1.INTRODUCTION

With the increase in the popularity of social networking, micro-blogging and blogging websites, a huge quantity of data is generated. We know that the internet is the collection of networks. The age of the internet has changed the way people express their thoughts and feelings. The people are connecting with each other with the help of the internet through the blog post, online conversation forums, and many more. The people check the reviews or ratings of the movies before watching that movie in theaters. The quantity of information is unreasonable for a normal person to analyze with the help of naive technique.

Sentiment analysis is mainly concerned with the identification and classification of opinions or emotions of each tweet. Sentiment analysis is broadly classified in the two types first one is a feature or aspect based sentiment analysis and the other is objectivity based sentiment analysis. The tweets related to movie reviews come under the category of the feature based sentiment analysis. Objectivity based sentiment analysis does the exploration of the tweets which are related to the emotions like hate, miss, love etc.

In general, various symbolic techniques and machine learning techniques are used to analyze the sentiment from the twitter data. So in another way we can say that a sentiment analysis is a system or model that takes the documents that analyzed the input, and generates a detailed document summarizing the opinions of the given input document. In the first step pre-processing is done.

To correctly classify the tweets machine learning technique uses the training data. So, this technique does not require the database of words like used in knowledge-based approach and therefore, machine learning techniques is better and faster. The several methods are used to extract the feature from the source text.

Feature extraction is done in two phases: In the first phase extraction of data related to twitter is done i.e. twitters specific data is extracted. Now by doing this, the tweet is transformed into

normal text. In the next phase, more features are extracted and added to feature vector. Each tweet in the training data is associated with class label.

This training data is passed to different classifiers and classifiers are trained. Then test tweets are given to the model and classification is done with the help of these trained classifiers. So finally we get the tweets which are classified into the positive, negative and neutral

2. LITERATURE SURVEY

2.1. Survey1:

Title: SENTIMENT ANALYSIS ON TWITTER USING STREAMING API

By: M.Trupthi, Suresh Pabboju, G.Narasimha

Abstract:

This project aims to provide an interactive automatic system which predicts the sentiment of the review/tweets of the people posted in social media using hadoop, which can process the huge amount of data. A precise method is used for predicting weighted sentiment polarity, which helps to improve marketing strategies. Real time tweets are considered as the rich sources of data for opinion mining and sentiment analysis.

Findings:

The first phase is Training in which the classifier is used. An input to this module 20,00,000 tweets were collected from several sources which are already classified and the job of this module is to build a classifier by training on the large data set. Nltk is used to remove the words with POS tags which are not useful to build the classifier.

The second phase is classification. Bag-of-words method is used in which the relationships between words was not considered at all for sentiment analysis and a sentence is simply considered as a collection of words. To determine the sentiment for the whole sentence, sentiment of every individual word was determined separately and those values are aggregated using some aggregation functions. The proposed system extracts the data from SNS services which is done using Streaming API of twitter. The extracted tweets are loaded into hadoop and it is been pre -processed using map reduce. This task is followed by classification which uses NLP and machine learning techniques. The classification used here is uni-word naive bayes classification.

The third phase is Application phase and User Interface. When a user gives a keyword it gives an analytical report which have number of positive, negative and neutral tweets.

2.2. Survey2:

Title: TWITTER SENTIMENT ANALYSIS

By: Afroze Ibrahim Baqapuri

Abstract:

This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Due to the large amount of usage we hope to achieve a reflection of public sentiment by analysing the sentiments expressed in the tweets. Analysing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream.

Findings:

The "Twitter Sentiment Analysis" is broken down into 5 phases. The first phase is DATA ACQUISITION, where data in the form of raw tweets is acquired using a python library, which provides a package for twitter streaming API. This API allows two modes of accessing tweets: SampleStream and FilterStream.

The second phase is "HUMAN LABELLING", where the tweets are labelled in four classes according to sentiments expressed/observed in the tweets: positive, negative, neutral/objective and ambiguous. Some guidelines have to be followed for labelling these classes.

The third phase is "FEATURE EXTRACTION", wherein we need to extract useful features from the training set which can be used in the process of classification. For this phase some text formatting techniques such as Tokenization, Removal of Punctuations, Stopwords removal, Lowercase conversion, Stemming and POS Tagging are used.

The fourth phase is "CLASSIFICATION", where data is divided into different classes according to some common patterns which are found in one class which differ to some degree with the patterns found in the other classes.

The final phase is " WEB APPLICATION', where a user provides with some keywords and the web application will perform the appropriate sentiment analysis and display the results for the user.

2.3. Survey3:

Title: SENTIMENT ANALYSIS OF TWITTER DATA USING MACHINE LEARNING

APPROACHES AND SEMANTIC ANALYSIS

By: Geetika Gautam & Divakar yadav

Abstract:

This paper contributes to the sentiment analysis for customers' review classification which is helpful to analyze the information in the form of the number of tweets where opinions are highly unstructured and are either positive or negative, or somewhere in between of these two.

Findings:

Initially, the twitter dataset is used and is analysed first. Later the raw dataset is pre-processed to improve the efficiency and the quality of the data.

After preprocessing the improvised data has some distinctive properties and hence the feature extraction method, extracts the aspect (adjective) from the dataset. So, by doing this the positive and negative polarity in a sentence is shown, which is used for determining the opinion of the individuals.

Finally Training and Classification of data is done and for that python, nltk and machine learning classifiers such as naive bayes classifier is used and the performance of classifier in terms of recall, precision and accuracy is measured.

2.4. Survey4:

Title: SCALABLE AND REAL TIME SENTIMENT ANALYSIS OF TWITTER DATA

By: Maria Karanasou, Anneta Ampla, Christos Doulkeridis and Maria Halkidi

Abstract:

The proposed system relies on feature extraction from tweets, using both morphological features and semantic information. For the sentiment analysis task, we adopt a supervised learning approach, where we train various classifiers based on the extracted features. Finally, we present the design and implementation of a real-time system architecture in Storm, which contains the feature extraction and classification tasks, and scales well with respect to input data size and data arrival rate.

Findings:

The sentiment analysis is a two phase approach consisting of an offline and an online process. In the offline process, a labeled dataset of tweets is preprocessed in order to extract useful features,

and then it is used to train a classifier. After training, the classification model is stored on secondary storage, in order to be loaded and used in the online phase.

In the online process, tweets from the Twitter stream are received, preprocessed and features are extracted. Then, each tweet (more accurately its representation using features) is given as input to the classifier, which has already loaded the classification model, and is able to predict the sentiment of the tweet.

The goal of the offline process is to build the model that is used to predict the sentiment of tweets. At the heart of the proposed system, 2 main modules are employed here: : (a) the preprocessing - feature extraction module (b) the classification module.

2.5. Survey5:

Title: SENTIMENT ANALYSIS ON TWITTER

By: Akshi Kumar and Teeja Mary Sebastian

Abstract:

This project proposes a paradigm to mine the sentiment from a popular real-time microblogging service, Twitter, where users post real time reactions to and opinions about "everything". In this paper, we expound a hybrid approach using both corpus based and dictionary based methods to determine the semantic orientation of the opinion words in tweets.

Findings:

Here, first the opinion words (combination of the adjectives along with the verbs and adverbs) in the tweets are extracted and then their orientation is found out i.e., deciding whether each opinion word reflects a positive sentiment, negative sentiment or a neutral sentiment. The corpus-based method is then used to find the semantic orientation of adjectives and the dictionary-based method is employed to find the semantic orientation of verbs and adverbs. The overall tweet sentiment is then calculated using a linear equation which incorporates emotion intensifiers too. The proposed system contains the following sub-systems:

(a) Pre-processing of Tweets: the transaction file that contains opinion indicators, namely the adjective, adverb and verb along with emoticons is prepared and the tweets are processed for better efficiency.

(b) Semantic Score of Adjectives : This is domain specific and therefore corpus based approach is used to quantify the semantic orientation of adjectives in the Twitter domain.

(c) Tweet Sentiment Scoring: Here the adverbs, verbs and adjectives are grouped and the adjective strength is calculated. To calculate the overall sentiment of the tweet, average the strength of all opinion indicators like emoticons, exclamation marks, capitalization, word emphasis, adjective group and verb group.

3. THEORETICAL BACKGROUND

3.1. SENTIMENT ANALYSIS

Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text. It involves the use of data mining, machine learning (ML) and artificial intelligence (AI) to mine text for sentiment and subjective information. This is a popular way for organization's to determine and categorize opinions about a product, service or idea.

Sentiment analysis systems help organizations gather insights from unorganized and unstructured text that comes from online sources such as emails, blog posts, support tickets, web chats, social media channels, forums and comments. Algorithms replace manual data processing by implementing rule-based, automatic or hybrid methods. Rule-based systems perform sentiment analysis based on predefined, lexicon-based rules while automatic systems learn from data with machine learning techniques. A hybrid sentiment analysis combines both approaches.

In addition to identifying sentiment, opinion mining can extract the polarity (or the amount of positivity and negativity), subject and opinion holder within text. Furthermore, sentiment analysis can be applied to varying scopes such as document level, paragraph level, sentence level and sub-sentence level.

3.2 TYPES OF SENTIMENT ANALYSIS:

Fine-grained sentiment analysis provides a more precise level of polarity by breaking it down into further categories, usually very positive to very negative. This can be considered the opinion equivalent of ratings on a 5-star scale.

1. Emotion detection identifies specific emotions rather than positivity and negativity. Examples could include happiness, frustration, shock, anger and sadness.

2. Intent-based analysis recognizes actions behind a text in addition to opinion. For example, an online comment expressing frustration about changing a battery could prompt customer service to reach out to resolve that specific issue.

3. Aspect-based analysis gathers the specific component being positively or negatively mentioned. For example, a customer might leave a review on a product saying the battery life was too short. Then, the system will return that the negative sentiment is not about the product as a whole, but about the battery life.

3.3 TWITTER:

Twitter is a social networking and microblogging online service that allows users to send and receive text-based messages or posts of up to 140 characters called "tweets." After the online sign-up process, users can post their tweets by using a computer or other Twitter-compatible device such as a smartphone, and can view tweets posted by other "followed" users.

Twitter has been used as a platform for a wide variety of purposes in many scenarios by different industries. It is used as the means for direct communication among social groups and organizations, especially with the use of hashtags, which enable a tweet to be viewed by all users who follow a given topic that starts with the hash (#) symbol.

3.4 TWITTER SENTIMENT ANALYSIS:

Sentiment Analysis is a technique used in text mining. Twitter Sentiment Analysis may, therefore, be described as a text mining technique for analyzing the underlying sentiment of a text message, i.e., a tweet. Twitter sentiment or opinion expressed through it may be positive, negative or neutral. However, no algorithm can give you 100% accuracy or prediction on sentiment analysis.

3.5 TWITTER STREAMING API:

An API , is the instruction set created for developers to interact with some type of technology. Twitter has an open API that allows external developers to develop technology which rely on

Twitter's data. To get real-time tweets, Twitter provides the Twitter Streaming API. The Twitter streaming API is used to download twitter messages in real time. It is useful for obtaining a high volume of tweets, or for creating a live feed using a site stream or user stream. Twitter's Streaming API pushes data as tweets happen in near real-time.

3.6 TWEEPY:

Tweepy, the Python client for the official Twitter API supports accessing Twitter via Basic Authentication and the newer method, OAuth. Twitter has stopped accepting Basic Authentication so OAuth is now the only way to use the Twitter API.

Tweety gives access to the well documented Twitter API. Tweepy makes it possible to get an object and use any method that the official Twitter API offers. The main Model classes in the Twitter API are Tweets, Users, Entities, and Places. Access to each returns a JSON-formatted response and traversing through information is very easy in Python.

4. SYSTEM ANALYSIS

4.1. Existing System:

The existing system works only on the dataset which is constrained to a particular topic. The existing systems also do not determine the measure of impact the results determined can have on the particular field taken into consideration and it does not allow retrieval of data based on the query entered by the user i.e. it has constrained scope.

In simple words, it works on static data rather than dynamic data. Unsupervised algorithms like Vector Quantization, are used for data compression, pattern recognition, facial and speech recognition, etc and therefore cannot be used in determining sentiment in twitter data. Apriori algorithm fails to handle large datasets and as a result can generate faulty results.

4.2. Proposed System:

In the proposed system, we will retrieve tweets from twitter using twitter API based on the query. The collected tweets will be subjected to preprocessing. We will then apply the supervised algorithm on the stored data. The supervised algorithm used in our system is Naïve Baye's Classifier. The results of the algorithms i.e. the sentiment will be represented in graphical manner (pie charts/bar charts/word cloud). The proposed system is more effective than the existing one. This is because we will be able to know how the statistics determined from the representation of the result can have an impact in a particular field.

5. FEASIBILITY STUDY

5.1. Economical Feasibility:

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

5.2. Technical Feasibility:

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and Web Logic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

5.3. Social Feasibility:

The affect that a proposed project may have on the social system in the project environment is addressed in the social feasibility. It may happen that particular category of employees may be short or not available as a result of ambient social structure. The influence on the social status of the participants by the project should be evaluated in order to guarantee compatibility. It must be identified that employees in the particular industries may have specific status symbols within the society. Social feasibility is one of the feasibility study where the acceptance of the people is

considered regarding the product to be launched. It describes the effect on users from the introduction of the new system considering whether there will be a need for retraining the workforce. It describes how you propose to ensure user co-operation before changes are introduced.

6. SYSTEM REQUIREMENT

6.1. Hardware Requirements:

The Hardware consists of the physical components of the computer that input storage processing control, output devices. The kind of hardware used in the project is mentioned below:

Operating System : Windows 10/8/7

RAM SIZE : 4GB

PROCESSOR : Minimum Core i3

DISK SPACE : 3GB (download and install)

6.2. Software Requirements:

Software is a set of programs to do a particular task. Software is an essential requirement of computer systems. The kind of software used in the project is mentioned below:

PROGRAMING FUNCTIONALITY : Python

IDE : PyCharm

PACKAGES : NLTK, PIP, Twitter, Wordcloud, Tweepy, Pandas

7. SOFTWARE ENVIRONMENT

7.1. About Python:

Python is one of those rare languages which can claim to be both simple and powerful. Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions- **Python 2 and Python 3**. Both are quite different.

7.2. Features Of Python:

Below are the features that describe about PYTHON programming language:

1. Simple.
2. Easy to Learn.
3. Free and Open Source.
4. High-Level Language.
5. Portable.
6. Interpreted.
7. Object Oriented.
8. Extensible.
9. Embeddable.
10. Extensive Libraries.

Simple

Python is a simple and minimalistic language. Reading a good Python program feels almost like reading English, although very strict English! This pseudo-code nature of Python is one

of its greatest strengths. It allows you to concentrate on the solution to the problem rather than the language itself.

Easy to Learn

As you will see, Python is extremely easy to get started with. Python has an extraordinarily simple syntax, as already mentioned.

Free and Open Source

Python is an example of a FLOSS (Free/Libré and Open Source Software). In simple terms, you can freely distribute copies of this software, read its source code, make changes to it, and use pieces of it in new free programs. FLOSS is based on the concept of a community which shares knowledge. This is one of the reasons why Python is so good - it has been created and is constantly improved by a community who just want to see a better Python.

High-level Language

When you write programs in Python, you never need to bother about the low-level details such as managing the memory used by your program, etc.

Portable

Due to its open-source nature, Python has been ported to (i.e. changed to make it work on) many platforms. All your Python programs can work on any of these platforms without requiring any changes at all if you are careful enough to avoid any system-dependent features.

You can use Python on GNU/Linux, Windows, FreeBSD, Macintosh, Solaris, OS/2, Amiga, AROS, AS/400, BeOS, OS/390, z/OS, Palm OS, QNX, VMS, Psion, Acorn RISC OS, VxWorks, PlayStation, Sharp Zaurus, Windows CE and PocketPC!

You can even use a platform like Kivy to create games for your computer and for iPhone, iPad, and Android.

Interpreted

A program written in a compiled language like C or C++ is converted from the source language i.e. C or C++ into a language that is spoken by your computer (binary code i.e. 0s and 1s) using a compiler with various flags and options. When you run the program, the linker/loader software copies the program from hard disk to memory and starts running it.

Python, on the other hand, does not need compilation to binary. You just run the program directly from the source code. Internally, Python converts the source code into an intermediate form called byte codes and then translates this into the native language of your computer and then runs it. All this, actually, makes using Python much easier since you don't have to worry about compiling the program, making sure that the proper libraries are linked and loaded, etc. This also makes your Python programs much more portable, since you can just copy your Python program onto another computer and it just works!

Object Oriented

Python supports procedure-oriented programming as well as object-oriented programming. In procedure-oriented languages, the program is built around procedures or functions which are nothing but reusable pieces of programs. In object-oriented languages, the program is built around objects which combine data and functionality. Python has a very powerful but simplistic way of doing OOP, especially when compared to big languages like C++ or Java.

Extensible

If you need a critical piece of code to run very fast or want to have some piece of algorithm not to be open, you can code that part of your program in C or C++ and then use it from your Python program.

Embeddable

You can embed Python within your C/C++ programs to give *scripting* capabilities for your program's users.

Extensive Libraries

The Python Standard Library is huge indeed. It can help you do various things involving regular expressions, documentation generation, unit testing, threading, databases, web browsers, CGI, FTP, email, XML, XML-RPC, HTML, WAV files, cryptography, GUI (graphical user interfaces), and other system-dependent stuff. Remember, all this is always available wherever Python is installed. This is called the Batteries Included philosophy of Python.

Besides the standard library, there are various other high-quality libraries which you can find at the Python Package Index.

7.3. What You Can Do With Python

From web development to data science, machine learning, and more, Python's real-world applications are limitless. Here are some projects that will assist you in finally putting your Python skills to good use. Below are few places where Python is used:

1. Automate this boring stuff.
2. Stay on Top of Bitcoin Prices.
3. Create a Calculator.
4. Mine Twitter Data.
5. Build a Microblog With Flask.
6. Build a Blockchain.
7. Bottle Up a Twitter Feed.
8. Play PyGames.
9. Choose Your Own Adventure.
10. Say "Hello World!" to Machine Learning.
11. Get Challenged.

8. SYSTEM DESIGN

8.1. Data Flow Diagram:

A data flow design is a graphical representation of the "flow" of data through an information system, modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system without going into great detail, which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design).

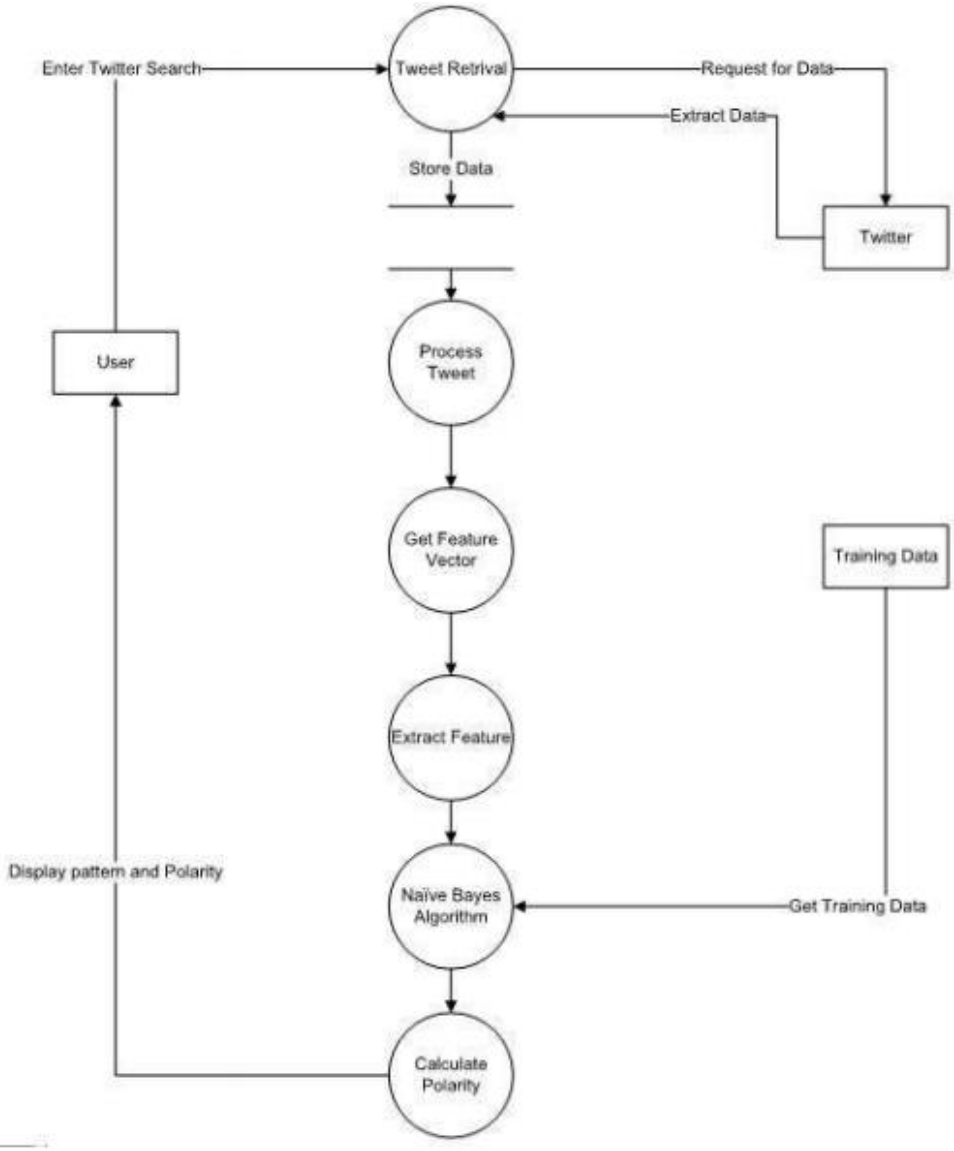

Figure 8.1 data Flow diagram

8.2. Flow Chart Diagram:

A flowchart is a diagram that depicts a process, system or computer algorithm. They are widely used in multiple fields to document, study, plan, improve and communicate often complex processes in clear, easy-to-understand diagrams. Flowcharts, sometimes spelled as flow charts, use rectangles, ovals, diamonds and potentially numerous other shapes to define the type of step, along with connecting arrows to define flow and sequence. They can range from simple, hand-drawn charts to comprehensive computer-drawn diagrams depicting multiple steps and routes.

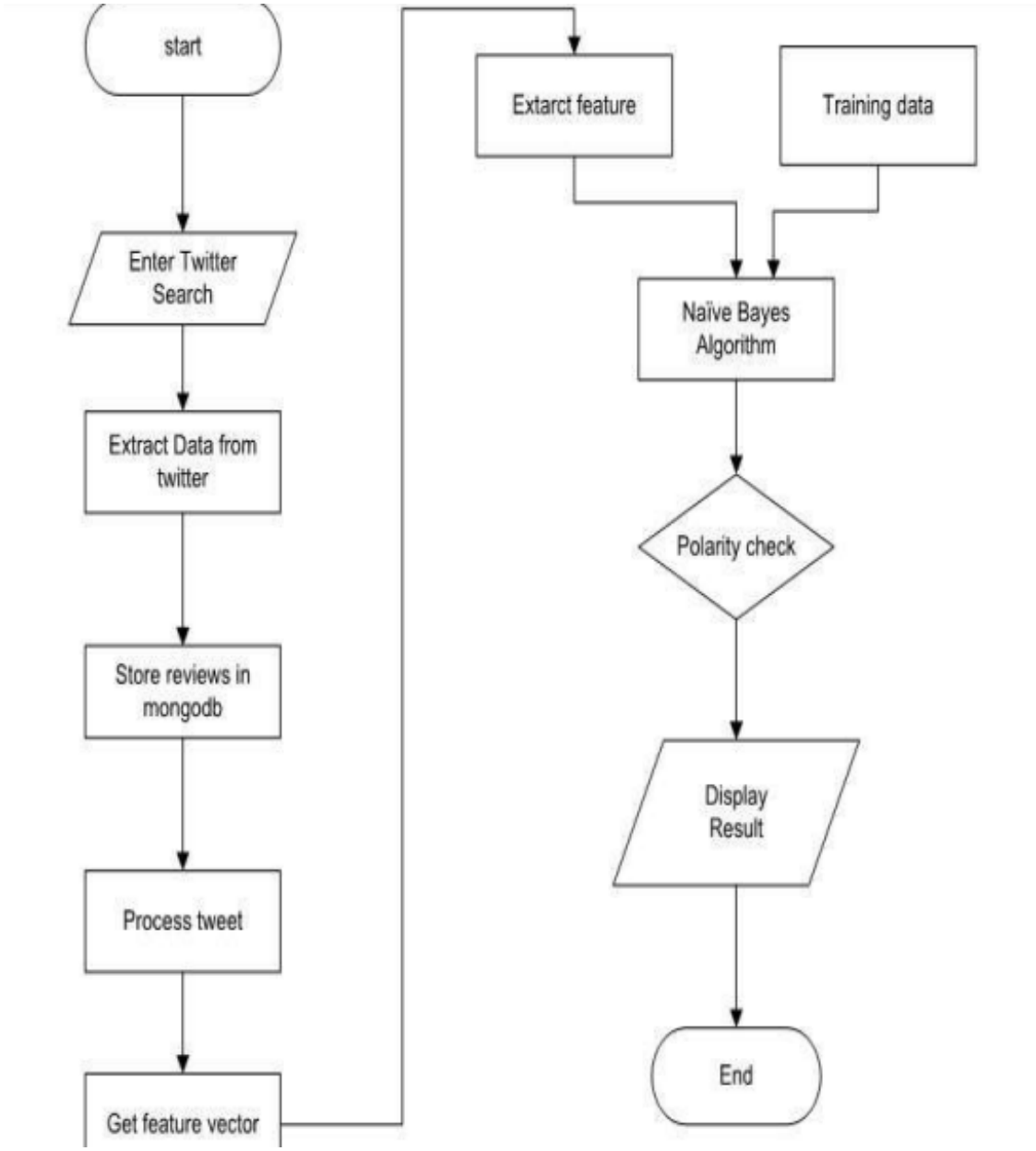

Figure 8.2 Flow Chart diagram

8.2. UML Diagrams:

UML is a way of visualizing a software program using a collection of diagrams. The notation has evolved from the work of Grady Booch, James Rumbaugh, Ivar Jacobson, and the Rational Software Corporation to be used for object-oriented design, but it has since been extended to cover a wider variety of software engineering projects. Today, UML is accepted by the Object Management Group (OMG) as the standard for modeling software development.

8.2.1. Use Case Diagram:

Use Case diagram is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well. The use cases are represented by either circles or ellipses. A use-case diagram can help provide a higher-level view of the system. It has been said before that "Use case diagrams are the blueprints for your system". They provide the simplified and graphical representation of what the system must actually do.

Use case diagrams are usually referred to as behaviour diagrams used to describe a set of actions (use cases) that some system or systems should or can perform in collaboration with one or more external users of the system (actors). Each use case should provide some observable and valuable result to the actors or other stakeholders of the system.

Use case diagrams are in fact two fold - they are both behaviour diagrams,because they describe behavior of the system, and they are also structure diagrams- as a special case of class diagrams where classifiers are restricted to be either actors or use cases related to each other with associations.

NOTATIONS:

USE CASE - Draw use cases using ovals. Label the ovals with verbs that represent the system's functions. It is a list of steps, typically defining interactions between an actor and a system, to achieve a goal. Use case is used to capture high level functionalities of a system.

Figure 8.3 Use case notation

ACTOR - An actor in a UML diagram represents a type of role where it interacts with the system and its objects. It is important to note here that an actor is always outside the scope of the system we aim to model using the UML diagram.

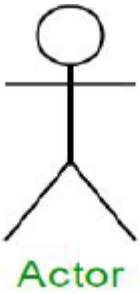

Figure 8.4 Actor notation

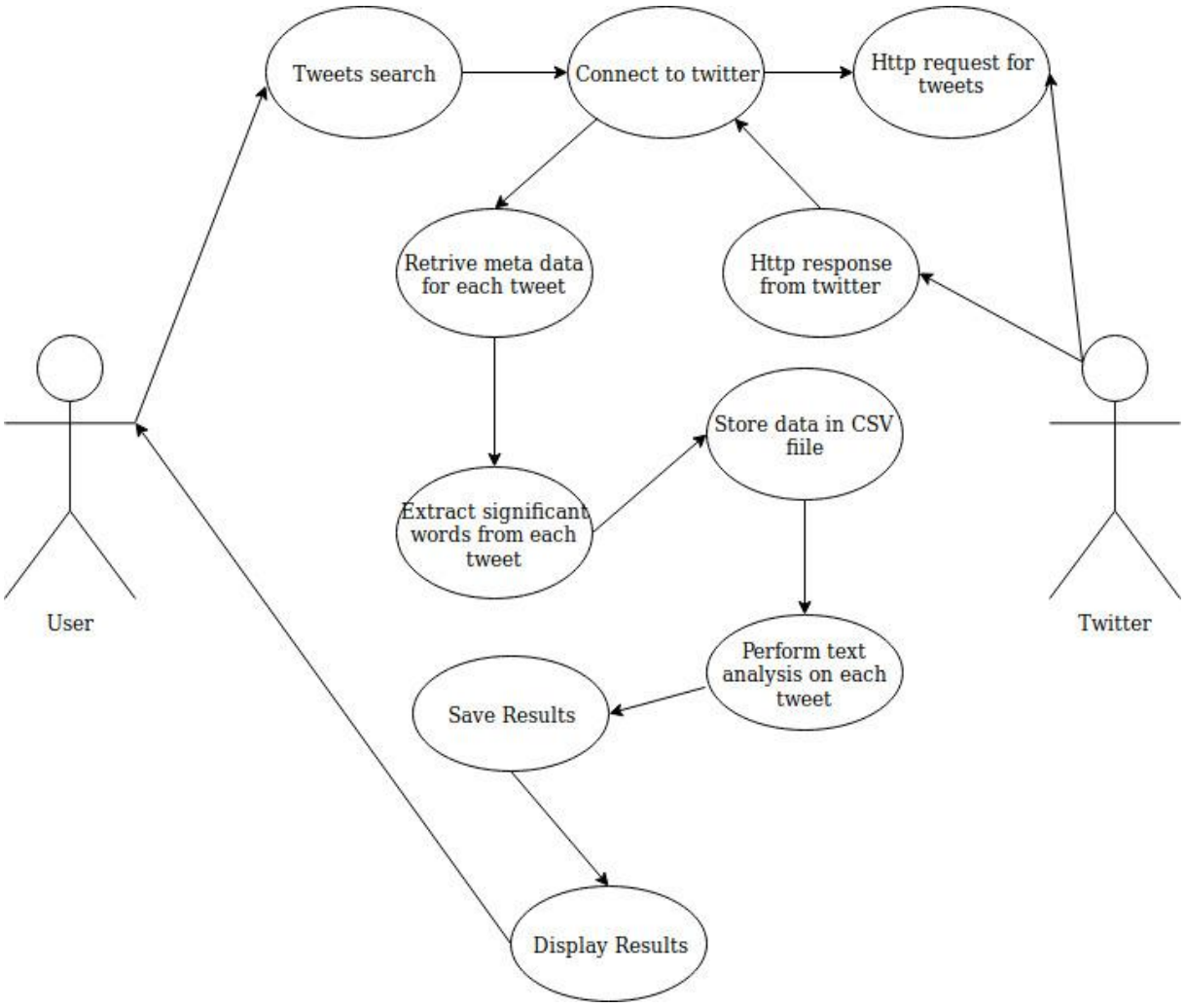

Figure 8.5 Use Case Diagram

8.2.2. Sequence Diagram:

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the

objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios.

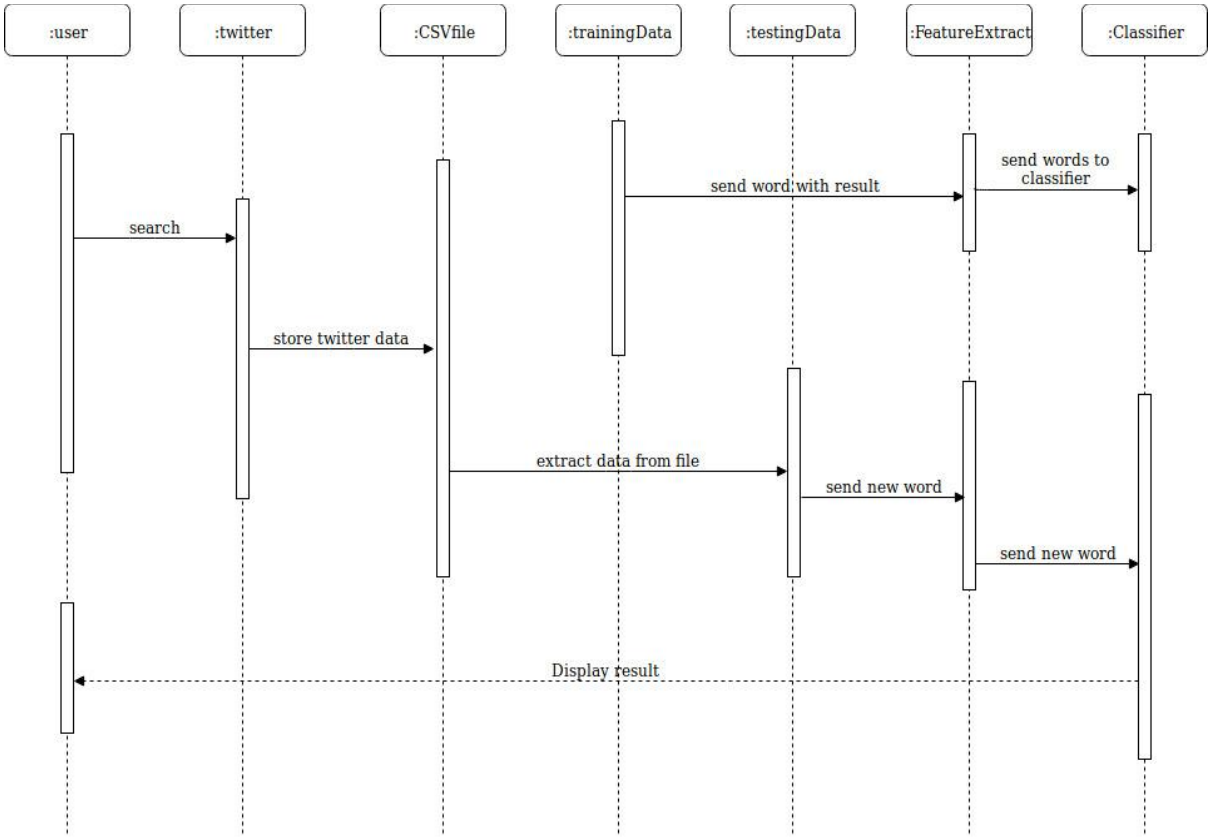

Figure 8.6 Sequence Diagram

8.2.2. Activity Diagram:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the UML, activity diagrams are intended to model both computational and organizational processes (i.e., workflows), as well as the data flows intersecting with the related activities. Although activity diagrams primarily show the overall flow of control, they can also include elements showing the flow of data between activities through one or more data stores.

NOTATIONS:

INITIAL STATE – The starting state before an activity takes place is depicted using the initial state.

Figure 8.7 Initial state notation

ACTION/ACTIVITY STATE – An activity represents execution of an action on objects or by objects. We represent an activity using a rectangle with rounded corners. Basically any action or event that takes place is represented using an activity.

Figure 8.8 Activity state notation

ACTION FLOW/CONTROL FLOWS – Action flows or Control flows are also referred to as paths and edges. They are used to show the transition from one activity state to another.

Figure 8.9 Control flow notation

DECISION NODE & BRANCHING – When we need to make a decision before deciding the flow of control, we use the decision node.

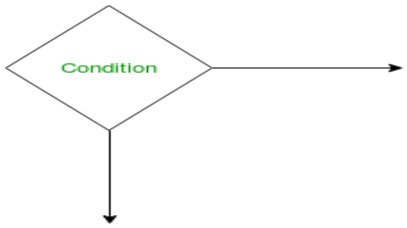

Figure 8.10 Branch notation

FORK – Fork nodes are used to support concurrent activities.

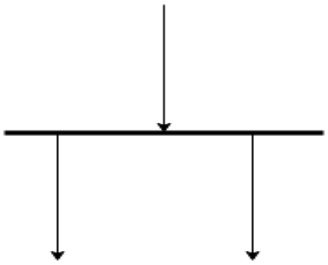

Figure 8.11 Fork notation

JOIN – Join nodes are used to support concurrent activities converging into one. For join notations we have two or more incoming edges and one outgoing edge.

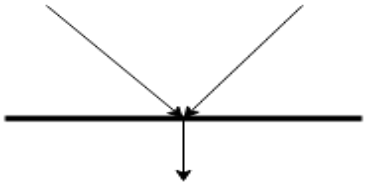

Figure 8.12 Join notation

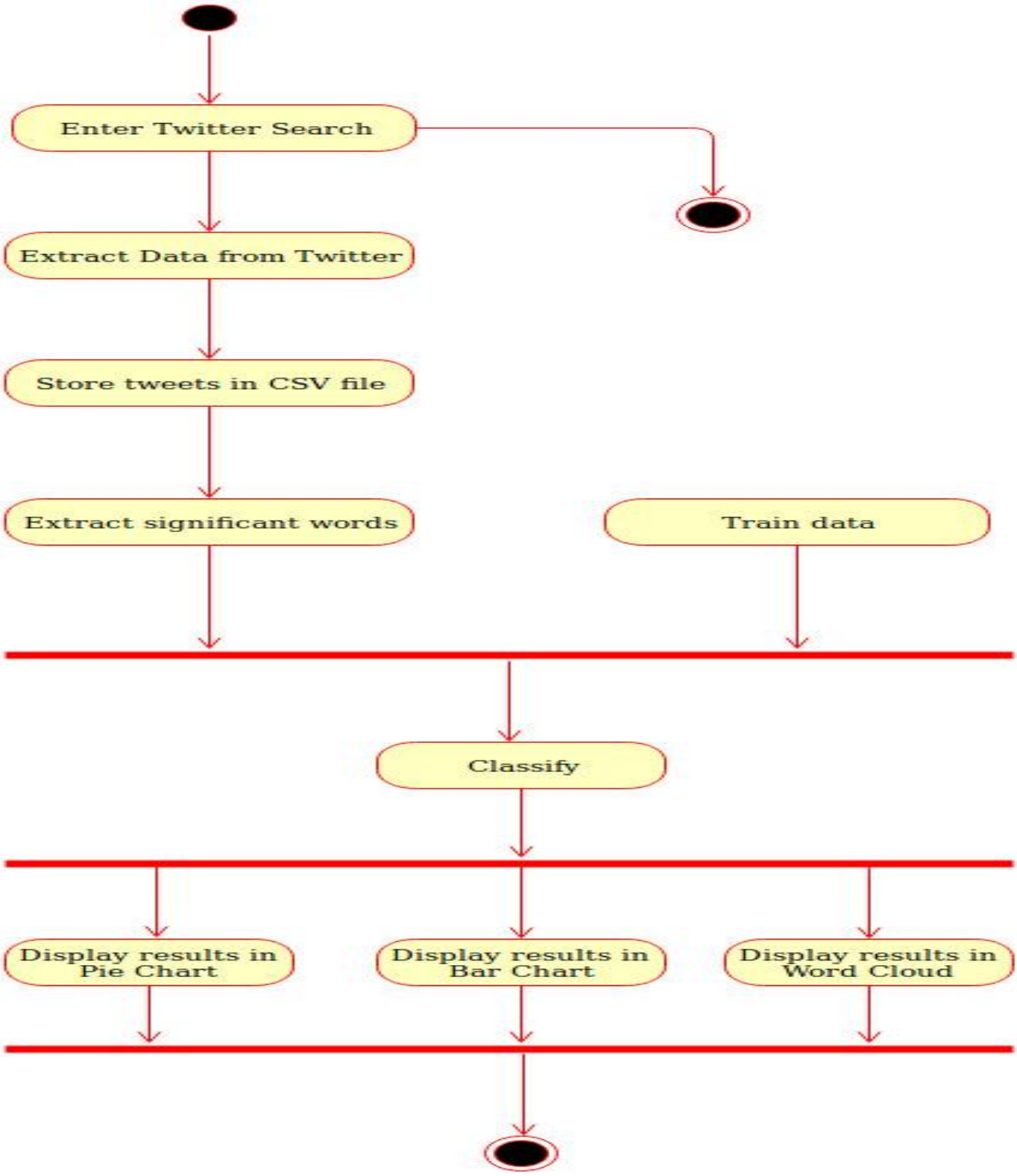

Figure 8.13 Activity Diagram

9. IMPLEMENTATION

9.1. CREATION OF DATA-SET:

- A data-set is created using frequently used words in twitter posts.
- The below tables shows data-set used for training the classifiers.
- A data-set is created by taking 430 words.

Data Set	Emoticons	Words
Positive	36	213
Negative	21	145
Neutral	1	14
Total	58	372

Table 9.1 Data set size

9.1.1 POSITIVE TOKENS:

Polarity	Emoji Name	Tokens
10	red_heart, yellow_heart, blue_heart, purple_heart, green_heart, sparkling_heart, heart_with_arrow	First-rate, legendary, powerful, must-see, award-winning, spectacular, highlet , outstanding, ovation, triumph, fabulous, oscar astounding, love, lovely, majestic, mind blowing, gorgeous
9	smiling_face_with_heart-eyes, face_blowing_a_kiss, kissing_face_with_closed_eyes, kissing_face, star-struck	Enjoyable, original, pleasant, magical, well-paced, blockbuster, awesome, appreciate, successful, amazing, brilliant, wonderful, impressive, nailed, pride, adorable, appealing, awestruck, charming
8	Laughing, satisfied, joy, smiling_face_with_smiling_eyes	Surprising, hilarious, recommended, superb, wow, excellent, superhit, congratulate, handsome, admire, admiring, deserved, delighted,

		magic, super, great, respect, luv, funniest, cool, enthusiast, professional, aspirations, astounding, astonished
7	grinning_face_with_smiling_eyes, grinning_face_with_big_eyes, grinning_face, beaming_face_with_smiling_eyes	Comical, fascinating, remarkable, family-friendly, fantastic, family-entertainer, win, victory, emphatic, flawless, splendid, Mashallah, excite, excited, happiest, Wow, enjoy, entertain, cheer, sunshine, accomplish, achievable, achievement, crazy, booming, brilliant, sensational, trendy
6	ok_hand, thumbs_up, victory_hand, clapping_hands	Thrilling, adventurous, interesting, super, perfect, favorite, classic, likely, influential, mighty, amazing, inspire, laugh, award, lucky, yay, reward, fun, benefit, energetic, funny, smart, biggie, affectionate, all-rounder, appreciative, assurance, fame, hit
5	slightly_smiling_face, grinning_face_with_sweat, smiling_face_with_halo, face_savoring_food	Sensitive, tender, charming, sweet, best, amazing, convincing, strong, celebration, help, prom, bless, cutest, happy, yummy, pretty, honest, pleasure, sweet, affinity, unbeatable, breathtaking, prestigious
4	winking_face, smiling_face_with_sunglasses	Narrative, playful, potent, thought-provoking, good, strength, well, nice, cheer, success, trust, relieve, proud, harmony, improve, satisfaction, encourage, attentive, attractive, darling, dashing, dazzling
3	winking_face_with_tongue, squinting_face_with_tongue, face_with_tongue	Imaginative, uproarious, intensive, solid, sufficient, encouraging, sublime, positive, champ, impress, interest, devote, strong, glad, attentive, spacious, affordable, creative, originality, world-famous

2	relieved_face	Cinematic, mysterious, effective, mystical, beautiful, peace, graceful, hope, heroic, attract, grand, fair, heal, heartfelt, bliss, joy, merry, fantasy, caring, energetic, impressive
1	smiling_face	Worthwhile, graphic, meaningful, encourage, harmony, play, freedom, everlasting

Table 9.2 Positive data set

9.1.2 NEGATIVE TOKENS:

Polarity	**Emoji Name**	**Tokens**
-10	Pouting_face, face_with_steam_from_nose, angry_face	Third-rate, dazzling, violent, Terrific, Utter-flop, drastic, awkward
-9	loudly_crying_face	Second-rate, confused, trite, flop, beseech, hate, hated
-8	oncoming_fist, thumbs_down	Flawed, disgusting, senseless, silly, moronic, abrupt, awful, backbiting, bitter, crap, crappy, hurt
-7	fearful_face, anxious_face_with_sweat, disappointed_face, worried_face	Brutal, ponderous, heavy-footed, matchless, adulterated, bane, conflict, defamation, defamations, defamatory, defame, defect, idiotic
-6	sad_but_relieved_face, persevering_face, confused	Controversial, sad, dominate, flatten, adverse, pathetic, burden, burdensome, contradict, contradiction, contradictory, unappealing, unattractive

	frowning_face_with_open_mouth, confused_face	Intuitive, instinctive, unthinking, abrasive, allegation, ashamed, arrogance, boggle, controversial, controversy, degrade, degrading, degradingly, disgusted, disgustful, disgusting, gross, shameful
-4	dizzy_face, face_screaming_in_fear, flushed_face	Uninteresting, outdated, worst, absurdly, absurd, antagonist, argumentative, bad, badly, depress Depressed, depressing, depressingly, depression, disaster, disastrous, flaw, hopeless
-3	expressionless_face	Offbeat, graceless, loss, accuse, annoy, angry, annoyance, bored, disappoint, disappointed, disappointment, dishearten, disheartening, embarrass, embarrassing, embarrassingly, embarrassment, falsehood, foul
-2	sad_but_relieved_face	Stupid, boring, tiresome, abnormal, aggressive, ambivalence, broken, complex, deny, disgrace, disgraceful, dissatisfaction, dissatisfactory, dissatisfied, drawback, helpless, waste, wasteful

-1	neutral_face	Low-budget, ordinary, abusive, afraid, assault, cheap, clumsy, deceiver, destruction, dirty, dislike, disliked, disturbing, disturbance, downgrade, downhearted, error, frustrated, frustrate, unsuccessful

Table9.3 Negative data set

9.1.3 NEUTRAL TOKENS:

Polarity	Emoji Name	Tokens
0	zipper_mouth_face	Hero, heroine , actor, actress, cast, crew, singer, songs, music, director, producer, war, theme, release

Table 9.4 Neutral data set

9.2. NATURAL LANGUAGE TOOLKIT:

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. It was developed by Steven Bird and Edward Loper in the Department of Computer and Information Science at the University of Pennsylvania. NLTK includes graphical demonstrations and sample data.

Natural language processing (NLP) is the automatic or semi-automatic processing of human language. NLP is closely related to linguistics and has links to research in cognitive science, psychology, physiology, and mathematics. In the computer science domain in particular, NLP is related to compiler techniques, formal language theory, human-computer interaction, machine learning, and theorem proving.

9.3. BAG OF WORDS:

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model has also been used for computer vision.

The bag-of-words model is commonly used in methods where the (frequency of) occurrence of each word is used as a feature for training a classifier.

Example:

Here are two simple text documents:

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

Based on these two text documents, a list constructed as follows for each document:

"John","likes","to","watch","movies","Mary","likes","movies","too"

"John","also","likes","to","watch","football","games"

Bag-of-words for each text document:

BoW1 = {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};

BoW2 = {"John":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};

Each key is the word, and each value is the number of occurrences of that word in the given text document.

9.4. TOKENIZING DATA:

Tokenization is a way to split text into tokens. These tokens could be paragraphs, sentences, or individual words. NLTK provides a number of tokenizers in the tokenize module.

9.4.1 TOKENIZATION OF WORDS:

We use the method word_tokenize() to split a sentence into words. The output of word tokenization can be converted to Data Frame for better text understanding in machine learning applications. Word tokenization becomes a crucial part of the text (string) to numeric data conversion.

Example:

Input to tokenizer:

"All work and no play makes jack a dull boy, all work and no play"

Output after word tokenizing:

['All', 'work', 'and', 'no', 'play', 'makes', 'jack', 'dull', 'boy', ',', 'all', 'work', 'and', 'no', 'play']

9.4.2 TOKENIZATION OF SENTENCES :

We use the method sent_tokenize() to split a text into sentences.

Example:

Input to tokenizer:

"All work and no play makes jack dull boy. All work and no play makes jack a dull boy."

Output after sentence tokenizing:

['All work and no play makes jack dull boy.', 'All work and no play makes jack a dull boy.']

9.5. PRE-PROCESSING DATA:

The idea of Natural Language Processing is to do some form of analysis, or processing, where the machine can understand, at least to some level, what the text means, says, or implies. The process of converting data to something a computer can understand is referred to as "pre-processing." One of the major forms of pre-processing is going to be filtering out useless data.

9.5.1 REMOVE NUMBERS:

Numbers are not relevant to sentiment analysis. Regular expressions are used to remove numbers from text.

Example:

Input:

“Box A contains 3 red and 5 white balls, while Box B contains 4 red and 2 blue balls.”

Output:

Box A contains red and white balls, while Box B contains red and blue balls.

9.5.2 CONVERT EMOTICONS TO TEXT:

Emoticons are introduced as expressive, non-verbal components into the written language, mirroring the role played by facial expressions in speech. Their role is mainly pragmatic, emoticons give a positive or negative sense to written sentences by a visual expression. According to this consideration, there is a relationship between the sentiment orientation of emoticons and messages. Emoticons have been distinguished in two main categories, i.e. positive and negative. Emoticons are converted to sentimental text.

9.5.3 REMOVE SYMBOLS, PUNCTUATIONS AND NON-ENGLISH WORDS

Symbols, punctuations and non-english words are not relevant to sentiment analysis. All these are removed.

9.5.4 STEMMING WORDS:

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

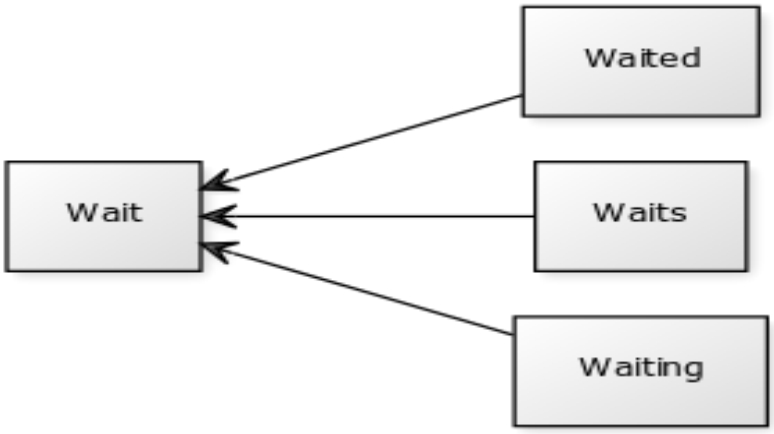

Figure9.1 Stemming word example

9.5.5 REMOVE STOP WORDS:

We can recognize ourselves that some words carry more meaning than other words. We can also see that some words are just plain useless, and are filler words. We use them in the English language, for example, to sort of "fluff" up the sentence so it is not so strange sounding. An example of one of the most common, unofficial, useless words is the phrase "umm." We would not want these words taking up space in our database, or taking up valuable processing time. As such, we call these words "stop words" because they are useless, and we wish to do nothing with them. There is no universal list of stop words in nlp research. Stop words can be filtered from the text to be processed.

9.6 CLASSIFICATION:

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. A data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, biometric identification, document classification etc.

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

9.6.1 NAVIE BAYES CLASSIFIER:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability.

9.6.2 SUPPORT VECTOR MACHINES:

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

9.6.3 DECISION TREES:

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

9.6.4 RANDOM FOREST:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

9.6.5 NEURAL NETWORKS:

A neural network consists of units (neurons), arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. Generally the networks are defined to be feed-forward: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase to adapt a neural network to the particular problem at hand.

9.6.6 NEAREST NEIGHBOUR

The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labelled points and uses them to learn how to label other points. To label a new point, it looks at the labelled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the "k" is the number of neighbors it checks).

REASONS TO CHOOSE NAIVE BAYES CLASSIFIER:

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.The main advantage of Naive Bayes classifier is that it analyses each feature independently. So it makes the use of all the features in feature vector. The Probability of Naïve Bayesian classifier is given as,

$$P\left(\frac{z}{b_j}\right) = \prod_{i=1}^{m} P(z_i/b_j)$$

Where the feature vector is represented by z and b is the class label (i.e. positive, negative, neutral). Another reason of using Naïve Bayesian classifier is that it is simple to use and can be scalable. This classifier as compared to all other classifiers:

1. Has high precision
2. Need less training data
3. Highly scalable.
4. It scales linearly with the number of predictors and data points.
5. Can be used for both binary and multi-class classification problems.
6. Handles continuous and discrete data.
7. Return's not only the prediction but also the degree of certainty, which can be very useful.
8. Easily updatable if new training data is received

10. SAMPLE CODING

10.1. TRAINING:

```
from nltk.classify import NaiveBayesClassifier
import pandas as pd
import glob

def word_feats(words):
   return dict([(word, True) for word in words])

def get_data_set(train_file, data_set):
   data_frame = pd.read_csv(train_file)
   for data in data_frame.itertuples():
      data_set = data_set + [(word_feats(data[1]), data[2])]
   return data_set

def classify_feature_set():
   path = r'./trainSets'
   filenames = glob.glob(path + "/*.csv")
   data_set = []
   for file in filenames:
      data_set = get_data_set(file, data_set)
   return data_set

def train():
   data_set = classify_feature_set()
   classifier = NaiveBayesClassifier.train(data_set)
   return classifier
```

10.2. STREAMING TWEETS:

```
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
import time
import csv
import sys

class StdOutListener(StreamListener):

    def __init__(self, api=None):
        self.api = api
        self.filename = 'dataSets/'+'data' + '_' + time.strftime('%Y%m%d-%H%M%S') +      '.csv'
        csv_file = open(self.filename, 'w')
        csv_writer = csv.writer(csv_file)
    def on_status(self, status):
        csv_file = open(self.filename, 'a')
        csv_writer = csv.writer(csv_file)

        if not 'RT @' in status.text:
            try:
                if status.is_quote_status:
                    text = [status.extended_tweet['full_text']]
                else:
                    text = [status.text]
                csv_writer.writerow(text)
            except Exception:
                pass

        csv_file.close()
```

```
        return
    def on_error(self, status_code):
        print('Encountered error with status code:', status_code)
        if status_code == 401:
            return False
    def on_delete(self, status_id, user_id):
        print("Delete notice")
        return
    def on_limit(self, track):
        print("Rate limited, continuing")
        return True
    def on_timeout(self):
        print(sys.stderr, 'Timeout...')
        time.sleep(10)
        return

def start_mining(queries):

    consumer_key = " "
    consumer_secret = " "
    access_token = " "
    access_token_secret = " "

    listener = StdOutListener()

    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, listener, tweet_mode='extended')
    stream.filter(track=queries)
    return listener.filename
```

```
def stream_tweets(hash_tag):
    tweets_file = unicode(start_mining(hash_tag).strip(codecs.BOM_UTF8), 'utf-8')
    return tweets_file

stream_tweets(["#IPL2019"])
```

10.3. CLASSIFICATION:

```
from nltk.tokenize import RegexpTokenizer
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import re
import emoji
import pandas as pd
import glob

tokenizer = RegexpTokenizer(r'\w+')
stopWords = set(stopwords.words('english'))
wnl = WordNetLemmatizer()


def word_feats(words):
    return dict([(word, True) for word in words])


def classify_feature_set():
    path = 'C:/Users/pravllika/Downloads/twitter/twitter/twitter/trainSets'
    filenames = glob.glob(path + "/*.csv")
    df = []
```

```
    for file in filenames:
        data_frame = pd.read_csv(file, index_col=None, header=0)
        df.append(data_frame)
        print(file)
    frame = pd.concat(df, axis=0, ignore_index=True)
    return frame


def get(word, data_frame):
    data = data_frame[data_frame['token'] == word]
    p = data['polarity'].tolist()
    if p:
        polarity = p[0]
        return polarity
    return 99

def data(classifier, tweets_file, frame):
    count = {"positive_tweets": 0, "negative_tweets": 0, "neutral_tweets": 0, "total_tweets": 0}
    positive_threshold = 2
    negative_threshold = -2
    all_words = ''
    data_frame = pd.read_csv(tweets_file, low_memory=False)
    count["total_tweets"] = len(data_frame)
    word_count = {-10: 0, -9: 0, -8: 0, -7: 0, -6: 0, -5: 0, -4: 0, -3: 0, -2: 0, -1: 0, 0: 0, 1: 0, 2: 0, 3:
0, 4: 0, 5: 0, 6: 0, 7: 0, 8: 0, 9: 0, 10: 0}
    for data in data_frame.itertuples():
        data = emoji.demojize(unicode(data[1], "utf-8"), delimiters=(" ", " "))
        data = re.sub(r'\d+', '', data)
        data = data.lower()
        words = tokenizer.tokenize(data)
        words_filtered = []
```

```
    for word in words:
        if word not in stopWords:
            word = wnl.lemmatize(word)
            words_filtered.append(word)
            all_words = all_words + word + ' '
            neg = 0
            pos = 0

    for word in words_filtered:
        class_result = get(word, frame)
        if class_result == 99:
            continue
            class_result = classifier.classify(word_feats(word))
        if class_result < 0:
            neg = neg + class_result
        if class_result > 0:
            pos = pos + class_result
        word_count[class_result] += 1
    polarity = float(pos+neg)/len(words_filtered)

    if polarity >= positive_threshold:
        count["positive_tweets"] += 1
    elif polarity <= negative_threshold:
        count["negative_tweets"] += 1
    else:
        count["neutral_tweets"] += 1
return count, all_words, word_count
```

10.4. VIRTUALIZATION:

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt
import numpy as np
import collections
activities = ['Positive Tweets', 'Negative Tweets', 'Neutral Tweets']


def bar_chart(sentiment_count):
    sentiment_count = collections.OrderedDict(sorted(sentiment_count.items()))
    activity = sentiment_count.keys()
    count = sentiment_count.values()
    y_pos = np.arange(len(activity))
    plt.barh(y_pos, count, color="red", align='center', alpha=0.5)
    plt.yticks(y_pos, activity)
    plt.xlabel('NUMBER OF WORDS')
    plt.ylabel('POLARITY OF WORD')
    plt.show()

def pie_chart(sentiment_count):
    del sentiment_count["total_tweets"]
    plt.pie(sentiment_count.values(), labels=sentiment_count.keys(), startangle=90,
autopct='%.1f%%')
    plt.show()

def word_cloud(words):
    cloud = WordCloud().generate(words)
    plt.imshow(cloud, interpolation='bilinear')
    plt.axis("off")
    plt.show()
```

```
def print_count(sentiment_count):
    print('Number Of Positive Tweets: ' + str(sentiment_count["positive_tweets"]))
    print('Number Of Negative Tweets: ' + str(sentiment_count["negative_tweets"]))
    print('Number Of Neutral Tweets: ' + str(sentiment_count["neutral_tweets"]))
```

11. TEST CASES

11.1. TEST CASE 1

TEST CASE FOR GETTING STREAMING TWEETS FROM TWITTER:

INPUT: ["#InternationalWomensDay", "#HappyWomensDay2019", "#WomensDay"]

OUTPUT:

#InternationalWomensDay https://t.co/Zcy2oLYty2

55% of our employees are women. Thanks for being part of this team. @panidor #InternationalWomensDay https://t.co/xSKGvqH6e8

@acadgild Ready #AcadgildWDQuiz #ContestAlert #InternationalWomensDay #Acadgild

#HappyWomensDay2019 #HappyInternationalWomensDay #InternationalWomensDay #WomensDay from #Kashmir. Down with India! https://t.co/2ZluqoeL91

″On this #InternationalWomensDay I would like to celebrate three things I adore, #Women #Wine and #Equality″

″happy #internationalwomensday to the most inspirational woman i know...

me.″

In honor of #InternationalWomensDay https://t.co/3eoO2Ochp0

@sarahjudge90 @LucyMPowell #Mcrbudget19 @ManCityCouncil #InternationalWomensDay #strongmcrwomen https://t.co/nV4vlgZTki

@UofGEngineering partner @KNTnano - great work! Together we are developing future talent such as @vikinginscience (pictured) who is on an @EPSRC @CDT_PIADS programme and is set for a long career in engineering #BalanceforBetter

#HappyWomensDay2019 https://t.co/VufV5xcDw7

″Could I ask when is #internationalblokesday

#InternationalWomensDay″

#HappyWomensDay2019 https://t.co/MYnUhUJ9HA

Happy #InternationalWomensDay to all the wonderful ladies blazing a trail #IWD2019 https://t.co/viQl2b5Gvu

Happy #WomensDay to all the wonderful women I know and am privileged to share moments with.

"PPP Senator Ma'am @SassuiPalijo speaking on Women's Day
#Senate.
#Womensda
2/3 https://t.co/ECGewWkBGJ"
Congratulations @AntoniaRomeoUK and all of the 100 women recognized with the Freedom of the City of London for #InternationalWomensDay
"Wishing you all a happy #InternationalWomensDay! Except for you, @theresa_may. You can f*ck right off. #IWD19"
"@SPIEGELONLINE @tagesschau @zeitonline @SZ @sternde @welt
@ZDF @RTLde @sat1 @tazgezwitscher @focusonline @DasErste
@DerSPIEGEL @tagesspiegel @WDR @euronews @DIEZEIT @cicero_online
@rponline @dwnews @SPIEGEL_Politik @derfreitag @3sat @dw_turkce"
"https://t.co/8KeGkrIFX8
#WomensDay
@Gurmeetramrahim ji
@derasachasauda″
@GayleLetherby #Auto/Biography #IWD19UoPFoMD #InternationalWomensDay #feminist #sociological #research #journey https://t.co/LnJUSx2kuq
#HappyWomensDay2019 https://t.co/mTLk7I93wh
@yazzcardPH Happy #InternationalWomensDay mga Yazzkada
″You don't hate men, you hate capitalism. #InternationalWomensDay″
#InternationalWomensDay https://t.co/QO8cHEZZIP
#InternationalWomensDay https://t.co/K0D0iVIWbL
″#HappyWomensDay2019
#womaninislam https://t.co/BMGjxL5hgO″
″#BackTo60 #50sWomen #InternationalWomensDay
#HappyWomensDay2019
https://t.co/7raq8YfTRf
Please RT and donate to support #50sWomen with no pension at the #JudicialReview .
Most grateful X″
#InternationalWomensDay https://t.co/0XOz7LhM6b

#TombRaider #HappyWomensDay2019 https://t.co/KfDKw2dYk9

We are at #Norbury Library celebrating #InternationalWomensDay! Come join in the festivities! #Croydon https://t.co/PBT6fw9L7Y

11.2. TEST CASE 2

TEST CASE FOR GETTING WORDS WITH POLARITY:

INPUT: Tweets with #118Movie

Atlast #kalyanram got a successful film after #pataas hope he will do more films like this in future. Exam season aina bagane collections vachay.minimum content unna audience encourage chestaru. #118Film SUPER HIT status.

"BTW, amazing work by whole team, special credits to #Kalyanram for his hard work and Dedication."

#118movie first half good.....classy ..#kalyanram best looks till date...

"#KalyanRam #118Movie collected around 2crores share worldwide on day 1, gets good reviews and 7 crores break even, what can be closing #NZvBAN #LetsConclave19 #F2 #Viswasam #ViswasamTelugu #Viswasam50thdayCelebration #SaahoSunday #Maharshi"

118 Movie Success Celebrations!! #KalyanRam #NivethaThomas

"What a movie #118 Tnx 4 making my day with wonderful suspense thriller, best performance - #KalyanRam @i_nivethathomas both stole the show"

"#118Movie #Nivethathomas #KalyanRam loved it anna, our Telugu industry going to the next level.#Nivethathomas what acting when prabha hospital scene. Two ultimate scenes one is interval bang and another church father shot. #KVGuhan your first movie is like you have lot of exp"

@NANDAMURIKALYAN & director @DirectorSriwass film on cards..!!! #KalyanRam

#KalyanRam's #118Movie 1st Week WW Collections... Day 1 - 2 cr+ Day 2 - 1.5 cr+ Day 3 - 2.2 cr+ Day 4 - 1.9 cr+ Day 5 - 1.1cr+ Day 6 - 0.7cr+ Day 7 - 0.55cr+ 1st Week WW Share = 10.5 cr+ 1st Week WW Gross = 16.2 cr+(aprx)

Lovely working stills from #118Movie. #118Film #118ThrillingBlockbuster @NANDAMURIKALYAN #KVGuhan @shalinipandeyyy @i_nivethathomas @smkoneru @EastCoastPrdns #KalyanRam #NKR #NTR #JrNTR

"Wishing the leading ladies of #118Movie - @i_nivethathomas (Adhya) and @shalinipandeyyy (Megha) , and all the women who make this world a better place, a very #HappyInternationalWomensDay #HappyWomensDay #HappyWomensDay2019 #KalyanRam #NKR #NTR #JrNTR #118ThrillingBlockbuster"

#118Movie Entered into 2nd Week At Box-office Running Successfully in All Areas #118ThrillingBlockbuster @NANDAMURIKALYAN @i_nivethathomas @shalinipandeyyy #KVGuhan @smkoneru @EastCoastPrdns @vamsikaka #118Film #KalyanRam #NKR #NTR #JrNTR

#118Movie is all set to have a successful 2nd week Too. Breakeven Don With in Just 4 Days #118ThrillingBlockbuster #KalyanRam #JrNTR

#118Movie is all set to have a successful 2nd week Too. Breakeven Don With in Just 4 Days #118ThrillingBlockbuster #KalyanRam #NKR #NTR #JrNTR

#118ThrillingBlockbuster ready to enter into 2nd week. #kalyanram #118Movie #118Film

"She is smiling at one point, serious as well || #Nivetha's funny cut on #118Movie LINK: https://youtu.be/3SYhOLG1UXk #118ThrillingBlockbuster #118Film #KalyanRam #Shalini"

#118movie i want keep fan wars aside and i want appreciate this movie...what a movie...gripping screen play especially2 characters #KalyanRam #Nivethathomas ..go and watch in theaters.. worth for 200 also

What a movie really enjoyed a lot #KalyanRam #ShaliniPandey #118Movie

#118Movie an exellent film..... #KalyanRam #ShaliniPandey and #Nivethathomas has done an extraordinary job. Happy sunday!!!!!!

"#118Movie superrrrrrrr BBB acting #Kalyanram bro,And thrilling and this movie first half is very horror thrilling.. climax so...double super... Congratulations #Kalyanram bro."

"#118Film (Telugu) is an edge of the seat thriller, very well made by our own DoP & Director #KVGuhan. #KalyanRam is superb in his role. Very engaging from start to end. Except for one aspect in 2nd half, the plot was very convincing. Do watch it to appreciate a good Screenplay"

Then #NTR Now #NKR Best Come Back Movies Ever #Temper #118Movie #118ThrillingBlockbuster #TrillingBlockBuster118movie @tarak9999 @NANDAMURIKALYAN #KalyanRam #JrNTR #Tarak

Show Tym With @kickVasimalla & 13 Other Friends #118Movie #KalyanRam #NKR #118Film #TrillingBlockBuster118Film #ThrillerBlockBuster118Movie

#KalyanRam 118 movie joins list of Disaster Read: https://goo.gl/iL3hKf SUBSCRIBE TO TOLLYWOOD VIDEO CHANNEL : https://goo.gl/DBvfV4

#118Movie is having dream run at box office.... Day 2 evening shows are fantastic all over AP/TG and performing better than Day 1. In few centres advance fulls for Night shows... #Kalyanram's best... #118ThrillingBlockbuster

#118Movie Completed Break Even & Become Clean HIT at Box-office With in 4Days #118ThrillingBlockbuster #118Film #KalyanRam #NKR #NTR #JrNTR @NANDAMURIKALYAN @i_nivethathomas @shalinipandeyyy #KVGuhan @smkoneru @EastCoastPrdns @vamsikaka

#118Movie collected close to 3 crore share in its first two day #118Film #KalyanRam

#118Film is a well made thriller with an intriguing content along with top notch production values... Decent direction with astounding background score made the film an interesting watch... #Kalyanram performed well... The flashback part would have been better... But overall

"Done watching #118Movie . Good psychological THRILLER based on DREAMS concept. @i_nivethathomas performance is very good congrats to director GUHAN and #KalyanRam congrats @smkoneru anna, for the winner."

Watching #118 movie it is quite interested. suspence thriller and cinematography and way of story driven is super. I think #kalyanram has a successful / good movie after long time. @NANDAMURIKALYAN

OUTPUT:

amazing - 9

good - 4

best - 5

good - 4

success - 4

celebration - 5

wonderful - 9

best - 5

director - 0

gross - -5

lovely - 10
successful - 9
successful - 9
well - 4
funny - 6
war - 0
appreciate - 9
play - 1
happy - 5
thrilling - 6
thrilling - 6
super - 8
well - 4
director - 0
superb - 8
convincing - 5
appreciate - 9
good - 4
best - 5
disaster - -4
fantastic - 7
best - 5
hit - 6
well - 4
astounding - 10
interesting - 6
well - 4
good - 4
good - 4
director - 0
super - 8

successful - 9

good - 4

11.3. TEST CASE 3

TEST CASE FOR CLASSIFYING 10 TWEETS:

INPUT: Random tweets

this is a fantastic movie

good movie

actors were cool

boring movie

i wasted my time by coming to this movie

mind blowing performance

awesome entertainer

worst climax

nice movie

feel good movie

OUTPUT:

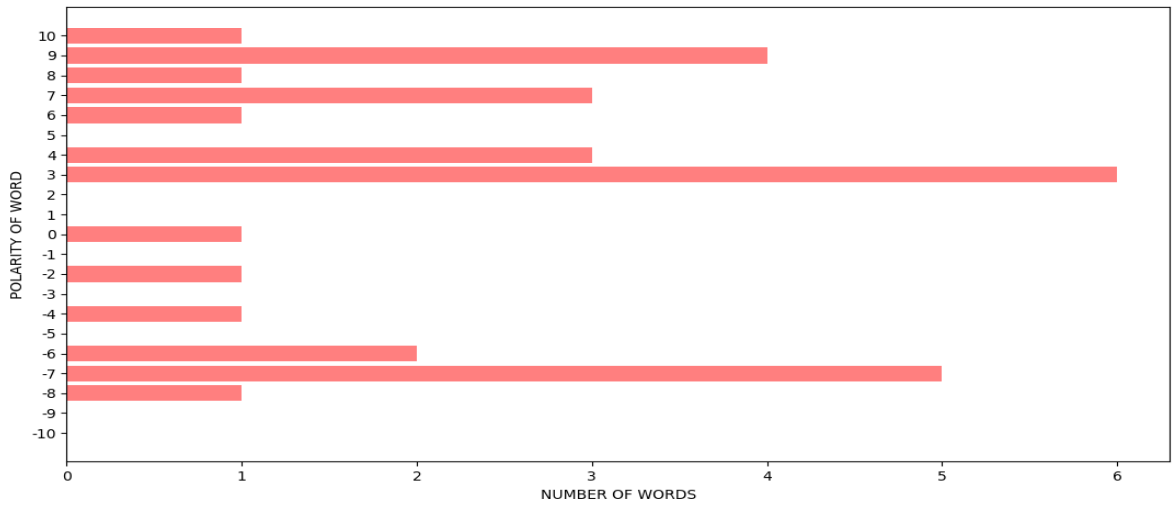

Figure 11.1 Word polarity barchart for 10 tweets

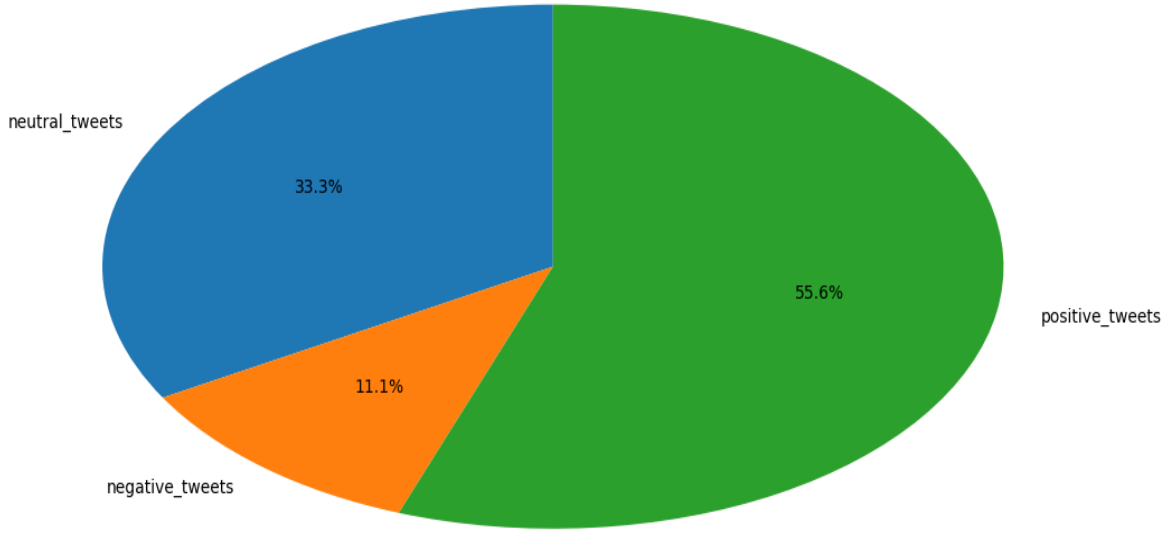

Figure 11.2 Classification pie chart for 10 tweets

Figure 11.3 Word cloud for 10 tweets

Number Of Positive Tweets: 5

Number Of Negative Tweets: 1

Number Of Neutral Tweets: 3

11.4. TEST CASE 4

TEST CASE FOR CLASSIFYING 25 TWEETS:

INPUT: Tweets with #118Movie

Atlast #kalyanram got a successful film after #pataas hope he will do more films like this in future. Exam season aina bagane collections vachay.minimum content unna audience encourage chestaru. #118Film SUPER HIT status.

″BTW, amazing work by whole team, special credits to #Kalyanram for his hard work and Dedication.″

#118movie first half good.....classy ..#kalyanram best looks till date...

″#KalyanRam #118Movie collected around 2crores share worldwide on day 1, gets good reviews and 7 crores break even, what can be closing #NZvBAN #LetsConclave19 #F2 #Viswasam #ViswasamTelugu #Viswasam50thdayCelebration #SaahoSunday #Maharshi″

118 Movie Success Celebrations!! #KalyanRam #NivethaThomas

″What a movie #118 Tnx 4 making my day with wonderful suspense thriller, best performance - #KalyanRam @i_nivethathomas both stole the show″

″#118Movie #Nivethathomas #KalyanRam loved it anna, our Telugu industry going to the next level.#Nivethathomas what acting when prabha hospital scene. Two ultimate scenes one is interval bang and another church father shot. #KVGuhan your first movie is like you have lot of exp″

@NANDAMURIKALYAN & director @DirectorSriwass film on cards..!!! #KalyanRam

#KalyanRam's #118Movie 1st Week WW Collections... Day 1 - 2 cr+ Day 2 - 1.5 cr+ Day 3 - 2.2 cr+ Day 4 - 1.9 cr+ Day 5 - 1.1cr+ Day 6 - 0.7cr+ Day 7 - 0.55cr+ 1st Week WW Share = 10.5 cr+ 1st Week WW Gross = 16.2 cr+(aprx)

Lovely working stills from #118Movie. #118Film #118ThrillingBlockbuster @NANDAMURIKALYAN #KVGuhan @shalinipandeyyy @i_nivethathomas @smkoneru @EastCoastPrdns #KalyanRam #NKR #NTR #JrNTR

″Wishing the leading ladies of #118Movie - @i_nivethathomas (Adhya) and @shalinipandeyyy (Megha) , and all the women who make this world a better place, a very #HappyInternationalWomensDay #HappyWomensDay #HappyWomensDay2019 #KalyanRam #NKR #NTR #JrNTR #118ThrillingBlockbuster″

#118Movie Entered into 2nd Week At Box-office Running Successfully in All Areas #118ThrillingBlockbuster @NANDAMURIKALYAN @i_nivethathomas @shalinipandeyyy

#KVGuhan @smkoneru @EastCoastPrdns @vamsikaka #118Film #KalyanRam #NKR #NTR #JrNTR

#118Movie is all set to have a successful 2nd week Too. Breakeven Don With in Just 4 Days #118ThrillingBlockbuster #KalyanRam #JrNTR

#118Movie is all set to have a successful 2nd week Too. Breakeven Don With in Just 4 Days #118ThrillingBlockbuster #KalyanRam #NKR #NTR #JrNTR

#118ThrillingBlockbuster ready to enter into 2nd week. #kalyanram #118Movie #118Film

"She is smiling at one point, serious as well || #Nivetha's funny cut on #118Movie LINK: https://youtu.be/3SYhOLG1UXk #118ThrillingBlockbuster #118Film #KalyanRam #Shalini"

#118movie i want keep fan wars aside and i want appreciate this movie...what a movie...gripping screen play especially2 characters #KalyanRam #Nivethathomas ..go and watch in theaters.. worth for 200 also

What a movie really enjoyed a lot #KalyanRam #ShaliniPandey #118Movie

#118Movie an exellent film..... #KalyanRam #ShaliniPandey and #Nivethathomas has done an extraordinary job. Happy sunday!!!!!!

"#118Movie superrrrrrrr BBB acting #Kalyanram bro,And thrilling and this movie first half is very horror thrilling.. climax so...double super... Congratulations #Kalyanram bro."

"#118Film (Telugu) is an edge of the seat thriller, very well made by our own DoP & Director #KVGuhan. #KalyanRam is superb in his role. Very engaging from start to end. Except for one aspect in 2nd half, the plot was very convincing. Do watch it to appreciate a good Screenplay"

Then #NTR Now #NKR Best Come Back Movies Ever #Temper #118Movie #118ThrillingBlockbuster #TrillingBlockBuster118movie @tarak9999 @NANDAMURIKALYAN #KalyanRam #JrNTR #Tarak

Show Tym With @kickVasimalla & 13 Other Friends #118Movie #KalyanRam #NKR #118Film #TrillingBlockBuster118Film #ThrillerBlockBuster118Movie

#KalyanRam 118 movie joins list of Disaster Read: https://goo.gl/iL3hKf SUBSCRIBE TO TOLLYWOOD VIDEO CHANNEL : https://goo.gl/DBvfV4

#118Movie is having dream run at box office.... Day 2 evening shows are fantastic all over AP/TG and performing better than Day 1. In few centres advance fulls for Night shows... #Kalyanram's best... #118ThrillingBlockbuster

#118Movie Completed Break Even & Become Clean HIT at Box-office With in 4Days #118ThrillingBlockbuster #118Film #KalyanRam #NKR #NTR #JrNTR @NANDAMURIKALYAN @i_nivethathomas @shalinipandeyyy #KVGuhan @smkoneru @EastCoastPrdns @vamsikaka

#118Movie collected close to 3 crore share in its first two day #118Film #KalyanRam

#118Film is a well made thriller with an intriguing content along with top notch production values... Decent direction with astounding background score made the film an interesting watch... #Kalyanram performed well... The flashback part would have been better... But overall

"Done watching #118Movie . Good psychological THRILLER based on DREAMS concept. @i_nivethathomas performance is very good congrats to director GUHAN and #KalyanRam congrats @smkoneru anna, for the winner."

Watching #118 movie it is quite interested. suspence thriller and cinematography and way of story driven is super. I think #kalyanram has a successful / good movie after long time. @NANDAMURIKALYAN

OUTPUT:

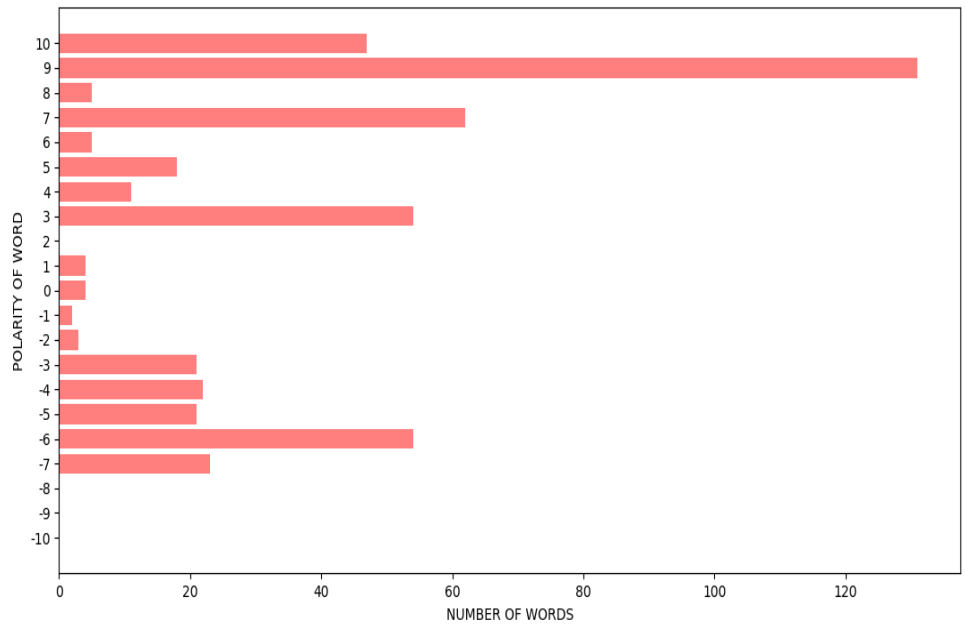

Figure 11.4 Word polarity barchart for 25 tweets

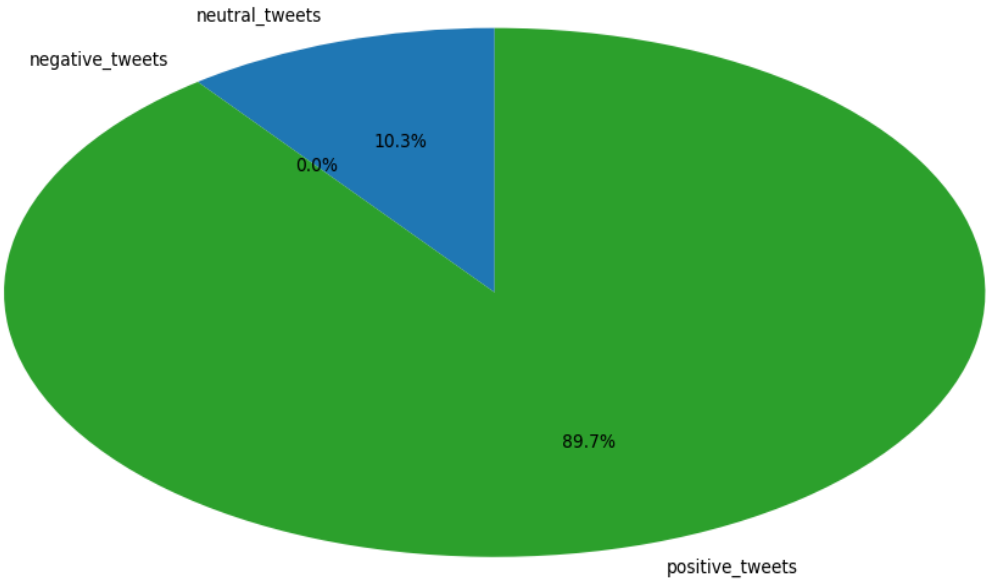

Figure 11.5 Classification pie chart for 25 tweets

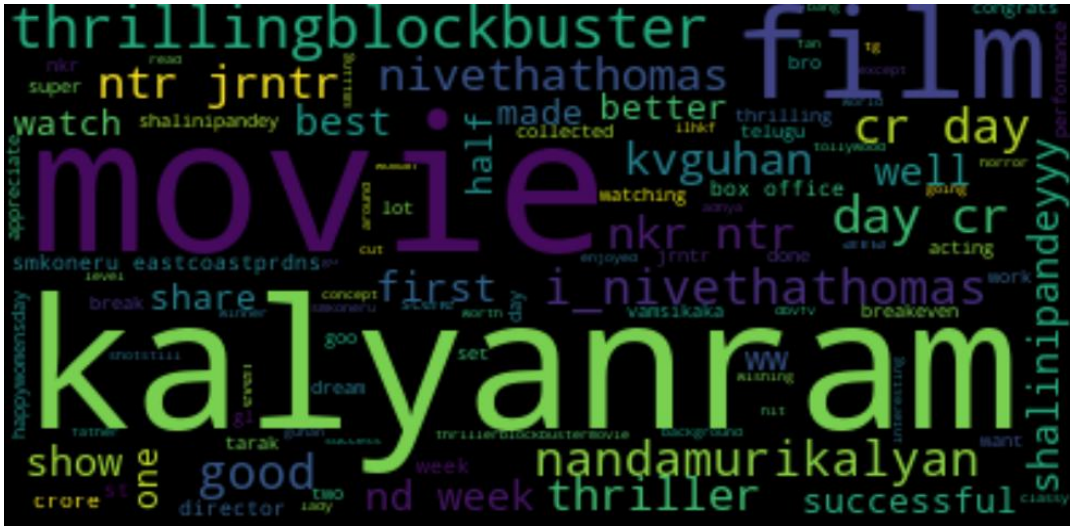

Figure 11.6 Word cloud for 25 tweets

Number Of Positive Tweets: 26

Number Of Negative Tweets: 0

Number Of Neutral Tweets: 3

11.5. TEST CASE 5

TIME ANALYSIS TO CLASSIFY TWEETS WITH DIFFERENT DATASETS:

"BatmanvSuperman", "JungleBook", "Deadpool", "Zootopia" movie review tweets are used for time analysis. An average time of 2.80s is taken to classify 100 tweets.

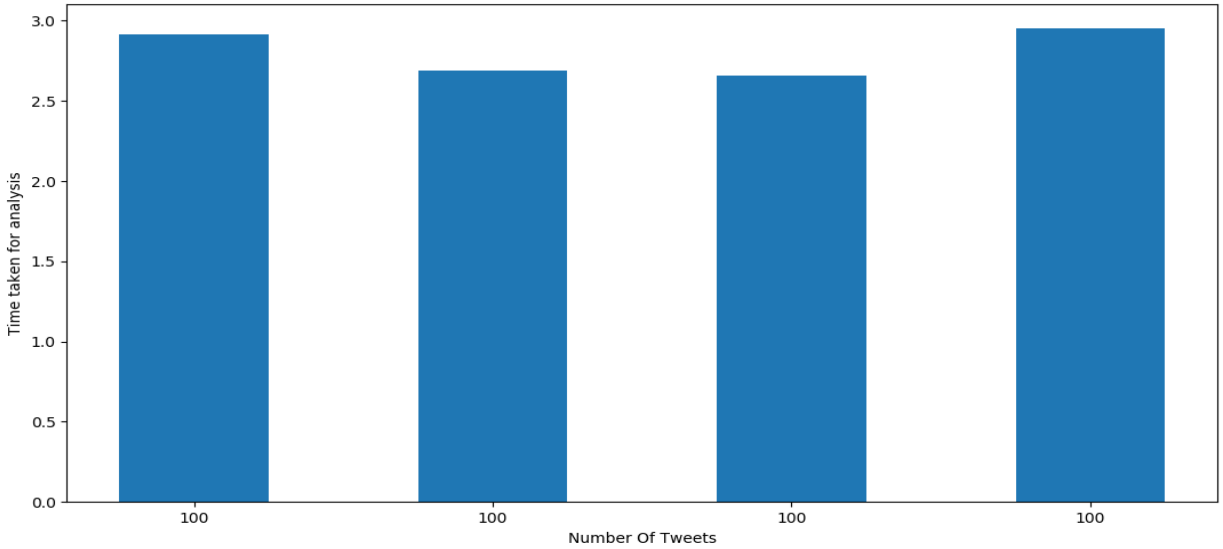

Figure11.7 Time analysis graph for different datasets

11.6. TEST CASE 6

TIME ANALYSIS TO CLASSIFY TWEETS WITH DIFFERENT SIZES:

"BatmanvSuperman", "JungleBook", "Deadpool", "Zootopia" movie tweets with each of 10, 50, 100 tweets are used for time analysis. Time for analysis is increased with increase in size of dataset

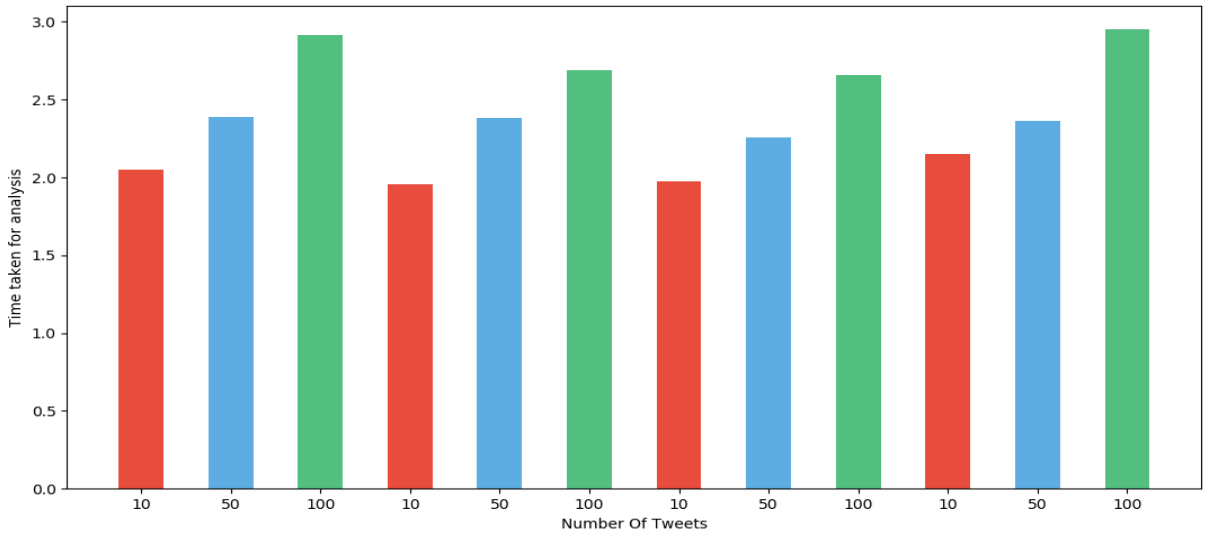

Figure 11.8 Time analysis graph for same datasets with different sizes

12. CONCLUSION

We have approached different ways of opinion mining and we got better accuracy than the existing methods. A new data set with 430 words is created for training the classifier. The method can still be improved by taking more number of words in data set which represent the sentiment which helps for achieving more accuracy. Time taken for analysis can be reduced using Hadoop distributed system.

13. BIBLIOGRAPHY

[1] IEEE Paper on Sentiment analysis on twitter using streaming API by M.Trupthi, Suresh Pabboju, G.Narasimha

[2] Twitter Sentiment Analysis by Aliza Sarlan, Chayanit Nadam, Shuib Basri3

[3] Sentiment Analysis of Twitter Data: A Survey of Techniques by Vishal A. Kharde, S.S. Sonawane

[4] Sentiment analysis of tweets using Machine Learning Approach by Ankita Gupta1, Jyotika Pruthi, Neha Sahu.

[5] Sentiment Analysis using Machine Learning through Twitter Streaming API by P.Akilandeswari, R.Harshita, Sumanth.KO.M

[6] https://pypi.org/project/emoji/

[7] https://www.webfx.com/tools/emoji-cheat-sheet/

[8] https://pythonspot.com/category/nltk/

[9] https://www.datacamp.com/community/tutorials/wordcloud-python,

[10] https://pythonspot.com/matplotlib-bar-chart/

[11] https://www.geeksforgeeks.org/graph-plotting-in-python-set-1/

[12] https://github.com/shalinc/ML-Sentiment-Analysis-of-Movie-Reviews-from-Twitter

[13] https://eskript.ethz.ch/filmstudies/chapter/describing-and-talking-about-a-film/

[14] https://www.words-to-use.com/words/movies-tv/

[15] http://docs.tweepy.org/en/v3.5.0/getting_started.html#api

[16] https://developer.twitter.com/en/docs/basics/getting-started