Sentimental Analysis for Social Networking Sites Using Different Techniques

A Dissertation Report
submitted

By

Balrajpreet Kaur

To

Department of Computer Science Engineering

In fulfilment of the Requirement for the

Award of the Degree of

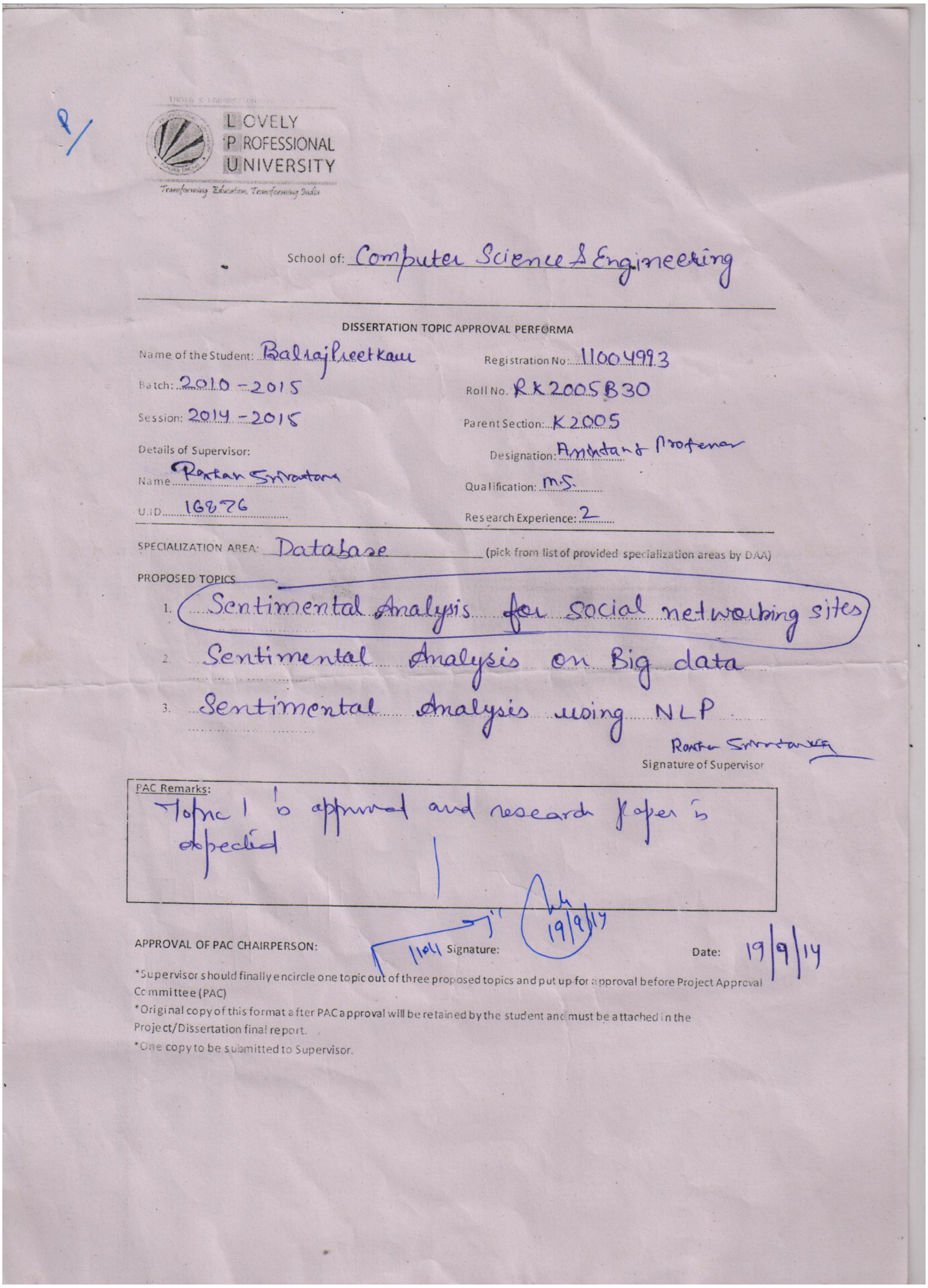
Master of Technology in Computer
Science and Engineering

Under the guidance of

Mr. Roshan Srivastava

(May,2015)

PAC FORM



LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: Computer Science & Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Balrajpreet Kaur

Registration No: 11004993

Batch: 2010 - 2015

Roll No. RK2005B30

Session: 2014 - 2015

Parent Section: K2005

Details of Supervisor:

Designation: Assistant Professor

Name Roshan Srivastava

Qualification: M.S.

U.ID 16876

Research Experience: 2

SPECIALIZATION AREA: Database (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Sentimental Analysis for social networking sites
2. Sentimental Analysis on Big data
3. Sentimental Analysis using NLP

Roshan Srivastava

Signature of Supervisor

PAC Remarks:

Topic 1 is approved and research paper is expected

19/9/14

APPROVAL OF PAC CHAIRPERSON: Signature: Date: 19/9/14

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to Supervisor.

ABSTRACT

Sentiment gives an idea about feeling of a person about something. Sentimental analysis is a type of opinion mining which is used to check the positivity and negativity of data. Sentimental analysis gives us feedback of people through which we can update the product. Sentimental analysis is needed for making updation in business brochure or product. There are different techniques used to check the polarity of data. We calculate the positive and negative score of words. These score help in improving the accuracy of sentiment analysis. We collect the data from social networking sites for analysing opinion of people on different product.Here in chapter 1 we have given introduction about sentiment analysis, its level and machine level techniques. In this chapter we also discuss about advantages of sentiment analysis. In chapter 2 we discuss about literature review about various paper. In chapter 3 we discuss aboutscope of study. In chapter 4 we discuss about objective of study. In chapter 5 we discuss about present work in which we discuss about research methodology. In chapter 6 we discuss about result and discussion about the implementation output. In chapter 7 we discuss about conclusion and future scope.

CERTIFICATE

This is to certify that Balrajpreet kaur has completed M.Tech dissertation proposal titled "Sentimental analysis for social networking sites using different techniques" under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfillment of the condition for the award of M.Tech in Computer Science and Engineering.

Date:

Signature of Advisor

Name: Mr.Roshan Srivastava

UID: 16876

ACKNOWLEGEMENT

The report has been written with the kind guidance and support of my mentor. The satisfaction and happiness that accompany the successful completion of any task would be incomplete without mentioning the names of people who made it possible, whose constant guidance and encouragement crowns all efforts with our success.

I would like to express my deep gratitude and thanks to my mentor **Mr.Roshan Srivastava,** Asst.Professor, as well as **Mr**. **Dalwinder Sir**, HOD, School Of CSE, Lovely Professional University, Phagwara, Punjab for their help and guidance throughout my dissertation work. I have received an enormous amount of valuable advice and knowledge from her that helps me a lot in my work. Thank you, **Mr.Roshan Srivastava**, for providing me motivating support in my research work.

DECLARATION

I hereby declare that the dissertation proposal entitled, "**Sentimental analysis for social networking sites using different techniques**" submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date:

Investigator
Regn No.:- 11004993

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Sentiment Analysis

There is a big influence of internet on our daily life. The internet provides us social media to connect to the world. We use Facebook, twitter, LinkedIn and more sites to get connected to people. We also give opinions and review about a process or product. Sentiment analysis is made up of two different keywords, first is sentiment which defines emotion, opinion and behavior about particular thing. Second one is analysis which defines as examination of structure i.e. breaking the sentences for getting result by which we can further improve the performance of a product or process. Both of these keywords give idea about any product. The main aim of sentiment analysis to get the reaction or behavior of a person about product or specific subject

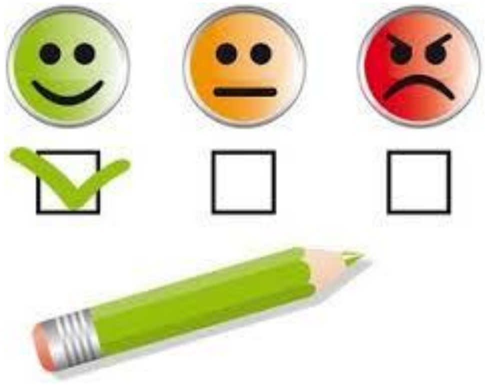


Figure 1 Different sentiments of people

1.2 Types of Opinion

- **Regular opinion:-**It is the opinion of people on some entities which are targeted.

- **Direct opinion:-**The sentences which are having direct meaning. For example, “the phone is really good”.

- **Indirect opinion:-**The sentences which are complex as compared to direct opinion. For example, “after taking this painkiller, I feel relief”.

- **Comparative opinion:-**Comparison of more than one entity. For example comparing the quality of two products.

1.3 Features Of Sentiment Analysis

- **Term Based Features:-**The term based features in text is determined by unigram, bigram and trigram. We can find the tokens from text by using unigram, bigram and trigram. The bigram gives the two words that occur together and trigram gives the three words from sentence that occurs together.

- **Part Of Speech (POS):-**Part of speech (POS) is very important part of sentiment analysis. It divides the words of sentence into grammatical form like adjective, adverb, noun, verb, pronoun etc. The part of speech provides the grammatical information of sentence. It is useful analysing the opinion and sentiments of people. Due to POS we can get the adjectives easily. The part of speech provides tokens containing labels of noun, pronoun, verb etc. We can use part of speech in different tool to get accurate result.

- **Syntax:-**In syntax we generally focus on the grammatical structure of the sentence. The syntax provides us the meaning of the sentence. There are two types of syntax:-

  - Simple syntax

  - Complex syntax

By the help of syntax we can determine the difference between the meanings of sentence containing similar words.

- **Negation:-**The negation is very important feature of sentiment analysis. It changes the meaning of sentence. For example, "I like this food "and "I don't like this food". The "don't" is a type of negation.it converts the positive sentence into negative sentence. The negation can change the meaning of complex sentences.

1.4 Different Level Of Sentiment Analysis

There are four types of sentiment analysis level :-

- **Document level of sentiment analysis:-**In document level the opinion is classified either in positive or negative. In document level we check single entity for result.
- **Sentence level of sentiment analysis:-**Sentence level refine the review of document level analysis. It also filters the sentence containing nothing.
- **Aspect based sentiment analysis:-**Aspect based sentiment analysis focus on all sentiment expression. It recognizes the aspects which contain the opinion. Aspect based sentiment analysis is basically used for getting the opinion which hide due to categorization of words i.e. The sentence which contain both positive and negative opinion.
- **Comparative sentiment analysis:-**In comparative sentiment analysis we get opinions by comparing it with similar product. In this we use sentiment lexicon acquisition which contains three approaches:-manual approach, Dictionary based approach and corpus based approach.

1.5 Sentiment Analysis Basic Working

The basic working of sentiment analysis includes collection of data from blogs, websites and reviews. The collection of data from different sources acts as a corpus. After that we process the documents by excluding the statements which are not having any sense. Next step is to apply different techniques for analysis for getting result.

From these techniques we can get the result of analysis with accuracy. These techniques can help in understanding opinion of complex sentences. We can also use unstructured data for analysis with these techniques.

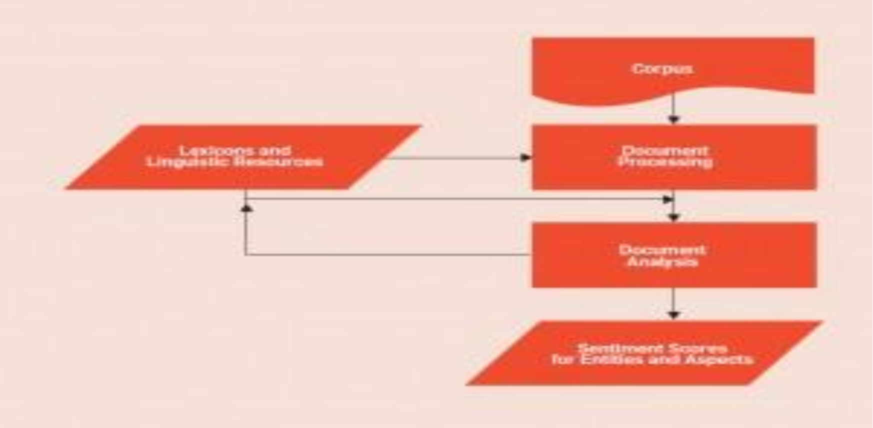

Corpus
Lexicons and Linguistic Resources
Document Processing
Document Analysis
Sentiment Scores for Entities and Aspects

Figure 2 Architecture of sentiment analysis

1.6 Machine Learning Techniques

The various machine learning techniques are discussed below.

1.6.1 Supervised Learning

In supervised learning the data contain relationship between data attribute and target attribute. The patterns discovered are based on these relationships. Supervised learning is used for data mining to find dependent variable. The supervised learning containing two sub techniques i.e. Naïve Bayes and Support Vector Machine (SVM).

- **Naïve Bayes:-**Naïve Bayes classifier is used for predicators which are independent of one another. It containing training the data and after that the finding the probability. Using this technique we can check the probability of words.
- **Support Vector Machine (SVM):-**We can use support vector machine to classify data into two types i.e. positive and negative. The SVM is mostly use where data is having only two classes. There is a hyper lane that separates the data from each other. For example if there is a data containing both positive and negative data then hyper lane divide the data into two parts by providing margin between them.

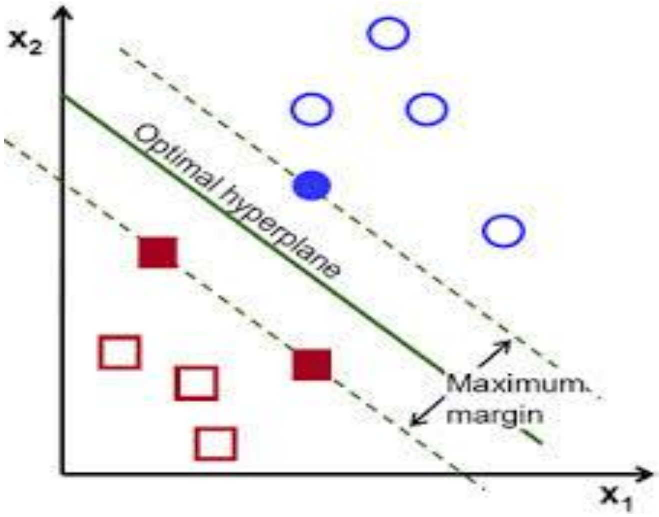


Figure 3 SVM Diagram

1.6.1 Unsupervised learning

The data that doesn’t depend on target attributes. Unsupervised learning is basically used for clustering. In unsupervised learning we collect more specific data. In unsupervised learning we are having positive, negative as well as neutral thoughts.

1.6.2 Semi-supervised learning

Semi-supervised learning is a learning in which we learn how to label and unlabelled the data. It is in between the supervised and unsupervised learning. It shows the labelled data and unlabelled data affect the earning behaviour. It is mostly used in machine learning and data mining because it can improve learning task using unlabelled data rather than labelled data which is expensive. It has three assumptions first the cluster assumption and second smoothness assumption. There so many methods for semi-supervised learning as following:-

- Graph method
- Using generative mode
- Heuristic method
- Using low density separation method

1.6.3 Natural language processing

Natural language processing is used for understanding human language. Natural language processing is used for morphological segmentation, syntactic analysis, summarization of data, tagging of data, speech recognition and information retrieval. In speech tagging the NLP tag the word into different categories such as noun, verb, adverb and adjective. In part of speech process tokenization make small-small tokens such as words, number etc. The main goal of NLP is to get efficient result in understanding human language.

1.6.4 Other techniques:-

- **Wordnet:-**It contains the group of words based on their synonyms relationship. For example, Develop synonyms are grow, become larger. It also contains relationship of Antonym like presence and absence.
- **Sentiwordnet:-**Sentiwordnet is a resource of lexical. It is the advance version of wordnet. It contains the score of word as positive, negative and neutral. The neutral score is in the form of object score.

  The score values are in the numerical form and it is very useful for getting the result with numerical value. The Sentiwordnet have three versions. The Sentiwordnet 3.0 is the latest version of Sentiwordnet which contain the synonym, anonym as well as the suffix and prefix also. The latest version is 20% more accurate from the other old version. The Sentiwordnet is available in many tools such as NLTK.

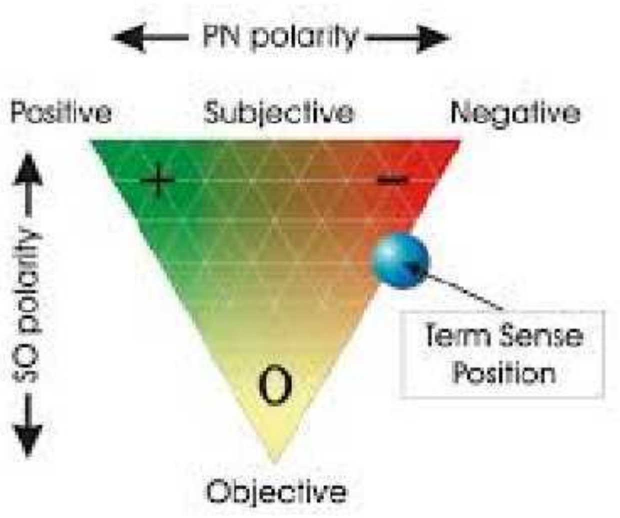


Figure 4 Sentiwordnet graphical representation

- **PMI:-**PMI is pointwise mutual information. It is used to find the words which are associated with each other. PMI tells about the co-occurrence of words which are together in the sentence. In some tools it is defined as function name as collocation.

  The formula of PMI is

$$PMI(x,y) = log(P(x,y)/p(x)p(y))$$

1.7 Challenges In Sentiment Analysis

- The opinion of people is in complex sentences.
- Humour and the different type of human speech is difficult to understand.
- Negation and changing topic.
- Sarcasm and different types of emotions.
- Difference in culture of people.
- The meaning of negative words in positive way is also difficult to analyse.
- Conversion of unstructured data into structured data.
- If only individual person is giving opinion then no analysis can be done.

1.8 Advantages Of Sentiment Analysis

- Sentimental analysis monitoring the brand of different companies and gives the analysed result.
- The sentiment analysis gives the rating of reviews.
- Sentiment analysis gives the idea about the changes which are needed by the users through this analysis.
- Sentiment analysis helps in decision making also.
- It also gives information about market strategies and also the sales of product.

CHAPTER 2
REVIEW OF LITERATURE

Marc Cheong et al[8] detected the hidden patterns from twitter messages. They used two techniques (a) the Cheong and Lee's content analysis framework and (b) SOM algorithm for visualization. For data gathering they used Twitter message corpus.

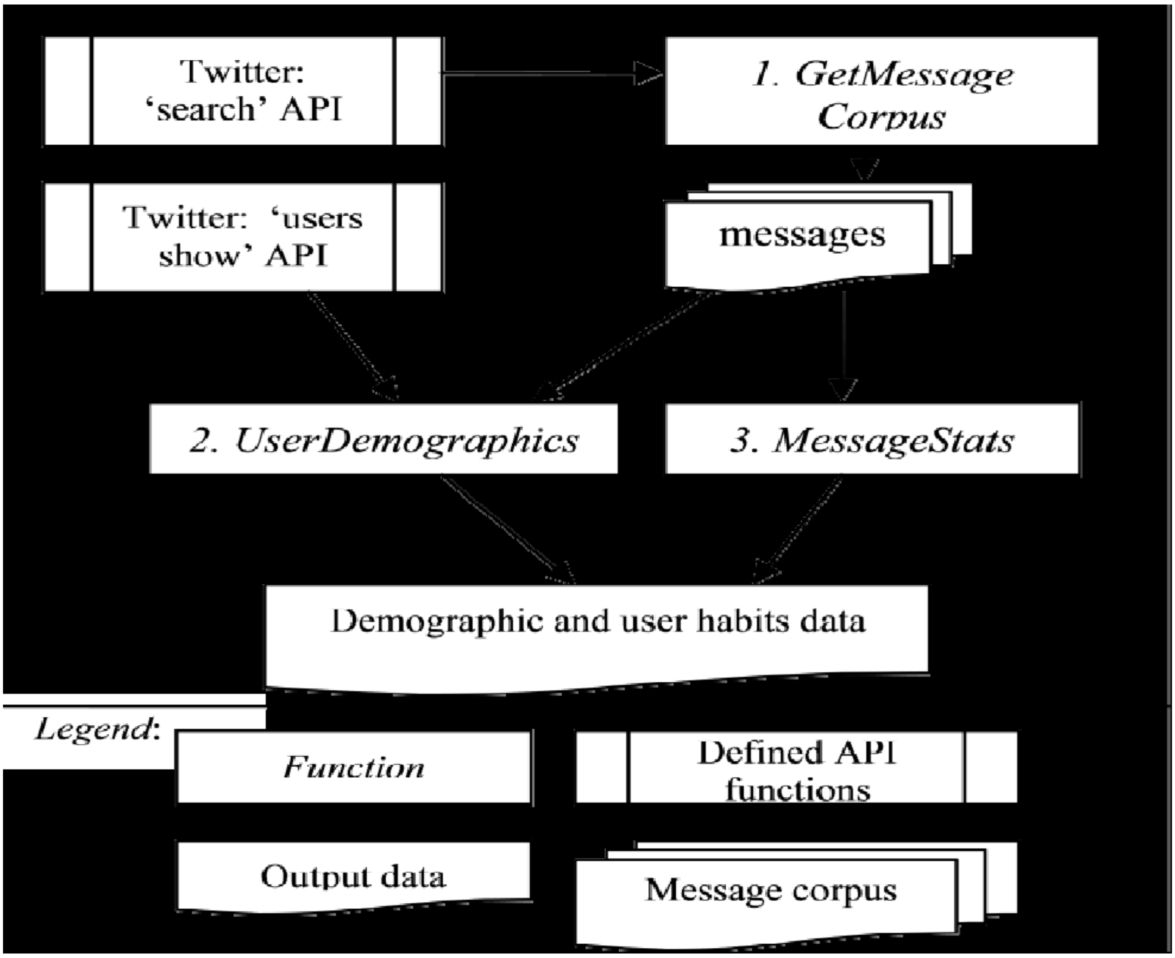


Figure 5 Cheong and Lee's framework.

They searched messages which helped them in getting useful data for clustering. They also used massage attributes and user attributes. In message attributes they determine the common message indicator, pictures and the type of device used by twitter user. In user attribute they determine the habits of user by checking their post, account age and friends following. For clustering and visualization they use SOM algorithm. SOM is a technique based on artificial neural network. In SOM clusters are made in the form of Map. In this paper the authors consider three topics:-

- Iran election issue in 2009.
- The iPhone OS 3.0 software launch.
- US President Obama's foreign policy.

In SOM different colors like red, blue, green and yellow are used for visualizing result.

Tetsuya Nasukawa et al[13] sentiment analysis approach is used to check the positive and negative opinion for specific subjects from a document. In this paper the positive indicate favourable opinion and negative indicate unfavourable opinion. NLP is used for capturing favourability. There are three identifications used:-

- Sentiment expression.
- Polarity and strength of the expression.
- Their relationship to the subject.

The authors used notation which consist of the following information:-

- **Polarity:-**good, bad, or neutral is denoted by g, b, or n, respectively, and sentiment transfer verbs are denoted by t.
- **Part of speech (POS):-**Contains adjective (JJ), adverb (RB) and verb (VB) are registered in our lexicon arguments such as subject (sub) and object (obj) that get sentiment from a sentiment verb or arguments that provide sentiment to and receive sentiment from a sentiment transfer verb.

In this paper the algorithm was used which contain upper limit of 50words. For identifying sentiment expression POS tagging was used. For POS tagging, author used Markov-model based tagger. In this paper the techniques fails to identify positive result if there is any negative word in the sentence. For example, "it is difficult to take a bad picture from this camera ". The result of this sentence polarity was negative according to above algorithm because the sentence containing word 'bad'.

Alena Neviarouskaya et al[1] proposed Sentiful is a lexicon based system for assigning scores to word for analysis of sentiment. In this paper the structure relations are purposed. The positive score and negative score tell about the polarity of a sentence. The author uses sentiword net for giving value to words on the basis of polarity.

For generating the sentiment lexicon the first step is to collect the content of word. There are four formulas's used to calculate positivity and negativity:-

$$Pos\_score = \left[\frac{\sum_{i=1}^{pos} Intensity(i)}{pos}\right], \quad (1)$$

$$Neg\_score = \left[\frac{\sum_{i=1}^{neg} Intensity(i)}{neg}\right], \quad (2)$$

$$Pos\_weight = \left[\frac{pos}{pos + neg}\right], \quad (3)$$

$$Neg\_weight = \left[\frac{neg}{pos + neg}\right], \quad (4)$$

There are three scores object score, positive score and negative score. The score range from 0.0 to 1.0 and the sum up to 1.0. If the positive score is greater than negative score the word is positive and if negative score is greater than positive score then the word is negative else if object score is greater and negative positive score are equal then the word is neutral.

There are four methods for Expanding sentiful**:-**

1.**Finding new lexical units through synonymy relation:-**In this method we derive new words from synonymy relation. In this authors consider the words which are having same meaning and assign sentiment score to them.

2. **Examining direct antonym relation**:-In this method authors consider the words which having opposite meaning. For example, carelessly and carefully. The authors calculate the positive and negative score of word through this relationship. By this method new words are added to WordNet.

3. **Examining Hyponymy relation:-**In this method the relations based on hierarchy between words. For example attainments => success -> winning. By this method we get a list of hyponyms from word net and it also remove the duplicate words from word net.

4. **Method to derive and score morphologically modified words:-** Morphologically words are derived through suffixes and prefixes. The meaning of many suffixes can change the meaning of word especially in term of sentiment.

In this paper the algorithm of derivation and scoring of new words give high accurate result. Through sentiful method we can expand sentiment lexicon and improve the result of polarity. The different tables in this paper help us in examine the words polarity .

Neethu Mohandas et al[9] focused on mood extraction from Malayalam text. The authors used NLP for part of speech tagging and PMI formula for SO calculation. The semantic orientation (SO) gives the result in numerical form through which we can check the polarity of sentence.

The method used in this paper

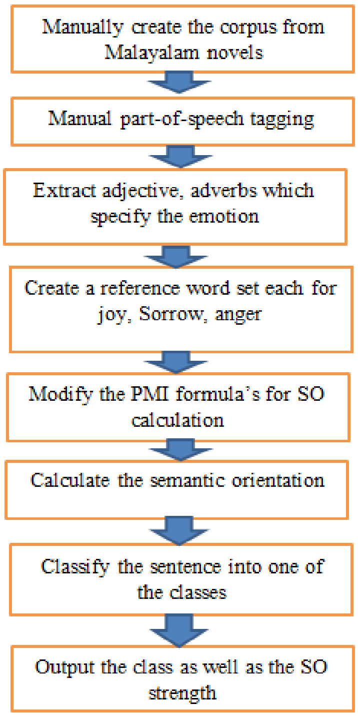

Manually create the corpus from Malayalam novels
Manual part-of-speech tagging
Extract adjective, adverbs which specify the emotion
Create a reference word set each for joy, Sorrow, anger
Modify the PMI formula's for SO calculation
Calculate the semantic orientation
Classify the sentence into one of the classes
Output the class as well as the SO strength

Figure 6 .Proposed method

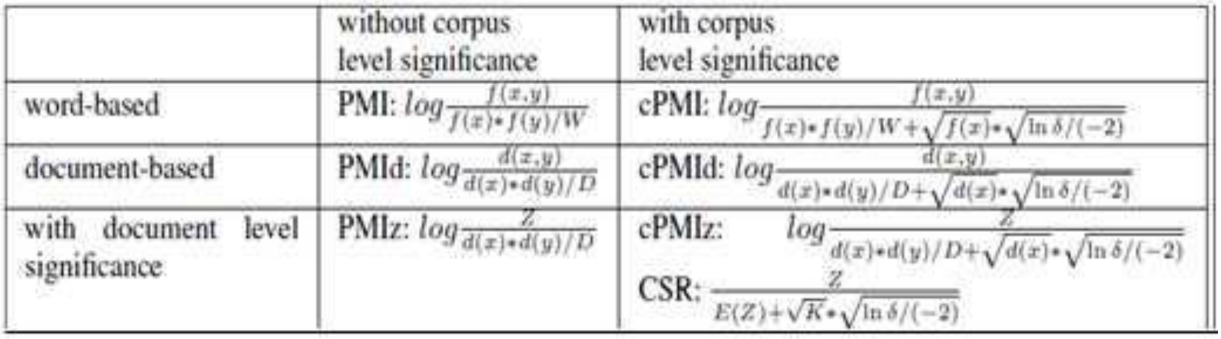
This paper gives idea to implement sentiment analysis on different languages such as Hindi, Punjabi, Marathi etc. By this method we can get result on applications contain different languages. We can calculate sentiment analysis of different movies which are based on area language and their comments are also in that language .

Peter D.Turney et al[11]used semantic orientation concept to calculate the positive and negative review. The author uses average semantic orientation to get result of review in positive and negative. In this paper the author explain pointwise mutual information (PMI) and latent semantic analysis technique.

The author done experiment by using these technique and calculate the result on the basis of words which they included in the database. The authors also compare the result from both techniques on different data and explain about technique with experimental result. They show result in the form of diagram also which is in graphical form.

Julia Kreutzer et al[5] give information about sentiwordnet .In this paper the author explains the applications and usage of wordnet. The authors also explain the structure of sentiwordnet. They also give examples of scores and scoring review of words. The authors also explain about the analysis and scoring technique. They also include the things which are recently updated in the sentiwordnet and also show the scores of positive, negative and objective word. The authors also explain the score of various word which doesn't provide any sentiments such as grandmother, brother etc. The author also explains the problems in Sentiwordnet.

OM P.Damini et al[10] explained the concept of PMI. PMI is basically used to measure the co-occurrence between two words. For calculating PMI by various methods they used different formula's which are shown in the following table.



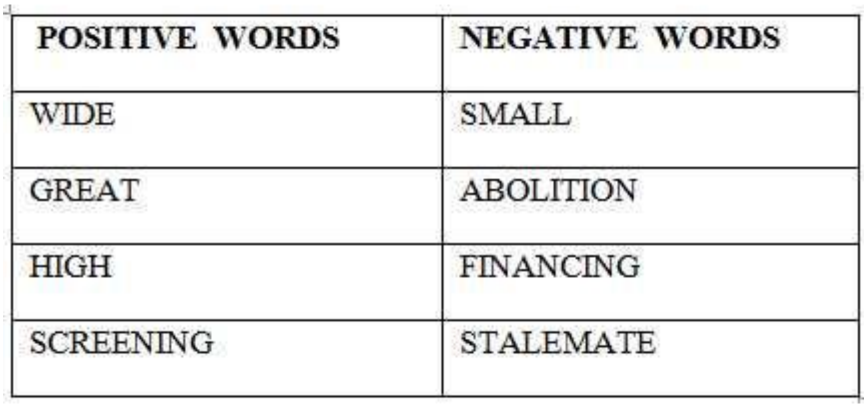
	without corpus level significance	with corpus level significance
word-based	PMI: $log\frac{f(x,y)}{f(x)*f(y)/W}$	cPMI: $log\frac{f(x,y)}{f(x)*f(y)/W+\sqrt{f(x)}*\sqrt{\ln\delta/(-2)}}$
document-based	PMId: $log\frac{d(x,y)}{d(x)*d(y)/D}$	cPMId: $log\frac{d(x,y)}{d(x)*d(y)/D+\sqrt{d(x)}*\sqrt{\ln\delta/(-2)}}$
with document level significance	PMIz: $log\frac{Z}{d(x)*d(y)/D}$	cPMIz: $log\frac{Z}{d(x)*d(y)/D+\sqrt{d(x)}*\sqrt{\ln\delta/(-2)}}$ CSR: $\frac{Z}{E(Z)+\sqrt{K}*\sqrt{\ln\delta/(-2)}}$

Table 1 Formula of PMI

They evaluate their result by correlating their result with gold standard dataset. The author used two types of co-occurrence i.e. document level and corpus level.

Keisuke Mizumoto et al[6] in this paper the stock market is analysed by the polarity dictionary. The dictionary is made up by using semi-supervised learning. The bootstrapping approach is used .the bootstrapping technique help in counting the words. After that they check how many times the number positive word is occurred more as compared to negative words or vice-versa.

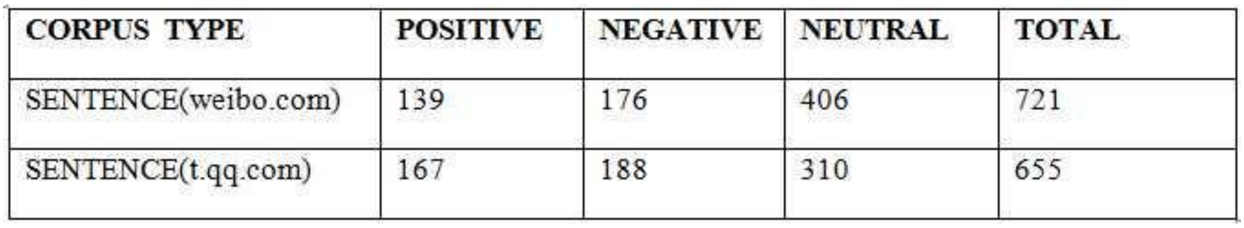
The stock market news data is collected first and after that we check the polarities of data from the data dictionary. They make graph on the basis of threshold for showing their results.



POSITIVE WORDS	NEGATIVE WORDS
WIDE	SMALL
GREAT	ABOLITION
HIGH	FINANCING
SCREENING	STALEMATE

Table 2 Table of word included in dictionary

Bin wen et al[2] in their paper explained about the PMI technique with semantic orientation. There are different pattern extracted by using different formulas. There are table which contain adjective, adverb, noun, verb with their orientation.



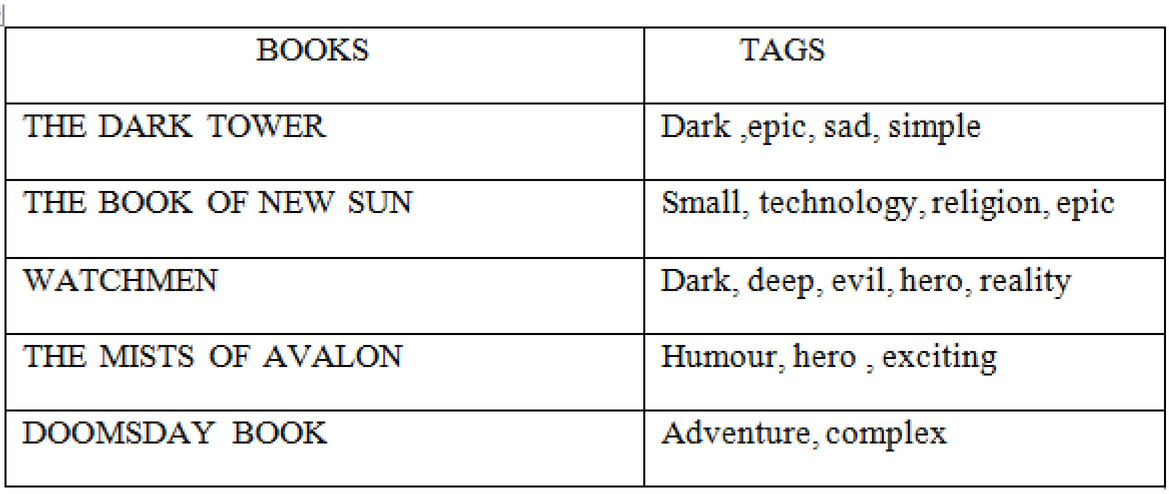
CORPUS TYPE	POSITIVE	NEGATIVE	NEUTRAL	TOTAL
SENTENCE(weibo.com)	139	176	406	721
SENTENCE(t.qq.com)	167	188	310	655

Table 3 Table of test data

The result is calculated on the basis of recall and precision.

Eric lin et al[4] in their paper explained the mining of online book review. The authors take the review of people which are available on the website. The analysis first step is data gathering and after that removing of those sentences which doesn't contain any meaning. After this step the mining of context is done by using weight formula.

There is a table contains the book heading with tags. The tags are labelled according to the features of books. The tags can be science, nature, innovation etc. There is a hierarchal clustering by using book name and tags. There is a clustering graph which shows the polarity of data. The peaks in the graph show the number of positive and negative comments in the review.

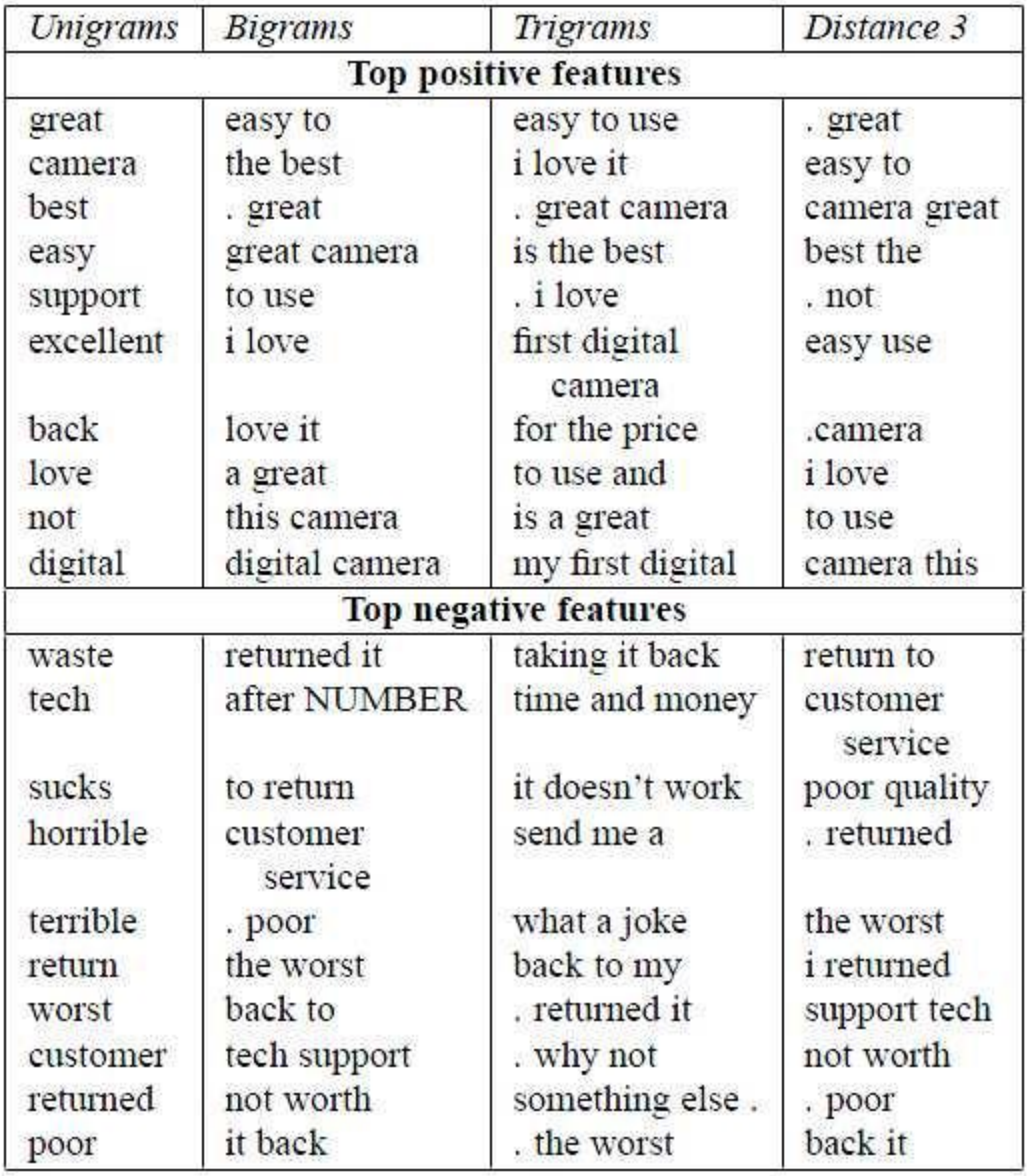


BOOKS	TAGS
THE DARK TOWER	Dark ,epic, sad, simple
THE BOOK OF NEW SUN	Small, technology, religion, epic
WATCHMEN	Dark, deep, evil, hero, reality
THE MISTS OF AVALON	Humour, hero , exciting
DOOMSDAY BOOK	Adventure, complex

Table 4 Table of books and tags

Samaneh moghaddan et al[12] in this paper the aspect based question answering method is used. The different type of question is asked from user to get answer and on the basis of that answer we can get the opinion of a person. The questions are simple. It it is based on the comparison of two products so that we can get the opinion of a person. There different type of question is used such as target, attitude etc. There is a table of correct and incorrect through which we can accurate result. The part of speech is also used in this paper.

Kushal Dave et al[7] found identifying the unique characteristics of mining problem and developed a method for automatically differentiating between positive and negative reviews of certain movie. Their classifier depicts information retrieval methods for extracting features and marking, for various metrics and heuristics the results vary depending on testing condition. The best methods found working better than traditional methods.



Unigrams	*Bigrams*	*Trigrams*	*Distance 3*
Top positive features			
great	easy to	easy to use	. great
camera	the best	i love it	easy to
best	. great	. great camera	camera great
easy	great camera	is the best	best the
support	to use	. i love	. not
excellent	i love	first digital camera	easy use
back	love it	for the price	.camera
love	a great	to use and	i love
not	this camera	is a great	to use
digital	digital camera	my first digital	camera this
Top negative features			
waste	returned it	taking it back	return to
tech	after NUMBER	time and money	customer service
sucks	to return	it doesn’t work	poor quality
horrible	customer service	send me a	. returned
terrible	. poor	what a joke	the worst
return	the worst	back to my	i returned
worst	back to	. returned it	support tech
customer	tech support	. why not	not worth
returned	not worth	something else .	. poor
poor	it back	. the worst	back it

Table 5 Table of features.

Lillian Lee et al[3] found movie reviews from various sources like movie streaming sites, rental websites as raw data, where they found that machine learning techniques can suppress the human defined baselines in artificial intelligence. They used Naive Bayes, support vector machines and maximum entropy classification, do not perform as efficient as sentiment classification as on traditional topic-based categorization of text.

CHAPTER 3
SCOPE OF STUDY

The scope of research incorporates various aspects of sentiment analysis in business intelligence. The sentiment analysis is applied on structured data as well as unstructured data. Sentiment analysis is used with big data also nowadays.

- We can apply sentiment analysis on big data to get analysis on unstructured data. Now days various companies uses Facebook, twitter and other social networking sites data for analysing the habits and opinion of people about any subject.

- The analysed result from sentiment analysis provides various advantages in decision making in business. The main focus of sentiment analysis is to get true sentiment of people about any subject. By applying various analysing tools and algorithm we can get more accurate result.

- The sentiment analysis can be more optimised by using efficient algorithm of finding patterns.

- The can make algorithm which use sentence level as well as aspect based sentiment analysis.

CHAPTER 4
OBJECTIVES OF STUDY

We introduce two main techniques for sentiment analysis. The first technique is PMI which is used to check the co-occurrence of two words. By this technique we can analyse which part of sentences have more positive result and which words are co-occur more. After that we use sentiwordnet formula for calculating the positive and negative score by using values which are present on sentiwordnet. By calculating the positive and negative values we can easily calculate semantic orientation. The techniques which we are using make result more accurate and can describe the polarity of data. The tagging of data is also useful which provide small - small parts of sentence in different categories which help in analysing the sentence.

CHAPTER 5
PRESENT WORK

5.1 Problem formulation

The major issue in sentiment analysis is to understand the opinion of a person. There are two types of sentences which are used by person first is the simple and second complex. It is difficult to understand the complex sentences.

The issue is the structured and unstructured sentences. If we cannot structured the data it is difficult to understand the data. The review of people is in the unstructured form, we need to first change it into structured from for getting result.

If there is a sentence which contains two negative words such as pain and killer and it is a positive word in the sentence then it is difficult to judge the meaning because both the words are negative.

It is quite difficult to understand the opinion of people completely but by using some methods and technique we can make it little bit easier to understand the sentiments of a person. The technique such as Sentiwordnet will provide the result in the numerical form. The solution of this problem is solved by using these techniques to get result.

5.2 Research Design

In this methodology we first collect the sentences from social networking sites such as Facebook and twitter. After that collocation function is used this function is already predefined in the Ipython. Through this function we get the words which occur together in the text. Now we use PMI method for counting the co-occurrence of words many times in the text manually.

After this we use tagging of text by using some lines of code. After executing the code we get the adjectives and adverbs from the text. After that we apply Sentiwordnet code for finding the value of words then we manually write the values and find the average of positive and negative value of words. After that we apply semantic orientation by subtracting negative values average from positive value average. If we get the positive value as result this means the text is positive either if negative result then the result is negative. We can get the numerical value result so that we can prove our result.

There are some points which is considered during this methodology:-

- If there are two negative values together then it became the positive value.
- If there is any word value which we don't get in the Sentiwordnet through python then we can check the value from website.
- If there is any positive word with not then its value is considered as negative value.
- If the complex sentence result is not match with true value then we take that problem in future scope.

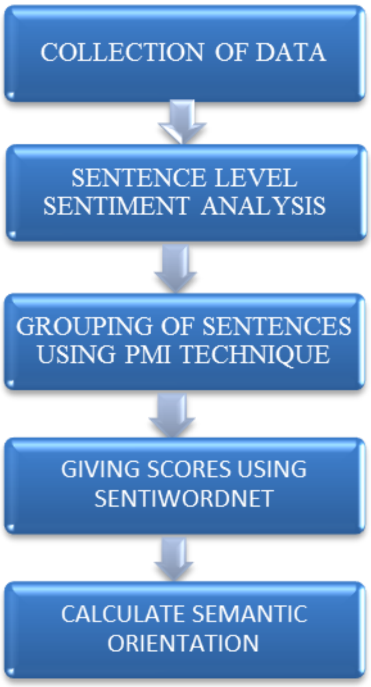

COLLECTION OF DATA
SENTENCE LEVEL SENTIMENT ANALYSIS
GROUPING OF SENTENCES USING PMI TECHNIQUE
GIVING SCORES USING SENTIWORDNET
CALCULATE SEMANTIC ORIENTATION

Figure 7 Research design

This research design is combination of different techniques. We can get result in the form of numerical value to check polarity of data.

CHAPTER 6
RESULT AND DISCUSSION

In this section the proposed method implemented result is shown. We implemented the methodology in Ipython which is included in Anaconda tool. We take comments of people on Samsung mobile through Facebook page of Samsung. We take 50 comments line. These comments included simple as well as complex sentences. We first remove those sentences which don't include any adjective and adverb. The sentences or comments are stored in a text file. We need to add file in a directory of python. We add that file in the python script. We can check the path for adding file by writing the file name in the file reading, it will give error of no such file is find out in this directory. From these errors we can easily calculate the path.

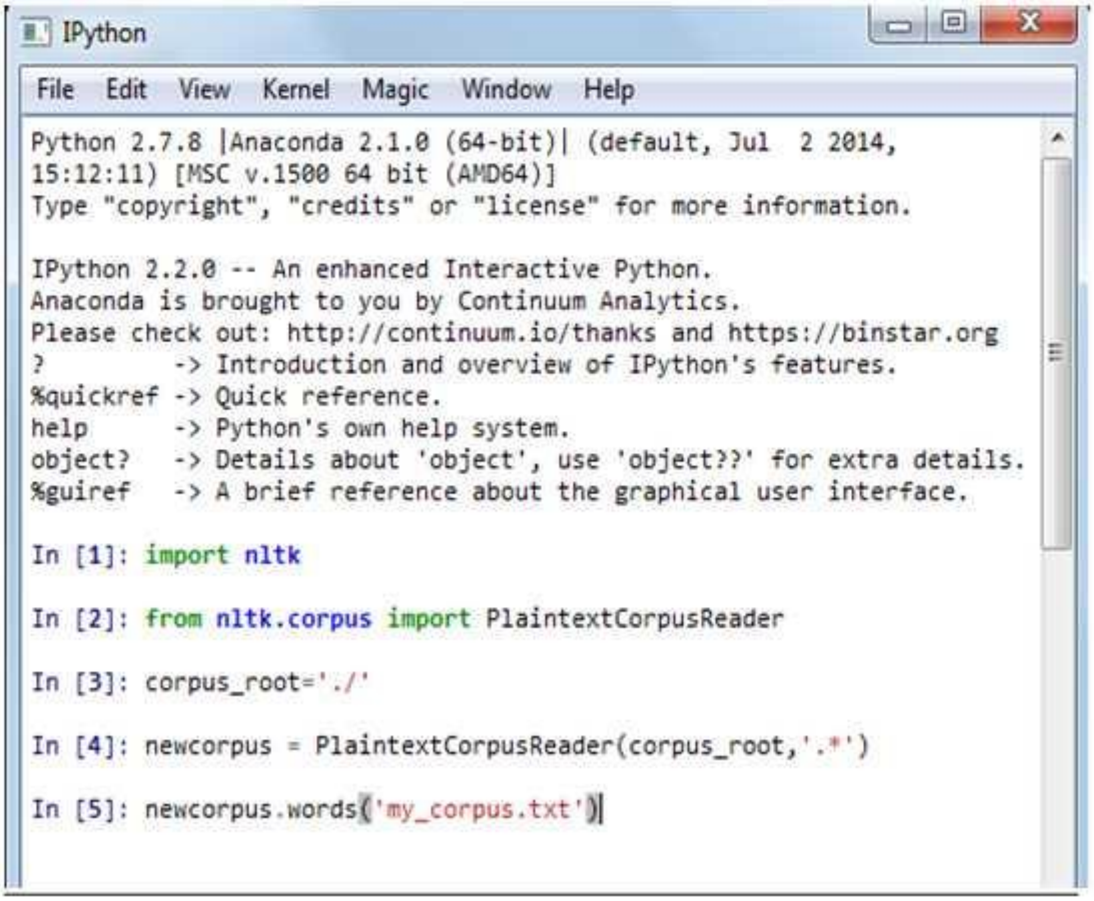
We can also count the specific word which we think repeated in the sentence by using collocation function. Using this function we can easily multiply the value in numerical form. We can also use word tokenization if we want to check the sentiments by taking single sentence.

We use part of speech tagging in the implementation so that we can easily find out the adjective, adverb, noun and pronoun. The part of speech gives us idea about word polarity in the sentence. In our implementation we manually write the adjectives and adverb to find their value in the Sentiwordnet.

It's difficult to do manually when there are huge amount of data. So we take small data which contain maximum 6 line so that we can easily find out the adjective and adverb used in Sentiwordnet. After writing these adjectives and adverb we can easily find their value from Sentiwordnet.

We need to write the specific word in code to get their positive and negative value. After that we collect the value of semantic orientation by applying the formula in which average positive subtract from average negative.

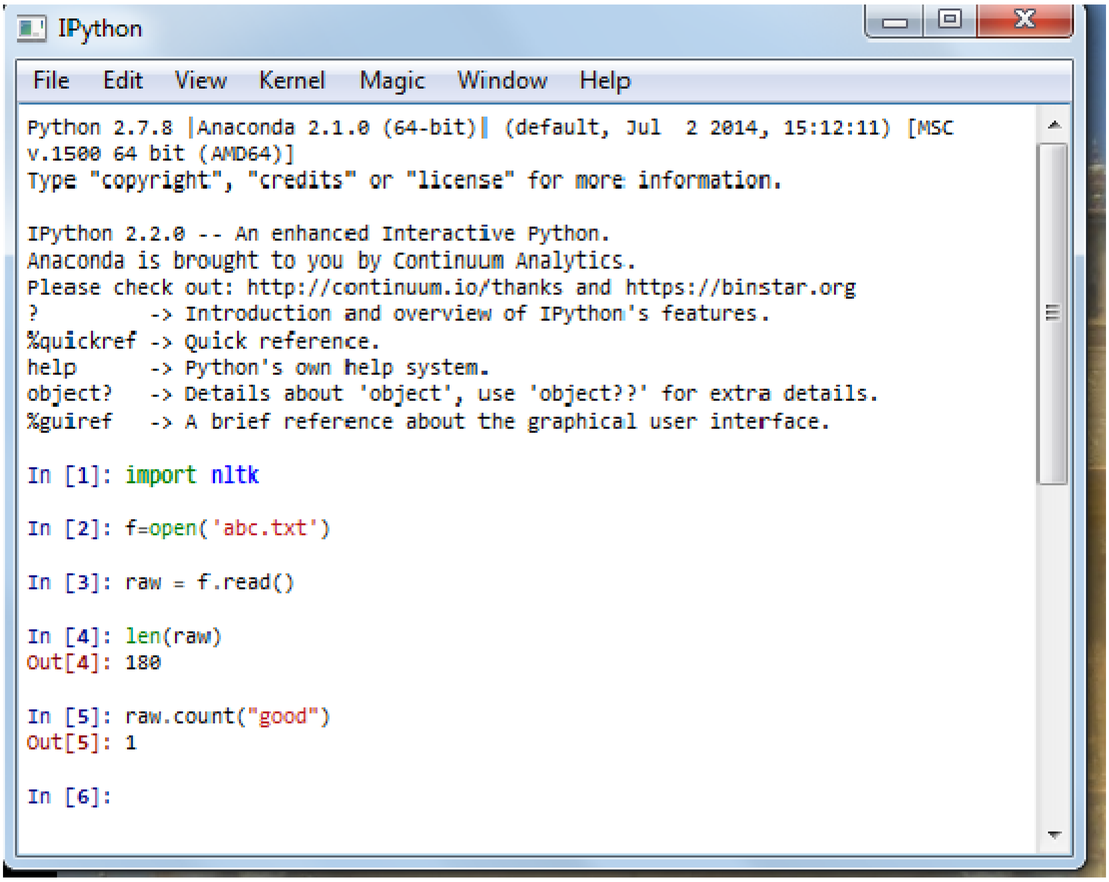
S.O = avg positive – avg negative



```
Python 2.7.8 |Anaconda 2.1.0 (64-bit)| (default, Jul  2 2014,
15:12:11) [MSC v.1500 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 2.2.0 -- An enhanced Interactive Python.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://binstar.org
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.
%guiref   -> A brief reference about the graphical user interface.

In [1]: import nltk

In [2]: from nltk.corpus import PlaintextCorpusReader

In [3]: corpus_root='./'

In [4]: newcorpus = PlaintextCorpusReader(corpus_root,'.*')

In [5]: newcorpus.words('my_corpus.txt')
```

Figure 8 Screenshot of add file in corpus

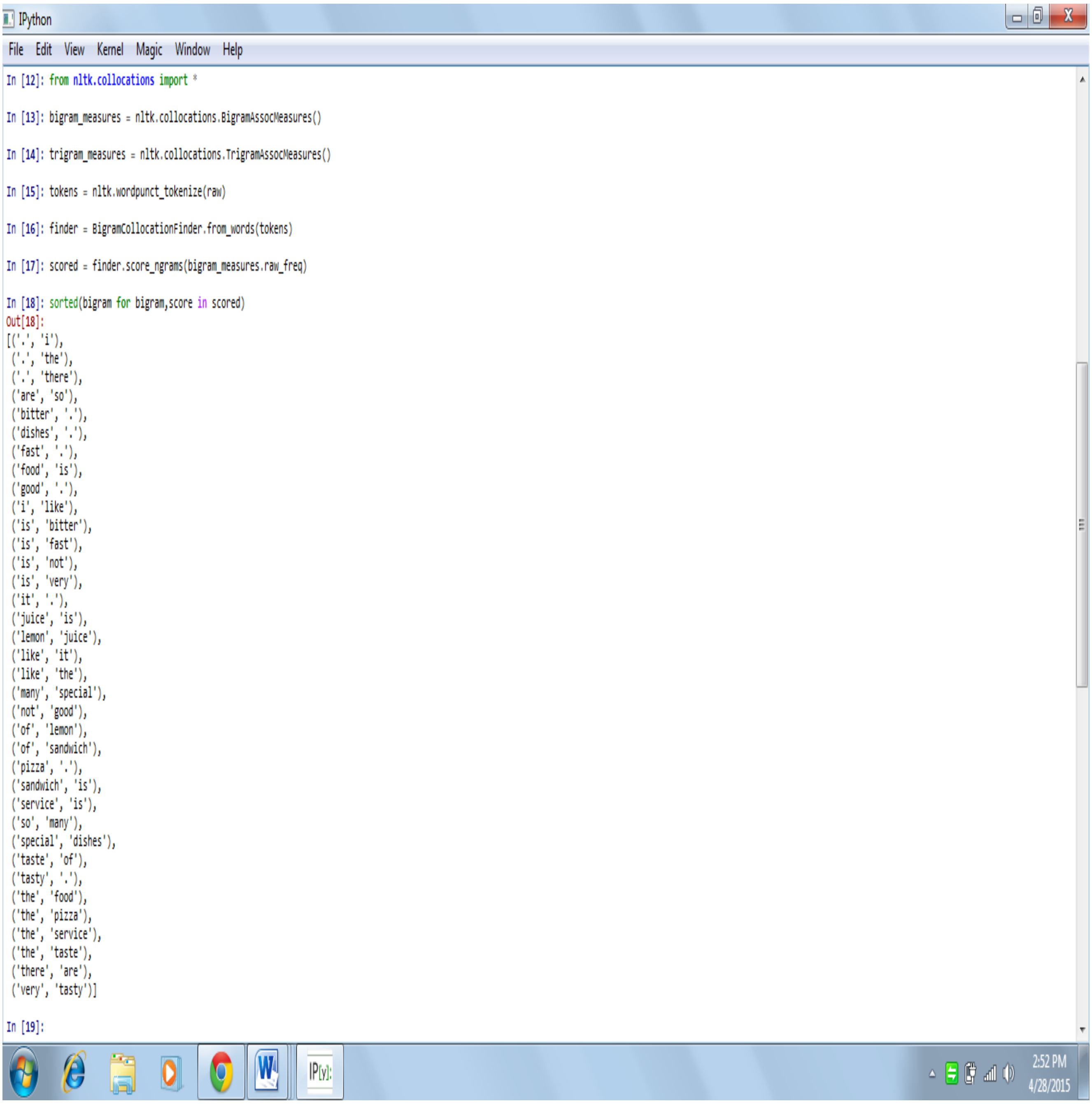
To add file in corpus first we add file in python scripts and after that we include this file in corpus by writing some line of commands in the above screenshot.



```
Python 2.7.8 |Anaconda 2.1.0 (64-bit)| (default, Jul  2 2014, 15:12:11) [MSC
v.1500 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 2.2.0 -- An enhanced Interactive Python.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://binstar.org
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.
%guiref   -> A brief reference about the graphical user interface.

In [1]: import nltk

In [2]: f=open('abc.txt')

In [3]: raw = f.read()

In [4]: len(raw)
Out[4]: 180

In [5]: raw.count("good")
Out[5]: 1

In [6]:
```

Figure 9 Screenshot counting words from text file

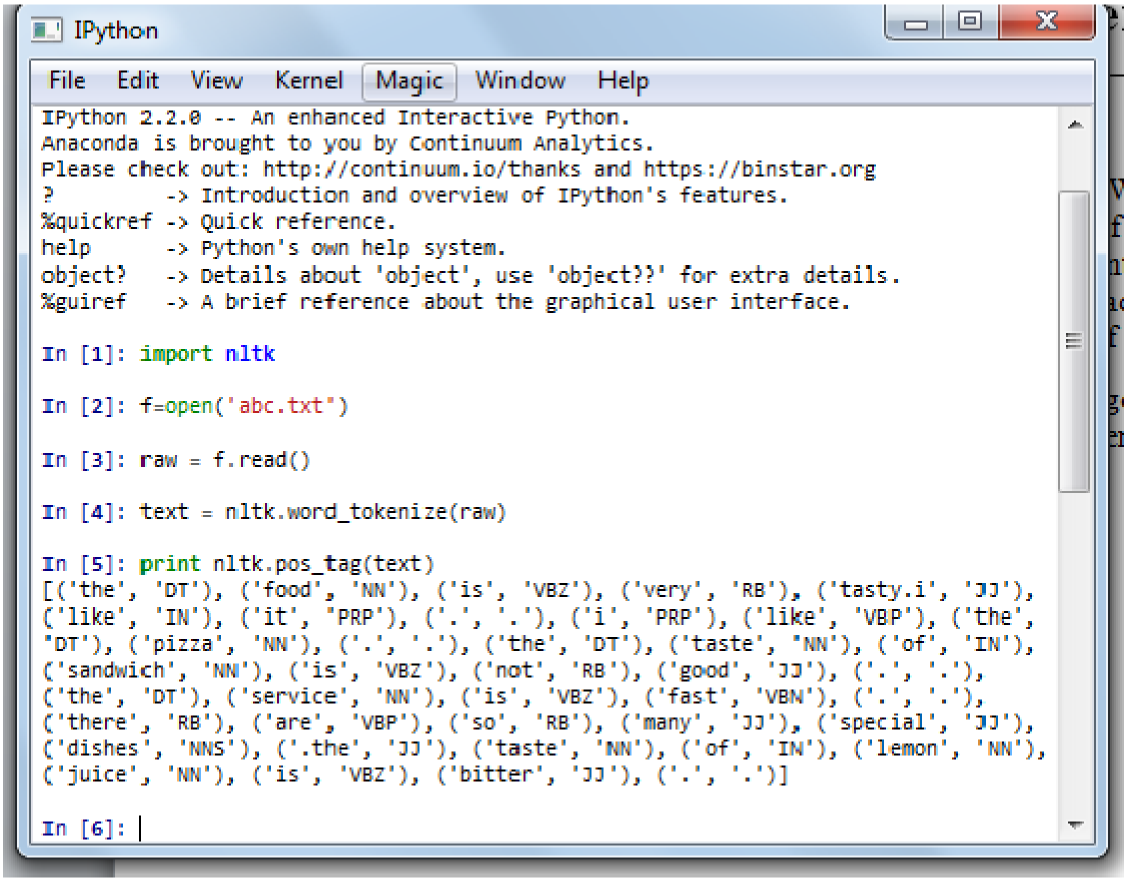
We can count the number words from that file which we are included in the corpus by writing the above screenshot commands. These commands help in finding the number of words occur in the text.



```
In [12]: from nltk.collocations import *

In [13]: bigram_measures = nltk.collocations.BigramAssocMeasures()

In [14]: trigram_measures = nltk.collocations.TrigramAssocMeasures()

In [15]: tokens = nltk.wordpunct_tokenize(raw)

In [16]: finder = BigramCollocationFinder.from_words(tokens)

In [17]: scored = finder.score_ngrams(bigram_measures.raw_freq)

In [18]: sorted(bigram for bigram,score in scored)
Out[18]:
[('.', 'i'),
 ('.', 'the'),
 ('.', 'there'),
 ('are', 'so'),
 ('bitter', '.'),
 ('dishes', '.'),
 ('fast', '.'),
 ('food', 'is'),
 ('good', '.'),
 ('i', 'like'),
 ('is', 'bitter'),
 ('is', 'fast'),
 ('is', 'not'),
 ('is', 'very'),
 ('it', '.'),
 ('juice', 'is'),
 ('lemon', 'juice'),
 ('like', 'it'),
 ('like', 'the'),
 ('many', 'special'),
 ('not', 'good'),
 ('of', 'lemon'),
 ('of', 'sandwich'),
 ('pizza', '.'),
 ('sandwich', 'is'),
 ('service', 'is'),
 ('so', 'many'),
 ('special', 'dishes'),
 ('taste', 'of'),
 ('tasty', '.'),
 ('the', 'food'),
 ('the', 'pizza'),
 ('the', 'service'),
 ('the', 'taste'),
 ('there', 'are'),
 ('very', 'tasty')]

In [19]:
```

Figure 10 Collocation of words

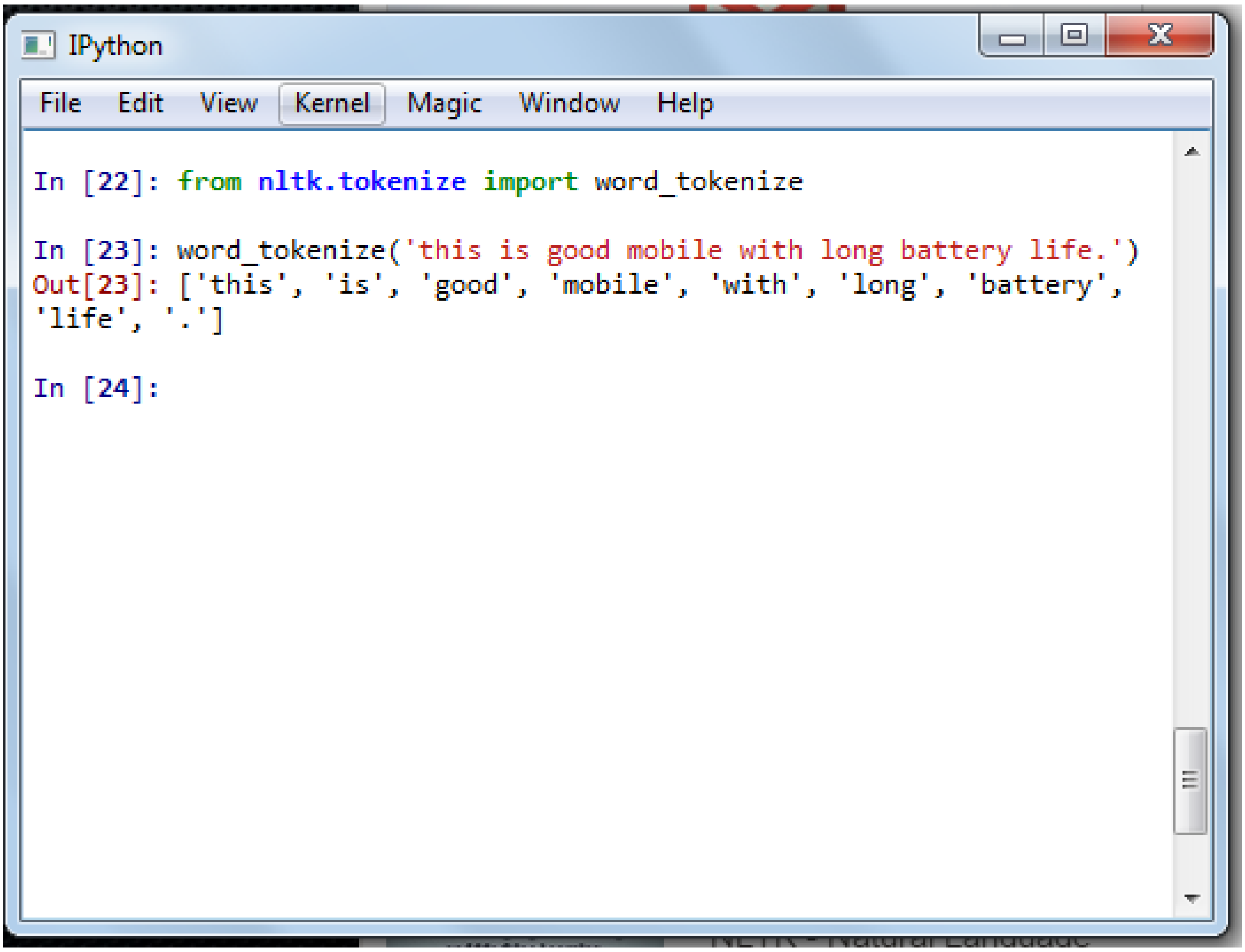
In the above screenshot we use collocation function. This function will provide the words which are occur together. For example “not good” word is shown in the above screenshot. The collocation helps us in understanding the sentiment of person by giving bigrams. As in the above screen shot the good is a positive word but when it is with not the good converted into negative word.



```
IPython 2.2.0 -- An enhanced Interactive Python.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://binstar.org
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.
%guiref   -> A brief reference about the graphical user interface.

In [1]: import nltk

In [2]: f=open('abc.txt")

In [3]: raw = f.read()

In [4]: text = nltk.word_tokenize(raw)

In [5]: print nltk.pos_tag(text)
[('the', 'DT'), ('food', 'NN'), ('is', 'VBZ'), ('very', 'RB'), ('tasty.i', 'JJ'),
('like', 'IN'), ('it', 'PRP'), ('.', '.'), ('i', 'PRP'), ('like', 'VBP'), ('the',
'DT'), ('pizza', 'NN'), ('.', '.'), ('the', 'DT'), ('taste', 'NN'), ('of', 'IN'),
('sandwich', 'NN'), ('is', 'VBZ'), ('not', 'RB'), ('good', 'JJ'), ('.', '.'),
('the', 'DT'), ('service', 'NN'), ('is', 'VBZ'), ('fast', 'VBN'), ('.', '.'),
('there', 'RB'), ('are', 'VBP'), ('so', 'RB'), ('many', 'JJ'), ('special', 'JJ'),
('dishes', 'NNS'), ('.the', 'JJ'), ('taste', 'NN'), ('of', 'IN'), ('lemon', 'NN'),
('juice', 'NN'), ('is', 'VBZ'), ('bitter', 'JJ'), ('.', '.')]

In [6]: |
```

Figure 11 Screenshot of part of speech

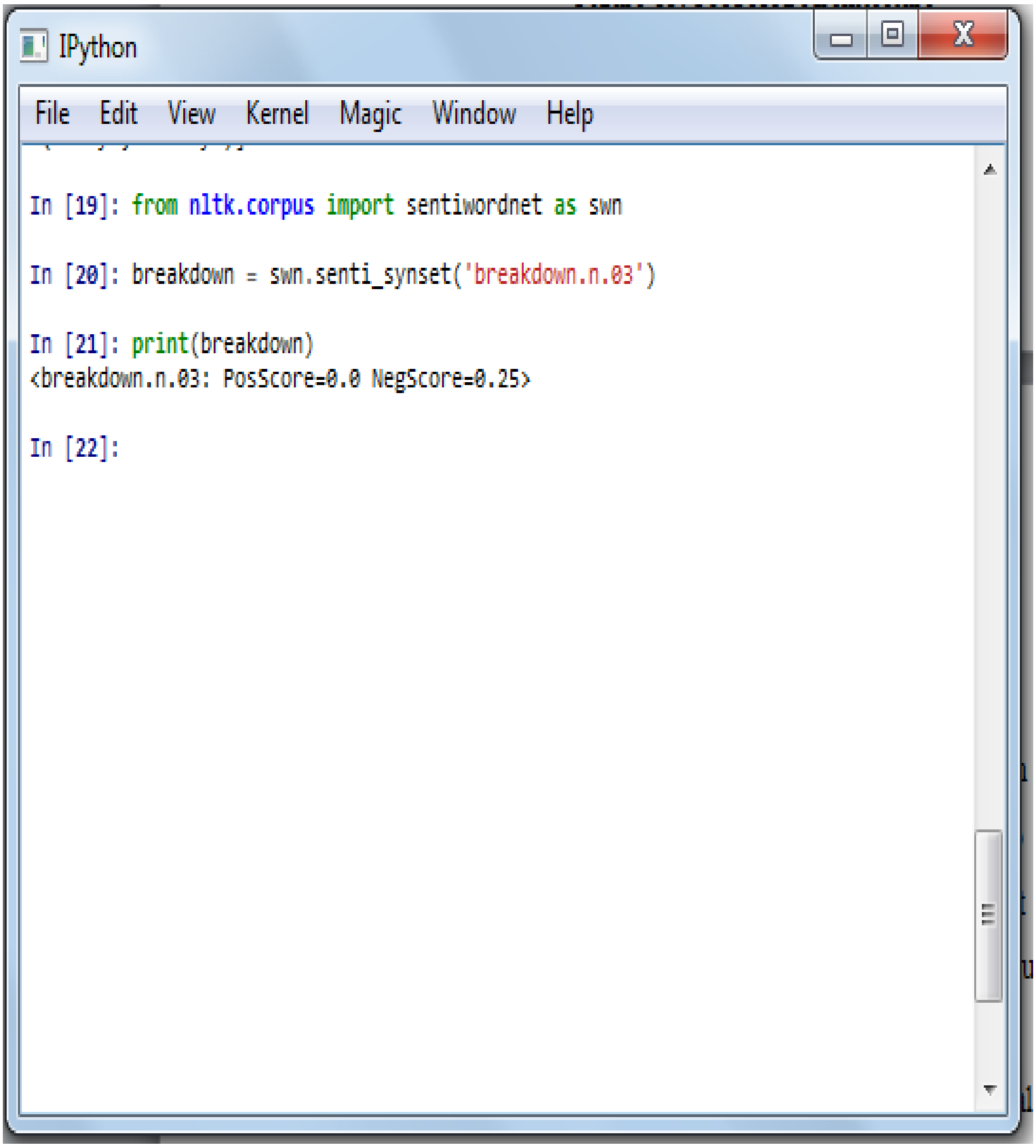
We use word tokenize for making tokens of word with their grammar. We can find adjectives, adverb, noun and pronoun. We can easily find out the adjectives and adverb to get their numerical value from Sentiwordnet. The adjectives and adverb are the most important part of sentence it helps in finding the sentiment of a person about particular thing.



```
In [22]: from nltk.tokenize import word_tokenize

In [23]: word_tokenize('this is good mobile with long battery life.')
Out[23]: ['this', 'is', 'good', 'mobile', 'with', 'long', 'battery',
'life', '.']

In [24]:
```

Figure 12 Screenshot of word tokenization

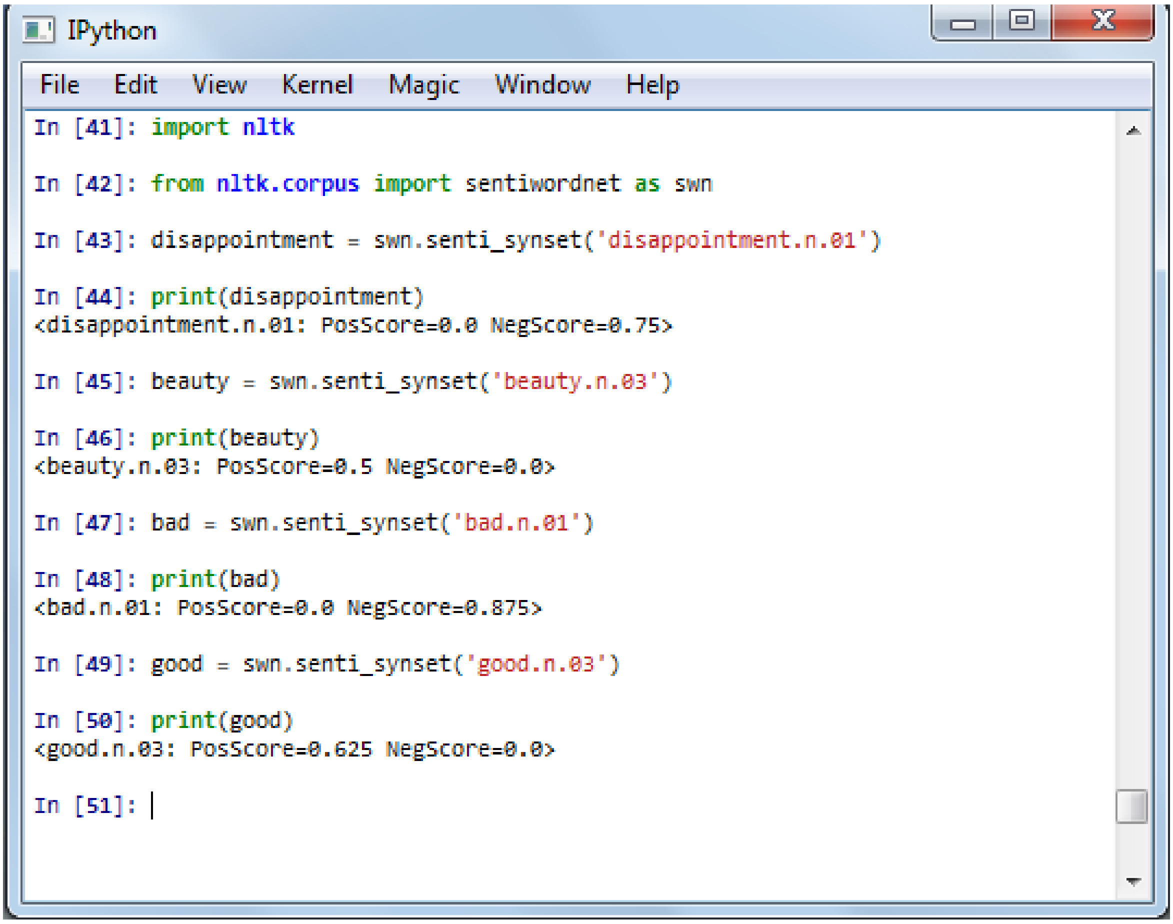
We can also tokenize sentence into word by using the above commands shown in the screenshots. This word tokenization helps in finding sentiments on short text. We can easily find out the words which tell about the polarity of data



```
In [19]: from nltk.corpus import sentiwordnet as swn

In [20]: breakdown = swn.senti_synset('breakdown.n.03')

In [21]: print(breakdown)
<breakdown.n.03: PosScore=0.0 NegScore=0.25>

In [22]:
```

Figure 13 Screenshot of Sentiwordnet value

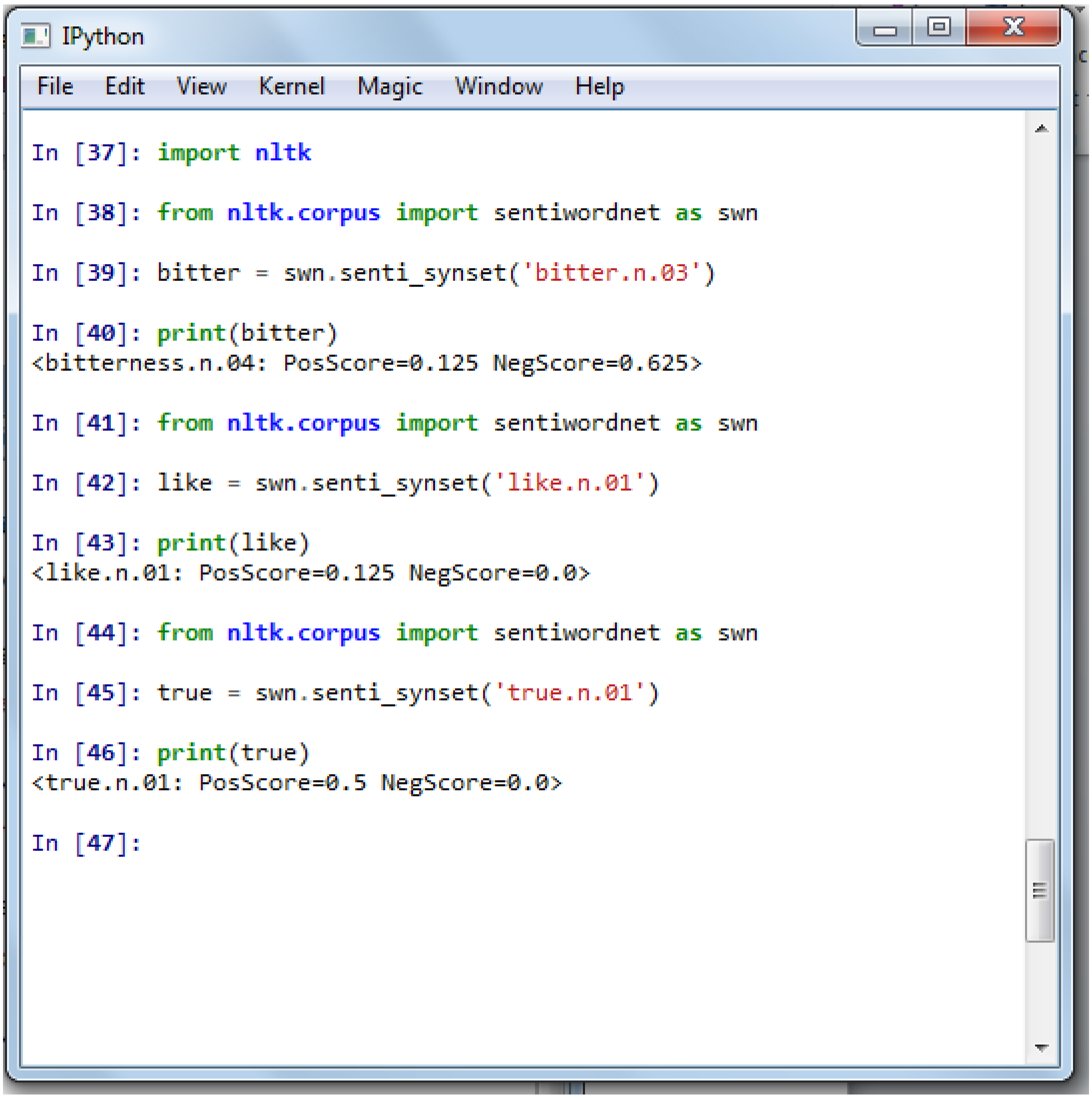
The above screenshot tell about how we can get the value of word from Sentiwordnet. We can write any word in place of breakdown to get their value from Sentiwordnet. After writing these codes we can get positive as well as negative value of the word.



```
In [41]: import nltk

In [42]: from nltk.corpus import sentiwordnet as swn

In [43]: disappointment = swn.senti_synset('disappointment.n.01')

In [44]: print(disappointment)
<disappointment.n.01: PosScore=0.0 NegScore=0.75>

In [45]: beauty = swn.senti_synset('beauty.n.03')

In [46]: print(beauty)
<beauty.n.03: PosScore=0.5 NegScore=0.0>

In [47]: bad = swn.senti_synset('bad.n.01')

In [48]: print(bad)
<bad.n.01: PosScore=0.0 NegScore=0.875>

In [49]: good = swn.senti_synset('good.n.03')

In [50]: print(good)
<good.n.03: PosScore=0.625 NegScore=0.0>

In [51]: 
```

Figure 14 Screenshot of positive and negative score

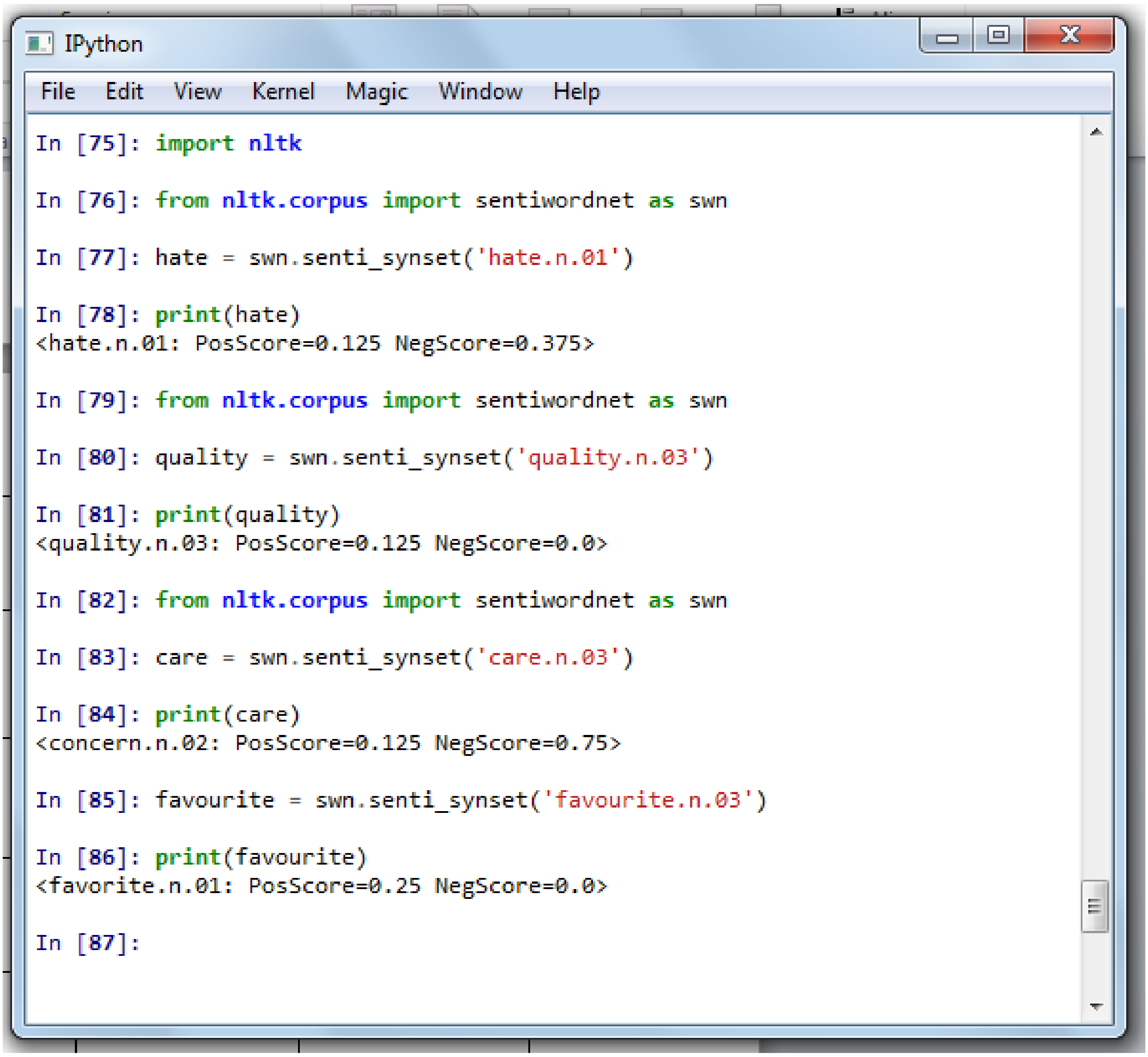
The above screenshot shows the positive and negative value of many words such as bad, good, beauty from these values we can easily calculate the semantic orientation.



```
In [37]: import nltk

In [38]: from nltk.corpus import sentiwordnet as swn

In [39]: bitter = swn.senti_synset('bitter.n.03')

In [40]: print(bitter)
<bitterness.n.04: PosScore=0.125 NegScore=0.625>

In [41]: from nltk.corpus import sentiwordnet as swn

In [42]: like = swn.senti_synset('like.n.01')

In [43]: print(like)
<like.n.01: PosScore=0.125 NegScore=0.0>

In [44]: from nltk.corpus import sentiwordnet as swn

In [45]: true = swn.senti_synset('true.n.01')

In [46]: print(true)
<true.n.01: PosScore=0.5 NegScore=0.0>

In [47]:
```

Figure 15 Screenshot of word value

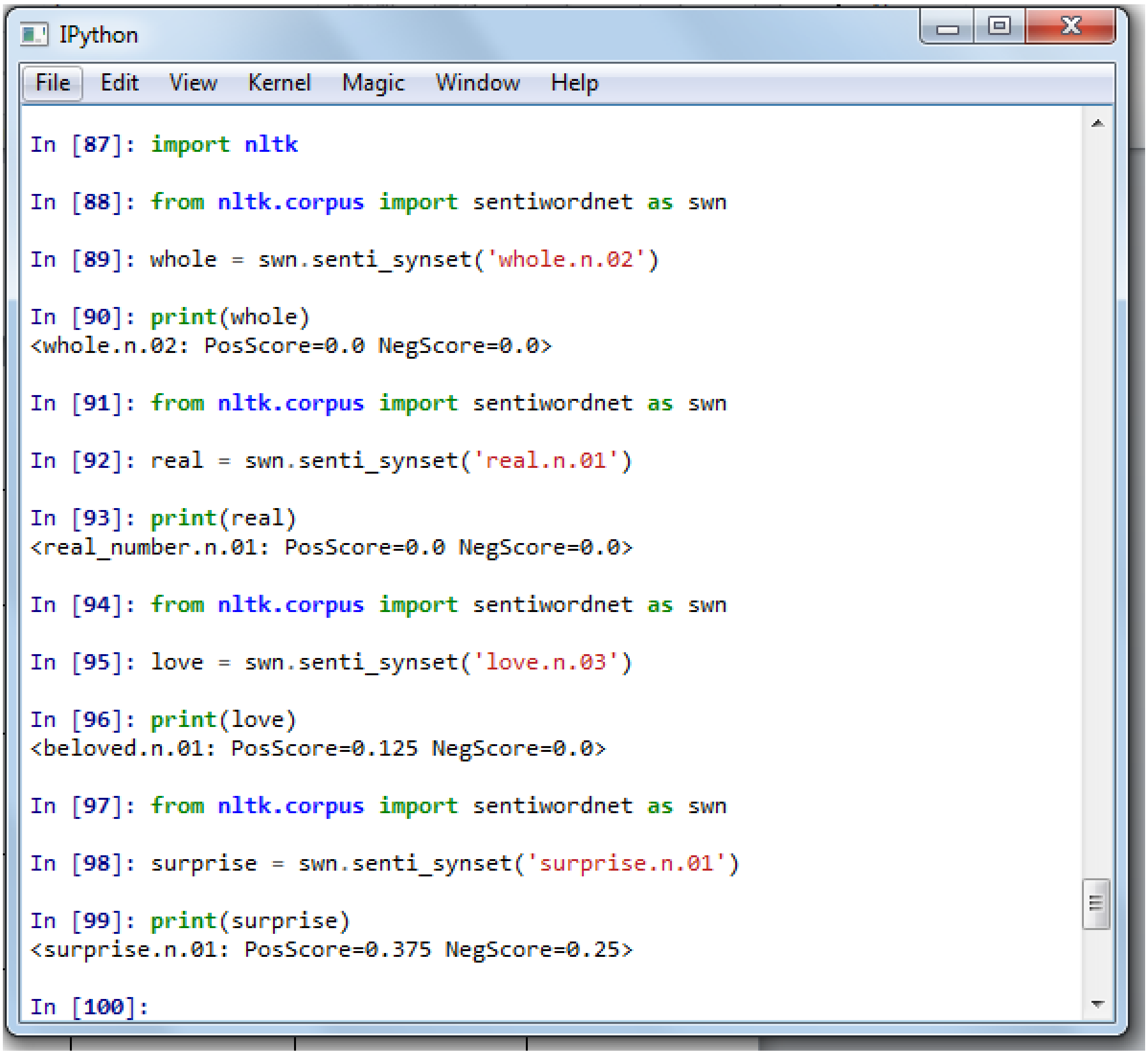
The above screenshot shows the positive and negative value of many words such as bitter, true, like from these values we can easily calculate the semantic orientation.



```
In [75]: import nltk

In [76]: from nltk.corpus import sentiwordnet as swn

In [77]: hate = swn.senti_synset('hate.n.01')

In [78]: print(hate)
<hate.n.01: PosScore=0.125 NegScore=0.375>

In [79]: from nltk.corpus import sentiwordnet as swn

In [80]: quality = swn.senti_synset('quality.n.03')

In [81]: print(quality)
<quality.n.03: PosScore=0.125 NegScore=0.0>

In [82]: from nltk.corpus import sentiwordnet as swn

In [83]: care = swn.senti_synset('care.n.03')

In [84]: print(care)
<concern.n.02: PosScore=0.125 NegScore=0.75>

In [85]: favourite = swn.senti_synset('favourite.n.03')

In [86]: print(favourite)
<favorite.n.01: PosScore=0.25 NegScore=0.0>

In [87]:
```

Figure 16 Screenshot of Sentiwordnet values

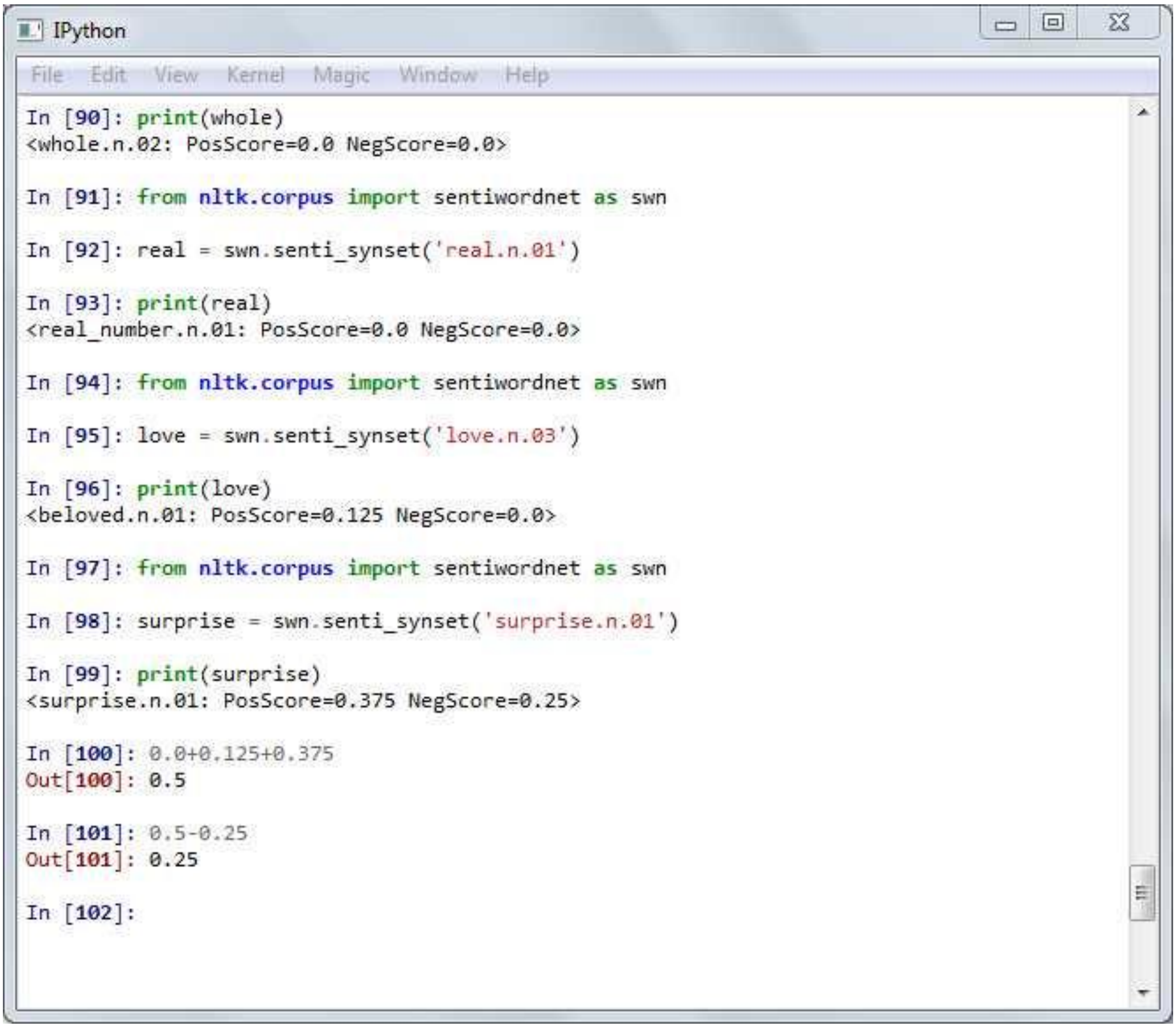
The above screenshot shows the positive and negative value of many words such as favourite, care, quality from these values we can easily calculate the semantic orientation.



```
In [87]: import nltk

In [88]: from nltk.corpus import sentiwordnet as swn

In [89]: whole = swn.senti_synset('whole.n.02')

In [90]: print(whole)
<whole.n.02: PosScore=0.0 NegScore=0.0>

In [91]: from nltk.corpus import sentiwordnet as swn

In [92]: real = swn.senti_synset('real.n.01')

In [93]: print(real)
<real_number.n.01: PosScore=0.0 NegScore=0.0>

In [94]: from nltk.corpus import sentiwordnet as swn

In [95]: love = swn.senti_synset('love.n.03')

In [96]: print(love)
<beloved.n.01: PosScore=0.125 NegScore=0.0>

In [97]: from nltk.corpus import sentiwordnet as swn

In [98]: surprise = swn.senti_synset('surprise.n.01')

In [99]: print(surprise)
<surprise.n.01: PosScore=0.375 NegScore=0.25>

In [100]:
```

Figure 17 Screenshot of Sentiwordnet values

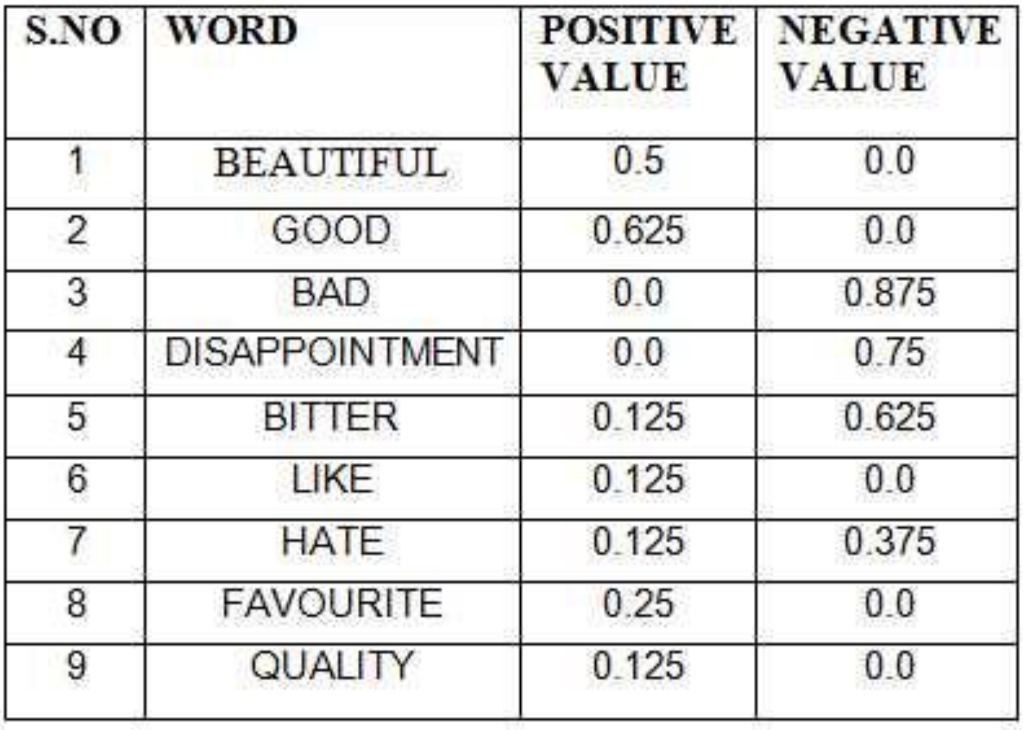
The above screenshot shows the positive and negative value of many words such as whole, real, beauty from these values we can easily calculate the semantic orientation.



```
In [90]: print(whole)
<whole.n.02: PosScore=0.0 NegScore=0.0>

In [91]: from nltk.corpus import sentiwordnet as swn

In [92]: real = swn.senti_synset('real.n.01')

In [93]: print(real)
<real_number.n.01: PosScore=0.0 NegScore=0.0>

In [94]: from nltk.corpus import sentiwordnet as swn

In [95]: love = swn.senti_synset('love.n.03')

In [96]: print(love)
<beloved.n.01: PosScore=0.125 NegScore=0.0>

In [97]: from nltk.corpus import sentiwordnet as swn

In [98]: surprise = swn.senti_synset('surprise.n.01')

In [99]: print(surprise)
<surprise.n.01: PosScore=0.375 NegScore=0.25>

In [100]: 0.0+0.125+0.375
Out[100]: 0.5

In [101]: 0.5-0.25
Out[101]: 0.25

In [102]:
```

Figure 18 Screenshot of semantic orientation

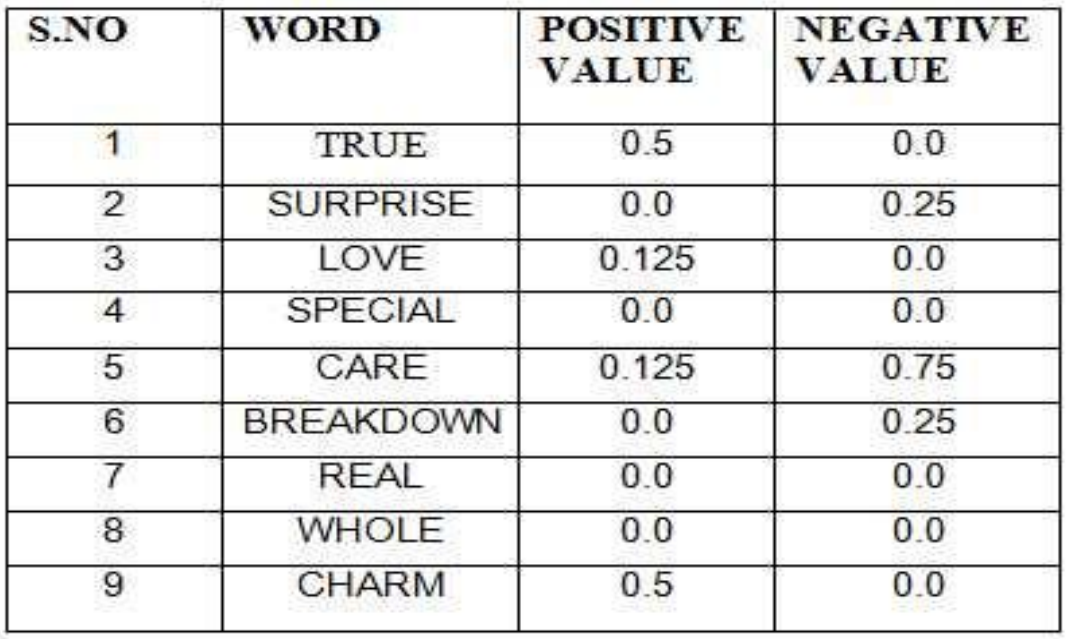
The table of words value is shown as positive and negative score. The 0.5 is the positive score and 0.0 is negative



S.NO	WORD	POSITIVE VALUE	NEGATIVE VALUE
1	BEAUTIFUL	0.5	0.0
2	GOOD	0.625	0.0
3	BAD	0.0	0.875
4	DISAPPOINTMENT	0.0	0.75
5	BITTER	0.125	0.625
6	LIKE	0.125	0.0
7	HATE	0.125	0.375
8	FAVOURITE	0.25	0.0
9	QUALITY	0.125	0.0

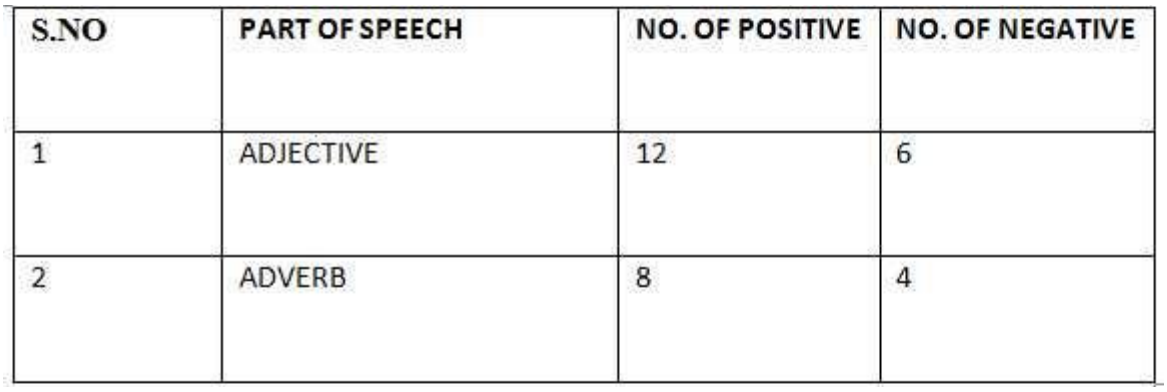
Table 6 Words value

The value in the table shows the positive and negative value of words in the Sentiwordnet. These values help in calculating the semantic orientation of opinion of people. We take average of the positive and negative value to calculate the result.



S.NO	WORD	POSITIVE VALUE	NEGATIVE VALUE
1	TRUE	0.5	0.0
2	SURPRISE	0.0	0.25
3	LOVE	0.125	0.0
4	SPECIAL	0.0	0.0
5	CARE	0.125	0.75
6	BREAKDOWN	0.0	0.25
7	REAL	0.0	0.0
8	WHOLE	0.0	0.0
9	CHARM	0.5	0.0

Table 7 Value of words

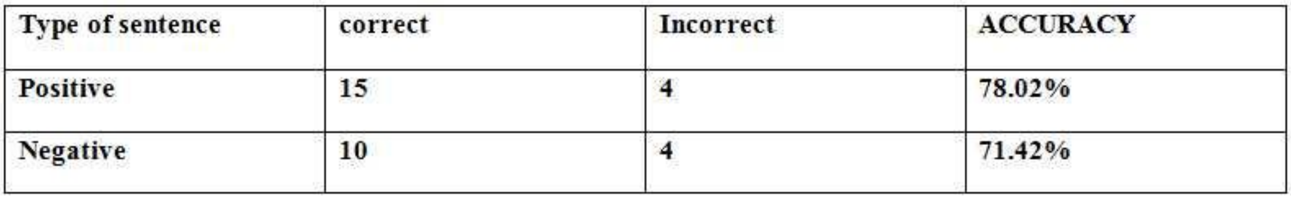


S.NO	PART OF SPEECH	NO. OF POSITIVE	NO. OF NEGATIVE
1	ADJECTIVE	12	6
2	ADVERB	8	4

Table 8 Table of number of POS

The above table shows the number of adjective and adverb with their positive and negative. We get these values from 50 comments. We get the 12 positive adjective and 6 negative adjective

Table of result



Type of sentence	correct	Incorrect	ACCURACY
Positive	15	4	78.02%
Negative	10	4	71.42%

Table 9 Table of result

The above table shows the correct and incorrect value of sentences that we analyze during implementation. We get 78.02% accuracy in positive sentences and 71.42% in negative sentences. The percentage of negative word is less because it is difficult to judge the negative sentence which contains both negative and positive word.

CHAPTER 7
CONCLUSION AND FUTURE SCOPE

Sentiment analysis provides us the feelings of people about any product or subject. Companies needed sentiment analysis for decision making and updation in product. Sentiment analysis provides favourable and unfavourable opinions about any product or any particular subject.

The techniques such as NLP, PMI and sentiful analysis provide us useful results. These techniques give us result in the numerical form or values. The result by these values helps us to analyse result and compare it with other techniques. The analysed result from sentiment analysis provides various advantages in decision making in business.

We will use tool for sentence level sentiment analysis so that we can remove those sentences which don’t contain any meaning. We can further incorporate these techniques for getting user views for new product design and analysis.

CHAPTER 8
REFERENCES

I. RESEARCH PAPERS

1. Alena Neviarouskaya, Helmut Prendinger, Mitsuru Ishizuka in" (2011) Sentiful: A lexicon for sentiment analysis" IEEE.

2. Bin Wen, Wenhua Dai,Junzhe Zhao (2012) "Sentence Sentimental Classification Based On Semantic Comprehension" IEEE.

3. Bo Pang, Lillian Lee and Shivakumar Vaithyanathan (2002)"Thumbs Up? Sentiment Classification Using Machine Learning Techniques" Proceeding EMNLP '02 Proceedings Of The ACL-02 Conference On Empirical Methods In Natural Language Processing - Volume 10 Pages 79-86.

4. EricLin,Shiaofen Fang,Jie Wang(2013) "Mining Online Book Reviews For Sentimental Clustering"IEEE.

5. Julia Kreutzer and Neele witte (2013)"Opinion Mining using Sentiwordnet"

6. Keisuke Mizumoto, Hidekazu Yanagimoto and Michifumi Yoshioka (2012) "Sentiment Analysis of Stock Market News with Semi-supervised Learning", IEEE Computer Society,IEEE/ACIS 11th International Conference on Computer and Information Science, p.325-328.

7. Kushal Dave,Steve Lawrencce,Davidm.Pennock (2003) "Mining The Peanut Gallery:Opinion Extraction And Semantic Classification Of Product Reviews,May.

8. Marc Cheong, Vincent Lee in (2010) "A Study on detecting patterns in twitter Intra-topic User and message Clustering" IEEE.

9. Neethu Mohandas,Janardhanan PS Nair, Govindaru V in(2012) "Domain Specific Sentence Level Mood Extraction from Malayalam Text" IEEE.

10.OM P.Damani (2013)”Improving pointwise mutual information (PMI) by incorporating significant co-occurrence” October.

11.Peter D.Turney, Michael L.Littman, ”Measuring praise and criticism: Inference of Semantic Orientation from association”.

12.Samanch Moghaddam,Martin Ester(2011)”AQA:Aspect Based Opinion Answering” IEEE.

13.Tetsuya Nasukawa, Jeonghee Yi in (2003) “Sentiment Analysis: Capturing favorability using natural language processing” October.

II. WEBSITES

1. www.edureka.co/blog/types-of-sentiment-analysis

APPENDIX

SOM- Self Organizing Maps

PMI -Pointwise Mutual Information

NLP- Natural Language Processing