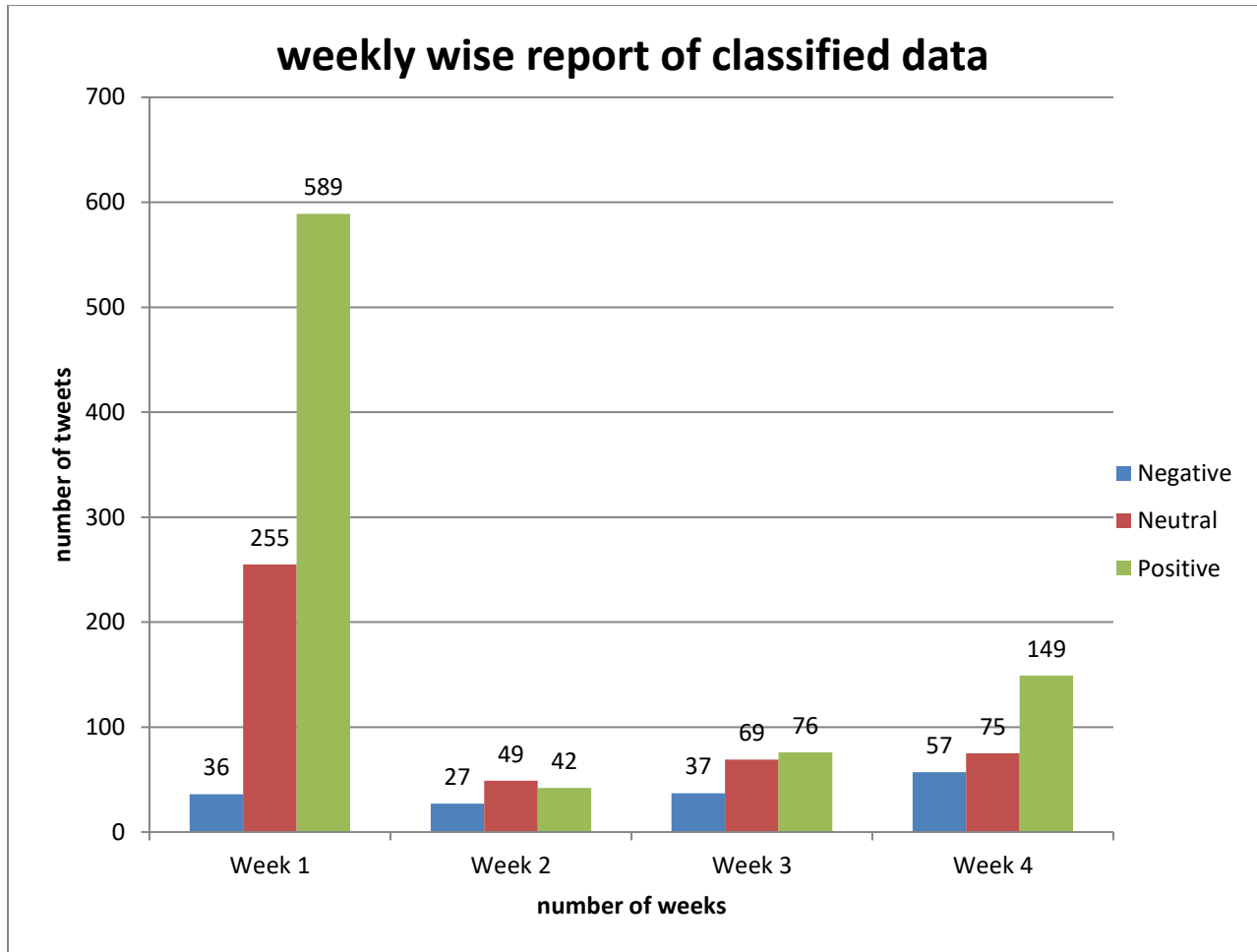


Figure 5.3 weekly wise report of classified data

This is the graph between the number of tweets and number of weeks. Which indicates that in each and every week, people has done how many tweets. Such as number of positive tweet and number of negative tweet, number of neutral tweet.

Chapter 6

Conclusion and Future Scope

In today's world, spacious amount of data is generated by various communication such as social media, organizations etc. these data may or may not be in structured form. Therefore to understand the polarity of data first we need to do the sentiment analysis of data. Opinion mining can be performed in various field such as marketing and customer feedback. large number of organizations are taking the valuable feedback of person and performing opinion mining on those data so that they could provide the better services to the customer and this data helps the organizations to enhance their future services. Furthermore, there are various scopes where we can perform the opinion mining such as sentence, paragraph, documents, sub sentences levels. In addition to this we took some sentiment classifiers such as support vector machine, naïve bayes, decision tree, K-nearest neighbor which performs best in terms of accuracy, precision, and recall. Out of these classifiers we conclude that DT performs best in finding accuracy of twitter dataset. It is best classifier on this dataset.

Basically our goal in this thesis is to find the public opinion and perform the opinion mining. Generally what happens, through tweets people express their thoughts, feelings etc. but we could not able to find the people thoughts and feelings. So by performing sentiment analysis on those tweets finally we can conclude how many person are in favor of this mission and how many person are against of this mission.

Future scope includes, we can make web application for our work. In addition to this we can improve our classifier system such that it could deals with sentences that conveys multiple meaning. Furthermore, we can add more classification categories so that we could get better results. We can also design system such that it can detect the images in tweets with the help of image processing.

Chapter 7

References

- [1] Gurkhe D., Pal N. and Rishit B. "Effective Sentiment Analysis of Social Media Datasets using Naïve Bayesian Classification." (2014).
- [2] Bouazizi, M., Ohtsuki, T.: Multi-Class Sentiment Analysis in Twitter: What if Classification is Not the Answer. *IEEE Access*. 6, 64486-64502 (2018).
- [3] Gautam, G., Yadav, D.: Sentiment analysis of twitter data using machine learning approaches and semantic analysis. 2014 Seventh International Conference on Contemporary Computing (IC3). (2014).
- [4] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." *International Journal of Engineering and Technology* 7.6 (2016): 1-7.
- [5] Mukherjee S., Malu A., Balamurali A.R, Bhattacharyya P. "TwiSent: A Multistage System for Analyzing Sentiment in Twitter".
- [6] Davidov D., Tsur O., Rappoport A." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys".
- [7] Neethu, M., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). (2013).
- [8] Pulkit Garg, Himanshu Garg, VirenderRanga "Sentiment Analysis of the Uri Terror Attack UsingTwitter" International Conference on Computing, Communication and Automation (ICCCA2017).
- [9] Prof. SudarshanSirsat, Dr.Sujata Rao, Dr.BhartiWukkadada"Sentiment Analysis on Twitter Data forproduct evaluation" *IOSR Journal of Engineering (IOSRJEN)* ISSN (e): 2250-3021, ISSN (p): 2278-8719PP 22-25.(2019)
- [10] Hetu Bhavsar, Richa Manglani" Sentiment Analysis of Twitter Data using Python"International Research Journal of Engineering and Technology (IRJET) Mar 2019e-ISSN: 2395-0056 p-ISSN: 2395-0072

- [11] Ali Hasan, Sana Moin, Ahmad Karim and ShahaboddinShamshirband” Machine Learning-Based Sentiment Analysis forTwitter Accounts” 2018 by the authors. Licensee MDPI, Basel, Switzerland.
- [12] Sahar A. El_Rahman, Feddah Alhumaidi AlOtaibi ,Wejdan Abdullah AlShehri “ Sentiment Analysis of Twitter Data”.
- [13] "India announces first manned space mission". Bangalore: BBC News. ^ Press Trust of India (25 April 2012). "Spaceflight stuck due to budget: CAG". *Times of India*. New Delhi. Retrieved 11 June2013.
- [14] Press Trust of India. "Human space flight mission off ISRO priority list". Retrieved 18 August 2013.
- [15] <https://indianexpress.com/article/what-is/what-is-gaganyaan/>
- [16] Priyadarshi, Siddhanta (23 February 2009). "Planning Commission Okays ISRO Manned Space Flight Program". *Indian Express*. p. 2.
- [17] Beary, Habib (27 January 2010). "India announces first manned space mission". Bangalore: BBC News.
- [18] E. Loper and S. Bird, “NLTK: the Natural Language Toolkit”, Proc. ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics ,vol. 1,pp. 63-70, 2002.
- [19] P. Pang and L. Lee, “Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval”, vol. 2(1-2), pp.1-135, 2008.
- [20] A. McCallum and K. Nigam, “A comparison of event models for Naive Bayes text classification”, Proc. AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48, 1998
- [21] G. Kontaxis, I. Polakis, S. Ioannidis, and E.P. Markatos. Detecting social network profile cloning. In Pervasive Computing and Communications Work- shops (PERCOM Workshops), 2011 IEEE International Conference on, pages 295300. IEEE, 2011.
- [22] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. ”Sentiment analysis algorithms and applications: A survey.” *Ain Shams Engineering Journal* (2014).

- [23] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B.Y. Zhao. Detecting and characterizing social spam campaigns. In Proceedings of the 10th annual conference on Internet measurement, pages 3547. ACM, 2010.
- [24] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In Proceedings of the 27th Annual Computer Security Applications Conference, pages 93102. ACM, 2011.
- [25] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schtze. Introduction to information retrieval. Vol. 1. Cambridge: Cambridge university press, 2008.
- [26] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing- Volume 10. Association for Computational Linguistics, 2002.
- [27] Sharma, Anuj, and Shubhamoy Dey. "A comparative study of feature selection and machine learning techniques for sentiment analysis." Proceedings of the 2012 ACM Research in Applied Computation Symposium. ACM, 2012