Dissertation on (Major Project-II)

"Application of Machine Learning Techniques in Sentiment Analysis"

Submitted in Partial Fulfillment of the Requirement
For the Award of Degree of

Master of Technology

In

Software Technology

By

Sandeep Gupta
University Roll No. 2K16/SWT/516

Under the Esteemed Guidance of

Dr. Rajni Jindal
Professor& Head of Department, Department of Computer Science & Engineering

2016-2020

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
DELHI – 110042, INDIA

ABSTRACT

There is ample amount of statements on social sites which can be inferred with the assistance of sentiment examination. This is very beneficial to find the public views. Sentiment Analysis includes catching of client's conduct, likes and dislikes from the created web content. There is no solid meaning of "Sentiments", yet by and large they are considered as musings, perspectives and frame of mind of an individual emerging basically dependent on the feeling rather than an explanation. Many of clients utilize social destinations to express their sentiment about brands, services, beliefs or opinions about things, political and religious views, emotions, personalities or places and people they interact with.

This data is mostly unorganized, slangs, etc. and therefore, text analytics and natural language processing are utilized to separate and group this data. Any Non-contextual and irrelevant contents are identified and discarded. The classification of sentiments will be performed on this data, which goes as follows: a training data set is created manually and based on this training data set sentiment analysis is performed on the twitter comments. Machine learning, for example, a hybrid Naive Bayesian classifiers are utilized for sentiment categorization with lexical word reference and natural language processing.

Table of Contents

Chapter One:

INTRODUCTION

1.1 Background

Social networking sites are considered as the significant sources of open perspectives and feelings on social issues prevailing at a given time. Sites, for example, Twitter reflect open perspectives through a large number of messages posted by their clients around the world. Sentiment investigation includes catching client conduct, different preferences from the created web content. There is no solid meaning of "Sentiments", however when all is said in done they are viewed as the considerations, thoughts, and frames of mind of an individual emerging based on feeling rather than a cause. In recent days it is found that popularity of sentiment analysis has increased as different stakeholders have understood the importance of analyzing the ulterior meaning of statements made on social media.

1.2 Problem Statement

There is ample amount of statements on social sites which can be inferred by means through sentiment examination. This is very beneficial to find public views. Sentiment analysis is tested tool to measure the opinions relating to product, performance of movies as shown in figure.1, election etc. It’s important to note that public opinion play a important role in evaluating several things like any place, product, or any person. Openings can be broadly classified into positive, negative either indifferent. Thus, with assistance of sentiment analysis tool can be easily find the what is general mood with respect to any thing.

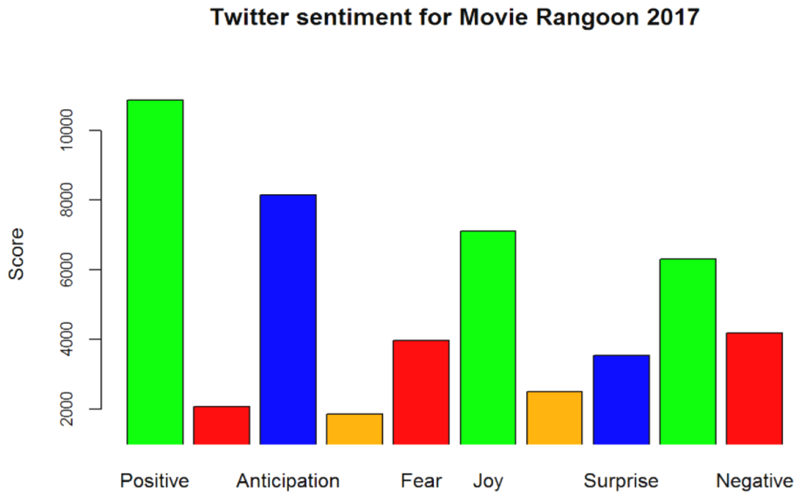


Figure 1: Twitter sentiment analysis of a movie

1.2.1 Analyzing Sentiments

Mining the data related to opinions of people Is a difficult undertaking due to availability of huge data which is getting generated on daily basis as result of communication taking place among the web users. The scope covers a number of discipline as it uses techniques relating to computational linguistics, Semantics, artificial intelligence, retrieval information, machine learning and natural language processing. In current analysis, effectiveness of applying AI procedures to the sentiment classification issue is analyzed. Figure 2 display a typical sentiment analysis process of Twitter data. The difficult part of this issue seems to be to identified it by customary point-based classification, while topics are often frequently identifiable by catchphrases alone, sentiment can be conveyed subtly. By streamlining learning and the accuracy of data, words and sentences can be accomplished on social media, for example, Twitter. At the root level, data tokens are used to distribute positive and negative parts of data.

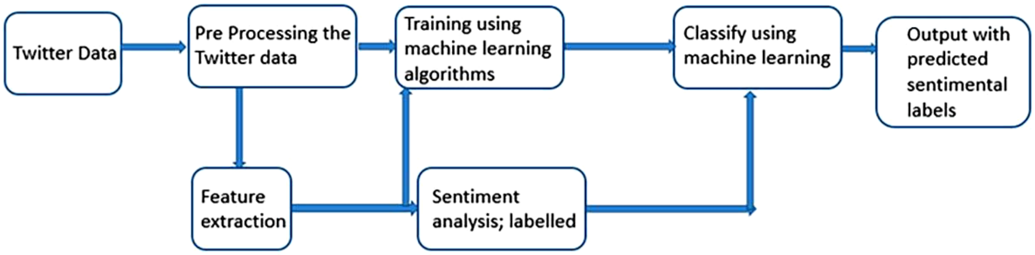

Twitter Data
Pre Processing the Twitter data
Training using machine learning algorithms
Classify using machine learning
Output with predicted sentimental labels
Feature extraction
Sentiment analysis; labelled

Figure 2: Twitter sentiment analysis process

1.3 Proposed work

To analyzing sentiment data analysis is a complex process involving separate steps of sentiment. Figure 3 shows the means associated with generic sentiment analysis approach. The first and foremost step is to gather data from the user tweets. This data is mostly unorganized, slangs, etc. and therefore, content analytics and natural language handling are used to solve and characterize this information. Any Non-contextual and irrelevant contents are identified and discarded.

The classification of sentiments will be performed on this data, which goes as follows: a training data set is created manually and based on this training data set sentiment analysis is performed on the twitter comments. Machine learning such as that a hybrid Naive Bayesian classifier is utilised with the lexical dictionary and process of natural language for the sentiment classification. The experiment is setup in python programming language, Twitter APIs and tweepy is used for extraction of tweets. The dataset contained movie tweets. The outcomes are classified as positive, negative and neutral sentiments.

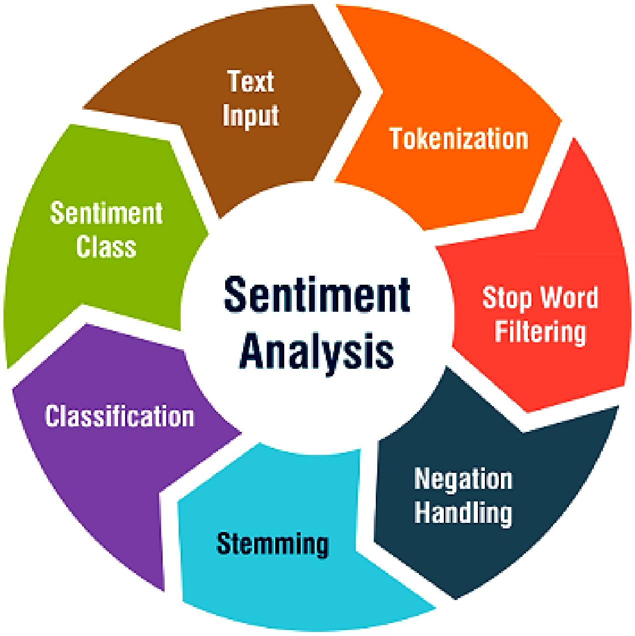


Figure 3: Forward process of sentiment analysis

The sentiment analysis technique generally depand on:

a) machine learning based techniques and

b) lexicon-based techniques.

Python 3.4 is used as programming language for the implementation of experimental setup.

1.4 Motivation

Human decision making is constantly affected by other's ideas, thinking and opinions. Today, exceptionally a lot of data is available in web journals and on-line archives. As a major aspect of the effort to all the more likely arrange this data for clients, analyst are viably exploring the issue of modified substance classification. With the proliferation of Web to applications such as blogs, forums, e-commerce website, survey sites, e-news and social systems, there came audits, remarks, suggestions, motion picture appraisals and feedbacks created by clients [1] [2].

The user created substance and surveys can be about for all intents and purposes anything including politicians, people, events, products, and so on. Microblogging has become an exceptionally famous specialized device among Internet clients [3]. Millions of messages are coming every day Facebook. With an exponential rise in social media usage to share emotions, thoughts and opinion, Twitter has become the gold-mine to analyze brand performance.

These opinions can be clubbed into categories such as negative, positive or neutral. Opinions found on Twitter are casual, honest and informative than what can be picked from formal surveys etc. [3, 4] Millions of users utilize social websites to express their sentiment about brands, services, political and religious views, beliefs or opinions about things, emotions, personalities or places and people they interact with. As many of the users post their opinions and views, micro blogging, websites become significant spruce of individuals' sentiments and opinions. Such data can be used efficiently for various analysis-protocols or monitoring marketing, criminal activities and social studies. This field of software engineering deals with examining and foreseeing the concealed data stored in the content. This hidden data gives significant bits of knowledge about user’s intention, taste and likeliness.

1.5 Problem Statement

The expanded examination and analysis of truncated data arouses the interest of sentiment examination. Which is originates from the social media as unstructured data. Given a lot of data containing various highlights and fluctuated conclusions, the goal is to remove articulations of supposition portraying an objective element and arrange it as positive or negative. The best test of sentiment investigations to plan appatilicon-explicit calculations along with methods which can dissect the human language etymology precisely. We propose a half and half methodology for sentiment examination that is a mix of an AI calculation (Naive Bayes) and an uncommon lexical word reference with NLP.

1.6 Thesis Organisation

Chapter 1: In the chapter one, we tried to give general idea of sentiment analysis while describing the methods. it is there be relevant in the processing of natural language.
Chapter 2: In this chapter, we discussed earlier work done in the field.
Chapter 3: In this chapter we have dealt with the concept of sentiment analysis. It also discusses that why we need this analysis and how do we perform. Levels of sentiment analysis, challenges and the proposed method is also discussed.
Chapter 4: Method of sentiment analysis and Machine learning technique has been elaborated.
Chapter 5: In this chapter Model Social media network: Twitter has been taken into consideration.
Chapter 6: Sentiment analysis classification has been elaborated use as basis on the research.
Chapter 7: This chapter has concluded the research.

1.7 Literature Reviews

In this section we summarize the findings from the literature review conducted to understand what metrics and features have worked well and how can they be adopted to sentiment analysis visualization. We also describe other related public tools that perform the similar task [5]. sentiment investigation is an extremely testing undertaking. From data recovery and regular language handling, voyagers have created various ways to deal with tackle this issue, prompting promising or good outcomes [6, 7]. Suchita and Sachin [8], looked at the SVM and Nave Base surveys of the two most much of the time utilized administered AI draws near. The outcomes propose that SVMs have a more noteworthy number of information focuses misclassified than Naïve Bayes and that Nave Byers drew closer SVM when audit numbers were lower.

One of the popular work in area of Twitter analysis of sentiment was done by Go et al. in 2009 [9]. They used SVM, MaxEnt and Naive Bayes and reported that SVM and

Naive Bayes were equally good and beat MaxEnt. The research basically used distant supervision technique to overcome the problem of manual annotation of large set of tweets. Tweets with “:)” emoticon were considered positive while tweets with “:(“ in the message were considered negative. This resulted in two defects. First, emoticons, which were considered important by other works in the area couldn’t be used to learn sentiments. Secondly, the authors were unsure if all tweets with “:)” are truly positive or can contain negative or sarcastic sentiments too. Therefore, the dataset used in the experiment was labelled noisy.

The same distant supervision procedure of tagging tweets positive or negative based on the emoticons it mentions was used by Pak and Paroubek [10] also used linear kernel SVM to run the experiment. Although, the results were not reported as accuracy metric. However, highlighting the significance of neutral class, the team also collected neutral tweets. These tweets were strictly objective and were collected from newspapers and magazines.

In this paper [11], the researcher has used the Naive Bayesian classifier in their attempt to analyze the sentence. In their research they tried to show with the help of their experiment by using Naive Bayesian classifier model that large data set can easily be analyzed.

Tumasjan et al. and Bollen et al. [12, 13] employed dictionaries to measure the quote level of pre-day dictionaries. Hu et al. [16] incorporated social cues into their uncontrolled sentiment analysis framework. They defined and integrated both emotion indication and correlation into a framework to learn parameters for their sentiment classifier.

In yet another experiment real time data retrieved from two different accounts of politicians were used [15]. In yet another similar research Twitter-streaming API (Application Programming Interface) [16, 17] were used to extract the data. Two sentiment examination named Sentivordnet [18, 19] and WordNet. [20] Both were utilized to score positive and negative. The Twitter gushing API was used to gather information by the creators of [21], for anticipating the political race of the Indonesian

President. The reason for existing was to utilize Twitter information to comprehend general assessment.

Fang et al. [22] on the other hand for SVM learning used both Domain Specific sentiment lexicon along with General Purpose Sentiment Lexicons. It helped them to get the product aspect and they utilize this system is use to identifying both related polarities and product aspects.

Zhang et al. And Trinh et al. [23, 24] utilized an enlarged dictionary based strategy for substance level sentiment investigation. In various papers [25, 26] creators portray the different devices utilized in sentimental examination of Twitter information. Since assessments in Twitter are miscellaneous, exceptionally disorganised and alongside these it incorporates positive, negative or unbiased in various circumstance, it is critical to dissect the sentiments. In these papers' creators utilized vocabulary based strategies for characterization which requires little exertion in individual named content report.

In yet another research, scholar while analyzing the sentiments objective sentences have been ignored [27, 28] .Authors came up with sentence classification. For this reason [15] [1] the scientists utilized SVM, Part of Speech and SentiWordNet innovation and Naive Bayes. They infer that the vector machine gives the best yield. Parikh and Movset [29] applied both models, a most extreme entropy model and a nave base bigram model, to characterize tweets. And found it that the classifier of the nave base worked obviously superior to the most extreme entropy model.

Chapter Two:

SENTIMENT ANALYSIS TOOLS

2.1 Sentiment Analysis

Sentiment analysis is the programmed mining of emotions through NLP (natural language processing), displacement, suppositions and discourse from content and database sources. It includes arranging conclusions in a book in classes, for example, "positive","negative" either "unbiased" [30]. This analysis recognizing, extricating, and ordering conclusions, emotions, and frames of mind identified with different subjects, as communicated in printed input [31] has gone. SA helps in accomplishing different objectives, for example, political development, showcase insight [31], estimation of consumer loyalty [32], open state of mind expectation about film deals and significantly more. For this, feelings are collecting from customers, which can be used for additional analysis and enrichment. Sentiment analysis gives perception Identified data with open view, as it investigates different audits and tweets. It is a verified apparatus. Box of CE exhibitions of movies and general races for foreseeing the occasions of numerous significant occasions [35].

Sentiment analysis is the marvel of separating contemplations either feelings from audits communicated online by a specific subject, region, or item. The motivation behind The analysis of sentiment is consequently to determine the expressive course of customer surveys.[36].

2.1.1 Why Sentiment Analysis

Most users Utilize social locales to express their beliefs, feelings, thoughts or opinions about various products, services, movies or places. Clients want to know other's opinion about an item before purchasing it. Business associations need to recognize what clients are stating about their item or administration that an association is giving, to settle on further choices. With significant advancements in interpersonal interaction on the web (i.e, Facebook, Twitter, Instagram, LinkedIn, Stumble, and so forth.), people and huge affiliations are centering general sentiment to settle on their choices. The working of supposition data on sites isn't simple, because of the immense number of sites that right now exist are as yet populated and need institutionalized technique to do as such. What's more, the content corpora on sites establish both futile and helpful information that is important for analysis. There are likewise different factors, for example, human mental limit and physical confinement that make people unfit to examine a lot of information. Hence, there is a requirement for a computerized feeling mining that will eventually help people in sentiment analysis.

A sentiment grouping should be possible at Sentence level, Document level and Feature level or Aspect level [37]. In Sentence level sentiment grouping arranges each sentence first as emotional or target and afterward characterizes into positive, negative or impartial category. In Document level the entire archive is utilized as an essential data unit to arrange it into positive either negative category. Perspective or Feature level sentiment grouping manages recognizing and separating item includes from the source information. [38-39].

Generally, various approaches are in sentimental analysis such as considering symbolic method, by machine learning method or using lexicon dictionary. In representative learning method, which is classified by some taking in systems, for example, gaining from similarity, revelation, models and from root learning. Figure 1 represents the basic concepts involved in sentiment analysis. In AI system it utilizes unaided adapting, pitifully directed learning and regulated learning. Alongside dictionary based and etymological technique, AI will be considered as one of the mostly utilized methodology in sentiment grouping. Figure 2 shows a typical flow diagram of sentiment analysis classification.

2.1.2 SA: How It Works?

A. **Subjectivity/Objectivity-**To execute sentiment analysis, we must first recognize the subjective text and objective text. Just emotional content holds the sentiments. Objective content contains just authentic data**.** Examples:

1. **Subjective:** Inception is a superb movie. (this sentence has a sentimental opinion “superb”, which talks about the movie and the writer’s feelings, thus it is subjective).

2. **Objective:** James Cameron is the director of titanic. (this sentence has no sentiment, it is a fact, a general information rather than an opinion or a view of some individual thus it is objective).

B. **Polarity-** Situated on sentiments expressed in the context subjective text clasified into three division :

Examples:

1. **Positive** : I love the new Transformers movie.
2. **Negative** : The acting and graphics of the film were terrible.
3. **Neutral** : I normally get tired by the evening. (This sentence contains the client's sentiments, musings, along these lines, it is emotional yet since it has no positive or negative extremity, it is impartial.)

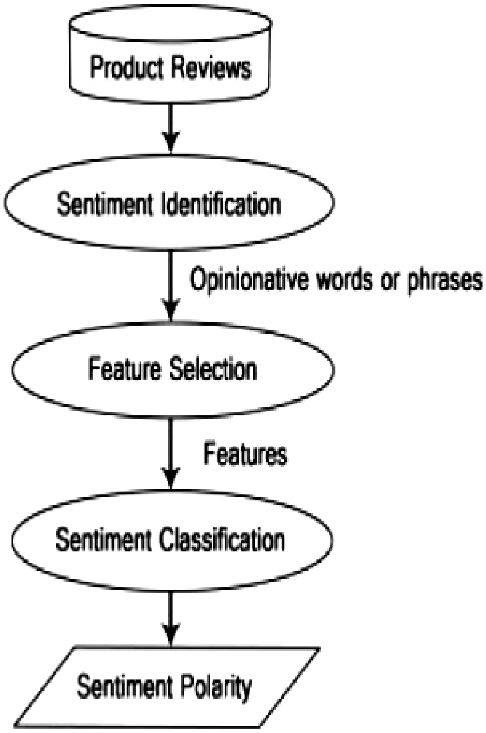


Figure 4: Flow diagram of classification of sentiment analysis

Though certain words infer similar meaning ad usually are used interchangeably but has different meaning.

Opinion: A conclusive remark that ignites other to speak out their mind

View: subjective Opinion that has different meaning

belirf: Intentional acknowledgment and intellectual affirmation

sentiment: Opinion speaking to one's emotions

2.2 Natural language processing

Natural language processing is the sub-field of AI that is centered around empowering PCs to comprehend and process human language.[40]. Association for Computational Linguistics. In principle, it manages the scope of procedures that figure, examine and speak to naturally happening writings at staggered investigation of dialects for the reason to make the machine procedure applications and like human language for various controls.NLP calculations rely profoundly on machine learning with the dominant part have been factual. More established execution of language-preparing undertakings regularly required hard coding of enormous arrangement of standards [41].

By utilizing machine learning, we can utilize ordinary learning calculations normally in factual derivation, to learn governs by investigating substantial corpora of genuine models.

2.2.1 Kinds of Natural Language Processing are:

1. Morphological preparing:
2. In this state of language arrangement, strings are broken in a set of tokens related to the word, sub-word and pronunciation frame. Commonly, a modification can occur by prefixes or postfixes as well as various changes.
3. Syntax and semantic investigation:

A processor that separates works principally dependent on semantic and syntax investigation. Have two employments of syntax investigation. One is to break it into a structure and other is to check if there is a sentence is all around framedthat gives syntactic connection between them. The equivalent can be accomplished by a parser utilizing a interpretation of word definitions and an arrangement of syntax rules. In a straightforward dictionary, each word has a syntax class. Theories are characterized by sentence structure which means that they can be involved in a wide variety of expressions.

4. Pragmatic investigation:

Translating the consequences of semantic investigation assuming from the purpose of perspective of a distinct circumstance is called pragmatic investigation.

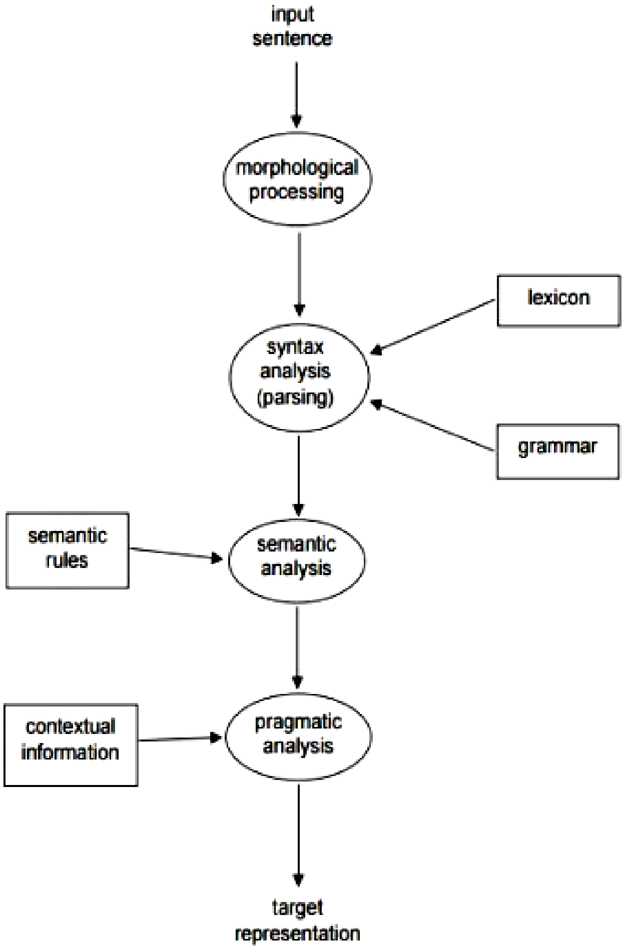


Figure 5: Flow-diagram of NLP steps

2.3 Sentiment analysis levels

When all is said in done, sentiment analysis has been explored primarily at following levels [42].

a) Document Level:

In Document level the entire archive is arrange into positive either negative category. It manages labeling singular archives with their feeling. The fundamental assignment in the record level is to characterize a completely supposition chronicle conveys a affermstive or negative opinion. At this level of assessment accept that every report imparts decisions on a singular component.

b) Phrase or Sentence Level:

Phrase-level Sentiment Analysis oversees naming solitary sentences with their specific decision polarities. Sentence level conclusion gathering orders sentence into positive, negative or unprejudiced category. The critical effort in the sentence level is to test whether each sentence provides a positive, negative, or neutral sentiment.This level of examination is highly related to spontaneity. Who assumes targeted sentences, which convey honest information from unique sentences that convey enthusiastic viewpoints and supposition. Sentence level and record level assessments don't found what unequivocally human delighted in and disdained.

c) Aspect Level:

The Aspect level arrangements with marking their sentiment to each word and, additionally distinguishing the element to which the sentiment is coordinated. There is a concern about separating and separating aspect level or feature level sentiment orders, which include source information. Strategies such reliance parser and talk structures are used in this. This procedure level performs better grained analysis. Rather than seeing language develops (sentences, sections, reports, expressions or provisos), The aspect level takes a wanderer directly into sentimentality.

d) Word Level:

The latest works have used the earlier extremity of words and expressions to characterize emotion at the level of sentence and collection. Word Sentiment Order has been used as highlighters for most part descriptors. There are two techniques to explain sentiment naturally at the word level are.:

1) Corpus-Based Approaches
2) Dictionary Based Approaches

Level of sentiment analysis shows in figure 5. Figure 6 and 7 show graphical comparison of a few methods for twitter classification.

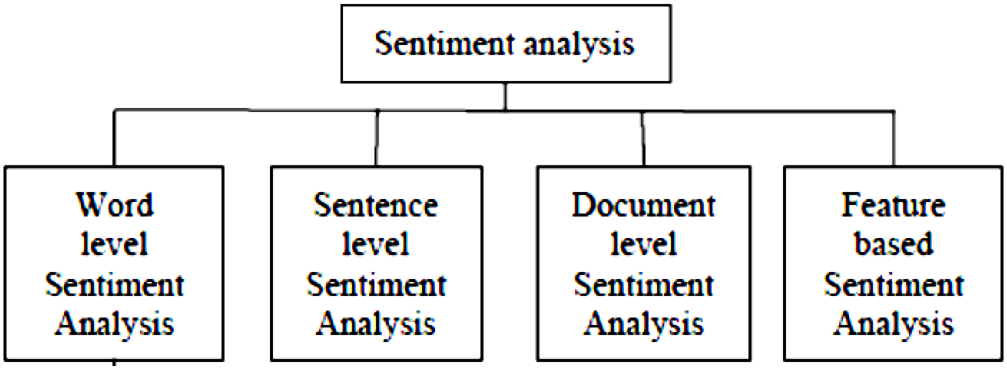

Sentiment analysis
Word level Sentiment Analysis
Sentence level Sentiment Analysis
Document level Sentiment Analysis
Feature based Sentiment Analysis

Figure 6: Sentiment analysis levels

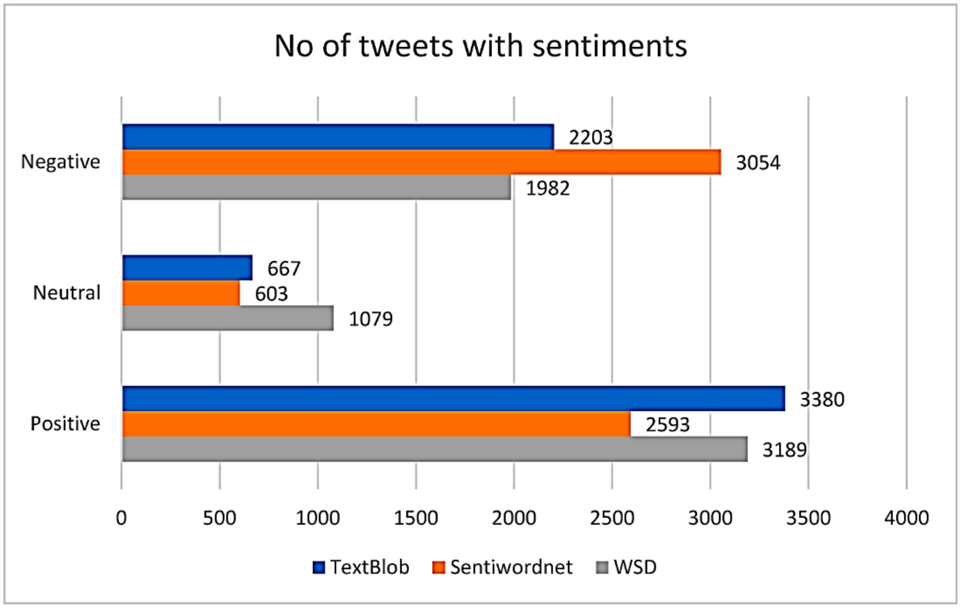


Figure 7: Sentiment classification based on word assessment

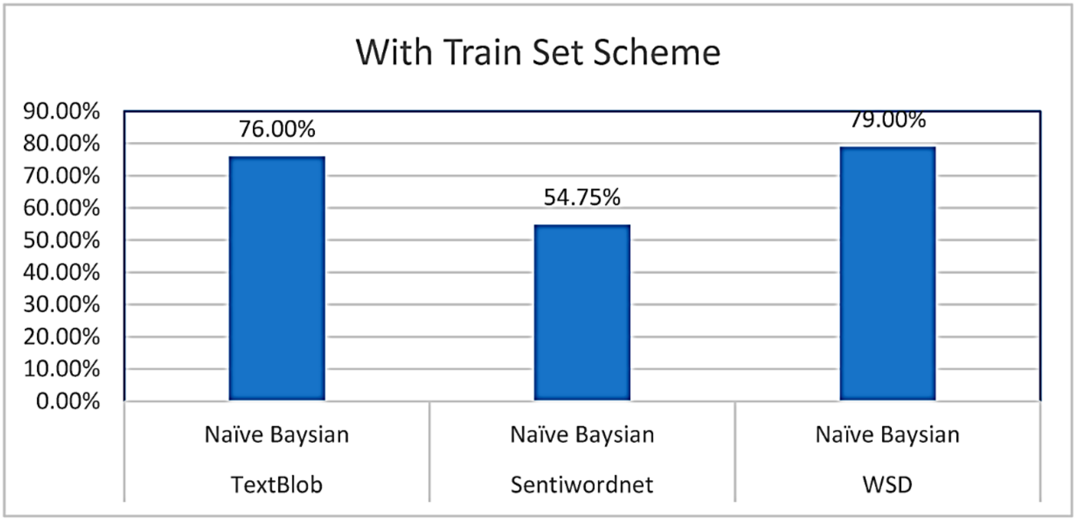


Figure 8: Performance of Naive Bayesian classifier with different analyzers

2.4 Challenges in Sentiment Analysis

Sentiment Analysis is an exceptionally testing undertaking. Following are a portion of the difficulties looked in Twitter's Sentiment Analysis.

1. Segmenting the abstract aspect in content: Emotional parts address feeling bearing substance. A comparative term can be considered enthusiastic in one case, or targeted in another. It makes it hard to perceive emotional bits of substance.
2. Domain dependence [24]: Multiple meaning of a word is yet another issue. More often the words use to have its meaning when it is used in a certain sentence. Therefore, it is the domain that decides the meaning.
3. Sarcasm Detection: The problem with a Sarcastic sentence is that positive words are used but ultimately bears a negative meaning.

3.5 Proposed Method

The proposed model (figure 9) is a prediction system that predicts the behavior of user, specifically twitter account holders on twitter, through which the behavior of each user and groups are predicted or their character is determined using their tweets. This work implements python and machine learning with Naive Bayes classifier for the purpose.

The system works as follows: the input of the system contains twitter commends that are collected from twitter tweets and after the data collection these tweets are pre-processed. The data will be well cleaned and irrelevant data might be removed after pre-preparing the cleaned information at that point exposed to sentiment analysis so as to compute the sentiment of the tweets. The sentiment is computed by using a training dataset which was deliberately created, based on the training dataset polarity is being calculated for each tweet. Finally, based on the polarity of these tweets they are classified as positive, negative or neutral.

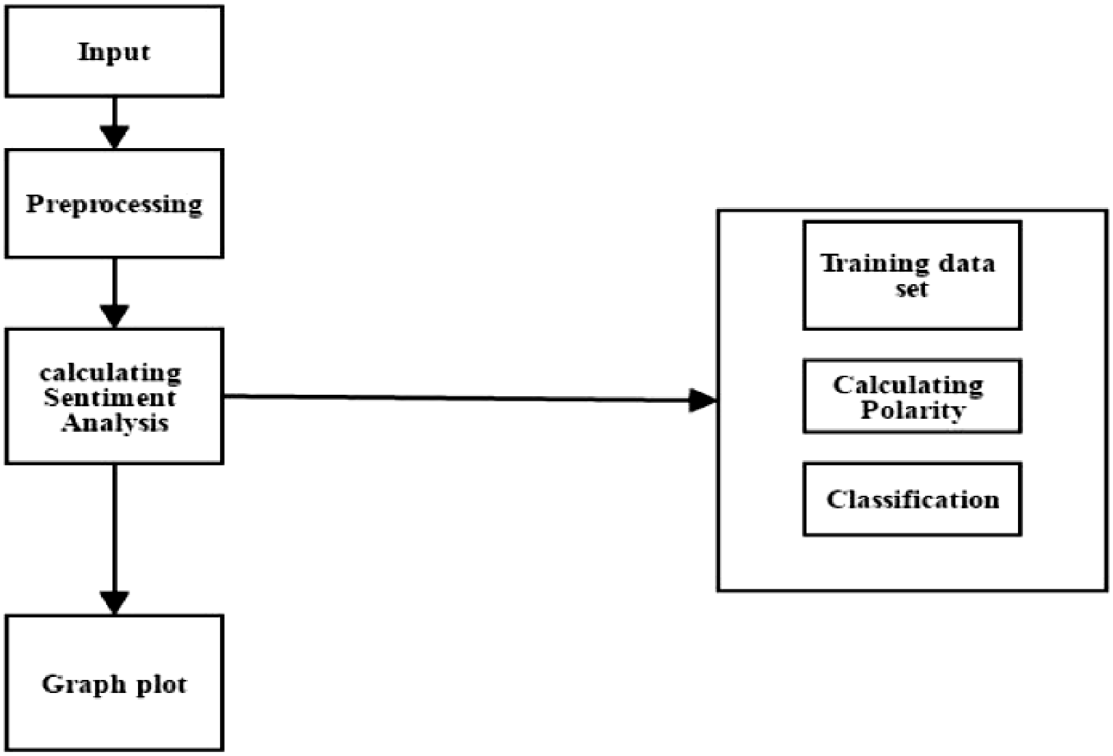


Figure 9: Architecture of sentiment analysis

Chapter Three:

METHODS FOR ANALYSIS

3.1 Methods for Sentiment Analysis

Sentiment analysis have two main technique: machine learning based and lexicon based. Machine learning usually takes the supervised or unsupervised approaches (as shown if figure 10). Some research considers have likewise joined these two techniques and achieved moderately better execution. There are still numerous investigates are proceeding to discover better choices because of its significance in this situation.

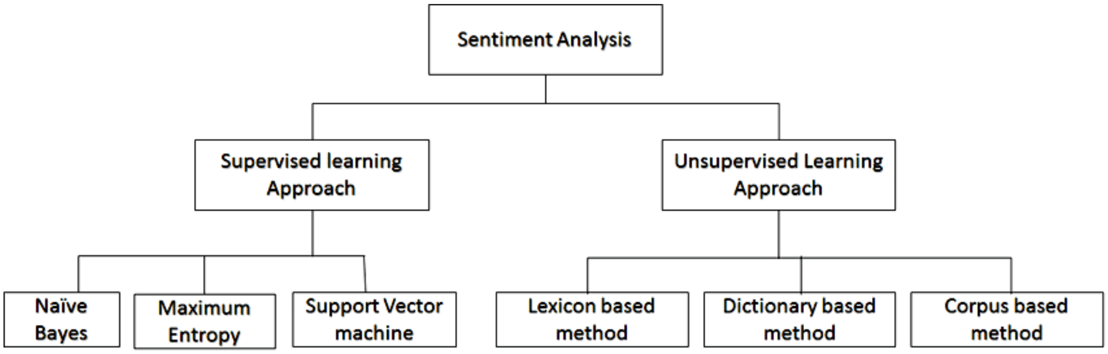


Figure 10: Methods of sentiment analysis

Table 1: Types of sentiment classifiers (with advantages and disadvantages)

Sentiment Classification Approaches		Features / Tecniques	Advantages and Limitations
Machine learning	Bayesian Networks Naive Bayes Classification Maximum Entropy Neural Networks Support Vector Machine	Term presence and frequency Negations Part of speech information Opinion words and phrases	Advantages the ability to adapt and create trained models for specific purposes and contexts Limitations the low applicability to new data because it is necessary the availability of labeled data that could be costly or even prohibitive
Lexicon based	Dictionary based Novel Machine Learning Corpus based Ensemble Approaches	Manual construction, Corpus-based Dictionary based	Advantages wider term coverage Limitations finite number of words in the lexicons and the assignation of a fixed sentiment orientation and score to words
Hybrid	Machine Learning and Lexicon based	Sentiment lexicon constructed using public resources	Advantages lexicon/learning symbiosis, the detection and measurement

		for initial sentiment Detection. Sentiment words as features in machine learning method.	low sensitivity to changes in emotion and subject area at the concept levelLimitations noisy reviews

3.1.1 Machine Learning Technique

Machine Learning is a part of computer science that enables frameworks to be consequently and develop by experience without being unequally modified or mediated by humans. Its fundamental point is to get PCs from experience naturally. Figure 11 shows different scopes and implementation of machine learning.

The machine learning approach to analyse sentiment generally has a location with directed grouping. A managed learning classifier utilizes preparation set to learn and prepare themself as for separating traits of content, and the presentation of the classifier is tried utilizing test dataset. In this, two courses of action of chronicles are required: planning and a test set. A readiness set is used by a modified classifier to get acquainted with the isolating characteristics of files, and a test set is used to check how well the classifier performs.

Different machine learning procedures have been adopted for ordering surveys. Machine learning begins with preparing a dataset. Some machine learning calculations such as maximum entropy (ME), nave bay (NB) and support vector machine (SVM) are commonly used for characterization of content (tweet). These algorithms have produced successful results. Table 1 represents a comparison table with advantages and disadvantages of different machine classifiers. Machine learning is a part of computer science that enables frameworks and is developed and developed by experience without being unequally modified or mediated by humans. Its basic point is to get PCs from experience naturally. Figure 11 shows the different scope and implementation of machine learning.

Machine learning approach to sentiment analysis usually has a location with directed grouping. A managed learning classifier uses scheduled preparation to learn and prepare itself for individual traits of content, and the presentation of the classifier is tried using a test dataset. It requires two arrangements of archives: preparation and a test set. A preparation set is used by the programmer classifier to become familiar with the different properties of the archives, and a test set is used to see how well the classifier performs.

Various machine learning procedures have been adopted for ordering surveys. Machine learning begins with gathering preparing dataset. A few machine learning calculations such as Support Vector Machines (SVM), Naive Baye's (NB) and Maximum Entropy (ME) are generally utilized for content characterization (tweets). These algorithms have produced successful results. Table 1 represents a comparison table with advantages and disadvantages of various machine classifiers.

3.1.2 Deep Learning

Deep learning was first introduced to G.E. In 2006 Hinton and AI are a piece of the process that connects with the Deep Neural Network. The neural system is influenced by the human mind and consists of a few neurons that make up a noteworthy system(figure 12). Deep learning is exceptionally powerful in learning strong highlights in an administered or unaided style. Deep learning systems are skilled for giving preparing to both administered and solo classes [43].

Deep learning has been as of late stretched out to incorporate distinctive different systems, for example, RNN (Recurrent Neural Networks), DBN (Deep Belief Networks), Recursive Neural Networks, CNN (Convolutional Neural Networks), DNN (Deep Neural Networks) and some more. Neural system contents are useful in content age, vector illustration, word depiction inference, sentence grouping, sentence display, and introduction.

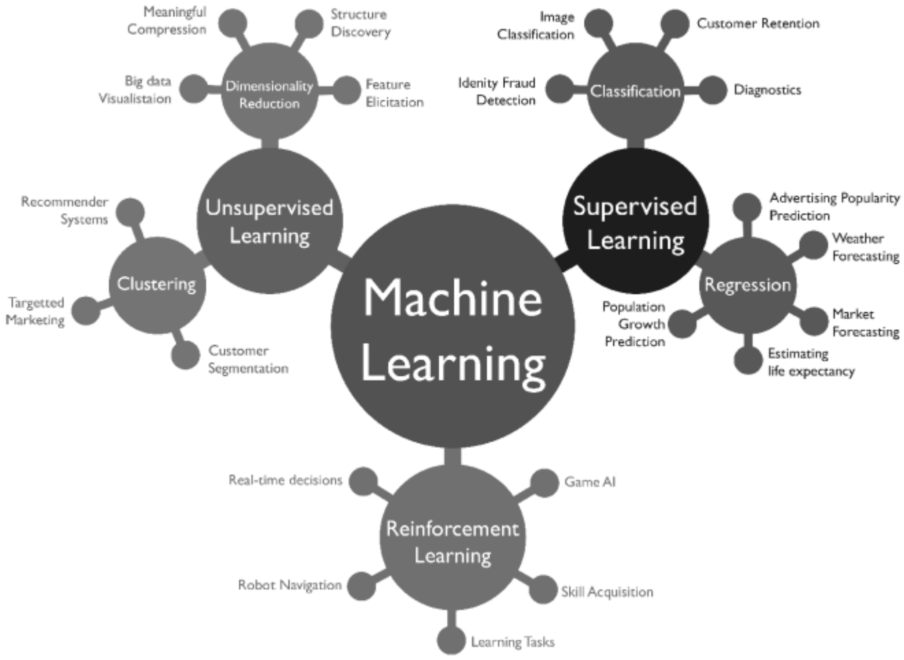


Figure 11: Branches and scopes of Machine learning

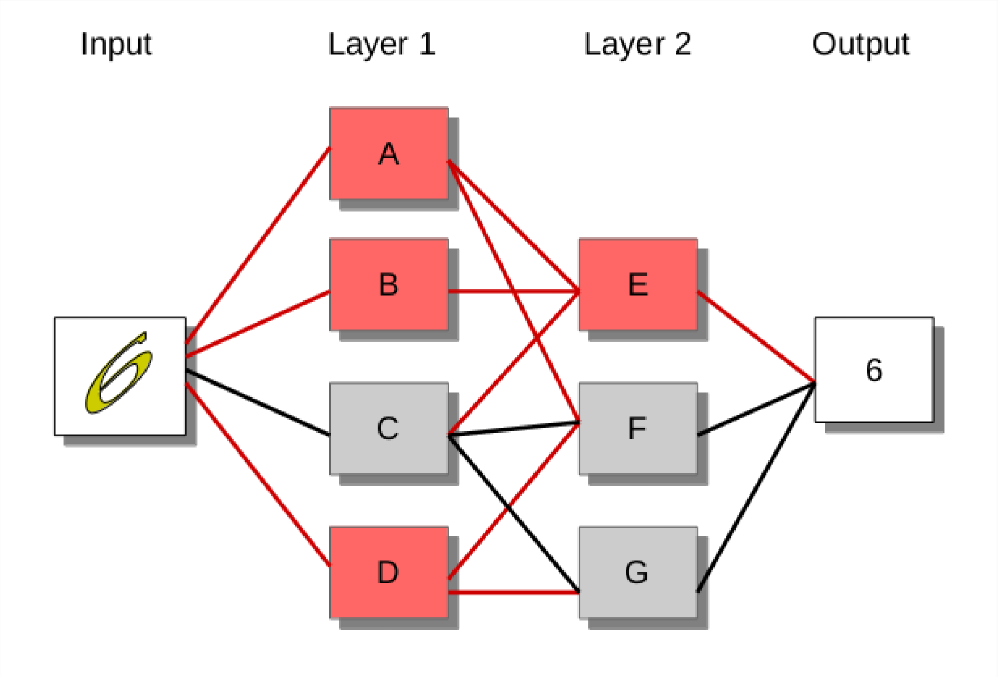


Figure 12: A 2 layered Neural Network

3.1.3 Supervised Learning

Machine learning begins with gathering preparing dataset. The subsequent stage is to develop a classifier on preparation data. When a supervised order strategy is chosen, a significant choice to make is include choice. Supervised machine learning method is partner with utilization of a stamped list of capabilities to hold some grouping capacity and incorporates learning of capacity from the analysis alongside its information and yield. Preparing informational index incorporates set of preparing models; every model comprises of couple of an info information just as anticipated yield. The exactness of expectations by the PC during training is also analyzed.

3.1.4 Unsupervised Learning

For this circumstance, no such ready input is provided, leaving PCs to find the yield with no info. Unsupervised learning is generally associated on esteem based data. It is used in increasingly perplexing endeavors. It uses another technique of cycle realized a significant making sense of how to land at a couple of finishes. A few models for unsupervised learning approach are bunch examination, desire amplification calculations. These calculations utilize Dictionary based way to deal with assemble nostalgic content.

3.1.5 Naive Bayes Method

It is a probabilistic classifier and is basically utilized when degree of planning set is low. In machine learning this Bayes hypothesis dependent test is in the group of probabilistic classifiers. The restrictive probability that the proof given by an opportunity X is Y is governed by the Bayes rule:

$$P(X/Y) = P(X)\ P(Y/X)/P(Y)$$

Thus, to discover the suppression status has been changed to below:

$$P(Sentiment/Sentence) = P(Sentiment)P(Sentence/Sentiment)/P(Sentence)$$

P (sentence/slant) is computed as the result of P (token/opinion), which is detailed by:

Tally (Thistokeninclass) + 1/Count (Alltokensinclass) + Count (Alltokens)

One and tally of all tokens are included here or Laplas Smoothie.

3.1.6 k-Nearest Neighbor method and weighted k-Nearest Neighbor method

K-NN technique depends upon the way that arrangement of an occurrence. Will have any degree like who adjacent it in the vector space. Some assembled facilities were inspected on a weighted k-nearest neighbor strategy. In which he weighted those components in the preparation set and he used these weights to count his perception of the content in the word like word. [44].

Inspiration Score = (1 Σ j score (pos) + 1 Σ k score (neg))/1 Σ s most extreme score

Here s = j + k, i.e. check both positive and negative simultaneously. In the weighted k-NN technique, as a subject of first importance, they interpose sentences and expel avoidance words from tweets they bring. A positive score is removed for each survey after the main survey. It is passed on for the second parsing and contributed nonpartisan survey. This score is changed when needed. This condition is better situational assurance and a yield document that includes the survey ID and resolves its positive score.

3.1.7 Lexicon based techniques

These process rely on judgment trees, for example, Naive Bayesian Classifier (NBC), Conditional Random Field (CRF), Single Dimensional Classification (SDC), k-Nearest Neighbors (k-NN), Sequential Minimal Optimization (SMO) and Hidden Markov Model (HMM) that are defined with techniques of conclusion colorization. Vocabulary Based methods chip away at a suspicion that the aggregate extremity of a sentence or reports is the total of polarity of a person's expressions or words. In the 2011-12 ROMIP and RCDL course the dictionary-based strategy was utilized [45, 46]. This strategy depends on passionate research for slant investigation lexicons for every space. Next, every area lexicon was renewed with evaluation expressions of suitable preparing gathering the highest weight known by the strategy for RF (Relevance Frequency) [47].

In unsupervised strategy, order is finished by looking at the highlights of a given content against assessment vocabularies whose opinion esteems are resolved preceding their utilization. Estimation dictionary contains arrangements of words and articulations used to express individuals' emotional sentiments and conclusions. For instance, begin with positive and negative word dictionaries, check the collection that requires searching. At that point if there are more positive word dictionaries in the collection, it is certain, otherwise it is negative. Dictionary-based procedures for sending checks are unsupervised learning because it is not required earlier preparing with the end goal to characterize the information. The essential strides of the vocabulary-based systems are laid out beneath:

1) Process each lesson(ie HTML tags, remove the noise character).
2) Total Text Sentence Score :: 0.
3) Interrupt the text For each token, check if it exists in a sentimental dictionary.
    a) In the event that token is available in dictionary,
        i. In the event that token is positive, at that point s ← s + w.
        ii. In the event that token is negative, at that point s ← s − w.
4) See total text sentiment score,
    a. In the event that s > edge, at that point characterize the text as positive.
    b. On the off chance that s < limit, at that point characterize the text as negative.

Construct a sentiment lexicon are in three ways: manual construction, dictionary-based methods and corpus-based methods. The manual development of the sentiment dictionary is a troublesome and tedious undertaking. The idea in lexicon-based techniques is to initially assemble a small arrangement of physically repressed words along known directions, and later develop this set by looking in the WordNet word context for their equivalent words and oppoite.

3.1.8 N-gram Sentiment Analysis

In areas of probability and phonology, N-gram is an inherited succession of things from a given set of materials or discourse. Depending on the application, things can be phone, syllable, letter, word or base set. The n-gram is usually gathered from a material or discourse corpus. When things are words, n-gram can be called herpes in the same way. In this they are thinking about the sentence all in all [48]. They are making utilization of four kinds of vocabularies to be specific conclusion express vocabulary, estimation quality dictionary, vocabulary with viewpoints and exemption lexicons.

3.1.9 Multilingual Sentiment Analysis

These days, there are options for customers to express their point of view in different dialects. To get better results, analyst ought to think about the posts in various dialects. A few inquires about [49, 50] clarified techniques inside multilingual system to do the undertaking of deciding the extremity of the content. It is finished utilizing a few NLTK (Natural Language Tool Kits). Herein, language is recognized first utilizing language models. After ID, the language is meant English utilizing standard interpretation programming.

3.1.10 Maximum Entropy Classifier

In case of Maximum Entropy Classifier, a set of weights are parameterized. This further can further be utilized for mixing the features which can be produced with the help of features through encoding. The encoding maps each match of list of capabilities and name to a vector. The ME classifier has a space that contains a system of classifiers known as exponential or log-straight classifiers. They work by extracting some system of information, directing them, and then using this overall example. On the occasion that this strategy is done in an unsafe manner. Exploring the co-occurrence of a word with positive and negative words at that point is used with the ultimate goal. The ME classifier is one of those models that does not expect autonomy highlights [51].

$$P_{ME}(c \mid d, \lambda) = \frac{\exp[\sum_{i} \lambda_i f_i(c,d)]}{\sum_{c} \exp[\sum_{i} \lambda_i f_i(c,d)]}$$

Where c is the square, d is the tweet and λi is the weight vector. Weight vectors vectors choose the significance of a component in the vectors arrangement.

Chapter Four:

PROPOSED APPROACH

4.1 The Model on social media network: Twitter

4.1.1 Twitter Sentiment Analysis

Twitter is an online informal communication administration and microblogging administration that empowers its clients to send and peruse content based messages named "tweets".Tweets are freely unmistakable naturally, however senders may message messages to a constrained group. Twitter is one of the largest microblogging administrations with over 500 million listed subscribers. Insights uncovered by the Info-designs Labs propose that on consistent schedule around a large portion of a billion tweets are imparted. There is an enormous mass of individuals utilizing twitter to express sentiments, which makes a intriguing and testing decision for sentiment examination. When so much consideration getting paid to twitter, why not screen and develop techniques to examine these sentiments [52]. Recent investigations have demonstrated [53,54] that with Twitter it is conceivable to get individuals' knowledge from their profiles rather than conventional methods for getting data about recognitions.

We zeroed on the data available on Twitter due to more authentic accounts of the people. It is assumed that the views expressed on twitter are far more honest and informative than other modes of surveys (figure 13). Also, in relation to Facebook, Twitter confines clients to give their minimized and finish assessments in 280 characters only [55].

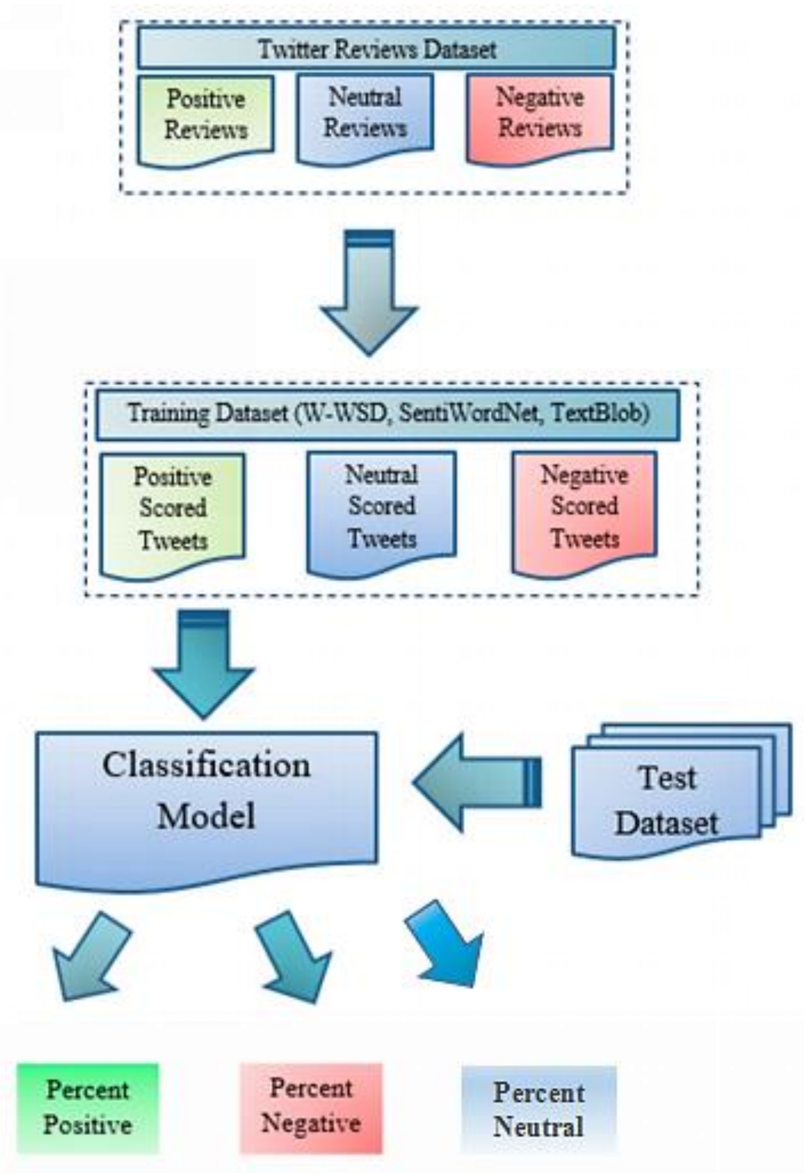


Figure 13: Flow diagram of sentiment analysis framework

4.2 Objective:

It is significant that the task of understanding the sentiment in a tweet is more complex that of any well formatted document. Tweets do not follow any formal language structure, nor do they contain words from formal language (i.e. no-vocabulary words). Often, punctuations and symbols are used to express emotions (smileys, emoticons etc.).

Classify tweets as a positive or negative sentiment aplying hybrid techniques and NLP under python program and check the performance.

In this work, we propose a hybrid approach that is a combination of natural language processing techniques. Machine learning algorithms (Naive Bayes) and a particular lexical dictionary patterns and features of the tweet predict and predict sentiment (if any). In particular, We create a computational model that can group a given tweet as positive, negative, or nonconventional that reflects sentiments. Polar tweets in a positive and negative category will communicate sentimentality. We slithered multi-sized datasets comprising of around 4 million tweets with an assortment of most well known catchphrases and so forth., for the preparation and testing purposes. We using a natural language toolkit for test the proposedhybrid nave bae approach and see that it beats the current methodologies conveying aggressive outcomes having 98.59% precision.

Twitter has following objectives which has been chosen in mind:

- It is an open access interpersonal organization.
- Twitter is an ocean of sentiments (constrained inside 280 characters, for example high sentiment thickness)
- Twitter makes the API easy to understand and API making it simpler to mine sentiments progressively.

4.3 Data collection:

Twitter enables specialists to gather tweets by utilizing a Twitter API. To obtain Twitter credentials (i.e, API secret, API key, access token secret and access token) one must have a Twitter account which can be obtained from the Twitter developer site. At that point introduce a twitter library to interface with the Twitter API. Twitter has built up its own language shows. Coming up next are instances of Twitter shows.

a. "RT" is a brief description for the retweet. That indicating the customer is releasing or reposting.

b. "#" represents hashtag is utilized to channel tweets as indicated by themes or classifications.

c. "@user1" speaks to that a message is an answer to client whose client name is "user1".

d. Colloquial and emoticons expressions or slang dialects are as often as possible used in tweets.

4.4 Proposed Machine Learning Model:

4.4.1 Hybrid Naive Bayes

The generality of both the lexical methodology (for its speed) and the AI approach (for its precision) is vague to the world. A lexical function is a direct result of preceding (ned highlights (for example dictionaries) that use it for various emotions. Having a dictionary for alluds at runtime reduces time usage faster. Literary approach Improving the list of abilities in the execution of. Certainly should be expanded, for example by classifying words with their frequencies. A vast dictionary of Bits should be given at runtime. It forms the presentation on the upper part of the framework and beyond. Su. Accordingly, there is a steady exchange oa between the performances. Time. Again, AI approaches light at its k . Uses restructuring, learning and tuning, provides vast information datasets, improves their exposure paths that in any case of any kind This allows the Ykshmta, it meets the requirements of time to learn a standout runtime performance tuning framework.

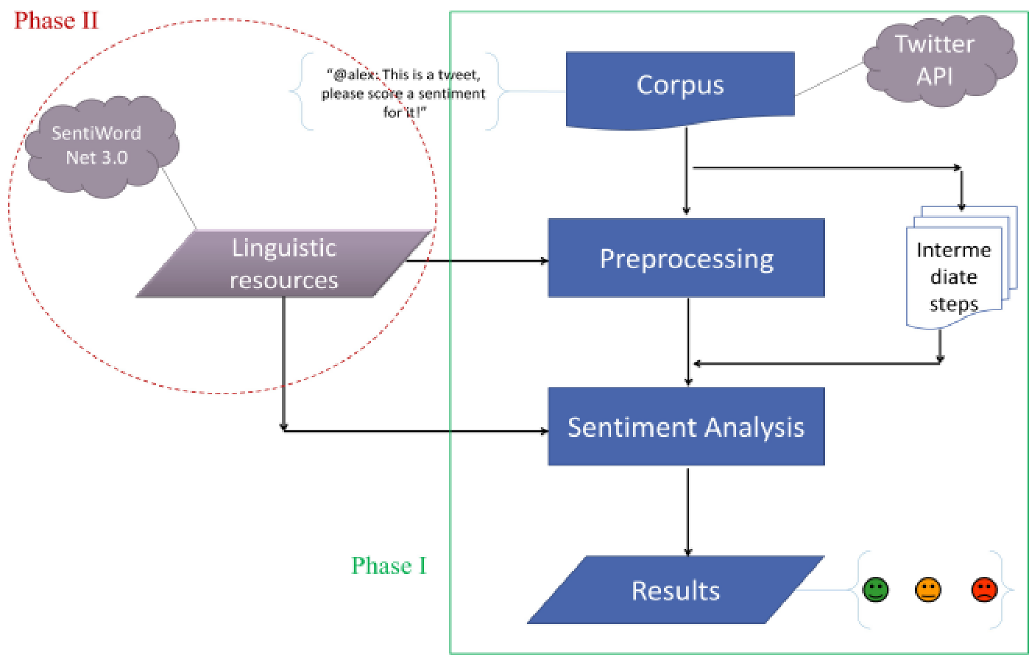


Figure 14: System architecture of hybrid naive bayes approach.

A Naive Bayesian classifier is a familiar observational technique often used for classification purpose. Fig. 14 represents a typical architecture of the hybrid nave base approach. Their classifier is named as naïve because it considers contingencies that are actually connected and not further based. Explain that d can be a tweet and c* is a category that is relegated to d, where is

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)},$$

where P(d) c plays no role in choosing c. In order to approximate the term P (d. C), Naive Bayes decomposed characteristics assuming that Fei are independently of d in the limit:

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

From the equation above, "f" is a "feature‟, the feature (Fi) count is denoted with Ni (d) and is available in d that speaks to a tweet. There, m means no. of highlights. Parameters P(c) and P(f/c) are processed from most extreme probability evaluates, and smoothing is utilized for inconspicuous highlights. To prepare and characterize utilizing Naive Bayes Machine Learning procedure, we can utilize the Python NLTK library. In spite of its spontaneity and the way its restrictive autonomy does not hold up under clearly evidenced circumstances, the nave base-based content classification will still perform as shocking in general.

4.4.2 Support Vector Machines

In event of conventional content categorization, Support vector machines have been appeared to be exceedingly effective. When all is said in done it beats other machine learning strategies. They are extensive edge, instead of probabilistic, classifiers, as opposed to MaxEnt and Naive Bayes. In the case of two-classification, the essential thought of finding a hyperplane behind the preparation system(figure 15), spoken to by vector, that not just isolates archive vectors in a those in the second single category, however for the detachment, or edge, is as vast as could reasonably be expected.

$$\vec{w} := \sum_{j} \alpha_j c_j \vec{d_j}, \quad \alpha_j \geq 0,$$

Where α j are acquired by tackling a double improvement issue. The d and j with the ultimate target α j are more support vectors than zero, since they are the main archive vectors. Examples of test classification essentially include finding which side of the hyperplane they fall on.

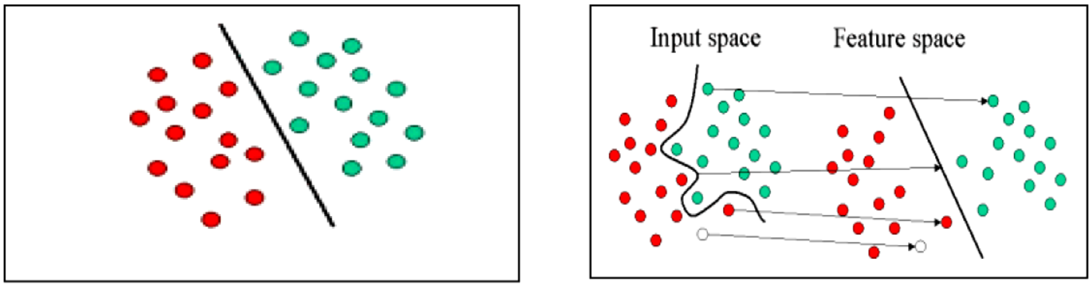


Figure 15: SVM in linear classification

Chapter Five:

CLASSIFICATION STEPS AND RESULTS

5.1 Sentiment Analysis classification

5.1.1 Document-level of sentiment analysis:

In layman understanding the opinions can be understood as the feeling or perception of an individual regarding anything. There are multiple ways to express their sentiments. It can be straightforward Yes or No. It can also be subjectively examined to find the actual meaning of sentiments. To measure the actual viewpoints, review framework can be utilized wherein on one side the rating of 4 or 5 can be assumed as yes and 1 or 2 can be assumed as No.

5.1.2 Sentence-level of sentiment analysis:

This strategy is used to give valuable information when we seek on the grounds that the extremity of sentence will be made flawless. In this level of conclusion investigation experience Sentences that contain evaluations and give audits as if it is negative or positive.

5.2 Preprocessing

The tweet submitted from Twitter contains a mix of URLs, and other non-sentimental information such as the hashtag "#", explanations "@" and retweet "RT". To get the n-gram highlight, we should pause the content information. Tweets for standard tokens made for formal and general content represent a problem. The accompanying figure16 shows different moderate handling highlight steps. Figure 17 shows block diagram of the classification based on dictionary.

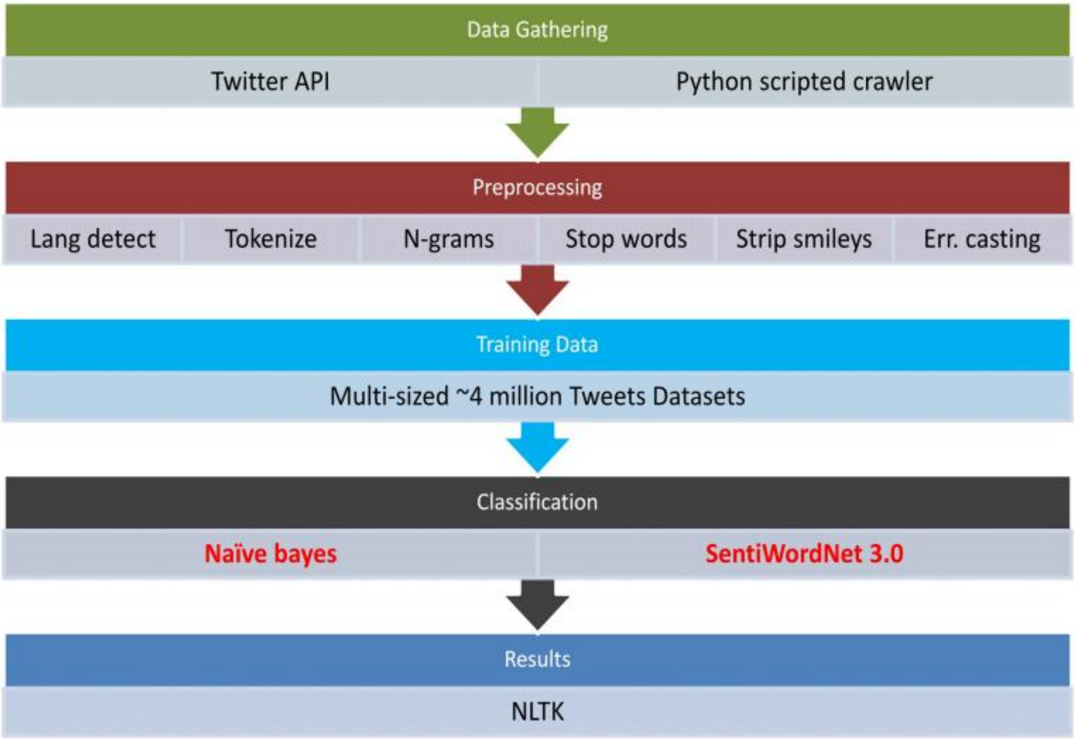


Figure 16: Process steps followed by hybrid Naive Bayes.

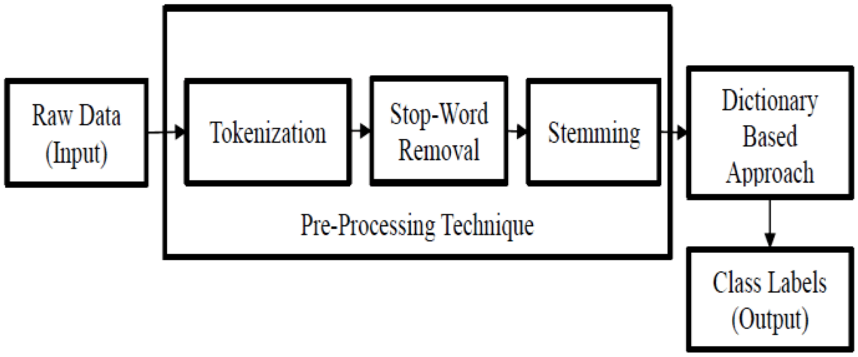


Figure 17: Classification process using Dictionary Based Approach

5.3 Data collection:

We need Twitter data for classification and training for Twitter API classification. For this reason, we use Twitter provided API's. Twitter gives two API's; REST and stream API. The API columns between the REST API and the stream API are:

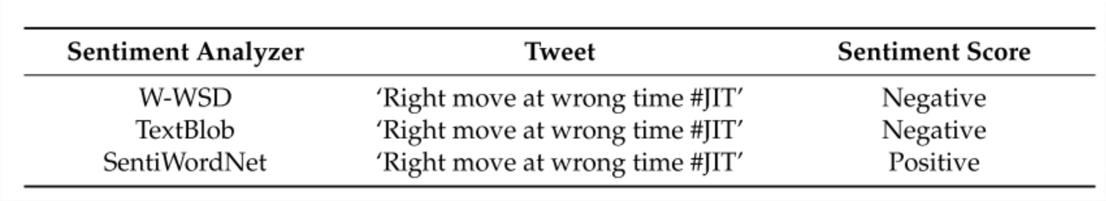
1. REST APIs enable access to Twitter information, for example, notices and client data paying little mind to time. Be that as it may,Twitter does not establish a week or more of accessible information. Accordingly, REST get to is restricted to information Twittered not before over seven days. In this manner, while REST API enables access to these amassed information, Streaming API empowers access to information as it is being twittered.
2. Streaming API bolsters seemingly perpetual association and gives information in practically continuous. The REST APIs bolster fleeting associations and are rate-restricted (any can download a specific measure of information [Per hour 150 tweets] yet not more everyday).

5.4 Polarity Calculation and Sentiment Analysis

There is a very large set of data available resembling the emotions of people on internet that too in a much unstructured format. In such scenario, analysis provided by Sentiment analysis will be very helpful. In this analysis, three classes are utilized one is negative, second is positive and the last is neutral. All the viewpoints are judged by giving the score in the range of −1 to 1. Consistently, all the negative words are given -1 score whereas all the neutral words are given 0 and all the positive are given + 1. A score of subjectivity appointed to every statement depends on what it is speaking to A objective meaning or a subjective significance; Subjective score scope additionally ranges from 0 to 1 where an incentive speaks for an objective close to 0 and close to 1 person. For distinguishing Extremism of political audits and individualism, and in order to give an accurate perspective of the most pricise analyzer for extensibility and subjectivity adding machine, we utilized TextBlow and the SentiWordNet analyzer.

Table 2 shows the results of the three thinker analyzer's test in which we can see the skewness imposed by every analyst. Capacity 3 recorded beneath figures the subjectivity and extremity of the handled tweets using every evaluation analyzer (SentiWordNet, W-WSD, TextBlob) with the use of Python code.

Table 2: Comparison of sentiment scores of different analyzers.



Sentiment Analyzer	Tweet	Sentiment Score
W-WSD	'Right move at wrong time #JIT'	Negative
TextBlob	'Right move at wrong time #JIT'	Negative
SentiWordNet	'Right move at wrong time #JIT'	Positive

For the model structure, we implemented a supervised machine-learning algorithm, nave base, on the preparation dataset. Analysts' approval steps through Knave Bay can be seen in the area below. These AI calculations (half and half Naïve Bayes) were applied on the preparation set to fabricate an investigation model. Given the developed model for each analyst, the test set was evaluated. After the test set evaluation, we recorded another analyst's accuracy under each model.

5.5 Naive Bayes Classifier Execution

The first step is to create the data files of the classifier wherein the below mentioned process need to be done.

1) Tweet file will be created which will have the sentiment analyzer.

2) For every analyzer, a model is created with the help of training set file.

Once first step is complete, the second step is to verify the model on a test set wherein below mentioned process need to be done.

1) we load the test set file first.

2) Apply string to the word vector with the letter with the following parameters: Stemmer: Snowball Stemmer, TF Transform: true, IDF Transform: true, Stop words Handler: rainbow, Tokenizer: Word Tokenizer,

3) Finally, the model executed on the test set.

4) In the end we save outcomes in the output file.

Once we come up with new words, they are attached in seed-list. There upon we go for another iteration. It ends only we are not left with any word. [8]. opinion words share indistinguishable introduction from their equivalent words and inverse introductions as their antonyms.

5.6 Experimental Setup

5.6.1 Recommended

Operating system Windows XP/7/8/10, Linux (Ubuntu 12.04 or above)
Processor type C2D/i3/i5/i7 (32/64 bits)
Min. Memory (RAM) ≥ 4 GB
HDD space >20 GB
Bandwidth HSI (1Mbps connection)
Third party software component NLTK 2.0, and SentiWordNet 3.0
Python (v2.7 or above) - (programming language implementation):

Python is a broadly useful, deciphered high-level programming language whose plan reasoning accentuates code clarity. Its sentence structure is clear and expressive. Python has a huge and far reaching standard library and in excess of 25 thousand expansion modules. We use python to build the backend of the test application. This and different modules executed are talked about later.

NLTK (3.3) - (Natural Language Processing Toolkit) validation and language processing modules:

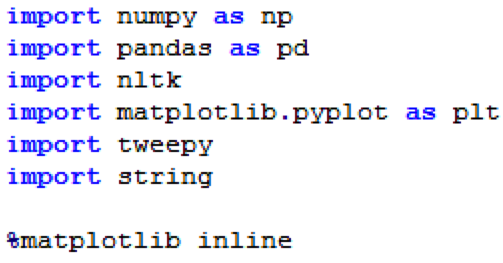
NLTK [56] is an open source language preparing module of human language in python. Made in 2001 as a piece of computational semantics course in the Department of Computer and Information Science at the University of Pennsylvania. NLTK gives inbuilt help to simple to-utilize interfaces more than 50 lexicon corpora.

SentiWordNet 3.0:

SentiWordNet is a literal asset to the mining feel. SentiWordNet concludes on every single point of WordNet's three conclusions: positivity, antisemitism, objectivity. It divides English words into sets of similar words called "syntets", gives short, common definitions, and records the various semantic relationships between these equivalent word sets. The reason for the current is twofold: creating a mixture of word context and thesaurus that is all the more inherently usable, and examining programmed content and helping man-made logic applications.

WordNet: WordNet [57,58]

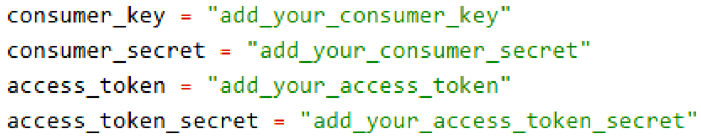
WordNet is a large lexical database of English. Things, action words, descriptive words, and intensities are assembled together in a set of psychological equivalent words (synonyms). Each communicates a different concept. Syntates are interconnected by methods for applied meaning and lexical relations. A later system of words and ideas that are actually related with the program can be traced. WordNet is likewise openly and freely accessible for download. WordNet's structure makes it a valuable tool for computational derivation and preparation of specific languages. WordNet externally takes after a thesaurus, in which it depends on their implication to the words. Install any of the few packages to import the Twitter APIs: Tweepy, Tkinter, Textblob. Nltk (natural language toolkit), and matplotlib (for plotting the results on graph). On linux OS these can be installed using the 'pip' command.



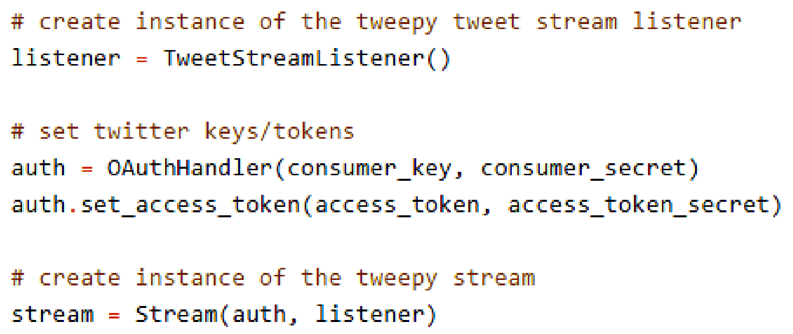
```
import numpy as np
import pandas as pd
import nltk
import matplotlib.pyplot as plt
import tweepy
import string

%matplotlib inline
```

Twitter Streaming API

To access the Twitter Streaming API, you have to enroll an application at http://apps.twitter.com. Once made, you ought to be diverted to the consumer key and consumer secret and create an access token under the "Key and Access Token" tab. Add these to another document called config.py:



```
consumer_key = "add_your_consumer_key"
consumer_secret = "add_your_consumer_secret"
access_token = "add_your_access_token"
access_token_secret = "add_your_access_token_secret"
```



```
# create instance of the tweepy tweet stream listener
listener = TweetStreamListener()

# set twitter keys/tokens
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

# create instance of the tweepy stream
stream = Stream(auth, listener)
```

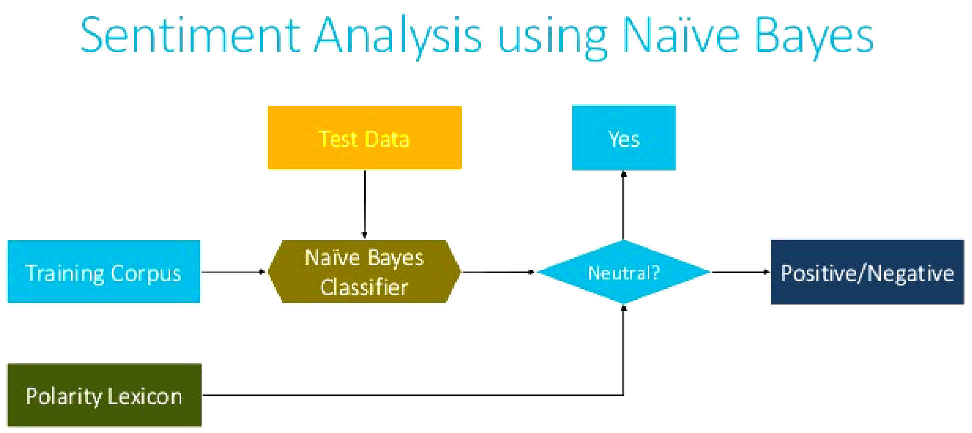


Figure 15: Flowchart of sentiment analysis architecture (using Naive Bayes Classifier)

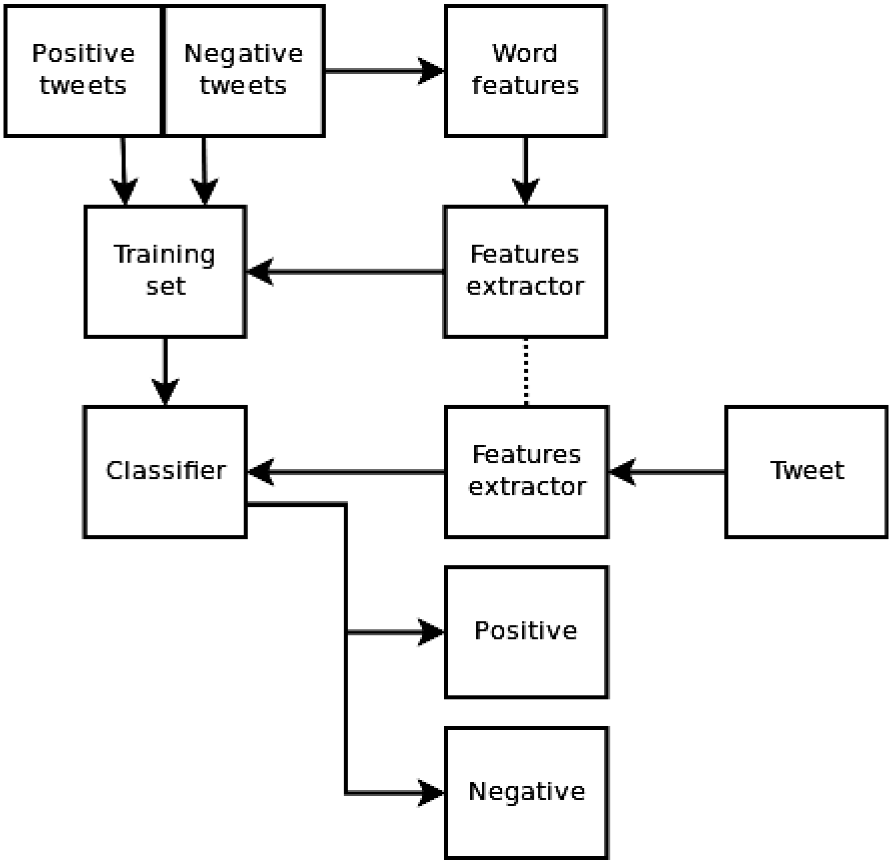
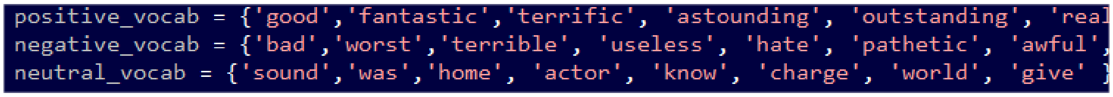


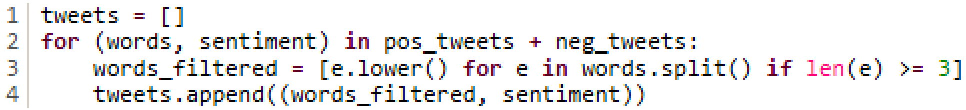
Figure 16: Block diagram for Twitter sentiment analysis

Defining the vocabulary: positive, negative and neutral



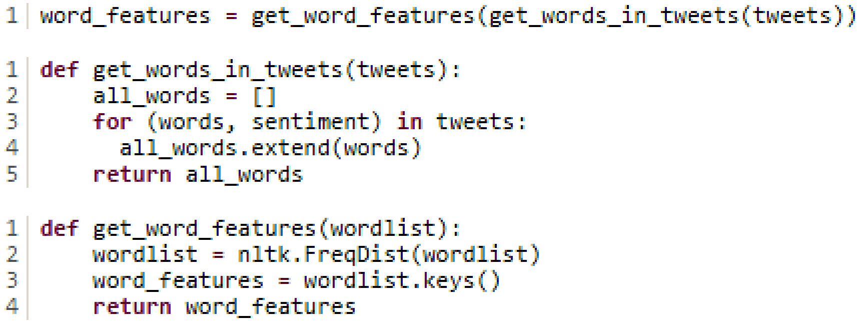
```
positive_vocab = {'good','fantastic','terrific', 'astounding', 'outstanding', 'real
negative_vocab = {'bad','worst','terrible', 'useless', 'hate', 'pathetic', 'awful',
neutral_vocab = {'sound','was','home', 'actor', 'know', 'charge', 'world', 'give' }
```

Eliminating words smaller than 2 characters

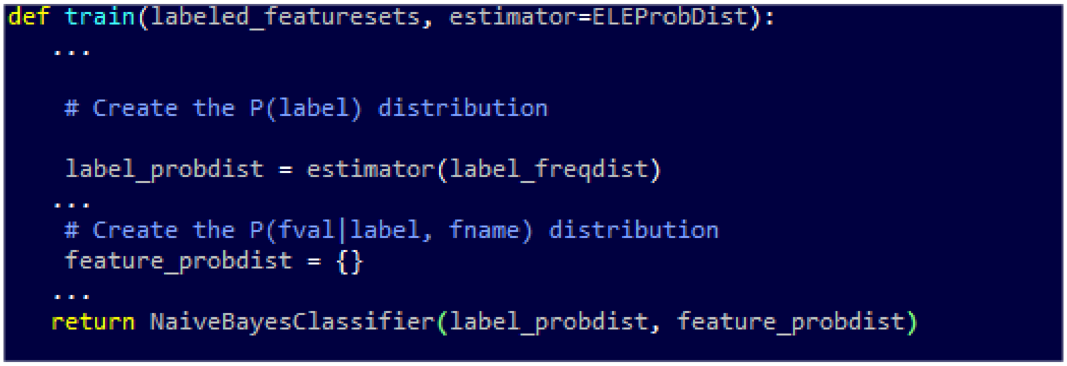


```
tweets = []
for (words, sentiment) in pos_tweets + neg_tweets:
    words_filtered = [e.lower() for e in words.split() if len(e) >= 3]
    tweets.append((words_filtered, sentiment))
```

Function to calculate the frequency of word occurrence



```
word_features = get_word_features(get_words_in_tweets(tweets))

def get_words_in_tweets(tweets):
    all_words = []
    for (words, sentiment) in tweets:
      all_words.extend(words)
    return all_words

def get_word_features(wordlist):
    wordlist = nltk.FreqDist(wordlist)
    word_features = wordlist.keys()
    return word_features
```



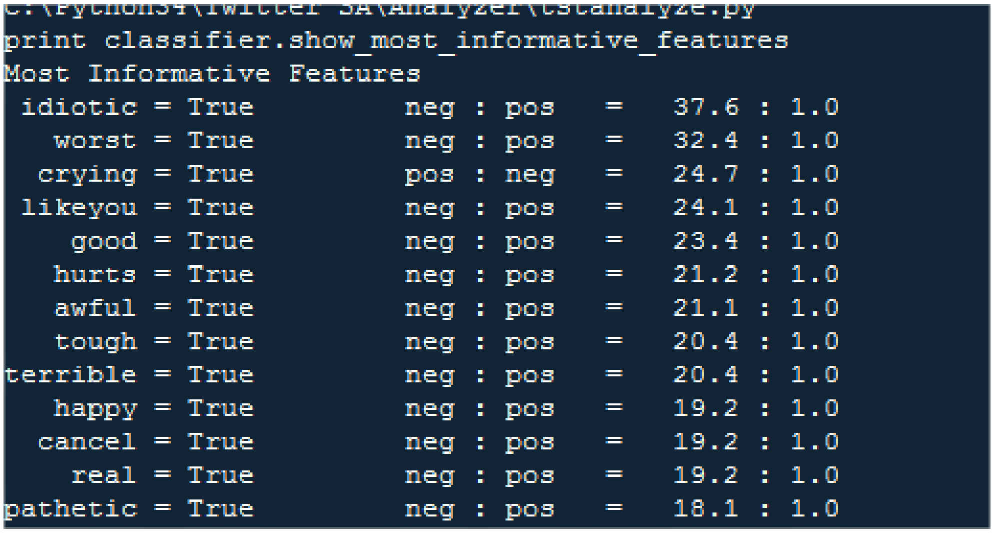
```
def train(labeled_featuresets, estimator=ELEProbDist):
   ...

    # Create the P(label) distribution

    label_probdist = estimator(label_freqdist)
   ...
    # Create the P(fval|label, fname) distribution
    feature_probdist = {}
   ...
   return NaiveBayesClassifier(label_probdist, feature_probdist)
```

Results SnapShot

Table 3: Comparison of performances of various analysis methods

	Method	Dataset	Accuracy	Author
Machine Learning	CoTraining SVM	Twitter	82.52%	Liu [59]
	SVM	Movie reviews	86.40%	Pang, Lee [61]
	Deep learning	Stanford Sentiment Tree-bank	80.70%	Richard [60]
Lexical based	Corpus	Product reviews	74.00%	Turkey
	Dictionary	Mechanical Turk (Amazon)	N/A	Taboada [62]
Cross-lingual	Ensemble	Amazon	81.00%	Wan, X. [63]
	Co-Train	Amazon, ITI68	81.30%	Wan, X.

	EWGA	IMDb movie review	>90%	Abbasi,A.
	CLMM	MPQA, N TCIR, ISI	83.02%	Meng [64]
Cross-domain	Active Learning	Book, DVD, Electronics, Kitchen	80% (avg)	Li, S
	Thesaurus			Bollegala [65]
	SFA			Pan S J[15]
Proposed model (Machine learning, Lexical based)	Naïve Bayes NLP	Twitter Movie reviews	92.18%	



```
c:\Python34\Twitter SA\Analyzer\tsanalyze.py
print classifier.show_most_informative_features
Most Informative Features
 idiotic = True          neg : pos   =   37.6 : 1.0
   worst = True          neg : pos   =   32.4 : 1.0
  crying = True          pos : neg   =   24.7 : 1.0
 likeyou = True          neg : pos   =   24.1 : 1.0
    good = True          neg : pos   =   23.4 : 1.0
   hurts = True          neg : pos   =   21.2 : 1.0
   awful = True          neg : pos   =   21.1 : 1.0
   tough = True          neg : pos   =   20.4 : 1.0
terrible = True          neg : pos   =   20.4 : 1.0
   happy = True          neg : pos   =   19.2 : 1.0
  cancel = True          neg : pos   =   19.2 : 1.0
    real = True          neg : pos   =   19.2 : 1.0
pathetic = True          neg : pos   =   18.1 : 1.0
```

Figure 17: Hybrid Naive Bayes feature accuracy on Twitter® dataset

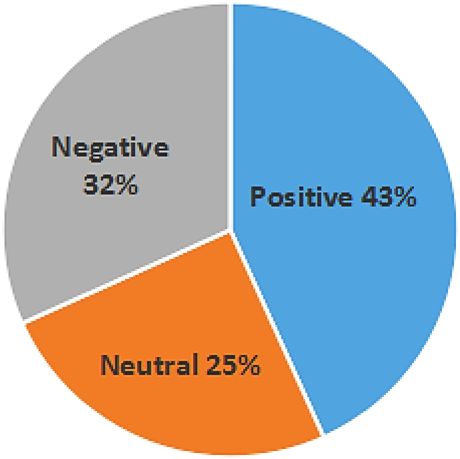


Figure 20: Result of Twitter sentiment analysis

Chapter Six:

CONCLUSION AND FUTURE SCOPE

In this theory, we have used oversein machine learning methods. For example, Naive Bayes classifier, Natural Language Processing and the Lexicon-based hybrid nave base classifier, to learn the sentiment of sentences in tweets. Validated outcomes have been found of around 4,000 tweets on twitter dataset. The performance results effectively show that hybridizing the current machine learning examination and lexical investigation strategies for assumption characterization yields similarly outflanking precise outcomes.

For every single data set utilized, we consistently recorded accuracy ≥ 92% - 94%. Clearly from the accomplishment of machine learning system using hybrid Naive Bayesian classifier, it is positively applied to other related opinion analysis applications such as business protocols, customer feedback services, financial sentiment analysis (stock market opinion mining) and product surveys etc.

Chapter Seven:

REFERENCES

BIBLIOGRAPHY

[1] H. Kaur and V. Mangat, "A survey of sentiment analysis techniques.," In IEEE, 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 921-925, 2017.

[2] J. Sankaranarayanan, H. Samet, Teitle and M. Li, "News in tweets," In Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems ACM, pp. 42-51, 2009.

[3] Montoyo, Andrés, P. MartíNez-Barco and A. Balahur, "Subjectivity and sentiment analysis," An overview of the current state of the area and envisaged developments., pp. 675-679, 2012.

[4] G. Hochmuth; , G. Magoulas, B. Lorica and S. Milstein, "Twitter and the micro-messaging revolution," Communication, connections, and immediacy--140 characters at a time, 2008.

[5] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis," In LREC, Twitter as a corpus for sentiment analysis and opinion mining., pp. 1320-1326, 2010.

[6] B. Liu, Y. Dai, X. Li, W. Lee and P. Yu, "Building text classifiers using positive and unlabeled examples.," In Third IEEE International Conference on Data Mining, (ICDM-2003),2003; 179-186, pp. 179-186, 2003.

[7] R. Jain, K. Chang; , S. Hoi and G. Li, "Micro-blogging sentiment detection by collaborative online learning.," In IEEE 10th International Conference on Data

Mining (ICDM-2010), pp. 893-898, 2010.

[8] Wawre, S. V., S. N and Deshmukh, "Sentiment classification using machine learning techniques," International Journal of Science and Research (IJSR), pp. 819-921, 2016.

[9] L. Huang, R. Bhayani; and A. Go , "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford., 2009.

[10] P. Paroubek and A. Pak, "Twitter as a corpus for sentiment analysis and opinion mining.," In LREc, pp. 1320-1326, 2010.

[11] A. Narayanan, I. Arora and A. Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model," In International Conference on Intelligent Data Engineering and Automated Learning, Springer, Berlin, Heidelberg, pp. 194-201, 2013.

[12] A. Tumasjan, T. Sprenger, P. Sandner and I. Welpe, "Predicting elections with twitter," ICWSM, pp. 178-85, 2010.

[13] J. Bollen, H. Mao and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.," ICWSM, pp. 450-3., 2011.

[14] X. Hu, Tang J, H. Gao and Liu, "Unsupervised sentiment analysis with emotional signals," In Proceedings of the 22nd international conference on World Wide Web, ACM., pp. 607-618, 2013.

[15] R. Jose and V. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using word sense disambiguation.," In 2015 IEEE International Conference on Control Communication & Computing India, pp. 638-641, 2015.

[16] T. Apps, "http://www.tweepy.org/," [Online]. Available: http://www.tweepy.org/

(accessed 2018)..

[17] A. Hasan, S. Moin, A. Karim and S. Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts.," Mathematical and Computational Applications, p. 11, 2018.

[18] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis.," In IEEE 24th International Conference on Data Engineering Workshop, 2008. (ICDEW), pp. 507-512, 2008.

[19] S. Baccianella, A. Esul and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.," LREC, pp. 2200-2204, 2010.

[20] G. Miller, "WordNet: An electronic lexical database.," MIT press, 1998.

[21] M. Ibrahim, O. Abdillah, A. Wicaksono and M. Adriani, "Buzzer detection and sentiment analysis for predicting presidential election results in a Twitter nation.," In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1348-1353, 2015.

[22] J. Fang and B. Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification," InProceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), pp. 94-100, 2011.

[23] L. Zhang, R. Ghosh and M. Dekhil, "Combining lexicon based and learning-based methods for twitter sentiment analysis.," HP Laboratories, 2011.

[24] S. Trinh, L. Nguyen and M. Vo, "Combining Lexicon-Based and Learning-Based Methods for Sentiment Analysis for Product Reviews in Vietnamese Language.," In International Conference on Computer and Information Science, Springer, Cham, pp. 57-75, 2017.

[25] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: a survey of

techniques," arXiv preprint, p. 1601.06971, 2016.

[26] D. Alessia, F. Ferri, P. Grifon and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," International Journal of Computer Applications, p. 125, 2015.

[27] N. Zainuddin and A. Selamat, "Sentiment analysis using support vector machine.," In International Conference on Computer, Communications, and Control Technology (I4CT 2014), pp. 333-337, 2014.

[28] A. Sarlan, C. Nadam and S. Basri, "Twitter sentiment analysis," n IEEE International Conference on Information Technology and Multimedia (ICIMU 2014), pp. 212-216, 2014.

[29] R. Parikh and M. Movassate, "Sentiment analysis of user-generated twitter updates using various classification techniques.," CS224N, 2009.

[30] A. Yeole, P. Chavan and M. Nikose, "Opinion mining for emotions determination," IEEE, pp. 1-5, 2015.

[31] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends® in Information Retrieval, pp. 1-35, 2008.

[32] L. YM and L. TY, "Deriving market intelligence from microblogs," Decision Support Systems, pp. 206-17, 2013.

[33] D. Bollegala, D. Weir and J. Carroll, "Cross-domain sentiment classification using a sentiment sensitive thesaurus," IEEE transactions on knowledge and data engineering, vol. 25, no. 8, pp. 1719-31, 2013.

[34] G. Miler, "WordNet: a lexical database for English.," Communications of the ACM., pp. 39-41, 1995.

[35] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using

machine learning techniques.," in In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 2002.

[36] A. Pak and P. Paoubek, "Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives.," In 2010, Proceedings of the 5th International Workshop on Semantic Evaluation (IWSE), Association for Computational Linguistics., pp. 436-439, 2010.

[37] J. Almeida and G. Pappa, "witter population sample bias and its impact on predictive outcomes: A case study on elections.," In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015.

[38] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis.," In IEEE 24th International Conference on Data Engineering Workshop, (ICDEW 2008)., pp. 507-512, 2008.

[39] X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu and H. Wang, "Cross-lingual mixture model for sentiment classification.," in nProceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012.

[40] X. Wan, "A comparative study of cross-lingual sentiment classification.," in InProceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society., 2012.

[41] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-based methods for sentiment analysis.," in Computational linguistics, 2011.

[42] J. Wu, J. Chuang, C. Manning, A. Ng, C. Potts, R. Socher and A. Perelygin, "Recursive deep models for semantic compositionality over a sentiment treebank.," in In Proceedings of the 2013 conference on empirical methods in natural language processing, 2013.

[43] C. Newton, "Twitter just doubled the character limit for tweets to 280.," The Verge., 2017.

[44] B. Badar, W. Kegelmeyer and P. Chew, "Multilingual sentiment analysis using latent semantic indexing and machine learning.," In IEEE 11th International Conference on Data Mining Workshops (ICDMW, 2011)., pp. 45-52, 2011.

[45] '. Z. I, P. Saloun, M. Hruzik and I. Zelinka, "Sentiment analysis, e-bussines and e-learning common issue.," In IEEE 11th International Conference on Emerging e-Learning Technologies and Applications (ICETA, pp. 339-343, 2013.

[46] M. Lan, C. Tan, J. SU and Y. LU, "Supervised and traditional term weighting methods for automatic text categorization," IEEE transactions on pattern analysis and machine intelligence, pp. 721-35, 2009.

[47] M. Klekovkina and E. Kotelnikov, "The automatic sentiment text classification method based on emotional vocabulary," Digital libraries: advanced methods and technologies, digital collections (RCDL-2012), pp. 118-23, 2012.

[48] I. Chetviorkin, P. Braslavskiy and N. Loukachevich, "Sentiment analysis track at ROMIP 2011," Dialog, 2012.

[49] S. Jiang, G. Pang, M. Wu and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," Expert Systems with Applications, pp. 1503-9, 2012.

[50] T. Koomsubha and P. Vateekul, "A study of sentiment analysis using deep learning techniques on Thai Twitter data.," In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE,, pp. 1-6, 2016.

[51] N. Joshi and S. Itkat, "A survey on feature level sentiment analysis.," International Journal of Computer Science and Information Technologies., pp. 5422-5, 2014.

[52] A. Maas, R. Daly, P. Pham, D. Huang and A. Ng, "Learning word vectors for sentiment analysis.," in n Proceedings of the 49th annual meeting of the association

for computational linguistics: Human language technologies, Association for Computational Linguistics., 2011.

[53] Shinde, Dinkar, Pooja and S. Rathod, "A Comparative Study of Sentiment Analysis Techniques.," IEEE, 2018.

[54] M. Devika, C. Sunitha and A. Ganesh, "Sentiment analysis: A comparative study on different approaches.," Procedia Computer Science., pp. 44-9, 2016.

[55] S. Vohra and J. Teraiya, "A comparative study of sentiment analysis techniques," Journal JIKRCE., pp. 313-7, 2013.

[56] F. Luo, C. Li and Z. Cao, "Affective-feature-based sentiment analysis using SVM classifier," In 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 276-281, 2016.

[57] M. Kamran, A. Noureen, A. Riaz, M. Ali and Q. Ain, "Sentiment analysis using deep learning techniques: a review," Int J Adv Comput Sci Appl, p. 424, 2017.

[58] D. Kang and Y. Park, "Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach.," Expert Systems with Applications., pp. 1041-50, 2014.

[59] H. Rui, Y. Liu and A. Whinston, "Whose and what chatter matters? The effect of tweets on movie sales.," Decision Support Systems, pp. 863-70, 2013.

[60] P. Gamallo, M. Garcia and Citius, "A naive-bayes strategy for sentiment analysis on english tweets.," In Proceedings of the 8th international Workshop on Semantic Evaluation (SemEval 2014), pp. 171-175, 2014.

[61] A. Tsakalidi, S. Sioutas, N. Nodarakis and A. Kanavos, "Large scale implementations for twitter sentiment classification.," Algorithms., p. 33, 2017.

[62] E. Looper, E. Klein and S. Bird, "Natural language processing with Python:

analyzing text with the natural language toolkit.," O'Reilly Media, Inc., 2009.

[63] C. Fellbaum, "A semantic network of English verbs. WordNet: An electronic lexical database.153," IEEE, pp. 153-78, 1998.

[64] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.," in In Proceedings of the 42nd annual meeting on Association for Computational Linguistics, 2004.

[65] H. Shen, X. Cheng, F. Li, F. Li and S. Liu, "Adaptive co-training SVM for sentiment classification on tweets.," in In Proceedings of the 22nd ACM international conference on Information & Knowledge Management, ACM., 2013.