SENTIMENTAL ANALYSIS OF MOVIE REVIEWS USING MACHINE LEARNING

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

Riya Dhiman -151429

Under the supervision of

(Dr. Amit Kumar)

to

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

CERTIFICATE

I hereby declare that the work presented in this report entitled **" Sentimental Analysis of Movie Reviews using Machine Learning"** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from January to May 2019 under the supervision of **Dr. Amit Kumar, Assistant Professor(Senior Grade).**

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Riya Dhiman, 151429

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Amit Kumar

Assistant Professor

Computer Science and Engineering

Dated-9/05/2019

ACKNOWLEDGEMENT

It is with profound gratitude that I express our deep indebtedness to our supervisor, **Dr. Amit Kumar ,Assistant Professor (Senior Grade)** without whose support and guidance it would not have been possible for us to successfully implement our project. His readiness for consultation at all the times, his educative comments, his concern and assistance even with practical things have been invaluable to us.

We are also highly grateful to all other staff members of the Department, Computer Science of Engineering for providing us the necessary opportunities for the completion of our project and owe our debt to them for their invaluable help and guidance.

Riya Dhiman, 151429

Table of contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Sentimental Analysis is a new variant in the research area. It is basically refers to as opinions or views of the different data that is being collected using surveys , comments and reviews over the web. .The data set we have used in our project from imdb movie rating. Through this we can classify the review as positive or negative, large amount of data is being generated daily. This will determine the polarity of reviews and these reviews will be classified as two types positive and negative.

CHAPTER-1

INTRODUCTION

1.1 What is Sentimental Analysis?

Sentiment Analysis deals with interaction between machine such as computers and natural languages used by human beings in short we can say training machine in accordance to the problem statement We can derive high quality information by using simple text entered by the user through different patterns and trends leading towards the output we expect by the various evaluations and interpretations, thus we categorize and cluster our text with in respect to the problem statement we are dealing with. After that we compute our output by using the data set we have using different models, algorithms, mathematical computations with comes under the category of computational linguistics. This field is mainly applied when we have to take reviews or survey from our customers on products or services. Sentimental Analysis is a new variant in the research area. It is basically refers to as opinions or views of the different data that is being collected using surveys , comments and reviews over the web.

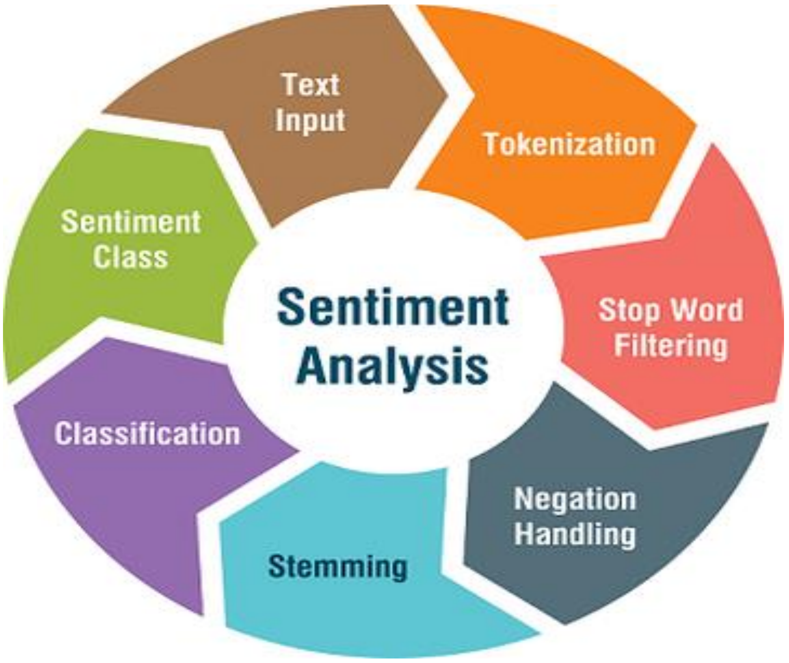


Fig 1.1 Sentimental Analysis

Large amount of data is being generated daily which is processed by using natural language processing, text analysis and computational linguistics. Opinion mining and sentiment analysis has been rapidly growing and exploring the views on different platforms through machine learning algorithms , techniques and tools. This research paper will determine the polarity of reviews and these reviews will be classified as two types positive and negative. Reviews of IMDB is used as source data and different algorithms like Naïve Bayes, K nearest neighbour, SVM(support Vector machines) and Ensemble classifier on different data streams has been provided.The reviews can be polar or neutral based on the how the customer felt by consuming that product or service.

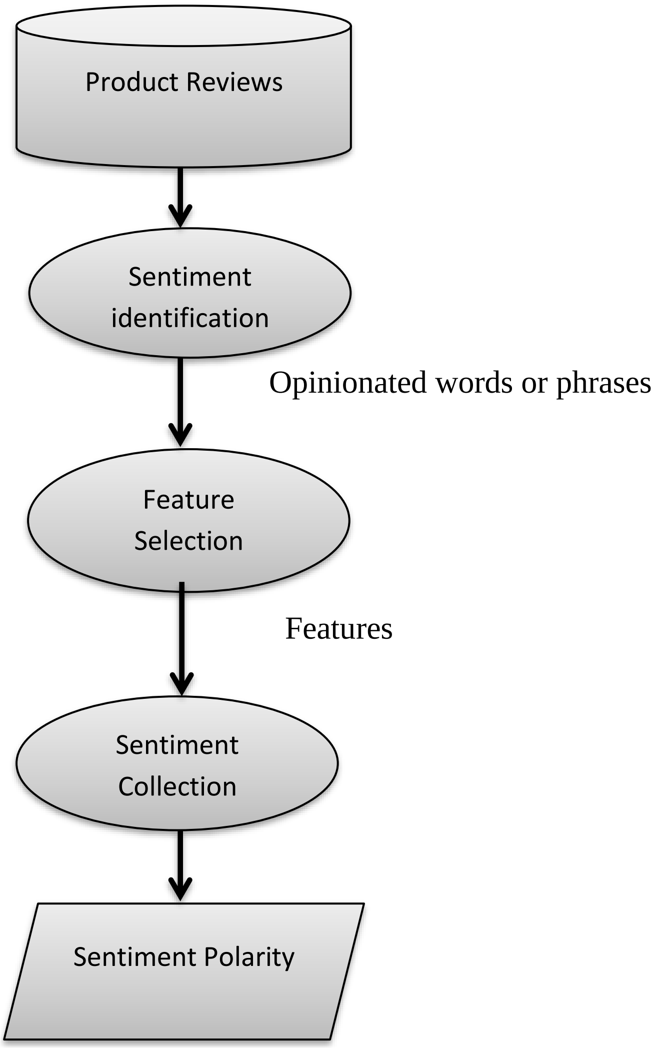


Fig 1.2 Process of determining polarities I

It also determines the attitude of a person on some topic or his\her emotional reaction about some incident, documents or events. Sentiment Analysis works on the principle of Machine learning, so we took the different concepts and algorithms of machine learning, tried joining them so that we can reach the destination of our project. With the increase in technology, various social media platforms such as twitter, facebook, instagram, linked in and many other

These platforms contain huge amount of data being generated daily in the form of tweets, blogs, posts, status etc. Sentimental analysis predicts the mood of these texts, tweets, reviews or posts which are available online on the platform by determining the polarity of emotions like happiness, affection, grief, anger and hatred.

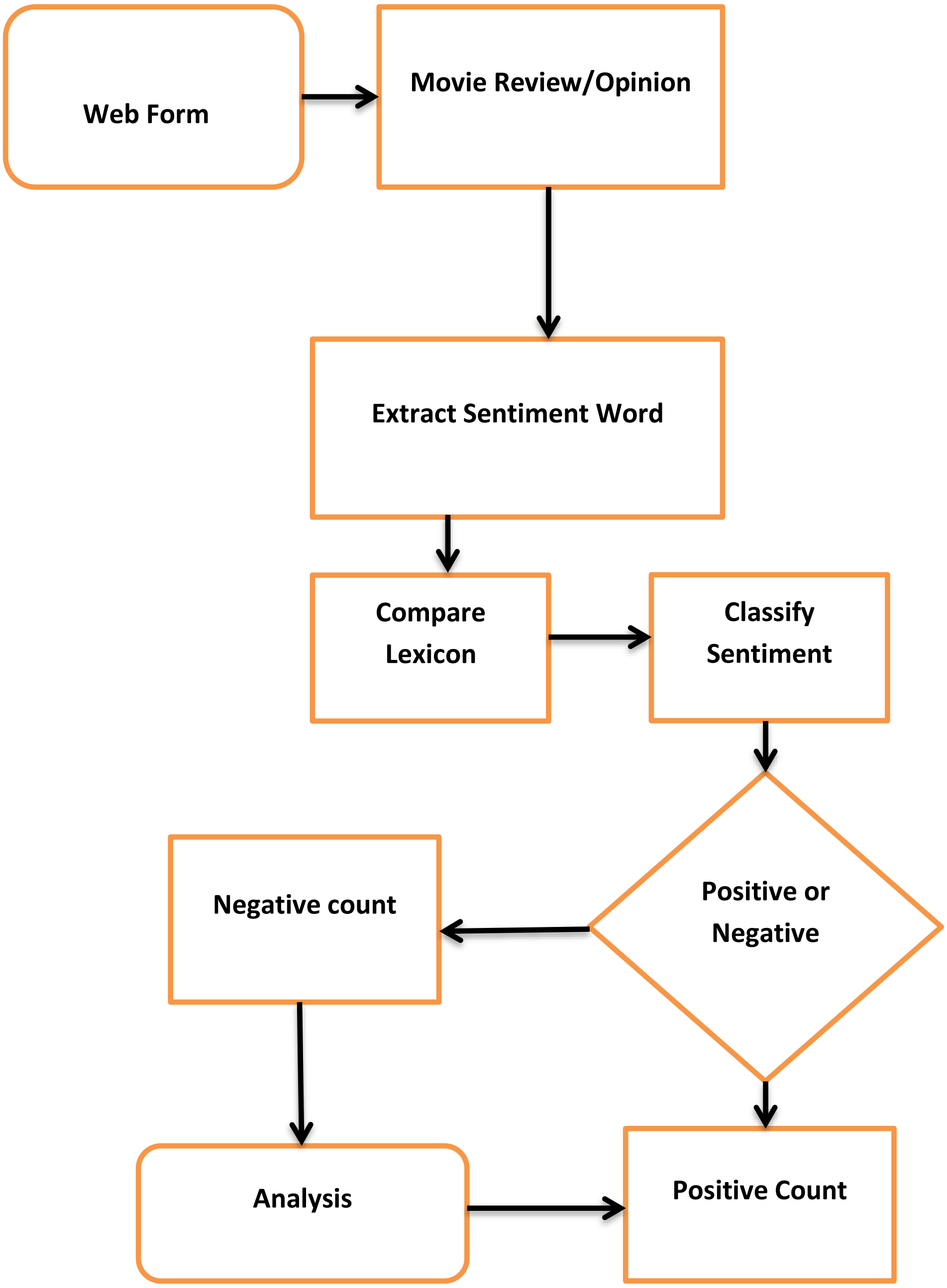


Fig 1.3 Process of determining polarities II

However, this task is not easy as people do not always express in the same way. Comments and reviews differ from person to person in terms of their regional languages, internet slangs, emoticons. Sentimental analysis is mainly concerned with identification and classification of opinions. It is broadly classified into two types first is knowledge based approach and other one is using machine learning techniques. Using first approach, it requires large database of predefined.

Fig 1.4 Polarity of Reviews

Emotions and an efficient knowledge representation for recognising opinions. But using machine learning techniques, we have train data and test data which will be used as a data set to develop a classifier. It is quite simpler also.Another task in sentiment analysis is subjectivity/objectivity identification where it focuses on classifying a given text (usually a sentence) into one of the two classes (objective or subjective). As the subjectivity of words and phrases may depend on their context and an objective document may contain subjective sentences (a news article quoting people's opinions), this problem can sometimes be more difficult than polarity classification.

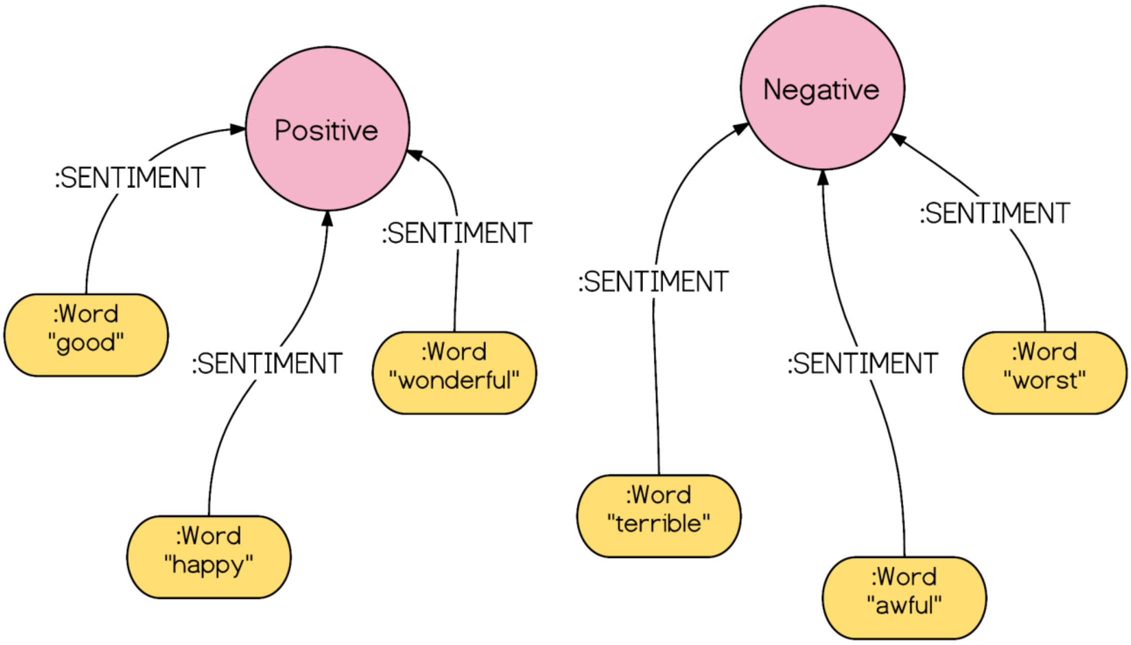


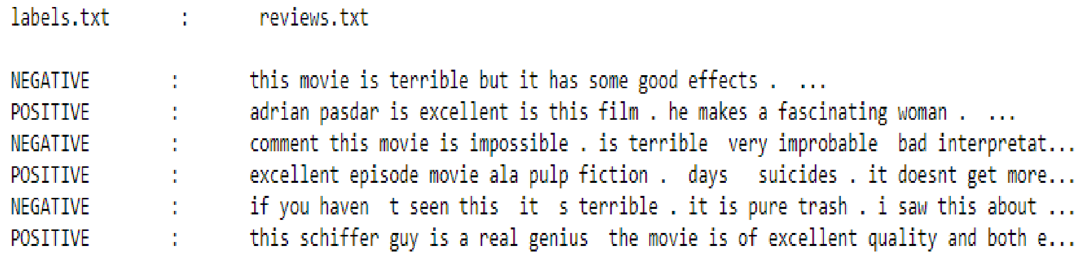
Fig 1.5 Determining Polarities

Existing approaches to sentiment analysis can be grouped into four main categories. They are keyword spotting, lexical affinity, statistical methods, and concept-level techniques. Keyword spotting classifies text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored. Lexical affinity improves keyword based approach by considering not only obvious affect words. It also assigns arbitrary words a probable “affinity” to particular emotions. Statistical methods influence on elements from machine learning such as latent semantic analysis, support vector machines, bag of words and Semantic Orientation. Unlike above mentioned purely syntactical techniques, concept-level approaches leverage on elements from knowledge representation such as ontologies and semantic networks so that they also able to detect semantics that are expressed in a subtle manner i.e. through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts that do so.

University of Stanford has proposed a novel approach of sentiment analysis. Most of the conventional sentiment prediction systems work just by looking at words in isolation, giving positive points for positive words and negative points for negative words and then summing up these points. In that approach, the order of words is ignored and important information is lost.

1.2 Problem Statement:

By using the concept of machine learning we want to get the reviews of various users about the various activities on social media and then categorize these reviews in accordance with the polarity, we want to know whether the attitude of the user for a particular issue is positive, negative and see whether the machine is able to detect that polarity and evaluate the correct output.



```
labels.txt      :       reviews.txt

NEGATIVE        :       this movie is terrible but it has some good effects .  ...
POSITIVE        :       adrian pasdar is excellent is this film . he makes a fascinating woman .  ...
NEGATIVE        :       comment this movie is impossible . is terrible  very improbable  bad interpretat...
POSITIVE        :       excellent episode movie ala pulp fiction .  days   suicides . it doesnt get more...
NEGATIVE        :       if you haven  t seen this  it  s terrible . it is pure trash . i saw this about ...
POSITIVE        :       this schiffer guy is a real genius  the movie is of excellent quality and both e...
```

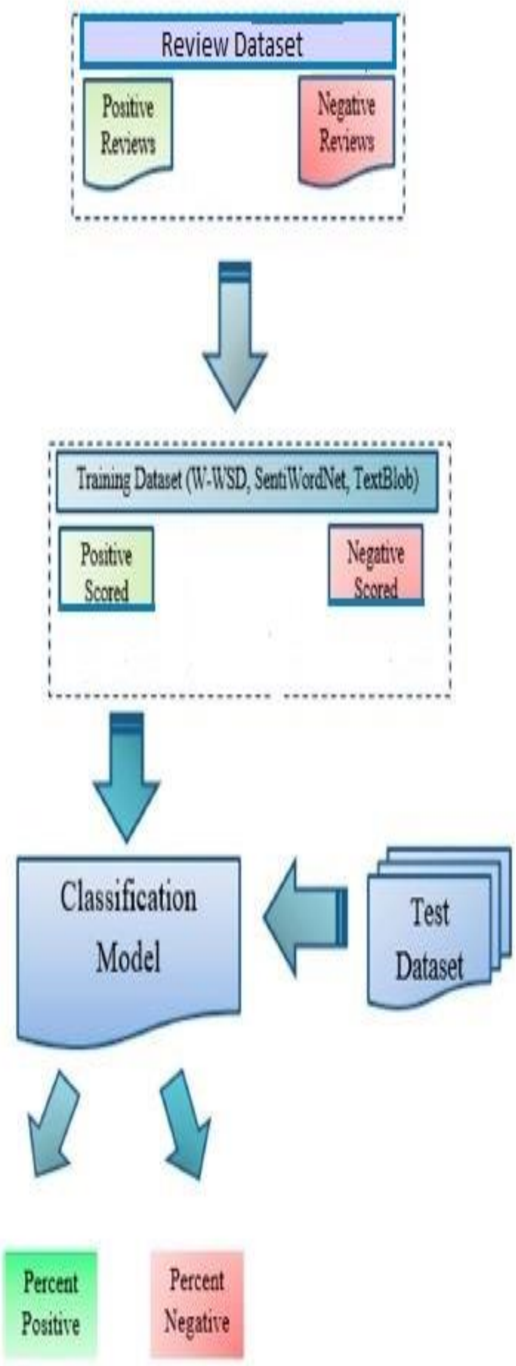


Fig 1.6 Classification model

1.3 Objectives:

To determine the thinking of people on a particular issue: this will tell about the extent to which people can go, thinking about a topic given to them and how they put forward their opinions, views and ideas plus how our machine will deal with this thought processing and how correctly it will categorize the same into positive, negative To determine how positive or negative are people about a particular issue: once we categorize the reviews and get the correct output we can calculate the percentage of people who are having positive reviews about the topic, same in case of negative and people who have mixed emotions on a issue, this is very useful in rating a product or service such as if we take an example of movie rating on imdB ratings: totally depends on the percentage of people who had positive and negative reviews about the movie. So the objective of our topic is same as mentioned above and tries to maintain higher accuracy rates in judging the opinions.

1.4 Sentimental Analysis various Uses:

a)Feature identification

Feature identification is one of the major application of sentimental analysis. For ex: the sentence "this movie has amazing plot and excellent characters". Selected features are "Plot" and "characters. They will be used to identify the features of a review.

b)Opinion identification

Determining the polarity i.e. positive and negative is also a very important task. For ex: the sentence "this movie has amazing plot and excellent characters" is of positive polarity because both opinion words are of positive polarity. Many words in natural language have similar meaning. We will combine them or group them as synonyms as a group of similar words together.

c) Synonyms grouping

Many words in natural language have similar meaning. We will combine them or group them as synonyms as a group of similar words together.

1.5 Phases:

a)Pre-Processing phase: Data set is first cleaned.

b)Feature extraction: Keywords are given token and this token is now put under certain analysis.

c)Classification phase: Based on different algorithms, then these keywords are put under certain category.

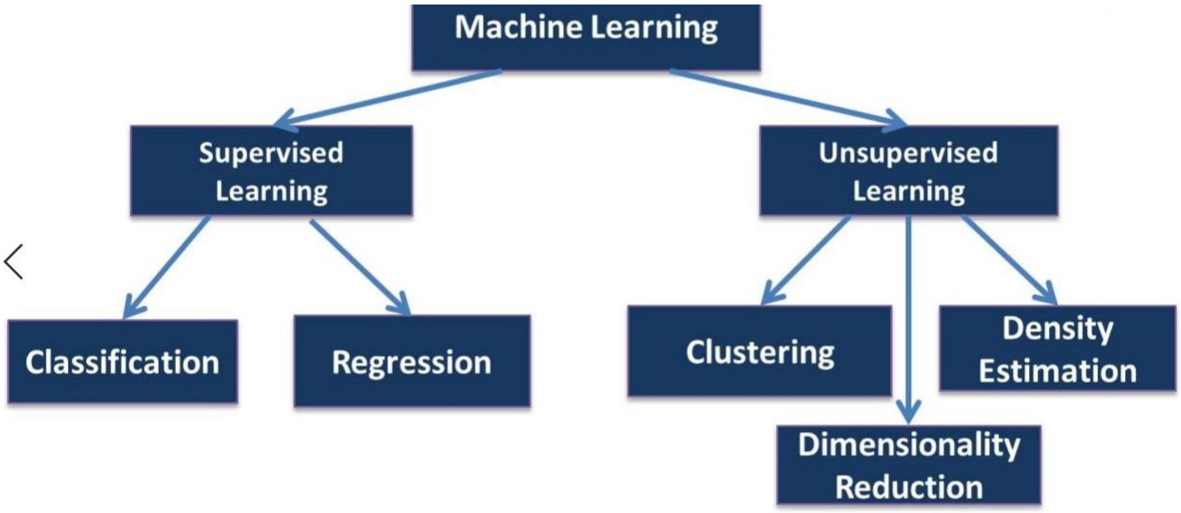


Fig 1.7 Division of machine learning

1.6 Framework

To commence our project we decided to follow the following steps:

- Project Title: Semantic Analysis
- Concept: Machine Learning
- Refers to: Programming of machine to process and analyse large amount of human language data.
- Text Analysis: Deriving of the high quality information from text.

1.7 Techniques:

***a)Knowledge based method*:** To look for the presence of unambiguous words such as happy, sad, nice, bored etc.

***b)Statistic based method*:** This method provides machine with the ability to learn as per human requirements.

c)Hybrid approach: Mixture of both the above procedures that is Knowledge based and Statistical methods. After studying all the three methods in deep we decided statistic approach for the project.

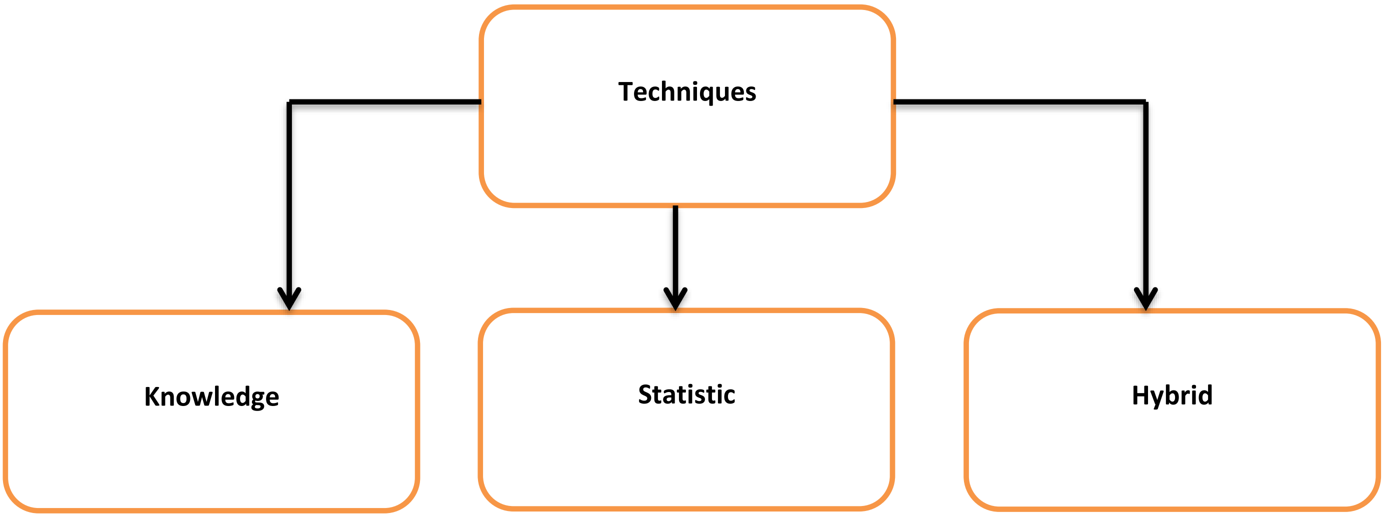


Fig 1.8 Various Techniques

CHAPTER-2

LITERATURE SURVEY

We have done our literature survey by reading various research papers, articles on internet about the concept of machine learning, the various terminologies and algorithms we can use to one by one make the various modules of our projects. So in this part of the report we will write about the various research paper we have studied about and the concepts they have used:

2.1 Capturing favourability using Natural Language Processing

Approach:

Derive some of the sentiments a document rather than classifying the entire document as favourable or unfavourable. It will help to analyse various competitions in the market, marketing skills and helps in risk management for the company.

If a statement contains two different statements which are contradicting each other it becomes difficult to analyse the polarity of the statement so instead of analysing the polarity of the whole text we have to analyse the polarity of each statement and provide the results to the users which lead to more accuracy in the output.

This research paper classified sentimental analysis into three domains: Various expressions, their strengths or polarity and its relationship with the project.

Distribution:

There are many words beside adjectives that define different sentiments such as noun, adverbs and verbs. Some verbs shows the polarity of the sentiment known as Sentiment verbs and some of them only transfer sentiments to and from the arguments which are known as Transfer verbs. Following are the entries which we have in our sentiment dictionary, 3513 in number and in 14 cases we have used regular expressions.

POS tags	Total	Positive	Negative	Neutral
Adverb	6	1	4	1
Sentiment verb	357	103	252	2
Transfer verb	109			
Noun	576	179	388	9
Adjectives	2456	969	1495	1

Table 2.1 Comparison of different polarities

Algorithm:

Lower limit: 5 words before and after the target, Upper limit: 50 words before and after the target. POS tagging is used when we have ambiguous words such as “like” which can be used as sentiment only when it is used a as verb not as a adjective or preposition .Syntactic parsing is used to find the relationship of the sentiment with respect to the subject. Markov-model-based tagger is used for POS tagging which is used in assigning the part of speech as text tokens based on the probability of labels for each word and that of the transition derived for a training corpus. The tagger is then used to identify unknown words such as numbers and treat them as noun, not counting them as sentiments .After the POS tagging part is done we use shallow parsing to see how the object and subject of the statement are bind to predicates. The above two methods are implemented using a Talent system that is based on the text architecture .Once we obtain the results from the shallow parser we analyse dependencies among different phrases in our text. Sentiment polarity for favourable statements is +1 and for unfavourable events is -1.

Capturing trends on Sentiments:

For performing the experiment the comparison is between the ratios of detected sentiments with missed sentiments by taking the data from camera reviews on web pages. The following table illustrates the comparison between sentiments detected by human being and the machine Human:

Human:

Polarity	Favourable	Unfavourable
Brand A	437	70
Brand B	169	65
Brand C	80	51
Brand D	39	41

Table 2.2.1 Comparison of favourability and unfavourability (Human)

System:

Polarity	Favourable	Unfavourable
Brand A	52	4
Brand B	22	5
Brand C	9	2
Brand D	3	1

Table 2.2.2 Comparison of favourability and unfavourability (System)

Conclusion:

This research paper successfully tried to show the relationship between different sentiments between the subject and objects of various statement, instead of finding the polarity of the whole document they used statement terms in order to attain high precision in the report. Accuracy: Initially with easy data set the accuracy came to be 95% and 20% recall but with the increase in difficulty and more data types the overall precision came out to be 75%.

2.2 Sentiment Analysis using subjectivity summarisation.

Approach: The use of efficient method like graph based computation by finding minimum cuts can be a great alternative rather than using basic principle of Machine learning to detect the various sentiments. This approach is basically focused on the physical proximity between the items to be classified. If we look from the vision of computer branch modelling the proximity using minimum cut graphs is proven to be more effective than any other ways.

Framework:

Classifying movie reviews as positive and negative there are three reasons for the same

1) It is a useful service

2) They are harder to classify than reviews of any other product or services

3) Correct label can be derived from the rating information.

Data:

1000 positive and 1000 negative reviews all written before 2002 with 20 reviews per author called as polarity dataset.

Algorithms:

Pair wise interaction deals with the individual vectors in each sentence and the equation for the same is: summation of all the ind2(x) vectors+ summation of all the ind1(x) vectors summation of all the associated vectors (xi, xk).When two disjoint subsets of a graph will meet at least a point that is known as cut of the graph and among these cuts the edges that will have the minimum weight is known as the minimum cut. When we create the graph all the predicates will be at the left with their polarity on the right side, the thickness of the edges will determine the strength associated with them. The graphs can be of two types depending on how symmetric they are 1) Directed graphs: they possess symmetry and 2) Undirected graphs: do not possess symmetry. With the subjectivity scores assigned to the sentences we should also provide them with the proximity score to rest of the sentences in the document and then find the value of N by using minimum cut.

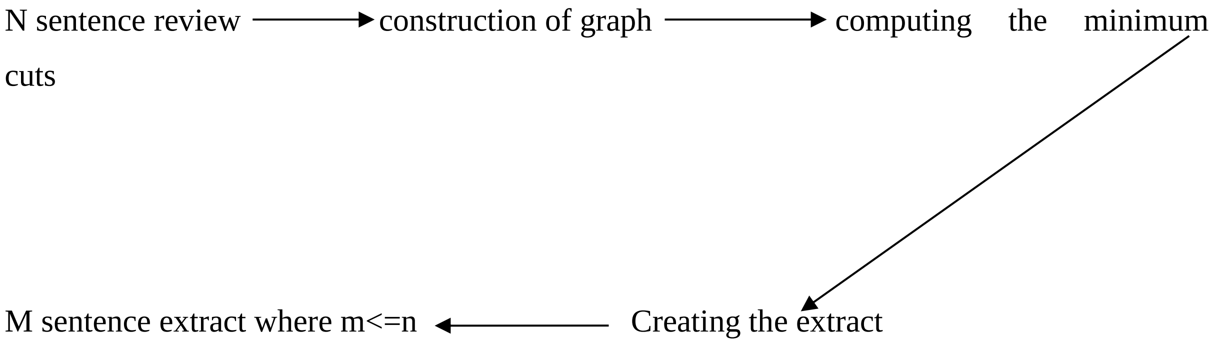


Fig 2.1 Minimum cut in shorter and cleaner way

Conclusion:

By the above classification we have shown that minimum cut is the shorter and cleaner way to represent the subjectivity detection and polarity classification, also it results in the developing of efficient algorithms that can be used in the process of Sentimental Analysis. This way has lead to a increase in accuracy of polarity classification with higher than only using SVM's and NB algorithms but they can be used as classifiers with associating them with minimum cut graphs. The maximum accuracy attained was 87.15% by this project.

2.3 Contextual Polarity in Phrase level Sentiment Analysis:

Approach

This paper adds the concept of neutral reviews with that of the polar reviews. The system gives a clue to each instance of the label but does not identifies the sentiment boundaries which can improve the performance.

Gold Standard

It is used to train and test the machine with respect to the manual annotations .1) If a clue instance is not in the subjective expression than it is said to be neutral.

Positive

Negative

Positive+ neutral=Positive

Negative + neutral=Negative

Study: 10 documents are used from the MPQA Corpus that contain 447 expressions

Agreement: 82% Kappa: 0.72

	Positive	Negative	Both	Neutral	Total
Positive	73	5	2	16	96
Negative	2	167	1	14	184
Both	3	0	3	0	6
Neutral	14	24	0	123	161

Table 2.3 Distribution of various categories

For considering these cases borderline and out of the study domain the value of agreement rises to 90% and kappa becomes 0.84.

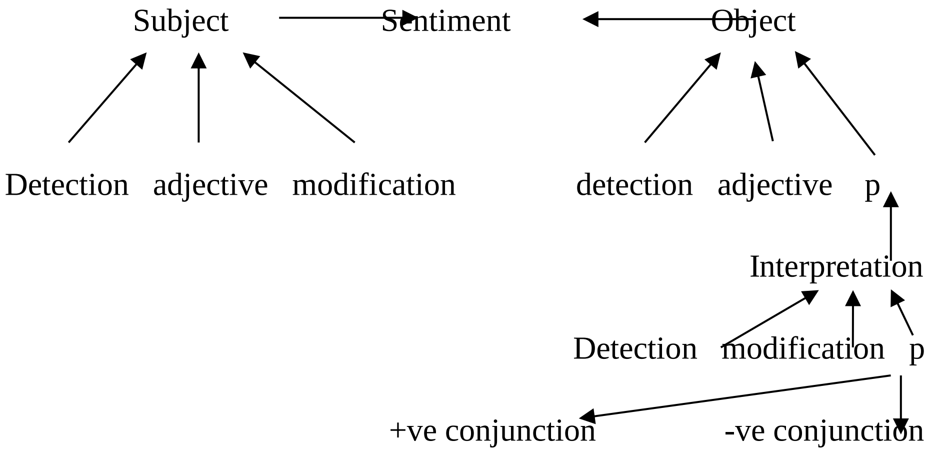


Fig 2.2 Framework associated with research paper

The above table shows the framework associated with this research paper , how the project was commenced and ended the flowchart which was followed by the research for attaining precision and successful results. The whole procedure by which they analysed the sentiments and categorized them into different polarities and neutral aspects is shown in the above flowchart.

Conclusion:

The above research paper classifies the statement into polar and neutral and then it classifies the polar into positive and negative thus we can use this approach on large amount of data which makes it better than other approaches. The accuracy attained was 65.7% using large set of data which the previous research paper do not deal with, thus making this approach more valuable in real life scenarios.

Websites:

We have used various websites to learn the concept of machine learning and python which are pre requisites of Sentimental Analysis and we have also read about the various other approaches and choose an approach for our project is discussed in this report.

Sources	Approach	Accuracy
Paper 1	Sentiments from statements	75%
Paper 2	Minimum cut graphs	87.15%
Paper 3	Large data including neutral	65.7%
Websites	-	100%

Table 2.4 Table Literature survey summary

In short we have described all that we have learnt for commencing our project all the research papers content and content from various websites with their accuracy rates and we took ideas from the above content used them in various modules of the project and leant a lot about the whole concept of sentiment analysis and how can we lead to the success of our journey of generating efficient outputs and maintain high precision rates of our project.

2.4 Gurshobit singh Brar and Prof. Ankit Sharma proposed a system form classifying a huge database of movie reviews. They used a web based API for sentiment analysis for movie reviews with ISON input to display results on any operating system. Their API can also be used for smartphones, laptops or clothes etc.

2.5 Humera Shaziy proposed a system for sentimental analysis using WEKA Tool. They have enhanced the earlier work done in sentiment categorization which analyzes opinions expressing both the positive and negative statement with an accuracy of 75%.

2.6 PalakBaid and Apoorva Gupta and Neelam Chaplot proposed system for recognition of polarity of tweets. They used Naïve bayes, K nearest neighbour and random forest machine learning algorithms. They achieved 81.45% accuracyusing naïve bayes, 55.30 using k nearest behaviour and 78.65% using random forest classifier.

2.7 Jsaiteja, G kiransai, M druvahumar and R manikandan proposed a system to classify the opinions form text. Their paper mainly focuses on comparative study of various machine learning techniques that are used to extract the sentiments from text. They have concluded that Naïve bayes and SVM can be considered as the benchmark for all the other algorithms.Cleaner the data , better the performance.

2.8 Ali hasan, sanamoin, ahmandkarim and shahaboddin shamshirband proposed a system for learning about election sentiments. Their system is lexicon based sentiment analysis, polarity is calculated on the basis of dictionary that consists of a semantic score of a particular word.

2.9 Bo Pang and Lillian Lee Proposed a system of sentimental analysis using subjectivity summarization based on minimum cuts. They examine both subjectivity detection and polarity classification and their realtion and showed that subjectivity detection can compress reviews into much shorter extracts.

CHAPTER-3

SYSTEM DESIGN

3.1 Machine Learning:

As the name suggests in this concept is based on training and testing, first step is to train the machine to take different reviews or opinions from the user and give useful outputs and then comes the second step of testing if the machine gives correct outputs of similar opinions or reviews.

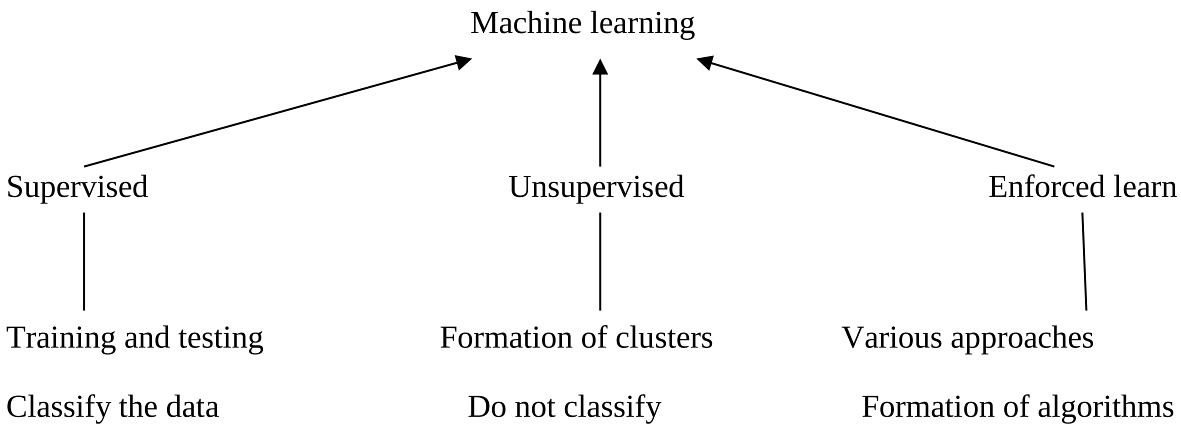


Fig 3.1 Flowchart showing the concept used for doing the sentiment analysis.

3.2 Language : Python

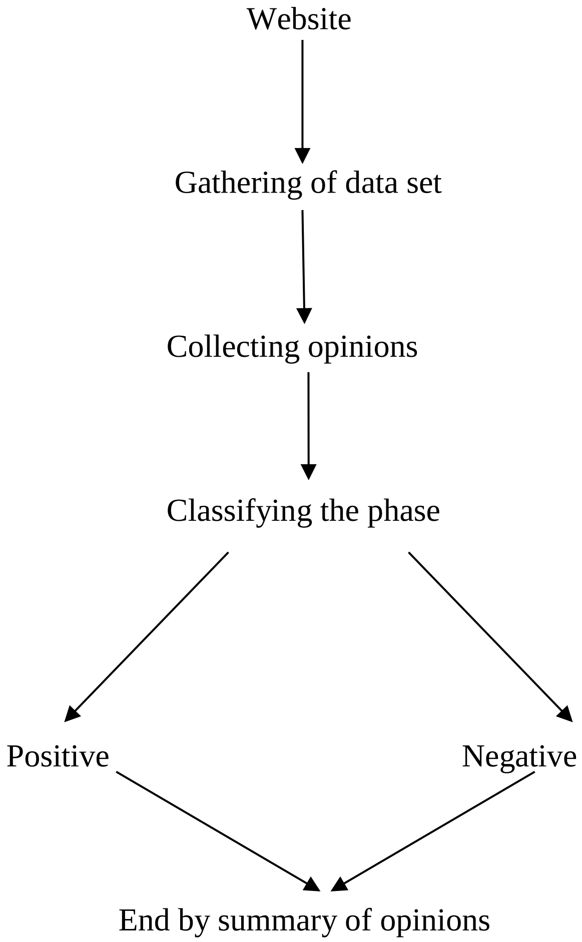


Fig 3.2 Basic Framework

3.3 Methodology:

a) Review line, paragraph or expression

b) Store each word in the array

c) Store the word count

d) Pass this word count through a neural network

e) Predict the Probability

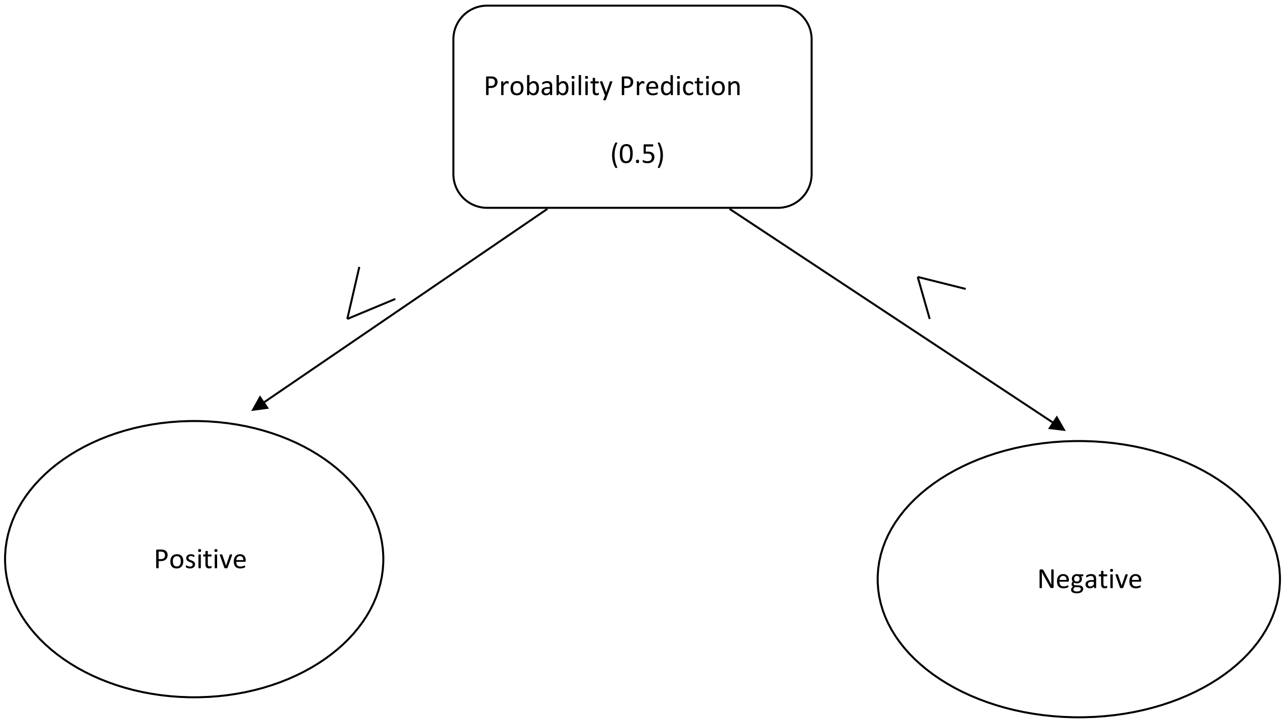


Fig 3.3: Expanded Framework

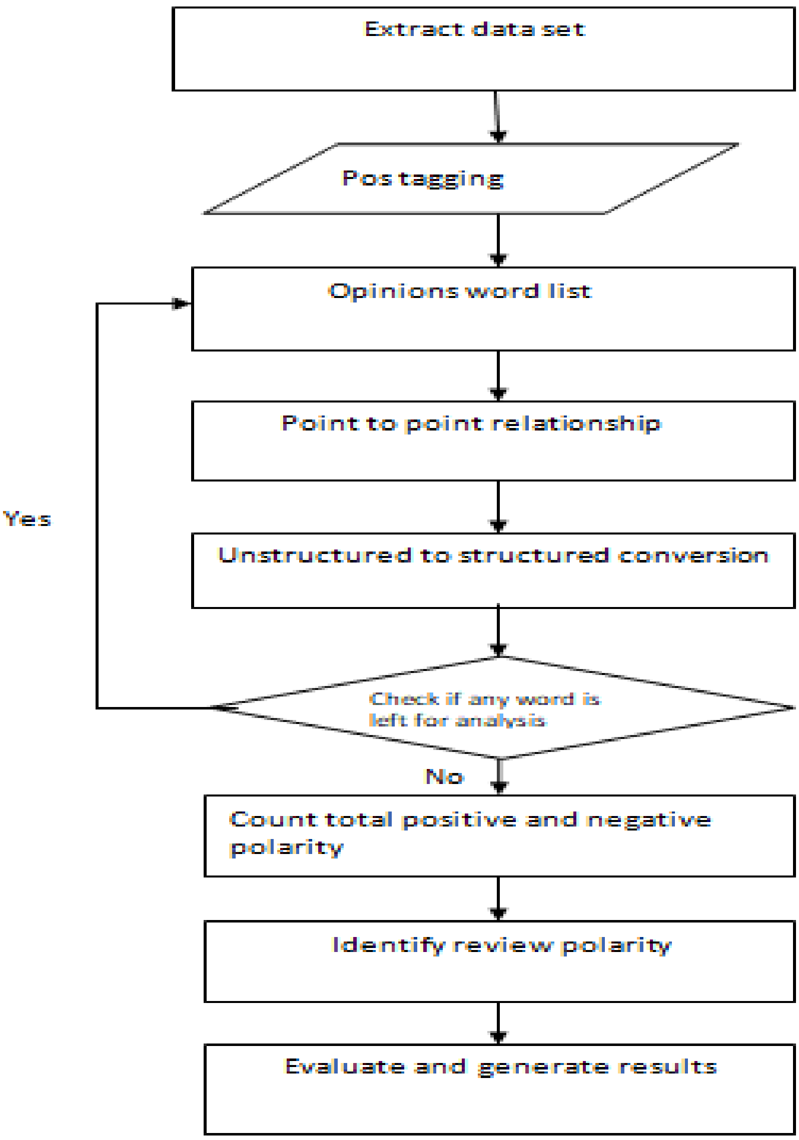

Extract data set
Pos tagging
Opinions word list
Point to point relationship
Unstructured to structured conversion
Check if any word is left for analysis
Yes
No
Count total positive and negative polarity
Identify review polarity
Evaluate and generate results

Fig 3.4: Flowchart of proposed System

If the predicted probability after passing through the neural network is greater than 0.5 the sentiment is positive and if it is less than 0.5 then the sentiment is negative this will totally depend on which algorithm we will use in the project to classify the sentiments which we will come to see in the next chapter, before which we should learn about the different tools used in or system design.

3.4 Tools:

a) Anaconda

This is one of the tool in our project work which helps us to provide a free platform for python and R programming and it is an open source distribution used when we have to apply the concept of machine learning on large data set. This open source tool is written in python and it was released 6 years ago in 2012.It has various applications such that it comes with packages which can be used to perform various function among which we have GUI, programming etc but we will use this tool only in our concept of machine learning as we have written our algorithm in python.

b) Jupyter Notebook

This tool is also an open source which contains certain facts, figures description and tables which can be easily read by humans as it is in the form of notebook and are used in various fields such as editing, marketing, analysis etc ,but we are using it as one of the tools for sentiment analysis using machine learning as this also includes python language.

c) Neural network

This is the most important part of our project, it is the main tool which works understands the human brain made by formation of different algorithms for training our system to read human mind and respond accordingly so that we can easily predict the probability of the opinions and events .It acts as a framework for many algorithms used in machine learning. All the words that we have stored in the array have to be passed through this network if we further have to predict the probability.

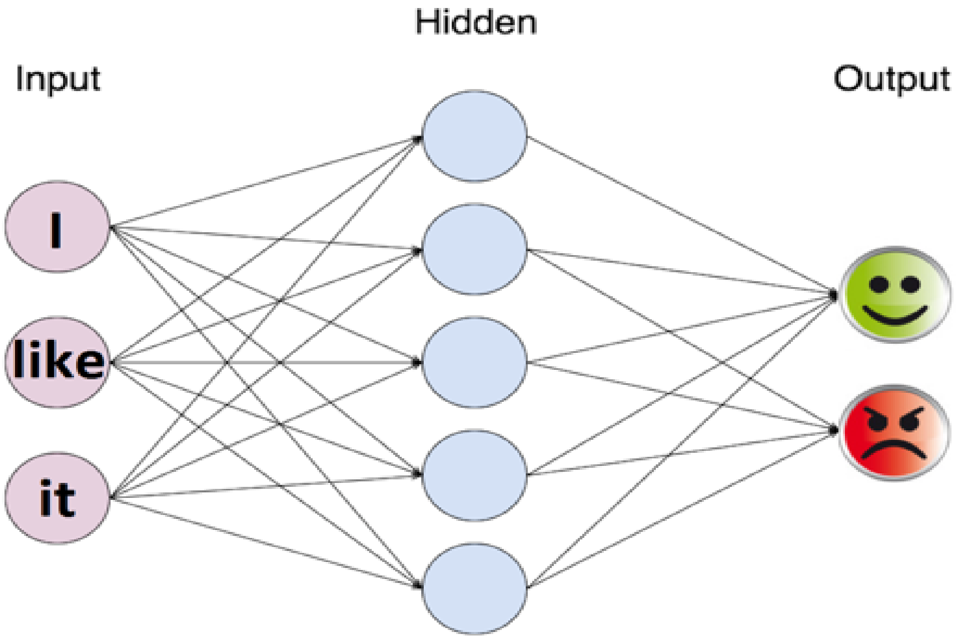
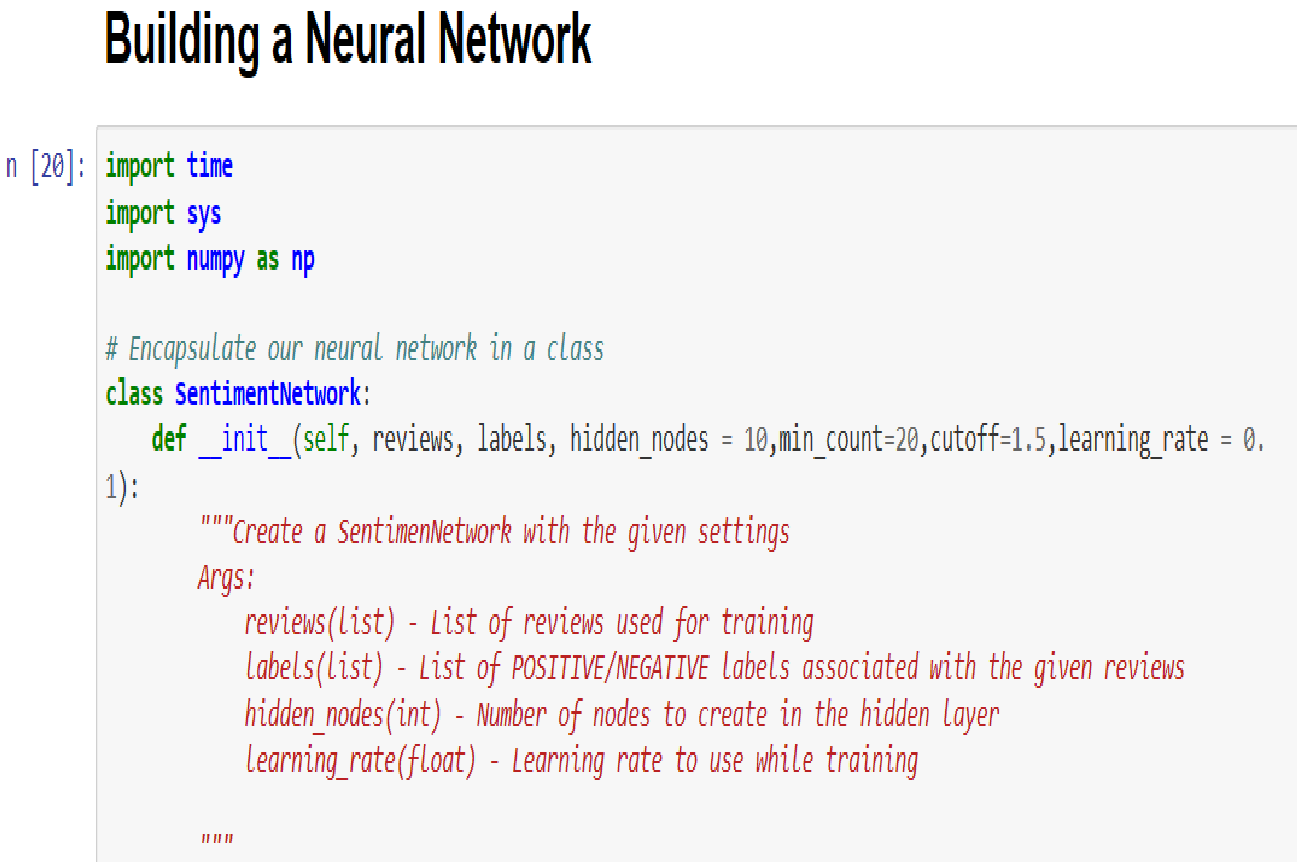


Fig 3.5 Classification of neural network



Building a Neural Network

```
n [20]: import time
        import sys
        import numpy as np

        # Encapsulate our neural network in a class
        class SentimentNetwork:
            def __init__(self, reviews, labels, hidden_nodes = 10,min_count=20,cutoff=1.5,learning_rate = 0.
        1):
                """Create a SentimenNetwork with the given settings
                Args:
                    reviews(list) - List of reviews used for training
                    labels(list) - List of POSITIVE/NEGATIVE labels associated with the given reviews
                    hidden_nodes(int) - Number of nodes to create in the hidden layer
                    learning_rate(float) - Learning rate to use while training

                """
```

3.5 Data set:

Determining how the movie was and as per different views and opinions from the end users.The data set we have used in our project from imdb movie rating.Through this we can classify the review as positive or negative by using neural network that helped us formation of various ML algoritms and determine the viewpoint of differnet users that whether the movie is liked by them or not.The given task was not that easy like it seems but we have tried our best to reach higher precision.

a) Data Input

There are two ways to give input to the movie review sentiment analyzer. One by providing a list of reviews in JSON file format. Or by providing the TMDB ID of Movie Title.

In Case if TMDB ID a TMDB JSON API is used to fetch and store reviews in MySQL Database.

b) Part of Speech Tagging

POS is used to disambiguate a sentence in order to extract features from a sentence [2]. In POS tagging each word is labeled. It is used to determine word position in the grammatical context. POS tagging helps to find out nouns, noun phrases, verbs and adjectives in a sentence. After POS Tagging there is a little chance selected word is a discarded word for feature selection and opinion words.

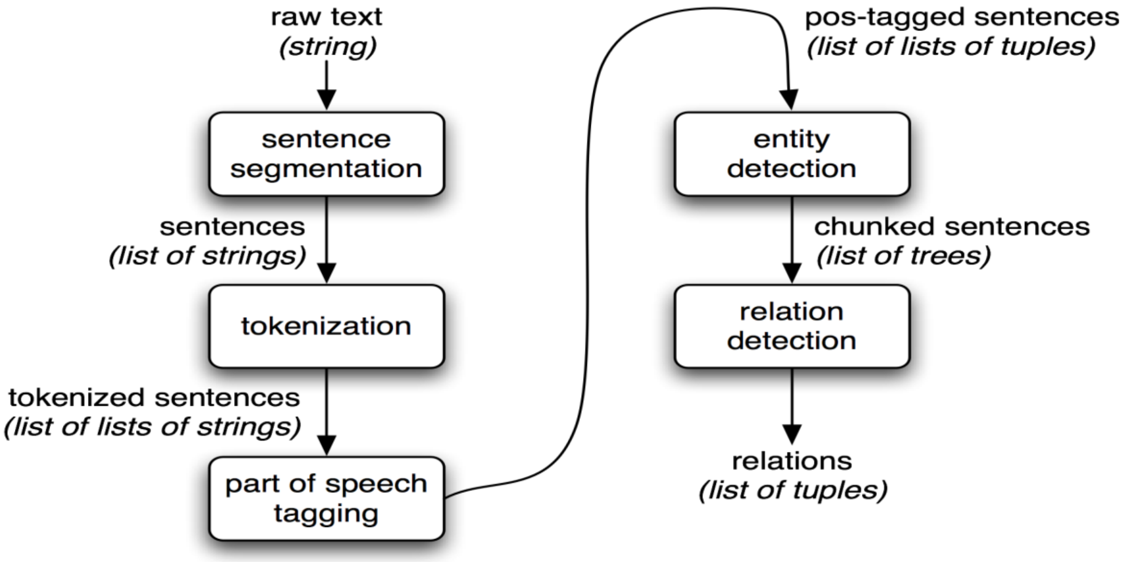


Fig 3.6 Part of Speech Tagging

c) Features and Opinion Words Extraction

All opinion words are selected from the sentence. The system extracts all nouns, noun phrases, verbs and adjectives from the movie review and compares with the existing list of words. These words are classified on basis of their polarity. For Example “good” word is of positive polarity. On the other hand, features are selected on basis of number times occurrence of opinion words. If opinion word is an occurrence in review higher than the threshold value then it is added features list. For this system API is trained only for movie reviews with keyword and phrases dictionary which includes “good acting”, “solid story” and “awesome action”.

d*) Identify Sentence Polarity*

After extracting all features and Opinion words, it is very easy to find the polarity of the sentence. Sentence polarity follows the same rules as arithmetic expressions. A negative sentiment contains all negative opinion words and positive sentiment contain all positive opinion words. A negative sentiment may contain a positive opinion word. For Example: “This movie Story is not good” sentence in a movie review. In this sentence, “good” opinion word is of positive polarity but “not” is a negative word. Therefore, the overall polarity of this sentence will be negative.

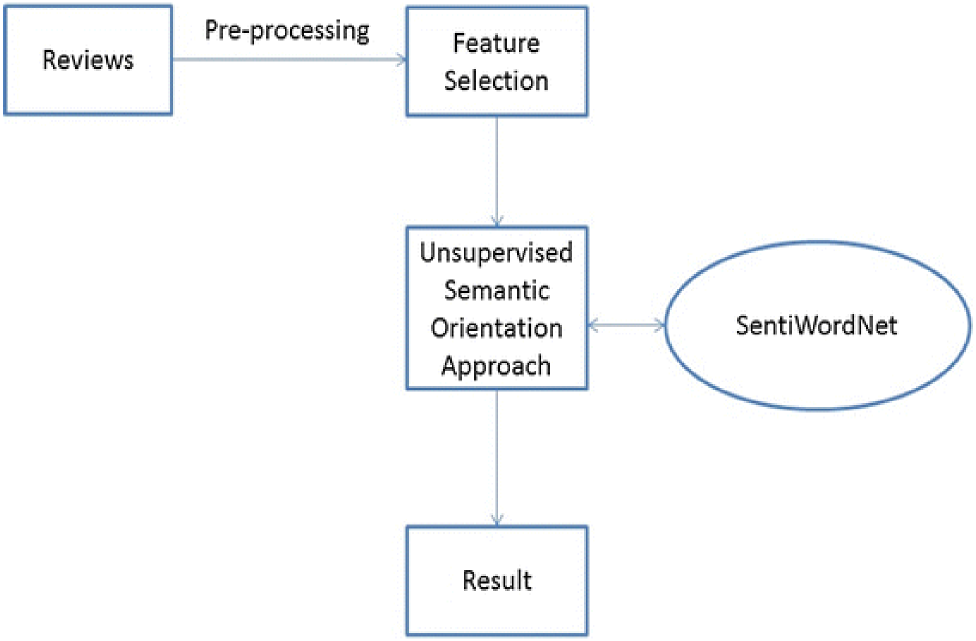


Fig 3.7 Sentence Polarity

e) Identify Review Polarity

Whole review polarity depends on a number of total positive or negative sentences found in a review. If the number of total positive sentences is greater than the number of total negative sentences then review polarity will be positive. Similarly, a review polarity will be negative if the number of total negative sentences is greater than the number of total positive sentences.

f) Storing word in an array

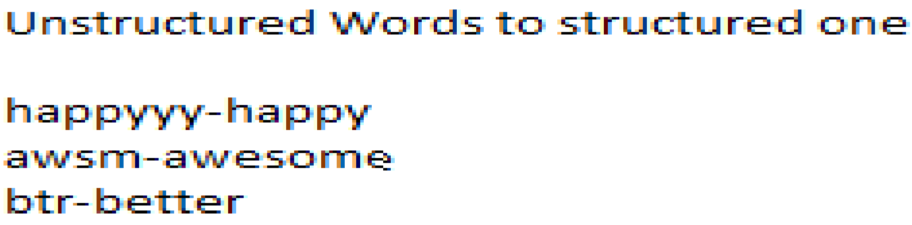
We have studied the various method to store words in an array so all the subjects, objects, predicators be it adjective, verb, adverb, noun is stored in a continuous array so that classification of sentiments become more easy and then we have to store this word count in a variable called count and then keep this iteration on-going so that each word of the statement is tested in accordance with the neural network and we can easily predict the probability using the prediction algorithm that will be discussed in the coming chapters.

g) Point to Point relationship

Information tool helps to determine or helps in finding the relationship with one another, formulates every word association in the document to the predefined adjectives for determining various sentiments that is related to the words.

h) Conversion of Unstructured to structured words

Movie review are generally constructed into'awsme' as 'awesome','happpy' is 'happy' ,'btr','better' in real. The final transformation of unconstructed to constructed words is determine by data set that is converted from unstructured to structured followed by addition of vowels.



Unstructured Words to structured one

happyyy-happy
awsm-awesome
btr-better

i) The Sentiment Directory

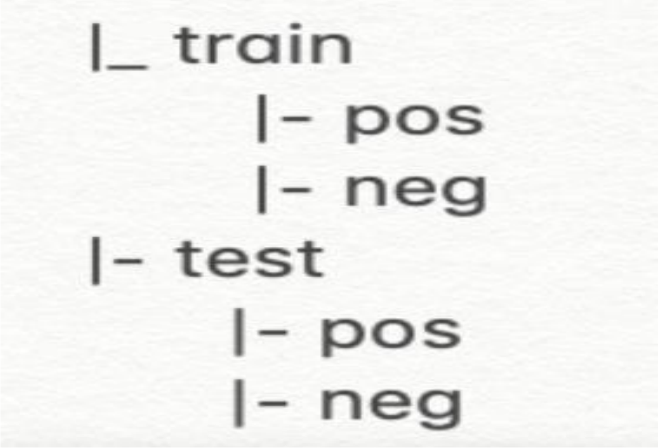
A tool which is used to create the words considered as sentiments. The use of a specific words are viewed such as "grand" can be utilized as a wide range of ways with every way having its own opinion according to the situation.

j) Testing Dataset

The dataset we further used to test our algorithm .The test contains 75,000 reviews. The goal is to find the classifier to determine and labels accurately which will further determine the polarity.

k) Trained Dataset

The tool which helps the learning algorithm ,obtaining different features from the dataset and classifying the reviews as positive or negative.



```
|_ train
      |- pos
      |- neg
|- test
     |- pos
     |- neg
```

l) Classification of Review

Once, review polarity is calculated. Review polarity percentage and polarity (Positive or Negative) classified [28] and saved for further analysis. With further analysis, box office collection can be predicted and overall performance of movie can also be predicted.

m) Evaluate

The need to evaluate our algorithm is there for understanding how it performs. For doing this we need to calculate the prediction accuracy which is nothing but the percentage of labels that were predicted correctly, higher will be the prediction accuracy better is the algorithm.

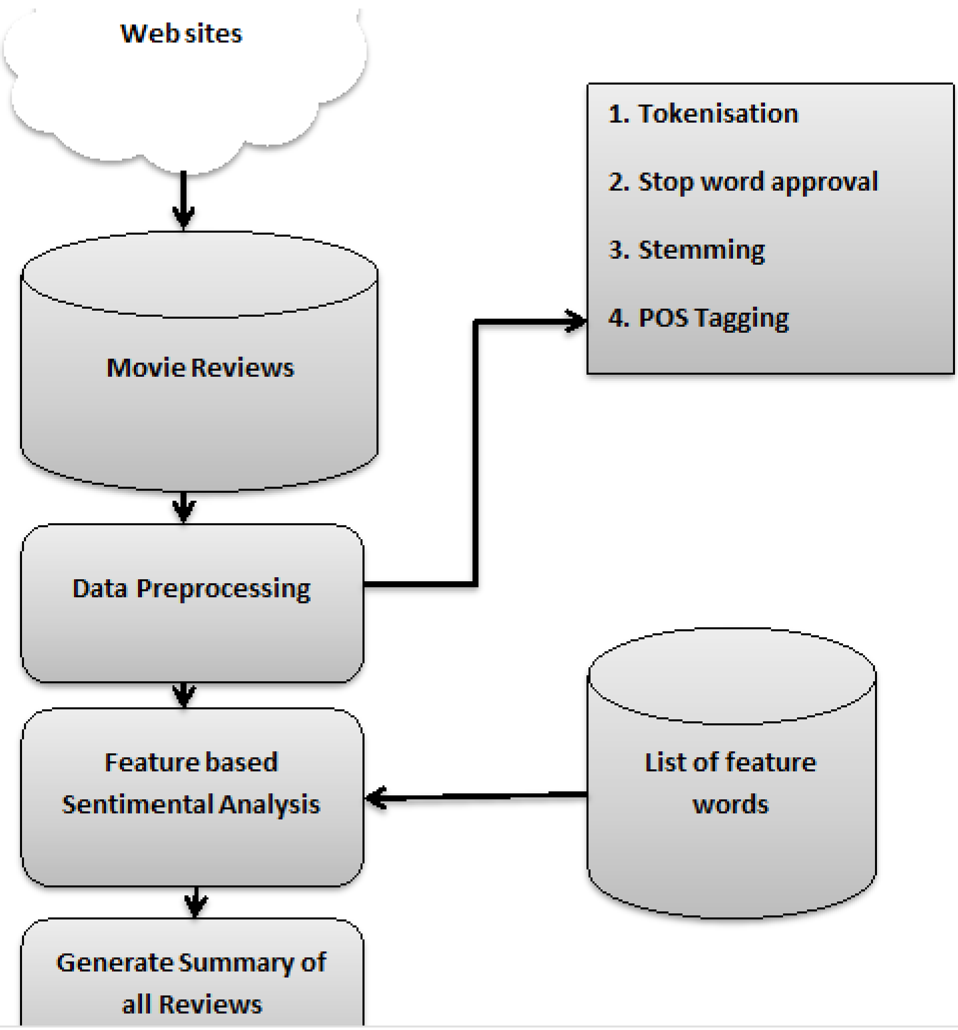
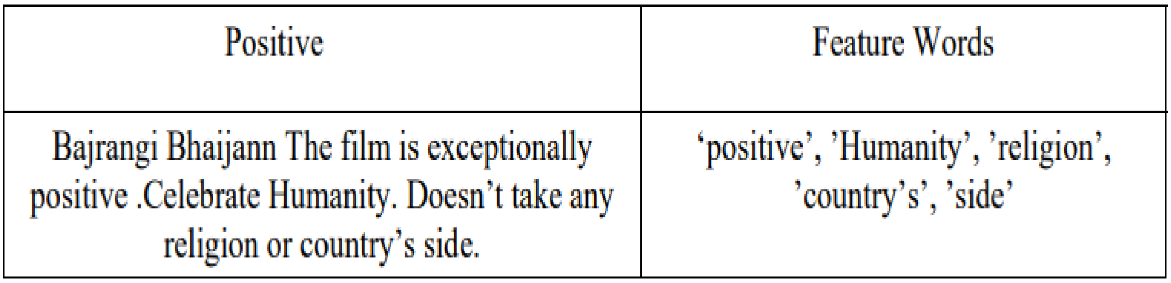
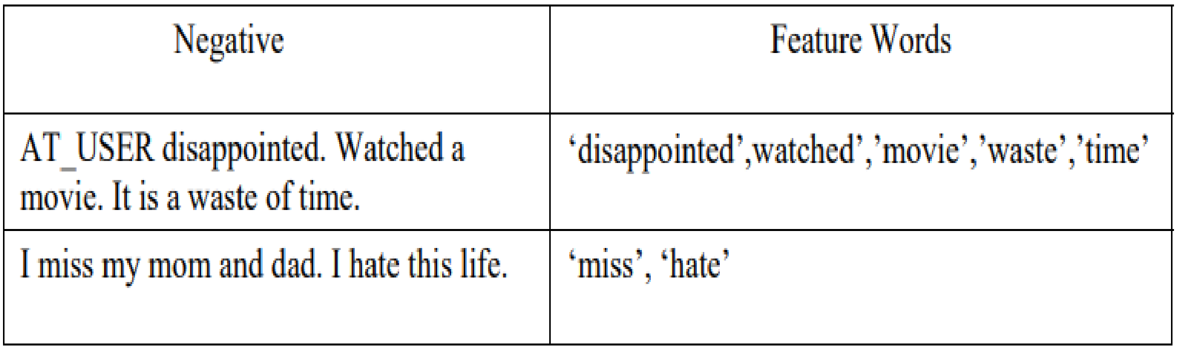


Fig 3.8 Determining sentimental model



Positive	Feature Words
Bajrangi Bhaijann The film is exceptionally positive .Celebrate Humanity. Doesn’t take any religion or country’s side.	‘positive’, ’Humanity’, ’religion’, ’country’s’, ’side’



Negative	Feature Words
AT_USER disappointed. Watched a movie. It is a waste of time.	‘disappointed’,watched’,’movie’,’waste’,’time’
I miss my mom and dad. I hate this life.	‘miss’, ‘hate’

3.6 Extracting Features:

The process which deals with the extraction of information containing and non redundant values present in our dataset which are then used to learn the procedure of machine learning algorithms to produce the classifier models. Once the cleaning up of 75,000 reviews was done from the training set then the creation of vocabulary using the word models was done which computes the frequency of occurrence of these words in the form of features which are then used for training the classifier. We used a kit to learn the feature extraction module for performing the actions. This module derives numeric features from the movie reviews which are in the format of text such as each string is converted into a 'tokens' which are provided with token IDs then the frequency of occurring of each token is calculated and then the tokens are organized based on how often they are occurring.

3.7 Deep Learning:

This is a subset of machine learning technique that teaches our system to perform actions that humans naturally, this is learning by giving examples. It is an important aspect of technology behind the invention of driverless cars by enabling them to understand about the stop signs and the difference between a pedestrian walking from a electric pole also it is the reason behind the voice control for devices like mobile, tablet, televisions, and speakers, gathering a lots of attention these days for a good reason and the achievement of results

Fig 3.9 Deep Learning

If we try to explain results of deep learning in a word we can totally use accuracy, achieving commendable accuracy levels than ever done by any other concept before. Helping customers meet the expectations of user. Recently the advancements in this concept is improving to the point where is outperforms humans in tasks like classifications, requiring a large amounts of labelled data also known as data set.When we combine the cluster it enables the developing teams in the reduction of training time for a network from, months to weeks or less.Deep learning is a more focused outline of machine learning, its workflow commence with pertinent way being manually derived from images. These characteristics are used to generate a model that classifies the substance in the image, by the help of this work flow appropriate features are involuntarily derived from images.

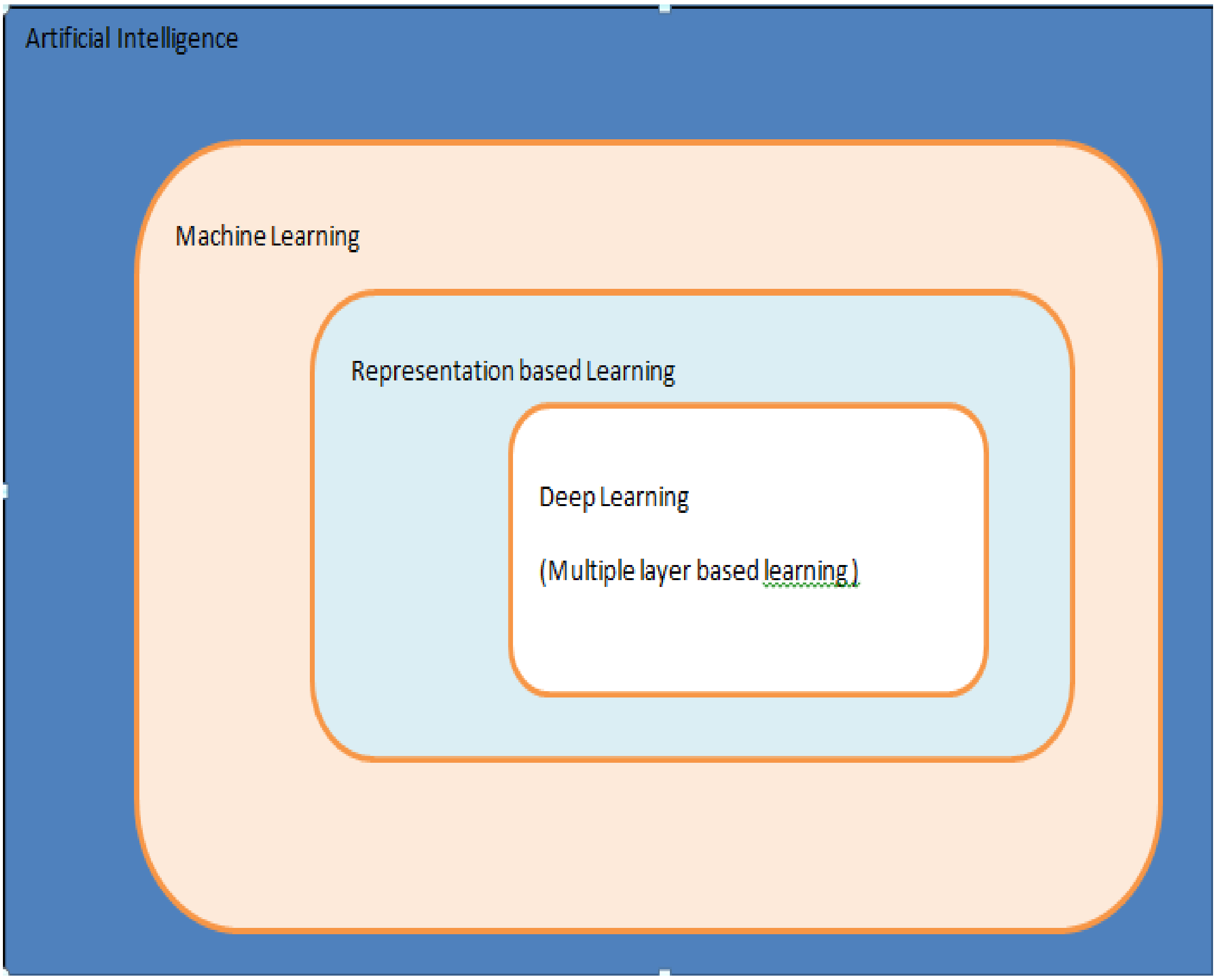

Artificial Intelligence
Machine Learning
Representation based Learning
Deep Learning
(Multiple layer based learning)

Fig 3.10 Deep learning mechanisms

3.8 SVM(Support Vector Machine):

SVM stands for support vector machine. It is differentiating classifier which is formally defined by a unravelling hyper plane or for a labelled training data this algorithm results in an best hyper plane which classifies the other fresh examples. When in two dimensional spaces this hyper plane looks as a line which divides the plane into two parts where in every class lies in its each side. This algorithm moderately separates the two classes and if any of the point is in the left of line cascade in a dark circle class and the one on right falls into blue square class this is the action that the SVM performs and also looks for the hyper-plane.

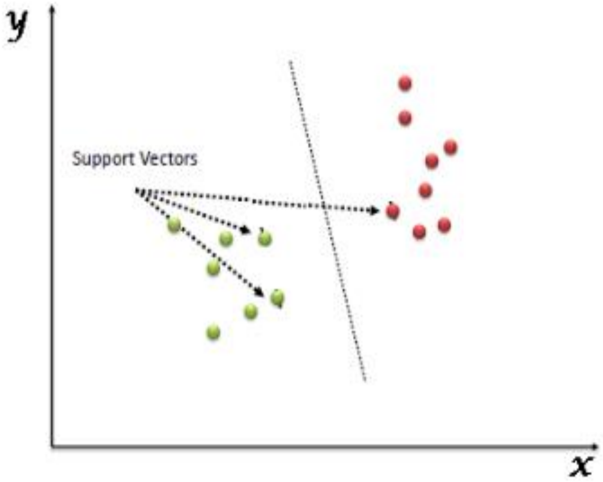


Fig 3.11 Support Vector Machine

3.9 Natural Language processing:

NLP works as follows:

- A human interacts with machine
- The machine takes the audio.
- Audio gets converted to text
- Processing of the text's data
- Data to audio conversion takes place.
- The machine responds to the human by playing the audio file.

Some of the applications where NLP is used are Language translation applications such as Google translate, Interactive Voice Response(IVR) applications, Assistant applications such as OK Google, Siri, Cortana and alexa. NLP is difficult to implement because of human nature. Even the single 's' in the language can be used for so many purpose in a word like 's' signifies plurality of items.

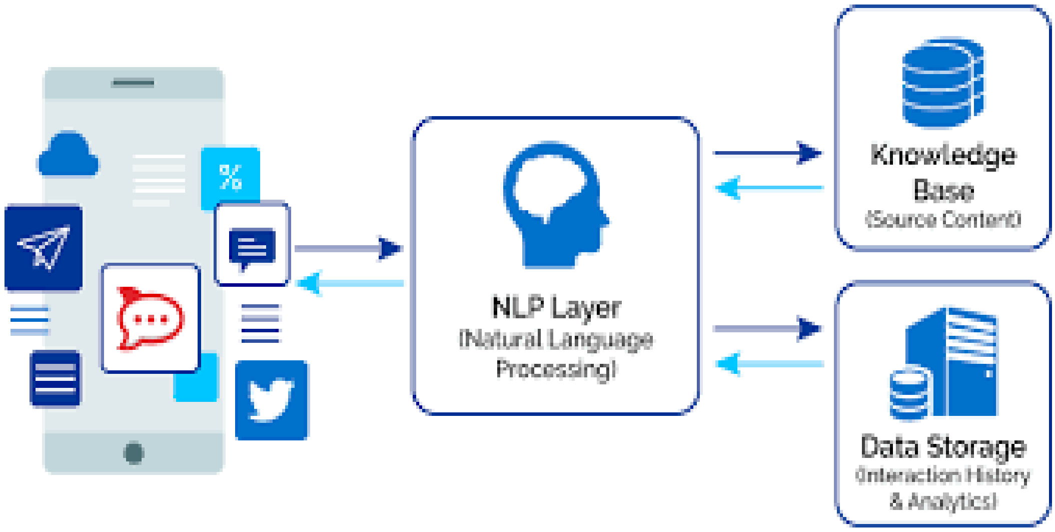

NLP Layer
(Natural Language Processing)
Knowledge Base
(Source Content)
Data Storage
(Interaction History & Analytics)
%

Fig 3.12 Natural Language Processing

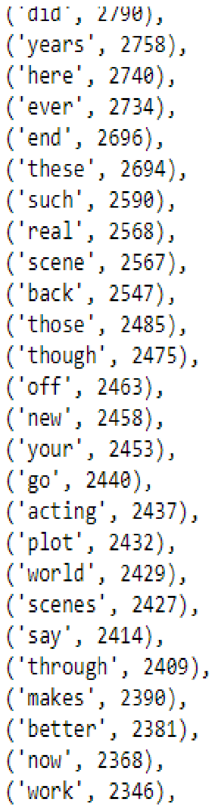
CHAPTER-4

ALGORITHMS

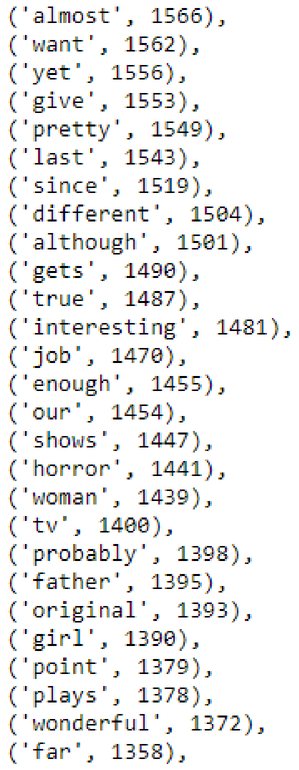
In our project we have used the subset of concept of machine learning that is deep learning and various algorithms associated with the same. In this chapter we will define those algorithms only.

4.1 Data set

The data set that we have used for our movies is from IMDB movie ratings as per different views and opinions from the end users about how the movie was. We will try to classify these reviews into positive and negative by using neutral network that has helped us in the formation of various ML algorithms and prediction about the viewpoints of different users that whether they like a movie or not. Even though the task is not that easy like it seems but we have tried our best to reach higher precision.



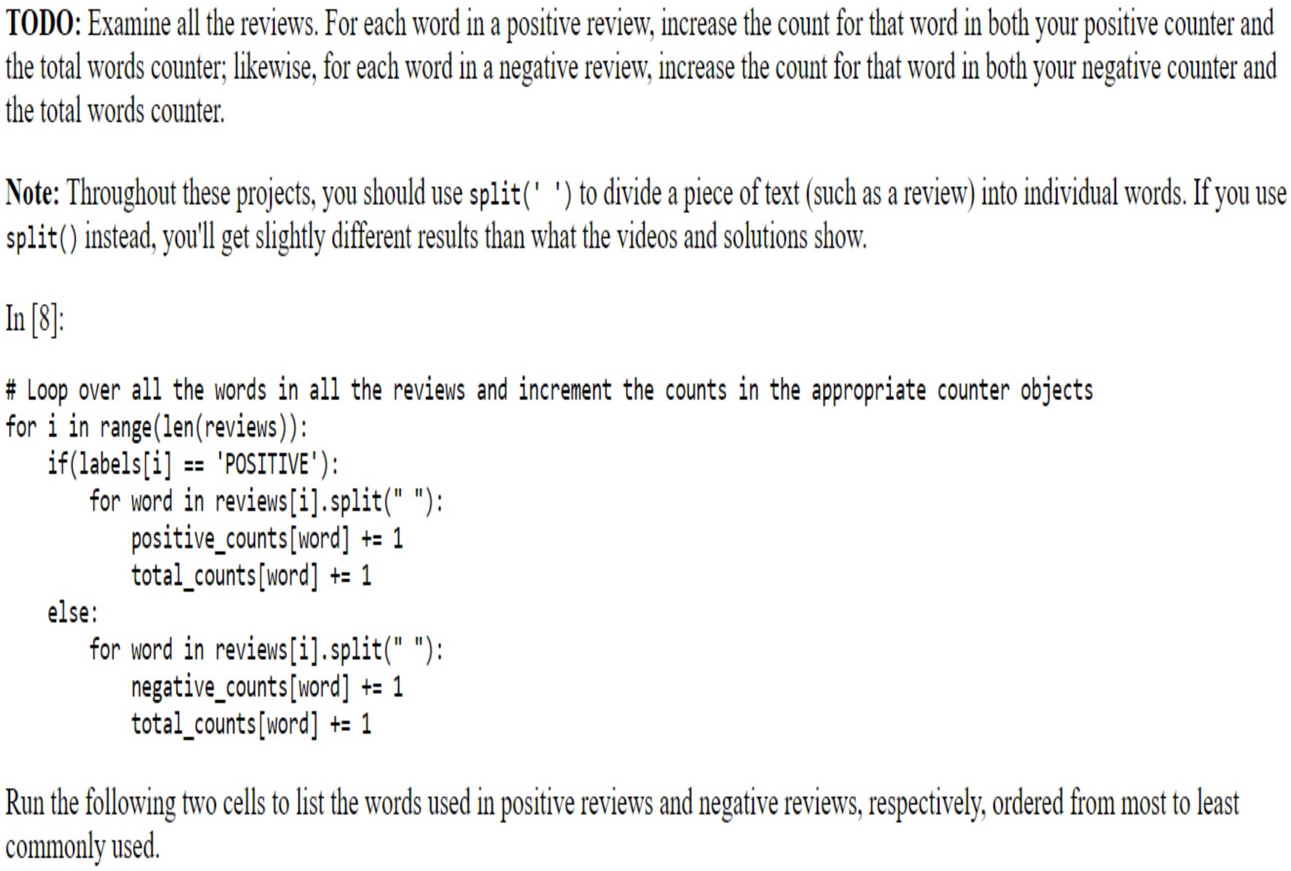
```
('did', 2790),
('years', 2758),
('here', 2740),
('ever', 2734),
('end', 2696),
('these', 2694),
('such', 2590),
('real', 2568),
('scene', 2567),
('back', 2547),
('those', 2485),
('though', 2475),
('off', 2463),
('new', 2458),
('your', 2453),
('go', 2440),
('acting', 2437),
('plot', 2432),
('world', 2429),
('scenes', 2427),
('say', 2414),
('through', 2409),
('makes', 2390),
('better', 2381),
('now', 2368),
('work', 2346),
```



```
('almost', 1566),
('want', 1562),
('yet', 1556),
('give', 1553),
('pretty', 1549),
('last', 1543),
('since', 1519),
('different', 1504),
('although', 1501),
('gets', 1490),
('true', 1487),
('interesting', 1481),
('job', 1470),
('enough', 1455),
('our', 1454),
('shows', 1447),
('horror', 1441),
('woman', 1439),
('tv', 1400),
('probably', 1398),
('father', 1395),
('original', 1393),
('girl', 1390),
('point', 1379),
('plays', 1378),
('wonderful', 1372),
('far', 1358),
```

4.2 Storing word in an array:

We have studied various method to store words in array to all the subjects, objects, predicators they may be verb, adjective, adverb ,noun are to be stored in a continuous array so that they can be easily classified and then we have to store this word count in a variable called count and keep the iteration going on, so that each word of the statement is tested through the neural network and determine it polarity accordingly.



TODO: Examine all the reviews. For each word in a positive review, increase the count for that word in both your positive counter and the total words counter; likewise, for each word in a negative review, increase the count for that word in both your negative counter and the total words counter.

Note: Throughout these projects, you should use `split(' ')` to divide a piece of text (such as a review) into individual words. If you use `split()` instead, you'll get slightly different results than what the videos and solutions show.

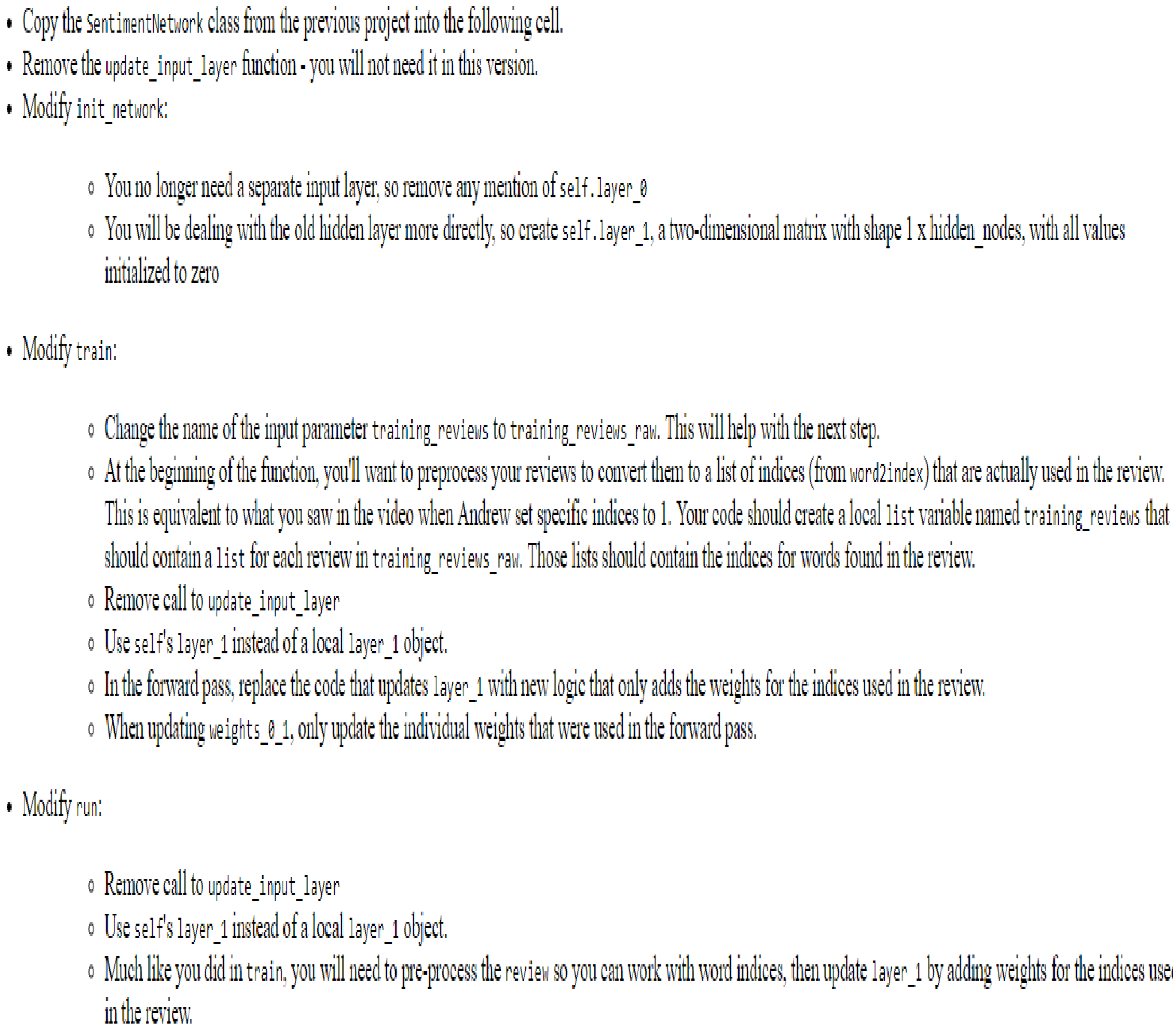
In [8]:

```
# Loop over all the words in all the reviews and increment the counts in the appropriate counter objects
for i in range(len(reviews)):
    if(labels[i] == 'POSITIVE'):
        for word in reviews[i].split(" "):
            positive_counts[word] += 1
            total_counts[word] += 1
    else:
        for word in reviews[i].split(" "):
            negative_counts[word] += 1
            total_counts[word] += 1
```

Run the following two cells to list the words used in positive reviews and negative reviews, respectively, ordered from most to least commonly used.

4.3 Binary classifying approach:

As trying various basic sentimental approaches or methods as described above ,we have performed several pre-processing steps in order to clean up the data which includes removal of the HTML and CSS which is done using python. Then comes not so needed punctuations which were removed from regular expression which we have studied in Theory of Computation an then the conversion from lower case and the removal of stop words for conversion of a cleaned words numeric figure vectors.



- Copy the `SentimentNetwork` class from the previous project into the following cell.
- Remove the `update_input_layer` function - you will not need it in this version.
- Modify `init_network`:
  - You no longer need a separate input layer, so remove any mention of `self.layer_0`
  - You will be dealing with the old hidden layer more directly, so create `self.layer_1`, a two-dimensional matrix with shape 1 x hidden_nodes, with all values initialized to zero
- Modify `train`:
  - Change the name of the input parameter `training_reviews` to `training_reviews_raw`. This will help with the next step.
  - At the beginning of the function, you'll want to preprocess your reviews to convert them to a list of indices (from `word2index`) that are actually used in the review. This is equivalent to what you saw in the video when Andrew set specific indices to 1. Your code should create a local `list` variable named `training_reviews` that should contain a `list` for each review in `training_reviews_raw`. Those lists should contain the indices for words found in the review.
  - Remove call to `update_input_layer`
  - Use `self`'s `layer_1` instead of a local `layer_1` object.
  - In the forward pass, replace the code that updates `layer_1` with new logic that only adds the weights for the indices used in the review.
  - When updating `weights_0_1`, only update the individual weights that were used in the forward pass.
- Modify `run`:
  - Remove call to `update_input_layer`
  - Use `self`'s `layer_1` instead of a local `layer_1` object.
  - Much like you did in `train`, you will need to pre-process the `review` so you can work with word indices, then update `layer_1` by adding weights for the indices use in the review.

4.4 Bag of Word:

This way to numerically represent the text was the simplest way. Let us take ext M we have to assign vector v(M) that belongs to our dataset, as V(M) is the frequency for the mth word in the text. a is the size of the dataset which is containing the word in the dataset and some are rare.After Learning about the bag of words, we will use different classifiers to classify our data.

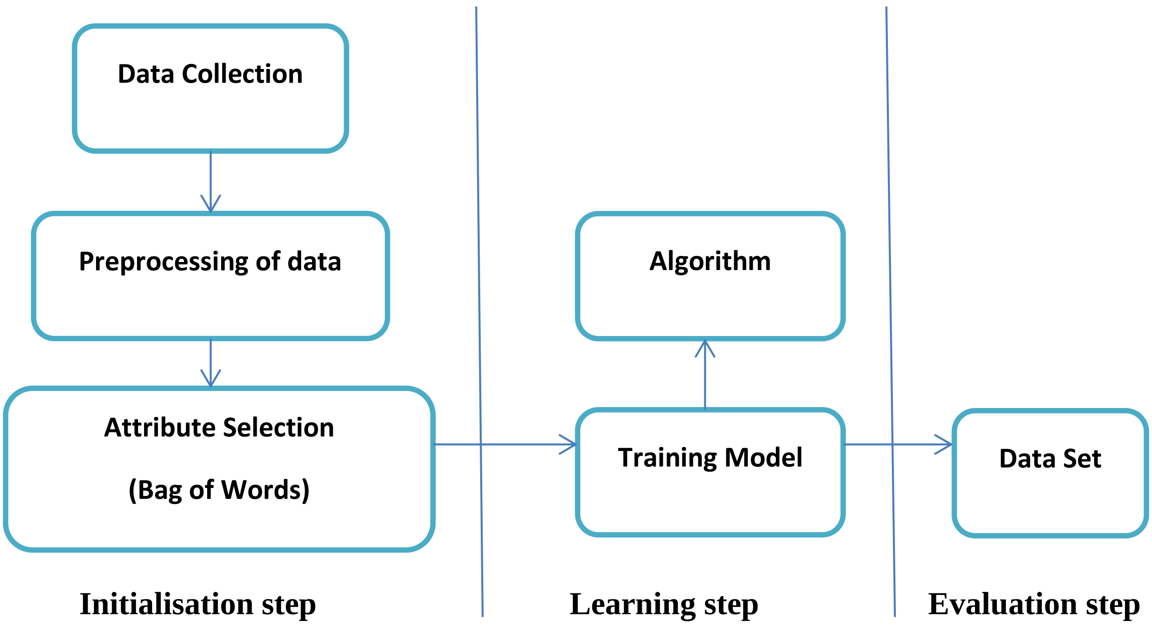


Fig : 4.1 Various steps Involved

4.5 Words to Vector:

This is the another way to numerically represent the texts by converting every word present in the text to its respective vector. One of th important aspect of the word to vector concept is, it is not dependent on sentiment analysis and does not require to have a labelled dataset. Therefore we have taken 75,000 words out of which 35k are labelled and 35K are unlabelled one.

4.6 Words to reviews:

The simplest way to assign a feature vector to the words of the data set which is also known as the review two stop words are frequent words they are ‘the’ and ‘is’ they do not have any determined sentiments. The Natural language toolkit package identifies the stop words by statistic analysis on some large data corpus, the two word vectors of all the words present in our dataset. As the stop words do not change the polarity of the paragraph so we can ignore them as they will affect our precision at a negligible amount.

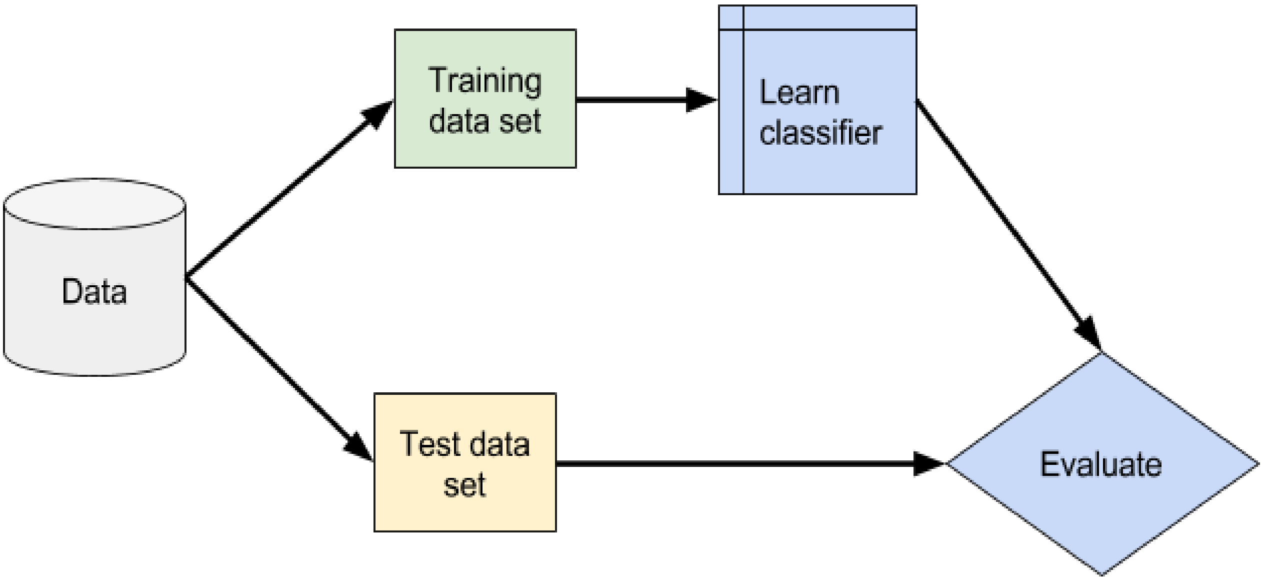


Fig 4.2 Framework of Train/Test Dataset classification

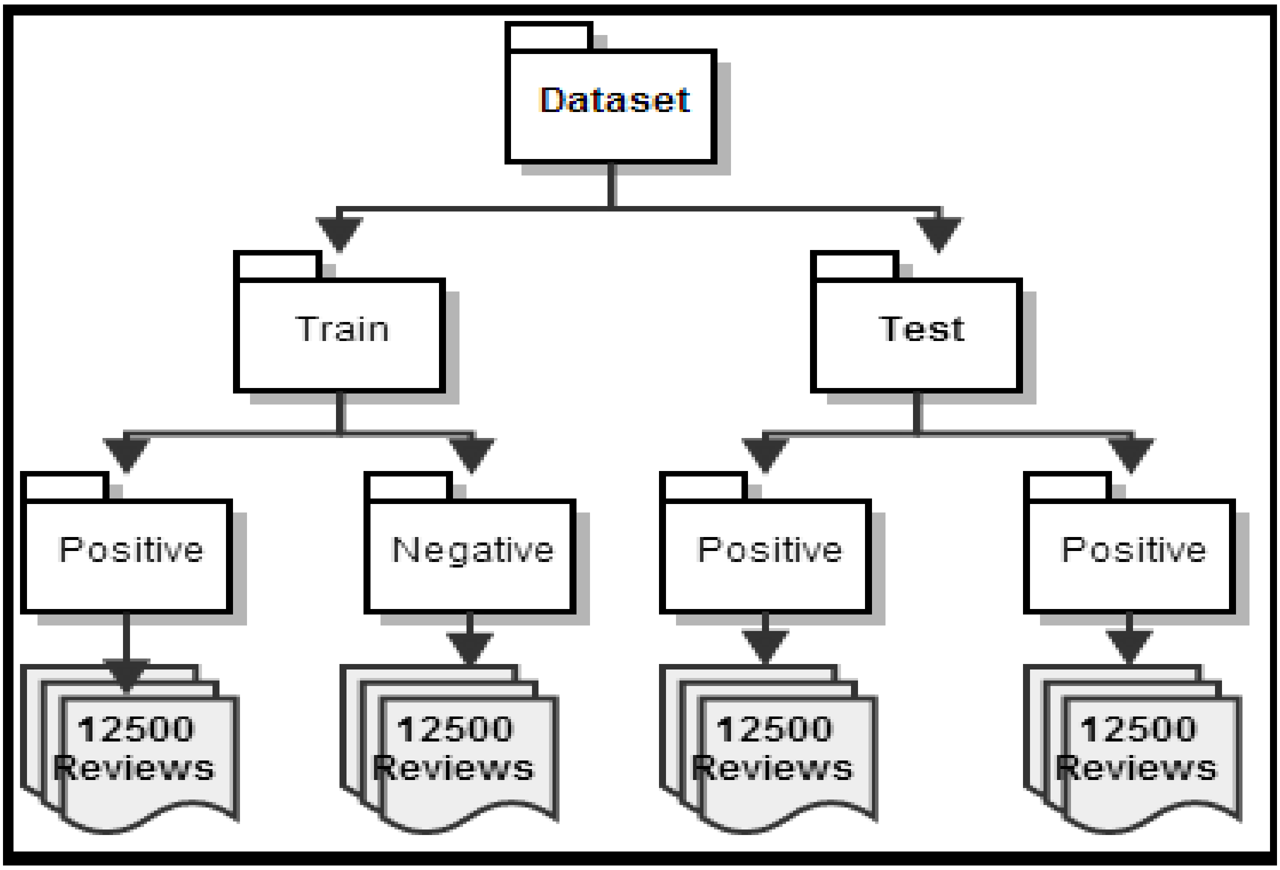
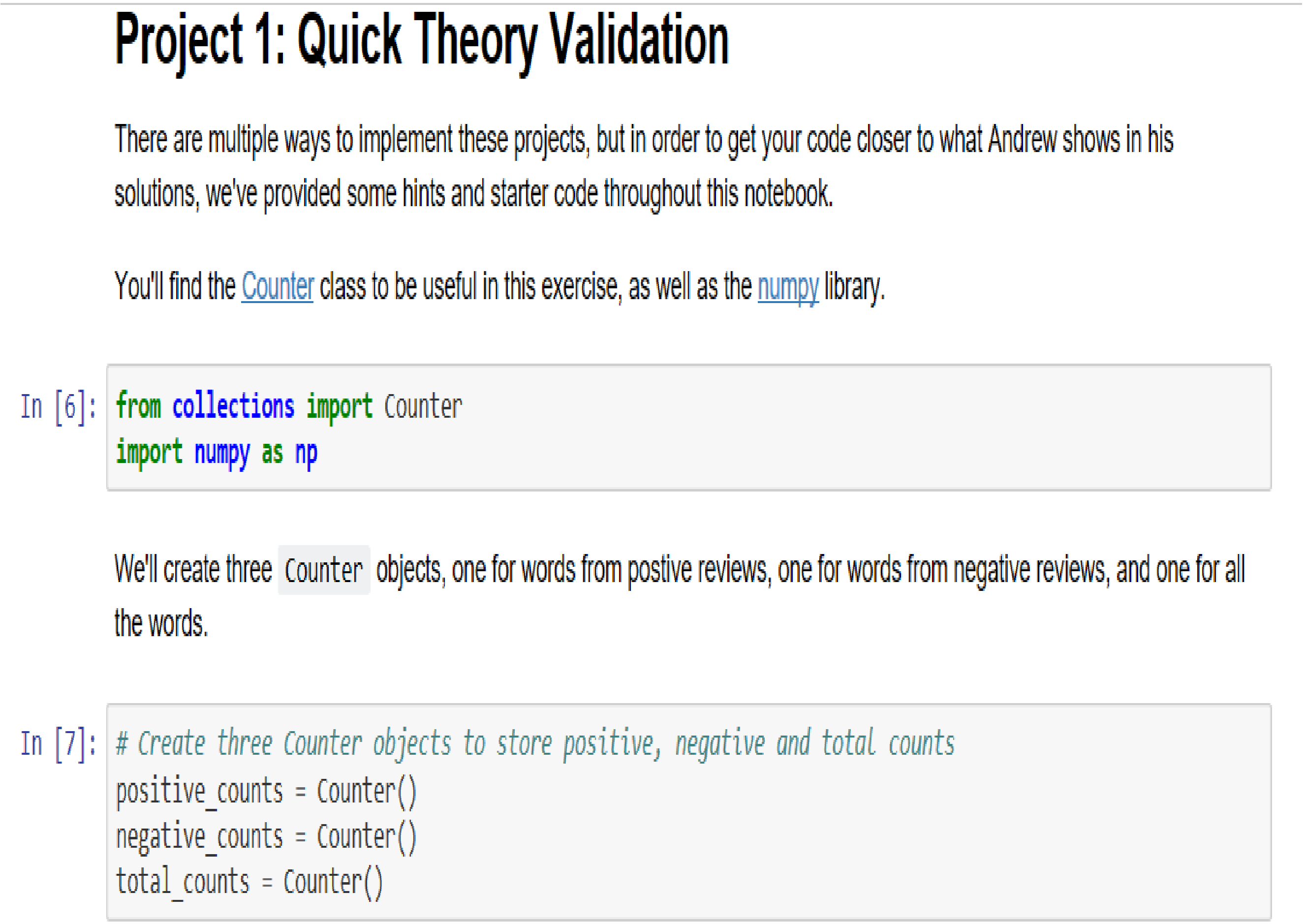


Fig 4.3 Expanded Framework of Train/Test Dataset classification

4.7 Libraries used:

a)NumPy: It is a library for python programming language in which ,adding support for large number of multi -dimensional arrays, matrices and high-level mathematical functions to operate on these arrays. Arbitrary data types can be defined. Though numPy can be easily and seamlessly integrate with large database.



Project 1: Quick Theory Validation

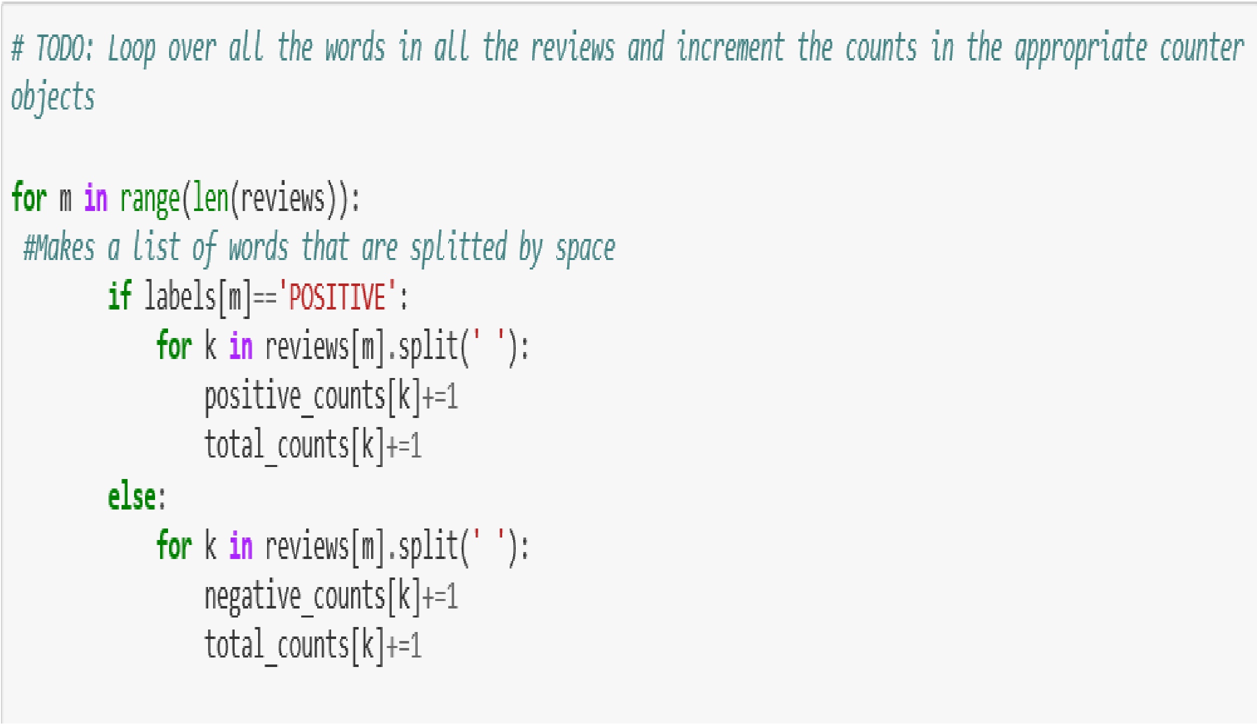
There are multiple ways to implement these projects, but in order to get your code closer to what Andrew shows in his solutions, we've provided some hints and starter code throughout this notebook.

You'll find the Counter class to be useful in this exercise, as well as the numpy library.

```
In [6]: from collections import Counter
        import numpy as np
```

We'll create three `Counter` objects, one for words from postive reviews, one for words from negative reviews, and one for all the words.

```
In [7]: # Create three Counter objects to store positive, negative and total counts
        positive_counts = Counter()
        negative_counts = Counter()
        total_counts = Counter()
```



```
# TODO: Loop over all the words in all the reviews and increment the counts in the appropriate counter objects

for m in range(len(reviews)):
 #Makes a list of words that are splitted by space
        if labels[m]=='POSITIVE':
            for k in reviews[m].split(' '):
                positive_counts[k]+=1
                total_counts[k]+=1
        else:
            for k in reviews[m].split(' '):
                negative_counts[k]+=1
                total_counts[k]+=1
```

b)Vowpal Wabbit:

Vowpal Wabbit is an open source fast out-of-core learning system library. It is specifically use for data mining and substantial as a well-organized scalable implementation of online machine learning and support for a number of machine learning reductions, importance weighting, and a selection of different loss functions and optimization algorithms. It also had descriptive documentation and tutorials to learn and to gain knowledge.

4.8 Applications:

a)Feature identification:

Feature identification is one of the major application of sentimental analysis. For ex: the sentence “this movie has amazing plot and excellent characters”. Selected features are “Plot” and “characters. They will be used to identify the features of a review.

b)Opinion identification:

Determining the polarity i.e. positive and negative is also a very important task. For ex: the sentence “this movie has amazing plot and excellent characters” is of positive polarity because both opinion words are of positive polarity. Many words in natural language have similar meaning. We will combine them or group them as synonyms as a group of similar words together.

c) Synonyms grouping:

Many words in natural language have similar meaning. We will combine them or group them as synonyms as a group of similar words together.

CHAPTER-5

TEST PLAN

5.1 Dataset

The dataset contains 75,000 reviews split evenly into 35k and 35k test sets.. The distribution of overall label is balanced. Afterwards, the given trained and tested dataset are considered as disjoint set of movies, In the given label trained/tested dataset ,a positive review has>=7 out 0f 10,and negative review has score<=4. In the unsupervised set, reviews of any rating are included and there are an even number of reviews > 5 and <= 5.

5.2 Files

There are[train, test] are the two top level directories corresponding to the training and test sets. In these directories they contain (pos,neg) positive and negative containing binary labels. These directories contains, reviews and are stored in various text files named follow where [id] is a unique id and [rating] is the star rating for that review on a 1-10 scale. For example, the file [test/pos/200_8.txt] is the text for a positive-labelled test set example with unique id 200 and star rating 8/10 from IMDb. The [train/unsup/] directory has 0 for all ratings because the ratings are omitted for this portion of the dataset.We also include the IMDb URLs for each review in a separate file. A review with unique id 200 will have its URL on line 200 of this file. Due the ever-changing IMDb, we are unable to link directly to the review, but only to the movie's review page.

In adding up to the appraisal text files, we include previously tokenized bag of words features that was used in our experiment.

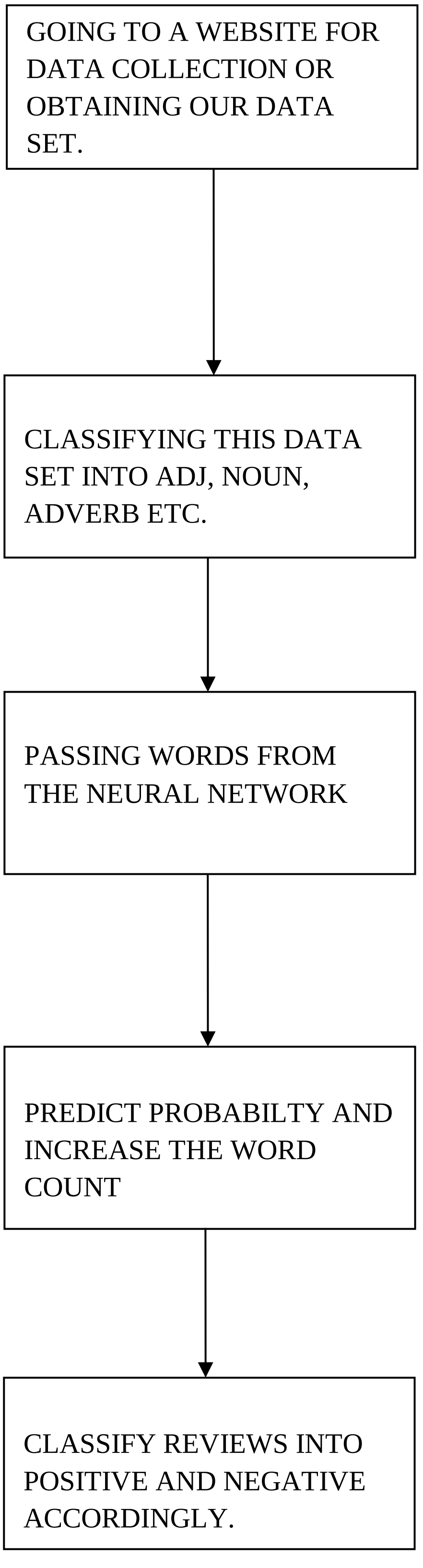

GOING TO A WEBSITE FOR DATA COLLECTION OR OBTAINING OUR DATA SET.
CLASSIFYING THIS DATA SET INTO ADJ, NOUN, ADVERB ETC.
PASSING WORDS FROM THE NEURAL NETWORK
PREDICT PROBABILTY AND INCREASE THE WORD COUNT
CLASSIFY REVIEWS INTO POSITIVE AND NEGATIVE ACCORDINGLY.

Fig 5.1 Flowchart for given process.

CHAPTER-6

RESULTS

The system is tested for 50 plus different movie titles each with max 10 reviews.It shows no of provided features, the accuracy of the system (Accurate Result Percentage), Error Percentage and False Negative Percentage and False Positive Percentage. False Negative means a positive polarity review considered as negative. Similarly, False Positive means a negative polarity review considered as positive. The average accuracy of this system for test review is 85%.

CHAPTER-7

CONCLUSION

In this paper, movie reviews are classified into positive or negative polarity. The system proposed by author in the paper can be used to classify a huge database of movie reviews.This will help movie producers to check the status of their movie. Future work, this API can be trained for other reviews like smartphones, laptops or clothes etc.

REFERENCES:

a) Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Capturing favourability using Natural Language Processing: In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86**.(2007).**

b) Janyce M. Wiebe, Theresa Wilson, and Matthew Bell. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis of the ACL/EACL Workshop on Collocation. **(2008).**

c) Richard M. Working Notes A Sentimental Education: Sentiment Analysis Using Subjectivity of the ACM SIGIR 2001 Workshop on Operational Text Classification, pages 1-6. **(2001).**

d)Rahate, R. S., & Emmanuel, M. Feature selection for sentiment analysis by using svm. International Journal of Computer Applications,Amolik, A., Jivane, N., Bhandari, M., & Venkatesan, movie reviews using machine learning techniques. International Journal of Engineering and Technology**(2013).**

(e) Hira, Z. M., & Gillies, D. F.A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, **(2015).**

(f)Sahu, T. P., & Ahuja, S.Sentiment analysis of movie reviews: A study on feature selection & classification algorithms. In Microelectronics, Computing and Communications(MicroCom), 2016 International Conference on **(2016).**

(g)Nagamma, P., Pruthvi, H. R., Nisha, K. K., & Shwetha, N. H. (2015, May). An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. In Computing, Communication & Automation (ICCCA), 2015 International Conference on **(2015).**

(h) Trupthi, M., Pabboju, S., & Narasimha, G. (2017, January). Sentiment analysis on twitter using streaming API. In Advance Computing Conference (IACC), 2017 IEEE 7th International(pp. 915-919). **(2017).**

(i) Ahmed, E., Sazzad, M. A. U., Islam, M. T., Azad, M., Islam, S., & Ali, M. H. (2017, March).Challenges, comparative analysis and a proposed methodology to predict sentiment from movie reviews using machine learning. In Big Data Analytics and Computational Intelligence (ICBDAC),2017 International Conference on (pp. 86-91). IEEE.

(j) In Electronics, Communication and Aerospace Technology (ICECA), 2017
International conference of (Vol. 1, pp. 6-11). IEEE. Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics.

(k) Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010).Krouska, A., Troussas, C., & Virvou, M. (2016, July). The effect of preprocessing techniques on
Twitter Sentiment Analysis. In Information, Intelligence, Systems & Applications (IISA), International Conference on (pp. 1-5), IEEE.

(l) Bahrainian, S. A., & Dengel, A. (2013, December). Sentiment analysis and summarization of twitter data. In Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on (pp. 227-234),IEEE.Ding, X. and B. Liu. Resolving Object and Attribute Coreference in Opinion Mining. In Proceedings of International Conference on Computational Linguistics (COLING-2010), 2010.

(m) Ding, X., B. Liu, and P. Yu. A holistic lexicon-based approach to opinion mining. In Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008), 2008.

(n) Ding, X., B. Liu, and L. Zhang. Entity discovery and assignment for opinion mining applications. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2009), 2009.

(o) Dragut, E., C. Yu, P. Sistla, and W. Meng. Construction of a sentimental word dictionary. In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2010), 2010.

(p) Du, W. and S. Tan. Building domain-oriented sentiment lexicon by improved information bottleneck. In Proceedings of ACM Conference on Information and Knowledge Management (CIKM-2009), 2009: ACM.