# Chapter 1

# Introduction

## 1.1 Introduction

Sentiment analysis, commonly called as opinion mining, is a Natural Language Processing and Information Extraction task that is used to analyse opinion of people, their sentiments, and appraisals, attitude, perspective and emotions towards an entity (For example. Services, products, movies, organizations, events, topics or trending issues etc.). There are several different names given to tasks related to it e.g., sentiment analysis, opinion mining, review analysis, subjectivity analysis, emotion analysis etc. Sentiment Analysis is a large research area that aims to obtain a person's feelings, whether they are positive or negative, expressed in reviews, comments, questions, blogs, by analysing a large numbers of documents. It is concerned with analysis of text containing opinions, sentiments and emotions. The main focus is on sentiment classification study which attempts to determine whether a subjective text contains positive or negative sentiments (i.e. classify it into positive and negative). Although, natural language processing (NLP) research has a long history, little research had been done about people's opinions and sentiments before the year 2000. Since then it has become an active research area. Research in area of sentiment analysis not only has a significant impact on NLP, but may also has a huge impact on economics, management sciences, social sciences and political science because they are all affected by people's views and opinions. Sentiment analysis concerns with the sentiments expressed in the review documents and do not care about the factual content in it. Sentiment classification is usually framed as a two way classification of positive or negative sentiment and has been applied to different levels: phrases level, sentence level, paragraph level, documents level and collection of documents. With the proliferation of Web applications, users are expressing their opinion and experiences on blogs, discussion boards, reviews, social networking websites etc. this trend has increased the demand of analysing this online content resulting into increase in the sentiment analysis research (Liu, 2012). Sentiment analysis is important for companies and users to know what people think about specific topic. Companies can improve their products, based on users' opinion. Users can take purchasing decision based on reviews about the product (Liu, 2012). For example, whenever user wants to purchase mobile phone, he refers the online reviews related to that mobile or reviews related to other companies mobiles phones and compares the phones on the basis of experiences of different users written in the reviews, however, he may online refer some of the reviews. Therefore it is required to have a system which can analysis thousands of reviews and can provide an aggregate opinion of users about any product like mobiles, automobile etc. In addition, sentiment analysis can be important in political or health fields as political parties may be interested in the opinion of common people about their policies or about their party, and in medical field it may be important to know that which doctor is specialised or which medicine has been proved successful on the basis of reviews written by various patients.

Sentiment analysis is to recognize whether a given natural language text express positive or negative polarity. Recognizing polarity in text requires polar words like "good", "bad", "excellent", "worst" etc. For example "*This is an excellent movie.*" shows positive polarity and "*This movie was worst watching.*" shows negative polarity. These polar words are key indicators for creating machine learning model for sentiment classification. There are various standard lexicon of positive and negative sentiments which are freely available on web, these lexicons are used by various researchers for sentiment analysis research, for example WordNet, SentiWordNet (SWN), opinion lexicon, general inquirer etc. which have positive and negative word that can be used to classify the text into positive and negative document. We can divide sentences into two classes or groups in context to subjectivity:

- Objective sentences that contain factual information and
- subjective sentences that contain explicit opinions, beliefs, and views about specific entities

For example, let suppose if we check reviews of some resort in Goa : then it may be like that spacious area, near beach (i.e. walking distance only) , services of waiters were good, have swimming pool, beds were of king size, spacious rooms having good room services. The only negative review would be that Wi-Fi is not working properly in rooms, it is only working in reception area.

So if we see over all review then it is more Positive.

Sentiment analysis systems must calculate a sentiment score for the whole review as well as analyze the sentiment of each individual aspect of the resort.

Sentiment analysis research can be broadly classified into two methods on the basis of the approach used for classification i.e. Machine learning based approaches and semantic orientation based approaches. Machine learning based methods works in four phases, (1) Feature extraction, (2) Feature representation and weighting scheme (3) Feature selection, and (4) Machine learning algorithms. And, semantic orientation based methods basically works in following phases, initially all the sentiment-rich words or phrases are extracted, then semantic orientation of these words are computed using some formula and finally semantic orientation of all the words in the document are aggregated for finding the polarity of the overall document.

Movie review polarity classification faces the various challenges like some real facts or sometimes people discuss about the movie or about the characteristics of the actors and actress in the movie review. These sentences or discussion “about the movie” is not important for movie review classification. These sentences are called objective sentences which do not contain any opinion or sentiment of the user.

Main objective of this thesis is twofold, first subjectivity detection is performed for each sentence of the review document, and further determining the polarity of the document. Two methods are explored for detection of subjectivity of a sentence to eliminate the objective sentences and further, two types of features are extracted i.e. Part-of-Speech (POS) based unigrams and two-word phrases that are important for sentiment classification. Since, sometimes, individual polar words are incapable of incorporating actual sentiment of the text. Individual words can have different polarity for different domains. For example, “unpredictable” word may have a negative polarity in automobile review, with phrase “unpredictable steering”, but it could have positive polarity for movie review with the phrase “unpredictable story” (Turney, 2002). Therefore, contextual information is important for sentiment analysis. Phrases are intuitively very effective in incorporating contextual and syntactic information. This thesis explores the methods for extracting phrases that are important for sentiment classification.

A vital part of the information era has been to find out the opinions of other people. In the pre-web era, it was customary for an individual to ask his or her friends and relatives for opinions before making a decision. Organizations conducted opinion polls, surveys to understand the sentiment and opinion of the general public towards its products or services. In the past few years, web documents are receiving great attention as a new medium that describes individual experiences and opinions. With proliferation of Web 2.0 [1] applications such as micro-blogging, forums and social networks came. Reviews, comments, recommendations, ratings, feedbacks were generated by users. Hence, with the advent of World Wide Web1 and specifically with the growth and popularity of Web 2.0 where focus shifted to user generated content, the way people express opinion or their view has changed dramatically. People can now make their opinion, views, sentiment known on their personal websites, blogs, social networking sites, forums and review sites. They are comfortable with going online to get advice. Organizations have evolved and now look at review sites to know how the public has received their product instead of conducting surveys. This information available on the Web is a valuable resource for marketing intelligence, social psychologists and others interested in extracting and mining views, moods and attitude [2].

There is a vast amount of information available on the Web which can assist individuals and organization in decision making processes but at the same time present many challenges as organizations and individuals attempt to analyze and comprehend the collective opinion of others. Unfortunately finding opinion sources, monitoring them and then analyzing them are herculean tasks. It is not possible to manually find opinion sources online, extract sentiments from them and then to express them in a standard format. Thus the need to automate this process arises and sentiment analysis [3] is the answer to this need.

Sentiment analysis or Opinion mining, as it is sometimes called, is one of many areas of computational studies that deal with opinion oriented natural language processing. Such opinion oriented studies include among others, genre distinctions, emotion and mood recognition, ranking, relevance computations, perspectives in text, text source identification and opinion oriented summarization [4]. Sentiment analysis has turned out as an exciting new trend in social media with a gamut of practical applications that range from applications in business(marketing intelligence; product and service bench marking and improvement), applications as sub component technology(recommender systems; summarization; question answering) to applications in politics. It has great potential to be used in business strategies and has helped organizations get a real-time feedback loop about their marketing strategy or advertisements from the reaction of the public through tweets, posts and blogs. For a new product launch it can give them instant feedback about the reception of the new product. It can gauge what their brand image is, whether they are liked or not. As the field of sentiment analysis is relatively new, the terminology used to describe this field of research is many. The terms opinion mining, subjectivity analysis, review mining and appraisal extraction are used interchangeably with sentiment analysis. Subjectivity analysis or subjectivity classification is focused on the task of whether the sentence or document is expressing opinions or sentiments of the author or just merely stating facts. Majority of the papers which use the phrase ―sentiment analysis‖ focus on the specific application of classifying reviews as to their polarity (either positive or negative) [4]. The term opinion mining was first noticed in a paper by Dave et al. [5]. The paper defined that an opinion mining tool would ―process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)‖. This definition has been broadened to include various other works in this area. The evolution of the phrase sentiment analysis is similar to that of Opinion Mining. We have used these terms interchangeably in this paper.

Recently a lot of interest has been generated in the field of sentiment analysis, with researchers recognizing the scientific trials and potential applications supported by the processing of subjective language. Some factors substantiated by research till date, that push the development of the research area, include, augmenting of machine learning methods in natural language processing and information retrieval, increase in World Wide Web to provide training datasets for machine learning algorithms and the realization of commercial and intelligent applications that the area provides. As an example of one of the latest applications of sentiment analysis, Twitter1, Inc. incorporated an advanced tweet-searching function based on sentiment direction, where users can search for positive or negative tweets on a particular topic.

This paper gives an overview of sentiment analysis, its basic terminology, tasks and levels and discusses practical and potential applications of sentiment analysis further expounding its significant open research directions. The paper is organized as follows: the first section introduces sentiment analysis and discusses its history. It is followed by a section which explains the basic terminology. Section 3 expounds how different Web 2.0 applications add dimensions to the sentiment analysis tasks, which are illustrated in section 4 followed by section 5 which explains the granularity at which these tasks can be performed. Section 6 explicates the current state- of- art and describes how machine learning has proved its worth as a technique used for solving the sentiment analysis tasks. Section 7 presents the various applications of sentiment analysis. Lastly, section 8 discusses the various issues that turn out as open problems to be addressed which urge researchers to make significant improvements to understand and work in the sentiment analysis domain.

## 1.2 Terminology of Sentiment Analysis

Formally stating Sentiment Analysis is the computational study of opinions, sentiments and emotions exprezssed in text. The goal of sentiment analysis is to detect subjective information contained in various sources and determine the mind-set of an author towards an issue or the overall disposition of a document.

Wiebe et al. [6] described subjectivity as the linguistic expression of somebody‘s opinions, sentiments, emotions, evaluations, beliefs and speculations. The words opinion, sentiment, view and belief are used interchangeablyq but there are subtle differences between them [4].

- Opinion: A conclusion thought out yet open to dispute (―each expert seemed to have a different opinion).
- View: subjective opinion (―very assertive in stating his views).
- Belief: deliberate acceptance and intellectual assent (―a firm belief in her party‘s platform).
- Sentiment: a settled opinion reflective of one‘s feelings (―her feminist sentiments are well-known).

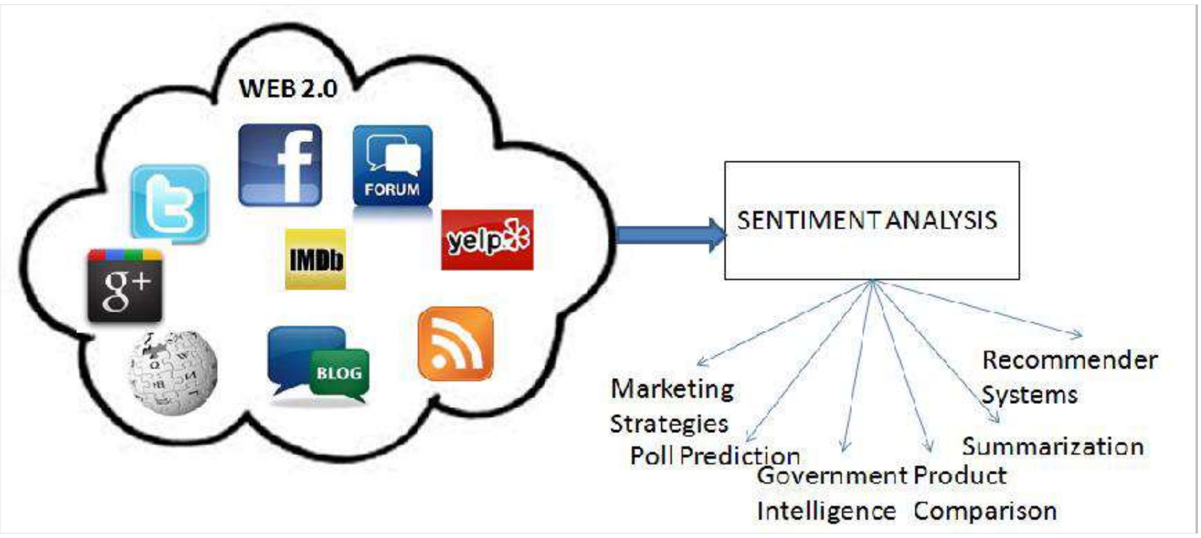
Sentiment analysis is done on user generated content on the Web which contains opinions, sentiments or views. An opinionated document can be a product review, a forum post, a blog or a tweet, that evaluates an object. The opinions indicated can be about anything or anybody, for e.g. products, issues, people, organizations or a service.

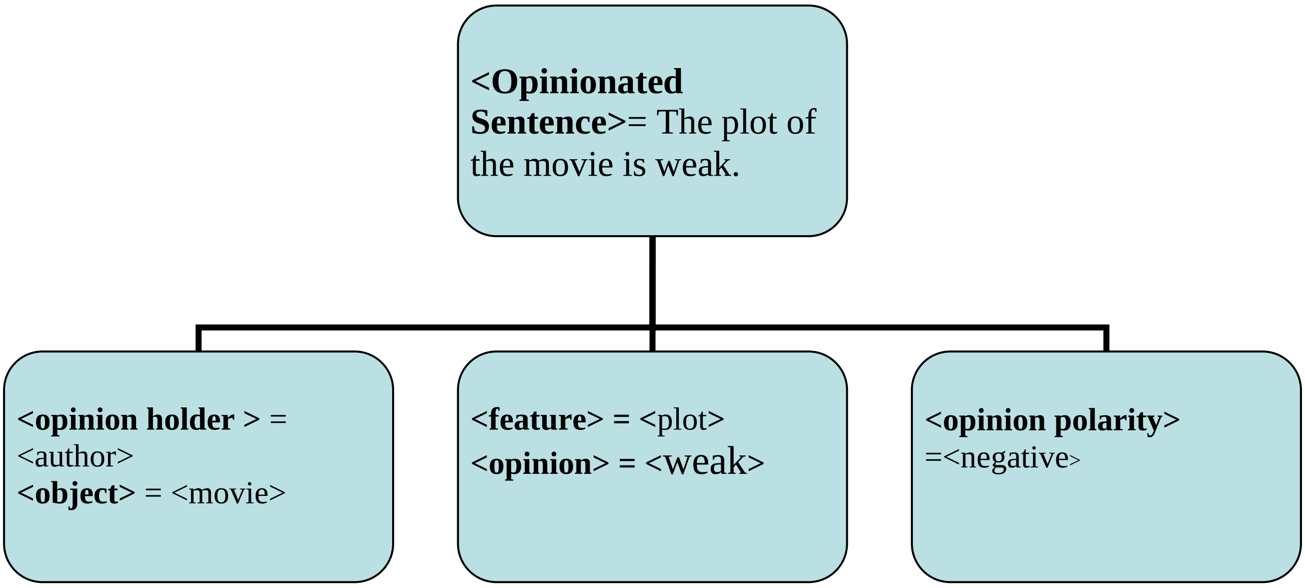
Lui [3] mathematically represented an opinion as a quintuple (o, f, so, h, t), where o is an object; f is a feature of the object o; so is the orientation or polarity of the opinion on feature f of object o; h is an opinion holder; t is the time when the opinion is expressed.

- Object: An entity which can be a product, person, event, organization, or topic. The object can have attributes, features or components associated with it. Further on the components can have subcomponents and attributes
- Feature: An attribute (or a part) of the object with respect to which evaluation is made.
- Opinion orientation or polarity: The orientation of an opinion on a feature f indicates whether the opinion is positive, negative or neutral. Most work has been done on binary classification i.e. into positive or negative. But opinions can vary in intensity from very strong to weak [7]. For example a positive sentiment can range from

content to happy to ecstatic. Thus, strength of opinion can be scaled and depending on the application the number of levels can be decided.

- Opinion holder: The holder of an opinion is the person or organization that expresses the opinion.





**<Opinionated Sentence>**= The plot of the movie is weak.

**<opinion holder >** = <author>
**<object>** = <movie>

**<feature>** = <plot>
**<opinion>** = <weak>

**<opinion polarity>** =<negative>

Fig1.1: Example corresponding to Terminology of Sentiment Analysis

## 1.3. Motivation

With the increasingly growth of web 2.0, sentiment analysis became a demanding and commercially supported research field, web 2.0 give its users to freedom to interact with each other in social media. So the social networking sites, blogs, wikis, video- sharing sites, hosted services, web applications have the large number of internet users and they are increasing exponentially year to year. Line chart of internet users from 2002 onwards as shown in Figure 1.1 (Source: http://www.internetworldstats.com/stats.htm). Estimated Internet users 2,405,518,376 on June,30 2012. Due to large number of people using blog and social networking sites, people are sharing their experiences about products, topics, etc. these online content can be very useful for people in taking purchasing decisions and for e-commerce companies in improving their products etc. One can be satisfied after viewing the reviews of product and can be able to judge the quality and usage of product in better way. Today e-commerce is in boom. Each website has so many products and has so many reviews about website and product. One tries for product only after watching the reviews of website i.e. the merchandise. Sentiment analysis helps in generating positive and negative reviews.

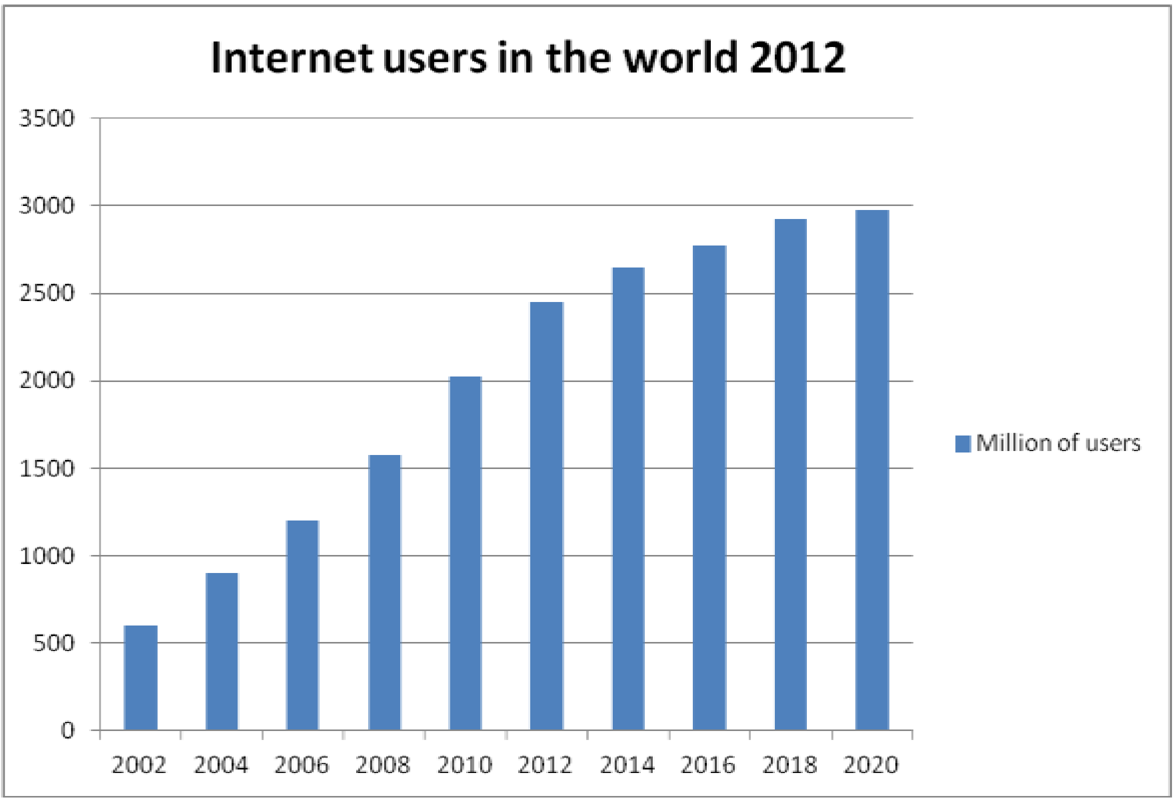


Figure 1.2: Internet users in the world

## 1.4. Research Goal

The main goal of this is to classify the given text into positive or negative polar document. This is done in broadly two phases, first is to identify the subjective portion of the text and eliminate the objective sentences from the review for sentiment classification. In second phase classy the remaining document into positive or negative class using semantic orientation based approach. Two methods are implemented for this purpose and compared there performances. In addition, two methods are implemented for identifying the polarity of a given document into positive or negative polar document.

## 1.5. Sentiment Analysis

The research field of sentiment analysis has been rapidly progressing because of the rich and diverse data provided by Web 2.0 applications. Blogs, review sites, forums, microblogging sites, wikis and social networks have all provided different dimensions to the data used for sentiment analysis.

**Sites Review.**

A review site is a website which allows users to post reviews which give a critical opinion about people, businesses, products, or services. Most sentiment analysis work has been done on movie and product review sites. The purpose of a review is to appraise a specific object, thus it is a single domain problem. Sentiment analysis on review sites is useful to both manufacturers and potential consumers of the product. The manufacturers can gauge the reception of a product based on the reviews. They can derive the features liked and disliked by the reviewer

**Blogs.**

The term web-log or blog refer to a simple webpage consisting of brief paragraphs of opinion, information personal diary entries, or links, called posts, arranged chronologically with the most recent first, in the style of an online journal. The bloggers post at hourly, daily or weekly basis which makes the interactions faster and more real-time. Different blogs have different styles of presentation, content material and writing techniques. Sentiment analysis on blogs has been used to predict movie sales, political mood and sales analysis.

### Forums

Forums or message boards allow its members to hold conversations by posting on the site. Forums are generally dedicated to a topic and thus using forums as a database allows us to do sentiment analysis in a single domain.

### Social Networks

Social networking is online services or sites which try to emulate social relationships amongst people who know each other or share a common interest. Social networking sites allow users to share ideas, activities, events, and interests within their individual networks. Social network posts can be about anything from the latest phone bought, movie watched, political issues or the individual‘s state of mind. Thus posts give us a richer and more varied resource of opinions and sentiments. Types of social network:

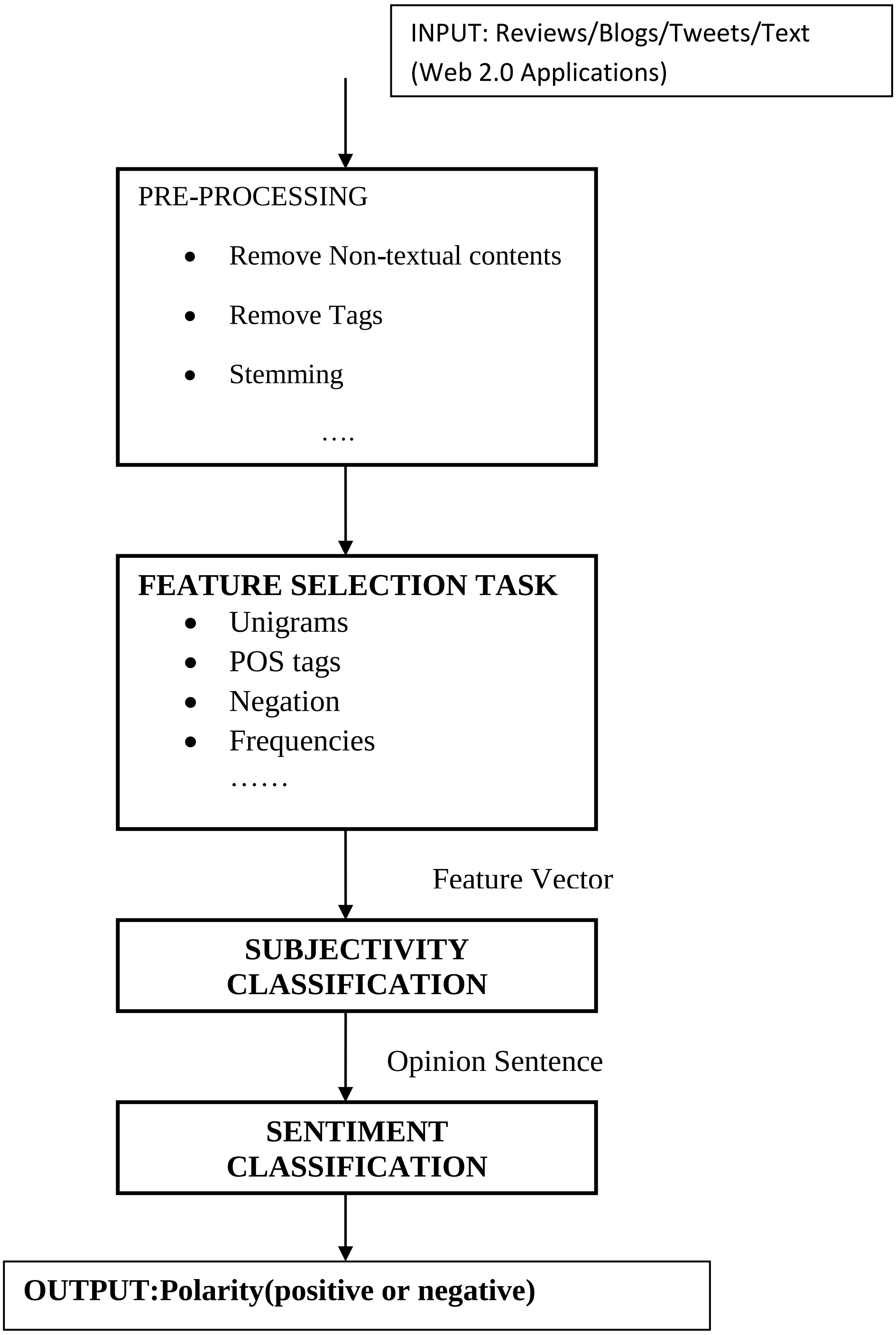
### 1) Twitter

Twitter is an online social networking and micro blogging service that enables its users to send and read text-based posts of up to 140 characters, known as "tweets. Sentiment analysis on twitter is an upcoming trend with it being used to predict poll results among various other applications.

### 2) Facebook

Face book is a social networking service and website launched in February 2004. The site allows users to create profiles for themselves, upload photographs and videos. Users can view the profiles of other users who are added as their friends and exchange text messages.

Social media is the new source of information on the Web. It connects the entire world and thus people can much more easily influence each other. The remarkable increase in the magnitude of information available calls for an automated approach to respond to shifts in sentiment and rising trends



## 1.6. Application of Sentiment Analysis

There are various applications of sentiment analysis which motivates research in sentiment analysis field. Some applications are given below.

### 1.6.1. Company Business Intelligence

Sentiment analysis provides companies a means to estimate their product by users review and they can improve their products based on user opinion. Users opinions are very much important to improve the quality. There were crores  of people who were using online system for shopping one cannot read each review. So sentiment analysis is the one which helps in finding the summarized reports.

### 1.6.2 Customer Decision

Today's people are very dependent on internet, they search required information on internet and take purchasing decision based on user reviews but there is a vast collection of text and user not have much time to read all the review, so sentiment analysis helpful to the customers to take decision.

### 1.6.3 Search Engines

Search engines also helpful for potential customers to make an informed selection of a product they want to buy. Such search engines include a sentiment classification subsystem that may not only present to a customer overall sentiment about a product, but also extract positive or negative reviews to illustrate advantages and shortcomings of a product.

### 1.6.4 Humanities and Media studies

Sentiment analysis also provides a range of possibilities for researchers in humanities whose studies involve analysis of large amount of human-generated data. For example, in media studies one might be interested to see if sentiments regarding the same events are shared in mainstream media and in social media.

### 1.6.5 Social Media and Political studies

Analysis of user-generated content may be very helpful in political studies. For example, monitoring of political debates in social media may help to estimates prospects of political candidates in elections or evaluate effectiveness of political campaigns. Many aspects of social studies may benefit from automatic analysis of sentiments expressed by people in ever-growing social networks.

## 1.7 Challenges of Sentiment Analysis

There are various challenges in sentiment analysis although various tools and techniques developed for sentiment classification however sentiment depends on writer's text. Here we discuss some issues with sentiment analysis.

### 1.7.1. Data Source

Generally dataset experimented for sentiment analysis is unstructured text from blog post, user reviews (about any product), chatting record, social networking sites, opinion poll. These text contain several noisy symbol (smiley, special symbol), casual languages and emotional symbols. For example, if we search "hungry" on twitter with arbitrary number of u's in the middle (e.g. "huuuungry", "huuuuuuuuuungry") then we get a non-empty result set. This is the major challenge for the sentiment analysis as could not judge the right result.

### 1.7.2 Negation Handling

Mostly users write their negative opinion by using "not" and "positive words" and negative opinion are classified as positive due to positive word. For example "*This is not a great movie.*" Contain positive word "great". There are some words other than not/never/n't which also shows negative polarity and difficult to handle them or some sentence contain implicit sentiment without presence of sentiment bearing word. For example "*No one likes these extra function*." and "*This news is too good to be true.*".
These example shows that these are the negative sentence but difficult to handle them. As no one can judge as they do not have words like "not" etc

### 1.7.3 Sarcasm detection and handling

A sentence may have an implicit sentiment even without the presence of any sentiment bearing words. For example:
(a) How can somebody not like this movie?

### 1.7.4 Subjectivity Detection

Writer’s opinion contains two parts in the movie reviews, first ‘opinion in the movie’ and second ‘about the movie’. First part contains user opinion so it is used to enhance the performance of Sentiment analysis, this part called subjective part of review other part is objective and it should be discarded. It is very difficult to identify subjective part of text. For example “*this is a love stories.*” present an objective sentence while “*I do not like this movie.*” depicts opinion about the particular movie.

### 1.7.5 Domain Dependency

Words polarity may be changed according to domain. For example “*The story was unpredictable.*” gives positive polarity in movie review domain and “*The steering of the car is unpredictable.*” gives negative polarity in automobile review documents. Therefore, polarity of same word can vary depending on the domain.

### 1.7.6 Entity Identification

A text may contain multiple entities, and text shows positive polarity for one entity then shows negative polarity for other entity. It is important to identify entity to which sentiment is directed. For example “*Samsung is better than nokia.*” return positive polarity for entity ‘Samsung’ and negative polarity for ‘Nokia’.

### 1.7.7 Dual sentiments

Sometimes review may contain both types of sentiment positive or negative, so identification of actual sentiment is very difficult.
For example “*I bought an iPhone a few days ago. It was such a good mobile. The touch-screen was awesome. The voice quality was also clear. Although the battery life was not long, that is ok for me. However, my mother was mad with because it was too expensive, and wanted me to return it to the shop*”.
Now these types of sentences having dual sentiments i.e. positive as well as negative could not explain clearly. It is very difficult to find positive or negative review out of it.

### 1.7.8. Thwarted Expectations

Sometimes the author deliberately sets up context only to refute it at the end. For example:

(a) This film should be nice.

(b) It sounds like a great plot,

(c) the actors were of first grade, and the supporting cast was also good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.

In each sentence author is praising the film but in last sentence (i.e. "it can't hold up") there is negativity of whole sentence.

### 1.7.9. Pragmatics

It is important to detect the pragmatics of user opinion which may change the sentiment thoroughly.

### 1.7.10 Multiple opinion in sentence

Single sentence can contain multiple opinions along with subjective and factual portions. It is helpful to isolate such clauses. It is also important to estimate the strength of opinions in these clauses so that we can find the overall sentiment in the sentence, e.g, ―*The picture quality of this camera is amazing and so is the battery life, but the viewfinder is too small for such a great camera*‖, expresses both positive and negative opinions

## 1.8 Current Trend and Techniques

Sentiment analysis can be broadly categorised among three types based on the size of text by which opinion is expressed.

### 1.8.1 Document level Sentiment analysis

In document level sentiment analysis, for a given document opinion is analysed, whether a given document express positive, negative or neutral polarity. There are three main approaches for identifying the polarity for a given review document, i.e. Machine learning based approaches, semantic orientation based approaches and lexicon based approaches.

### 1.8.2 Sentence level Sentiment Analysis

There are two tasks are analysed under sentence level, first is whether sentence belongs to Subjective category or Objective category and second is, if sentence is subjective then whether it belongs to positive polarity or negative polarity.

### 1.8.3 Entity level Sentiment Analysis

A sentence contain multiple entities, where entities return different polarities. For example "*Although the service is not that great, I still love this restaurant.*" return positive sentiment for restaurant but negative sentiment for service. Sentence level analysis is highly challenging, and Entity level is even more difficult.

## 1.9 Contribution of thesis

Contributions of this thesis to research in sentiment analysis are as follows.

1. Two methods are evaluated to identify the subjective portion of the review document for better classification results.
2. Two methods are evaluated to classify the review documents based on one-word and two-word features i.e. unigrams and POS based phrase patterns.
3. Proposed approach for sentiment analysis incorporates the two concepts i.e. subjectivity detection and polarity classification. In the literature, semantic orientation based polarity classification approaches does not use subjectivity detection, which is the one of the main objective of the proposed approach.

## 1.10 Roadmap of Thesis

This thesis is organised as follows.

Chapter 2 describes the related works reported in the literature with our proposed work in sentiment analysis.

Chapter 3 described about the Resources used in the proposed method.

Chapter 4 presents the Methodologies used in the proposed work.

Chapter 5 discusses about the Dataset, Experimental setup and results.

Chapter 6 presents the conclusions and future work.

# Chapter 2

# Related Work

Semantic orientation based approaches has been explored for Sentiment analysis in the literature. Generally, semantic orientation approaches which use phrases can be categorised into three phases. In the first phase, subjective sentences are extracted, in the second phase phrases are extracted using various methods, and in the third phase, semantic orientation is calculated using different methods. Further, by aggregating the semantic orientation values of all the phrases in the text, overall sentiment polarity is calculated.

Turney (2002) proposes an unsupervised method for identifying the polarity of a movie review document. Initially, they extracted two-word phrases using fixed POS based patterns, then semantic orientation of those phrases are computed using Point-wise Mutual Information (PMI) method. Finally, overall polarity of the document is recognized by aggregating the semantic orientation of all the phrases. Mukras et al. (2008) proposed the method for automatically identification of POS based pattern for extraction of polar phrases. They applied various feature selection methods namely Information Gain (IG), Chi-Squares (CHI), and Document Frequency (DF) for identification of important phrase patterns. Further, they applied PMI method for calculation of semantic orientation of phrases. Fei et al. (2004) constructed some phrase patterns with adjectives, adverbs, prepositions, conjunctions, noun, and verbs. Further, semantic orientation of these phrases is calculated using unsupervised method. Zhang et al. (2007), they extract multi-word features by post- processing approach and apply SVM classifier for text classification task. Okuno (2011) use post- processing approach for phrase extraction for Japanese documents, they first extract n-gram feature and counting their frequency and remove n-grams of lower frequency. Further select rule based features from n-grams. Then recalculate the frequency of features.

Adjectives, adverb, verb and noun semantic score are extracted from SentiWordNet and overall semantic orientation of document is computed by averaging semantic score of adjective, adverb, verb and noun (Ohana et al. 2009; Shukla et al. 2011). Lee et al. (2004), they apply min-cut method for subjective sentence extraction, and further classification algorithms like Naive Bayes (NB), Support Vector Machine (SVM) are applied for polarity classification. Bhattacharyya et al. (2005), used WordNet based method for the effective incorporation of linguistic information for subjective sentence extraction, further they also used Support Vector Machine classifier for polarity classification. Baharudin et al. (2011) , They classify sentence in subjective and objective using lexical approach and select subjective sentences, further they use bag-of-word model for feature selection and compute sentence semantic score using SentiWordNet or WorldNet dictionary. They update the polarity of each term using sentence structure and contextual feature in the sentence. Martineau et al. (2009) use bag of word model for feature creation, and Delta Term Frequency Inverse Document Frequency (TFIDF) technique is used for weighted word score, further they used support vector machine classifier for sentiment classification.

> Pang et al. (2002) extracted unigrams and created feature vector, which us given to various Machine Learning method for creating classification model. Their experimental results show that unigrams performs quite well with Support Vector Machine (SVM) for sentiment classification. Yang et al. (2011), They extract lexicon-based and semantics-based sequential features at various level: word, phrase, sentence. Further wrapper-based approach used for best feature selection and finally Support Vector Machine (SVM) algorithm is applied for sentiment classification. Misailovic et al. (2009) use movie review comment as a dataset and classification of text by subjectivity and polarity. Bag of word model used for feature selection and handle syntax and semantic property of feature, and handle negation. Further classification of text done by supervised learning: naive bayes, maximum entropy, decision tree and un-supervised learning: K-means clustering.

Bhattacharyya et al. (2008), detect subjective sentence and apply negation handling, further sentiment score based pruning is performed for better results. Sentiment score based pruning removes all non-sentiment words, and followed by information gain pruning remove specific stop word and noisy word. Tf-idf based feature vector is created for classification mainly containing sentiment word only. Further, SVM classifier is applied to this feature vector and review documents are classified in positive and negative category.

Kaji et al. (2007), extracted polar sentences from the Japanese HTML documents using language structural clues. Next, phrases are extracted from polar sentences using dependency parser. Further semantic score of each polar phrase is computed using Chi-square and PMI method. Bakliwal et al. (2011) use n-gram based method. They extract trigram, bigram and unigram and compute review score using their own scoring function further they apply naive bayes or SVM for classification of review. Yuanbin et al. (2009) extract dependency relation features and apply SVM classifier for sentiment classification. Penstein-Rose et al. (2009), they convert dependency relation features in to composite back off features, they generalize the feature by back off head word of dependency relation, and further SVM machine classifier is applied for text classification. Li and zhu et al.(2011), they use dependency parser for extraction of dependency relation further apply pruning algorithm to discard unnecessary dependency features and  noisy relation. Further, support vector machine classifier is used for sentiment classification task.

# Chapter 3

# Resources used

Proposed work in the area of sentiment analysis uses some of the standard publically available resources. Descriptions of these resources are presented in this chapter. Mainly two standard commonly used resources are used in the proposed work i.e. SentiWordNet and Part-of-Speech (POS) tagger. Used resources are described in subsequent subsection.

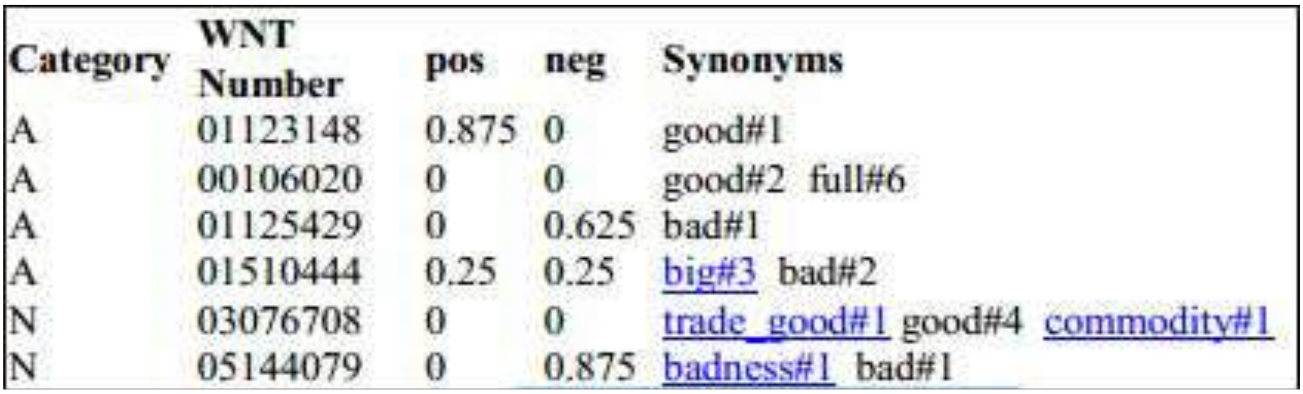
## 3.1 SentiWordNet

SentiWordNet is a tool used for finding the polarity score of any opinion word. It is based on the WordNet. Words in SentiWordNet are divided in four categories Adjective, Adverb, Verb and Noun. (http://sentiwordnet.isti.cnr.it ). It is a WordNet like lexicon which contain words with three scores as given below:

1. Positive score
2. Negative score
3. Objective score

For every word, positive, negative and neutral score are having values between 0.0 to 1.0 and addition of all the score i.e. positive score, negative score and objective score for a word is 1. The objective score of 1.0 denotes that it is a neutral word and don't denote any of the opinion. Any subjective opinion word will have the nonzero positive or negative score which is useful for us to find the feature wise rating of reviews [9].

A small fragment of SentiWordNet is shown in Figure 4.1.



| Category | WNT Number | pos | neg | Synonyms |
|---|---|---|---|---|
| A | 01123148 | 0.875 | 0 | good#1 |
| A | 00106020 | 0 | 0 | good#2 full#6 |
| A | 01125429 | 0 | 0.625 | bad#1 |
| A | 01510444 | 0.25 | 0.25 | big#3 bad#2 |
| N | 03076708 | 0 | 0 | trade_good#1 good#4 commodity#1 |
| N | 05144079 | 0 | 0.875 | badness#1 bad#1 |

Figure 3.1: SentiWordNet Fragment

## 3.2 Stanford POS tagger

A Part-Of-Speech Tagger (POS Tagger) is software that reads text from the reviews files and assigns parts of speech to each word, such as noun, verb, adjective, adverb, preposition etc., and the tag of the word is appended to the word in the tagged text. We can use the tagger in three modes: tagging, training, and testing. Tagging allows us to use a pre-trained model developed by the Stanford group to assign part of speech tags to unlabeled text sentences. Training allows us to develop a new model on the basis of a set of tagged data that we provide. Testing allows us to see how correctly a tagger tags the input sentences by tagging the labeled data and evaluating the results against the correct tags given by us to the sentences. We are using the tagging mode to assign part of speech tags to unlabeled text. [5]. Part of speech tagging done by POS tagger, for example "*This is a good movie.*" after applying POS tagger this sentence becomes "*This_DT is_VBZ a_DT good_JJ movie_NN.*". POS tagger used 36 types of tags (http://www.nlp.stanford.edu/).

# Chapter 4

# Methodology

Proposed approaches used in this thesis are discussed in this chapter. Initially dataset are pre-processed by Part-of-Speech (POS) tagging, Further, negation handling is applied. Then subjective sentences are selected using two methods based of SentiWordNet and Bayes algorithms i.e. discarding objective sentences from the documents. Further, one-word and two-word features are extracted i.e. unigrams and POS based fixed patterns. POS based rules are able to extract sentiment-rich phrases which incorporates contextual information from the text. After extraction of sentiment-rich phrases, semantic orientations of all these phrases are computed using Point-wise Mutual Information (PMI) method. Unigrams are extracted using POS i.e. adjectives, adverbs, verbs etc. and semantic orientations of unigrams are also computed using PMI method and with the help of SentiWordNet. Finally, overall semantic orientation of the document is determined by aggregating the semantic orientation of all the phrases and unigrams in the document. Proposed approach demonstrated in Figure 4.1.

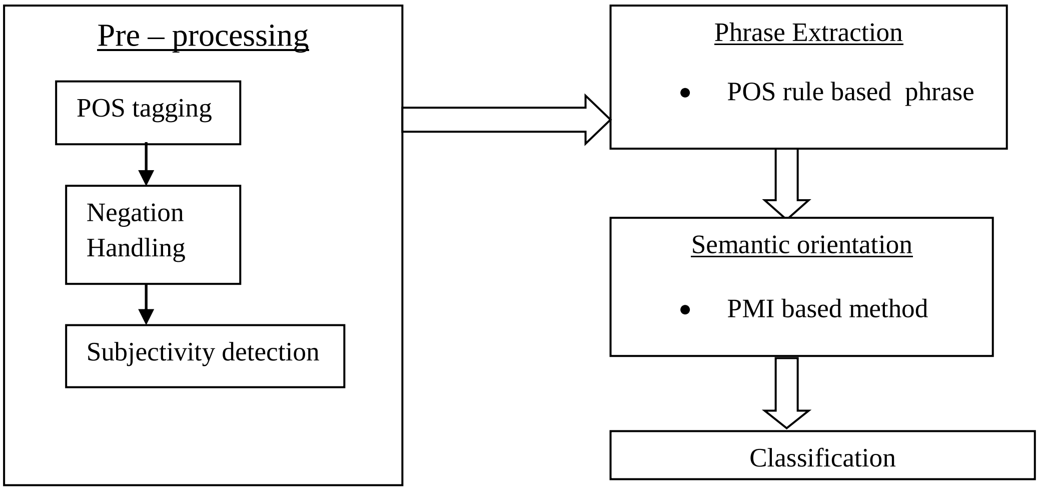


Figure 4.1: Proposed Approach

## 4.1 Negation Handling

Some negative sentiment sentences contain positive words, for example "*this is not a great movie.*", and word "great" is a clue for positive polarity so there is a need of negation handling. In the proposed work, negation handling is performed by removing all the stop-words which comes after not/n't/never and combine not/n't/never with non stop-word. Stop words are those words which do not contribute anything in identifying the sentiments like a, the, an etc. The given sentence after negation handling would be converted into this sentence "*this is NOT_great movie.*".

## 4.2 Subjectivity Detection

Subjective sentences are extracted from documents because mostly movie review documents contain two types of sentences, one which talks about the actors or plot in the movie and other which express the sentiment about the movie. Sentiment analysis and opinion mining task is interested in the "*about the movie*" part of review, and the sentences which express "about the movie" part are called subjective sentence. Generally, people express "about the movie" part by strong adjectives. For example: "there *is a great deal of corny dialogue and preposterous moments*" contain adjectives great, corny and preposterous. Average Subjective score these sentences are determined using various methods, for example sentence if the average subjective score of these adjectives is 0.710 and objective score is 0.290, then as subjective score is greater than objective so sentence declared subjective. One more example "*There is also a local detective who is conducting her own personal investigation*" contain adjectives "local", "own" and "personal" which gives subjective score 0.167 and objective score 0.833, as in this sentence objective score is greater than subjective score, therefore these types of sentences are discarded. In this thesis, two methods are explored for determining if a given sentence is a subjective sentence or objective sentence i.e. Naïve Bayes based subjectivity detector and SentiWordNet.

### 4.2.1 Naïve Bayes classifier based subjectivity detector

To determine if a given sentence is subjective or objective, a learning model is developed based on Naive Bayes classifier on the subjective dataset. Subjective dataset contains 5000 subjective and 5000 Objective sentences (http://www.cs.cornell.edu/people/pabo/movie-review-data/). This dataset is used for subjectivity classification. Naïve bayes classifier initially computes the probability that a given instance belongs to which class, and then it labels the instance whose probability is highest. Similarly, in this method, probability that a give sentence belongs to subjective class or objective class is computed and further sentence is labeled according to which probability (probability that given sentence is subjective or objective) is high. Process to determine the probability that a given sentence belongs to subjective or objective class is determined in two phases, in the first phase a lexicon is build which has subjective and objective score of words computed using subjectivity dataset. And in the second phase, for every sentence of testing movie review, subjective and objective score is retrieved from the lexicon and further average subjective and objective score of all the words in the sentence are computed, finally based on these average score, it is determined that a given sentence is subjective or objective. Detailed description of this process is described in following subsections.

1. Initially, adjectives/adverbs are extracted from subjective dataset using POS tagger and frequency of adjective/adverb in subjective and objective sentences i.e. f(w,sub) and f(w,obj) are computed respectively. Further, probability that a given sentence belongs to subjective sentences p(w,sub) is computed as given in Equation (1) and similarly probability that a given sentence belongs to objective sentences p(w,obj) is computed as given in Equation (2), and built a lexicon which contain adjectives/adverb and their subjective and objective probability.

$$p(w,sub) = \frac{f(w,sub)}{f(w,sub) + f(w,obj)} \qquad \textbf{(1)}$$

$$p(w,obj) = \frac{f(w,obj)}{f(w,sub) + f(w,obj)} \qquad (2)$$

2. For all the sentence in the testing review documents POS tagging is performed. Then adjectives/adverbs are extracted from a sentence of a document. Next, subjective and objective score are retrieved for all these words from lexicon build in first phase.

Finally, average subjective and objective scores of all adjective/adverb of a sentence are computed using Equation (3), (4).

$$Sub(s) = \frac{\sum_i p(w_i, sub)}{n} \tag{3}$$

$$Obj(s) = \frac{\sum_i p(w_i, obj)}{n} \tag{4}$$

For each sentence subjective Sub(s) and objective Obj(s) scores are computed, and classify the sentence into objective by two ways.

(1) If Sub(s) > Obj(s) then sentence is considered as subjective else sentence is objective and discarded the objective sentence.

(2) Sort the sentences by their average subjective score and select top 80% or 85% of sentences.

### 4.2.2 SentiWordNet Method

SentiWordNet based method works retrieves the polarity scores for each word from the standard polarity lexicon. Subjective and objective scores of each sentence are computed using Equation (5) and (6) with the help of SentiWordNet (Amitavadas et al., 2009). Various methods are explored to compute the subjective and objective score of a sentence to investigate the performance of only adjectives, adjectives with adverbs and adjectives with adverbs, noun, and verbs. Since, adjectives are intuitively the most important in expressing sentiments; therefore, the more weights are given to the adjectives as compared to other part-of-speech words. For example, “this is a very nice movie”. In this example, “nice” is the adjective which is the most important word in this sentence that is conveying information that a given sentence is expressing some sentiment and is likely to be a subjective sentence.

If a word is an adjective then subjective and objective scores are computed using Equation (5) and (6).

$$sub(w_i) = \alpha * (|pos(w_i)| + |neg(w_i)|) \tag{5}$$

$$obj(w_i) = \alpha * (1 - sub(w_i)) \tag{6}$$

And if a word is an adverb, verb, noun then subjective and objective scores are computed using Equations (7) and (8).

$$sub(w_i) = \beta * (|pos(w_i)| + |neg(w_i)|) \tag{7}$$

$$obj(w_i) = \beta * (1 - sub(w_i)) \tag{8}$$

Here pos($w_i$) is a positive score, and neg($w_i$) is a negative score of a given word retreived from SentiWordNet. α and β are constants where α > β, in our experiment we take α = 2 and β = 0.5 . If a sentence contains n words then subjective and objective scores of the sentence are computed using Equations (9) and (10).

$$sub_{score} = \frac{\sum_{i=(1,n)} sub(w_i)}{n} \tag{9}$$

$$obj_{score} = \frac{\sum_{i=(1,n)} obj(w_i)}{n} \tag{10}$$

After determining the average subjective and objective scores for the sentences, two cases are taken for further processing.

(1) If $sub_{score} > obj_{score}$ then sentence is considered subjective else considered as objective and then objective sentence are discarded.

(2) Sort the sentences by their average subjective score and select top 80% or 85% of sentences.

## 4.3 Feature Extraction

Feature extraction is the most important step in sentiment analysis, as if features extracted are good indicators of expressing sentiments then accuracy would be high. In the proposed approach, two types of features are extracted i.e. one-word features and two-word features called unigrams and phrases respectively.

### 4.3.1 Unigrams

Unigrams are the features that are extracted by tokenizing all the words in the documents. Each individual word is considered as unigram. In unigram as features adjectives, adverbs, nouns and verbs part-of-speech are considered as mostly these part-of-speech words convey the sentiments.

### 4.3.2 Phrases

Phrases are very essential for extraction of syntactic, contextual information which is very important for sentiment analysis. For example, attaching an adverb like "*very*" with a polar adjective "*good*" will increase the intensity of the word "*good*". This information may be useful for sentiment classification. In addition, phrases are capable of capturing contextual information like "*not good*", "*unpredictable story*", "*amazing movie*" etc. In the proposed approach, phrases are extracted which contain contextual and syntactic information that is important for sentiment analysis i.e. POS based fixed pattern.

#### 4.3.2.1 POS based fixed patterns

Individual words can have different polarity for different domains. For example, "unpredictable" word may have a negative polarity in automobile review, with phrase "unpredictable steering", but it could have positive polarity for movie review with the phrase "unpredictable story" (Turney, 2002). Therefore, contextual information is important for sentiment analysis. Phrases are intuitively very effective in incorporating contextual and syntactic information. Phrases extracted using POS based fixed patterns are good indicators of sentiment information. Adjectives and adverbs are subjective in nature (Turney's, 2002 & Vasileios,1997) . In Pattern, all the phrases are extracted in which one member of the two word phrase is either adjective or adverb. These patterns are given in Table 4.1.

Table 4.1.Rules for extracting phrases

| S.no. | First Word | Second Word |
|---|---|---|
| 1 | JJ | NN/NNS |
| 2 | RB/RBR/RBS | JJ |
| 3 | JJ | JJ |
| 4 | NN/NNS | JJ |
| 5 | RB/RBR/RBS | VB/VBD/VBN/VBG |

The JJ/JJR/JJS tags indicate adjectives, the NN/NNS tags are nouns, the RB/RBR/RBS tags are adverbs, and the VB/VBG/VBP tags are verbs.

Part-of-Speech (POS) rule based phrase extraction process is demonstrated as follows. For a given sentence i.e. “this movie is very nice and has awesome cinematography.”. then, Part-of-speech (POS) tagged sentence of this example sentence is determined using stanford POS tagger as follows, “this_DT movie_NN is_VBZ very_RB nice_JJ and_CC has_VBZ awesome_JJ cinematography_NN ._.” In this Part-of-Speech tagged sentence, all the phrases are extracted using the rules given in Table 4.1. Phrases extracted are “very_RB nice_JJ” and “awesome_JJ cinematography_NN”.

## 4.4 Semantic orientation

After extraction of features, semantic orientation of each feature is determined using some formula. Computation of semantic orientation of the sentiment-rich features is based on the assumption that if a phrase is occurring frequently and predominantly in one class (positive or negative), then that phrase would have high polarity. If a phrase has high positive value that indicates that phrase has occurred more with positive sentences i.e. positive words. Point-wise Mutual Information (PMI) is generally used to calculate the strength of association between a phrase and positive or negative sentences. It is defined as follows (Kaji, 2007).

$$PMI\ (c,pos) = \ log_2 \frac{P(c,pos)}{P(c)P(pos)} \qquad \text{....(11)}$$

$$PMI\ (c,neg) = \ log_2 \frac{P(c,neg)}{P(c)P(neg)} \qquad \text{....(12)}$$

Here, *P(c,pos)* is probability of a phrase that it occurs in positive documents i.e. frequency of a phrase in positive documents divided by total number of positive documents. *P(c,neg)*is the probability that a phrase occurs in negative document i.e. frequency of a phrase divided by total number of negative documents. Polarity value of the phrase is determined by their PMI value difference (Turney, 2002) .In this paper, semantic orientation of a phrase is computed using equation 15.

$$SO(c) = \ PMI(c,pos) - PMI(c,neg) \qquad \text{... (13)}$$

$$SO(c) = \ log_2 \frac{\frac{P(c,pos)}{P(pos)}}{\frac{P(c,neg)}{P(neg)}} \qquad \text{... (14)}$$

$$SO(c) = \ log_2 \frac{P\left(\frac{c}{pos}\right)}{P\left(\frac{c}{neg}\right)} \qquad \text{... (15)}$$

## 4.5 Semantic orientation aggregation

After determination of semantic orientation of each feature using Point-wise Mutual Information (PMI) in the document, overall positive or negative semantic orientation of the document is determined by summing up the semantic orientation of all the words in the document. Finally, overall semantic orientation of the document is determined as positive if the aggregated semantic orientation value of the document is greater than zero else it is determined as negative.

# Chapter 5

# Experiments and Results

This chapter discusses the results and experimental setup used for the evaluating the effectiveness of the proposed method. This chapter is organised as follows, Initially, Experimental setting and dataset used are described, and Next evaluation measure used for comparing the performance of various methods are presented. Further, all the results for the proposed methods are discussed in detail.

## 5.1. Dataset Used

To evaluate the proposed approach, publically available benchmark dataset is used i.e. Movie review dataset provided by Cornell university (Pang et al. 2004). This dataset contains 2000 movie reviews containing 1000 positive and 1000 negative reviews (http://www.cs.cornell.edu/people/pabo/movie-review-data/). For all the experiments, 700 review documents from each class are randomly selected for training and remaining 300 review documents from each class are used for testing the propose approach for sentiment analysis.

## 5.2. Evaluation Method

The accuracy evaluation criterion for classification is presented in a confusion matrix (Table 5.1). True positive (TP), True negative (TN), False positive (FP), and False negative (FN) are the four different possible outcomes of average semantic score of feature based classifier. TP means that a review is classified to a positive class when this review really belongs to the positive class. TN means that a review is classified to a negative class when this review belongs to a negative class. Both TP and TN are correct classifications. FP means that a review is incorrectly classified to a positive class when this review belongs to a negative class. FN means that a review is incorrectly classified to a negative class when this review belongs to a positive class. Accuracy is evaluated using Equation (16).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad \text{….. (16)}$$

Table 5.1: Confusion matrix of accuracy evaluation criteria

| | | Predicted semantic orientation | |
|---|---|---|---|
| | | Positive review | Negative review |
| Actual sentiment orientation | Positive review | TP | FN |
| | Negative review | FP | TN |

## 5.3 Experimental results and discussions

All the experiments performed in the proposed approach for sentiment classification can be categorised into two experiments, objective of the first experiment is to investigate the best feature i.e. unigrams or phrases for sentiment classification. Therefore, subjectivity detection is not applied and whole document is considered for classification of the given review document into positive or negative polar document. The main objective of the second experiment is to investigate the impact of elimination of objective sentences from the review document. Two methods are applied for identification of subjective sentences as discussed in previous chapters.

**Experiment 1:** Initially, Stanford POS tagger is applied on document corpus; further negation handling is applied on corpus. Further, without applying any subjectivity detection two types of features are extracted i.e. unigrams and POS pattern based phrases. In case of unigrams, only those unigrams are extracted which are POS tagged by adjective (JJ), adverb (RB), verb (VB) and noun (NN). And, phrases are extracted which conform to the predefined patterns as given in Table 4.1. Further, semantic scores of these POS based unigrams and phrases are computed by using Equation (15). Further, lexicons of unigrams words and phrases with their semantic orientations are built. Now, at the time of testing a review document, semantic orientations of all the POS tagged unigrams are retrieved from this lexicon. Finally, the semantic orientations of all the words of the testing documents are aggregated for determining the polarity of the overall document as positive or negative. Similarly, overall semantic orientation of the document is determined with the POS pattern based phrases. Experimental results for the effectiveness of the unigrams and phrases for the sentiment analysis are shown in Table 5.2.

Table 5.2: Accuracy for unigrams and phrases for movie review dataset

| | Positively correctly classified | Negatively correctly classified | Total correctly classified | Accuracy In (%) |
|---|---|---|---|---|
| Unigrams | 277 | 163 | 440 | 73.33 |
| Phrases | 248 | 209 | 457 | 76.17 |

Experimental results for unigrams and phrases for sentiment analysis are shown in Table 5.2. Experiments performed only on unigrams which are POS tagged by JJ/RB/VB/NN and it correctly classify 277 (300) positive and 163 (300) negative documents and accuracy computed based on only unigram is 73.33 %. With phrases are features accuracy is increased from 73.33% to 76.17%**.** From the experimental results, it is clear that phrases are better than unigrams for sentiment classification.

**Experiment 2:** Subjectivity detection is performed to evaluate impact of subjectivity detection on sentiment classification. First of all objective sentence are discarded using both basic subjectivity detector and SentiWordNet method. Further, phrases are extracted using POS based rules as given in Table 4.1. Next, semantic orientations of all these phrases are computed using Equation (15) and built the phrase lexicon. Further, at the time of classifying a new test document all the phrases are extracted with POS based rules, and then overall polarity is determined by averaging semantic scores of all the phrases in the document. To investigate the effect of elimination of objective sentences, different numbers of sentences are eliminated by considering 80%, 85% and 90% of sentences. Accuracies for both types of subjectivity detection methods with various setting of experiments are reported in Table 5.3.

Table 5.3: Accuracy of phrase with subjectivity detection for movie review dataset

| | Sub(s) > Obj(s) | Top 80% | Top 85% | Top 90% |
|---|---|---|---|---|
| Naïve bayes based Subjectivity detector | **77.33%** | 73.00% | 75.17% | 75.83% |
| SentiWordNet based on Adj. | 76.00% | 74.33% | 75.33% | 77.13% |
| SentiWordNet based on Adj./Adv. | 76.00% | 77.00% | 77.83% | 77.83% |
| SentiWordNet based on Adj./Adv./Verb/Noun | 76.33% | **78.00%** | 77.17% | 77.17% |

Experimental results show that phrase without subjectivity detection shows an accuracy of 76.17%, and accuracy after subjectivity detection by different method reported in Table 5.3, basic subjectivity detector give an accuracy improvement of 77.33% (+1.52%) over 76.17% and by SentiWordNet method accuracy improves up to 78% (+2.40%) over 76.17%. Experimental results show that SentiWordNet method of subjectivity detection performs better than Naïve bayes based method for sentiment classification.

# Chapter 6

# Conclusion and Future Work

Sentiment Analysis research deals with the extraction of the opinion expressed by people about specific topic from the text review documents. Sentiment analysis research is very important for users as well as e-commerce companies as users may use the online reviews for their purchasing decisions and companies can use these reviews for improving their products. Generally, real facts and discussion about the plot, actors, and actress etc. are mixed within the reviews these sentences do not contribute in extraction of sentiments from the text. These sentences are called objective sentences. Elimination of which may enhance the performance of the sentiment analysis. Objective of this thesis is two-fold. First is to explore the best features that conveys better sentiments and second is to investigate a method which can improve the performance of the sentiment classification by eliminating the objective sentences. Therefore, initially experiments are performed to explore which feature are best i.e. unigrams or phrases. And, in further experiments two subjectivity detection methods are employed to know which method is best.

Experimental results show that phrases are better in capturing the sentiment from the documents it is due to the fact that phrases can incorporate the contextual information unlike unigrams. Next, by adding subjectivity detection into simple semantic orientation based methods for sentiment analysis increases the performance. Further, SentiWordNet based method performs better than naïve bayes based subjectivity detection method with adjectives, adverbs, nouns and verbs.

In future, more various sophisticated features can be explored that can incorporate more syntactic and long distance relation among words in the sentences, because these type of feature can be useful for sentiment analysis. Further, the proposed methods may be explored on various other datasets with different domains. In addition, the proposed methods may be tested on reviews written in non-English language.

# References

1. B. Pang, L. Lee, “thumbs up? Sentiment classification using machine learning technique”, EMNLP 2002, pp. 79-86.

2. B. Ohana, B. Tierney, “sentiment classification of review using sentiwordnet”, 2009,

3. A. Shukla ,”Sentiment analysis of document based on annotation” , IJWesT vol. 2, Nov. 2011.

4. Peter D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews” ,ACL july 2002, pp.417-424.

5. N. Kaji, M. Kitsuregawa, “Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents”, ACL 2007, PP. 1075-1083.

6. B. Pang, L. Lee ,”Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts”,2005.

7. T. Wilson, J. Wiebe ,”Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis”,2006.

8. A. das, S. bandyopadhyay, “Subjectivity detection in English and Bengali”, 7th international conference on NLP. 2009.

9. Z. Fei, J. Liu, G. Wu, “Sentiment classification using phrase pattern”, IEEE,2004.

10. Tun Thura Thet, Jin-Cheon Na, Christopher S.G. Khoo, Subbaraj Shakthikumar, “Sentiment Analysis of Movie Reviews on DiscussionBoards using a Linguistic Approach”, ACM,2009.

11. A. Bakliwal, P. Arora, A. Patil, V. Verma. “Towards enhanced opinion classification using NLP techniques”, IJCNLP 2011, pages 101-107.

12. P. Bhattacharyya, S. Verma , “incorporating semantic knowladgefor sentiment analysis”, ICON 2008.

13. A. Khan, B. Baharudin, " Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs”, Emerging Tech, 2011, pp 539-552.

14. E. Riloff, S. Patwardhan, J. wiebe, “Feature subsumption for opinion analysis”, 2006

15. M. Tchalakova, D. Gerdemann, D. Meurers “Automatic sentiment classification of product review using maximal phrase based analysis” , ACL-HLT 2011, pp. 111-117.

16. Y. Okuno, “Phrase extraction for Japanese predictive input method as post-processing”, WTIM, 2011., pp. 48-52.

17. W. Zhang, T. Yoshida, X. Tang, “Text Classification using Multi-word Features”, IEEE, 2007, pp. 3519-3524.

18. Y. Wu, Q. Zhang, X. Huang, L. Wu, “Phrase dependency parsing for opinion mining”, ACL & AFNLP, 2009, pp. 1533-1541.

19. J. Martineau, T. Finin, “Delta TFIDF: An Improved Feature Space for Sentiment Analysis”, AAAI 2009.

20. B. Liu, “Sentiment Analysis and Opinion Mining”. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers 2012.

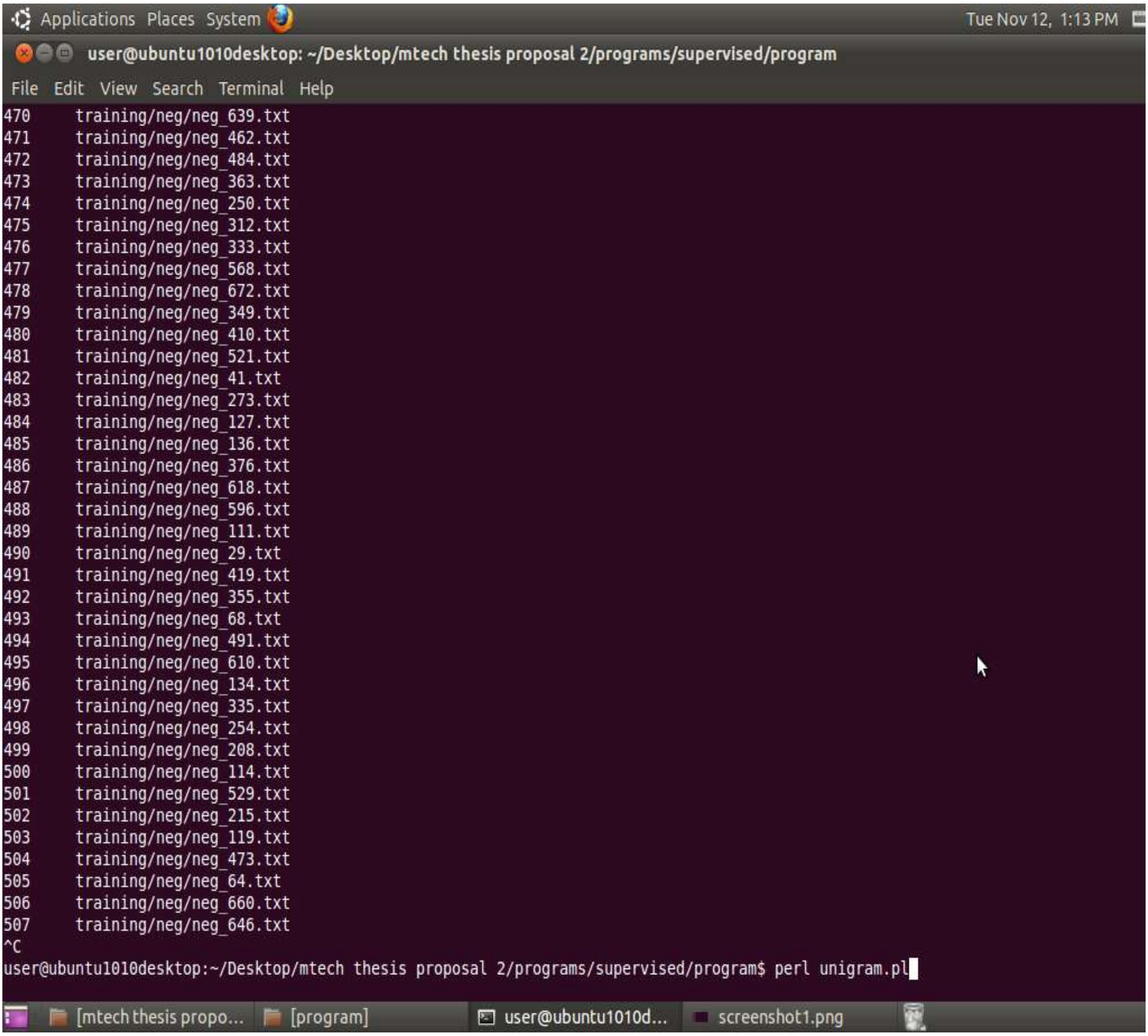
# Appendix

There are some examples that shows the output process of Sentiment analysis.

Experiment 1:

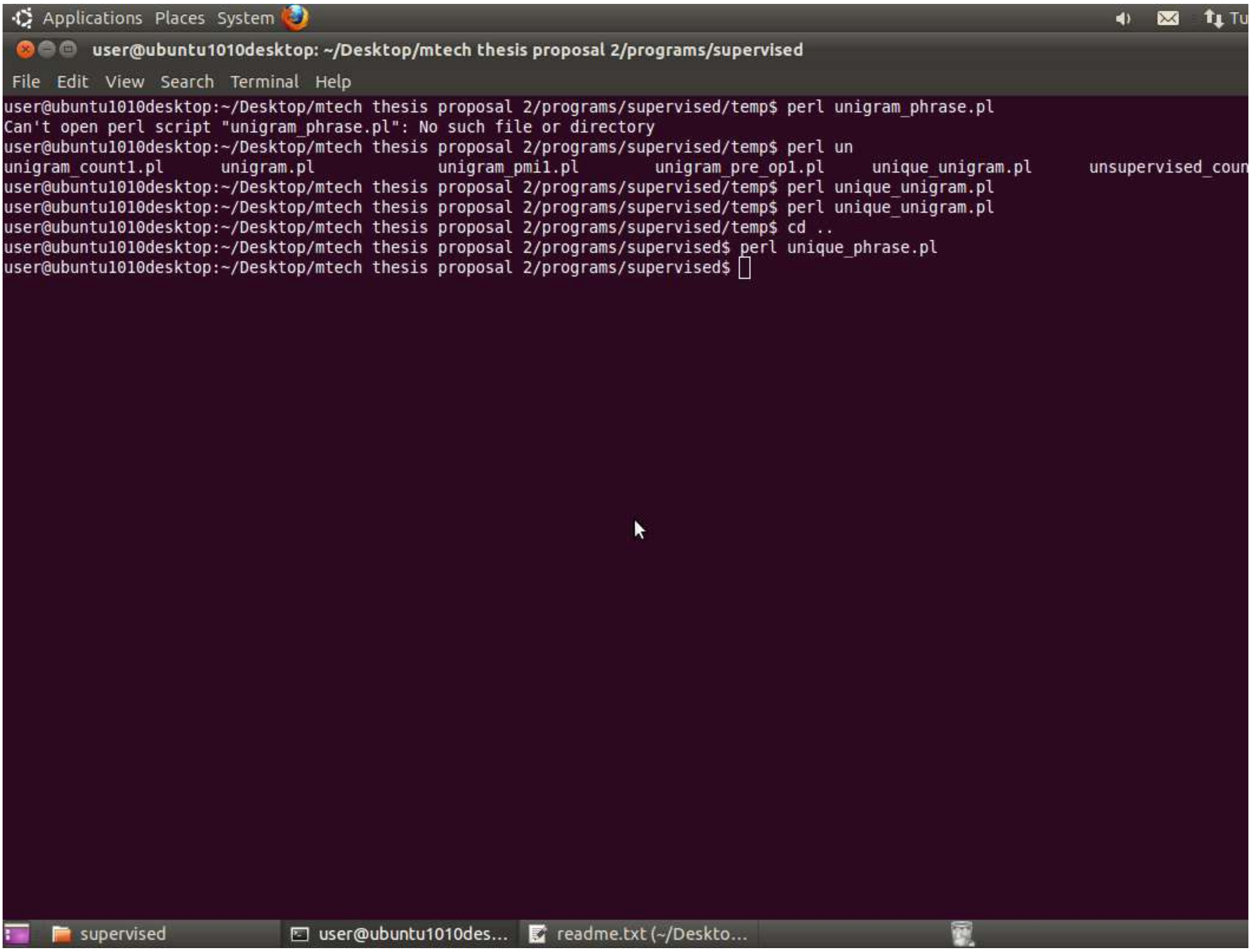
Unigrams:-

1. Below the figure shows the output of statement “Perl unigram.pl”. Unique Keywords will be generated (i.e. adjectives , adverbs, nouns etc.)  gen_phrase.txt will be created.


Applications Places System
Tue Nov 12, 1:13 PM
user@ubuntu1010desktop: ~/Desktop/mtech thesis proposal 2/programs/supervised/program
File Edit View Search Terminal Help
[mtech thesis propo...
[program]
user@ubuntu1010d...
screenshot1.png

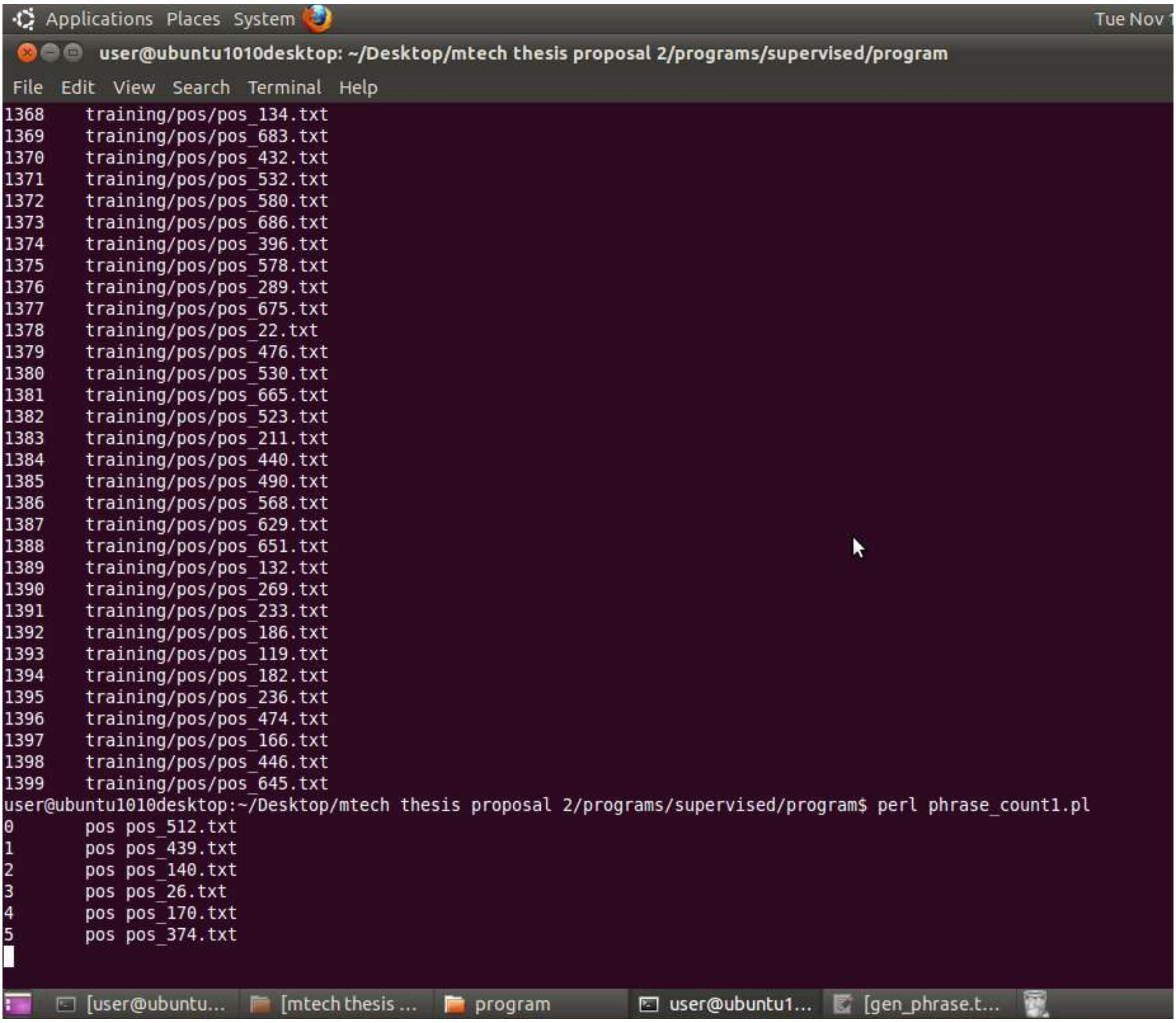
```
470     training/neg/neg_639.txt
471     training/neg/neg_462.txt
472     training/neg/neg_484.txt
473     training/neg/neg_363.txt
474     training/neg/neg_250.txt
475     training/neg/neg_312.txt
476     training/neg/neg_333.txt
477     training/neg/neg_568.txt
478     training/neg/neg_672.txt
479     training/neg/neg_349.txt
480     training/neg/neg_410.txt
481     training/neg/neg_521.txt
482     training/neg/neg_41.txt
483     training/neg/neg_273.txt
484     training/neg/neg_127.txt
485     training/neg/neg_136.txt
486     training/neg/neg_376.txt
487     training/neg/neg_618.txt
488     training/neg/neg_596.txt
489     training/neg/neg_111.txt
490     training/neg/neg_29.txt
491     training/neg/neg_419.txt
492     training/neg/neg_355.txt
493     training/neg/neg_68.txt
494     training/neg/neg_491.txt
495     training/neg/neg_610.txt
496     training/neg/neg_134.txt
497     training/neg/neg_335.txt
498     training/neg/neg_254.txt
499     training/neg/neg_208.txt
500     training/neg/neg_114.txt
501     training/neg/neg_529.txt
502     training/neg/neg_215.txt
503     training/neg/neg_119.txt
504     training/neg/neg_473.txt
505     training/neg/neg_64.txt
506     training/neg/neg_660.txt
507     training/neg/neg_646.txt
^C
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised/program$ perl unigram.pl
```

2. After that we have to discard the repeated words. For this we run unique_phrase.pl which will create gen_phrase1.txt. The output of statement perl unigram_phrase.pl is shown below:


Applications Places System
user@ubuntu1010desktop: ~/Desktop/mtech thesis proposal 2/programs/supervised
File Edit View Search Terminal Help
supervised
user@ubuntu1010des...
readme.txt (~/Deskto...

```
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised/temp$ perl unigram_phrase.pl
Can't open perl script "unigram_phrase.pl": No such file or directory
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised/temp$ perl un
unigram_count1.pl      unigram.pl             unigram_pmi1.pl        unigram_pre_op1.pl     unique_unigram.pl      unsupervised_coun
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised/temp$ perl unique_unigram.pl
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised/temp$ perl unique_unigram.pl
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised/temp$ cd ..
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised$ perl unique_phrase.pl
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised$ 
```

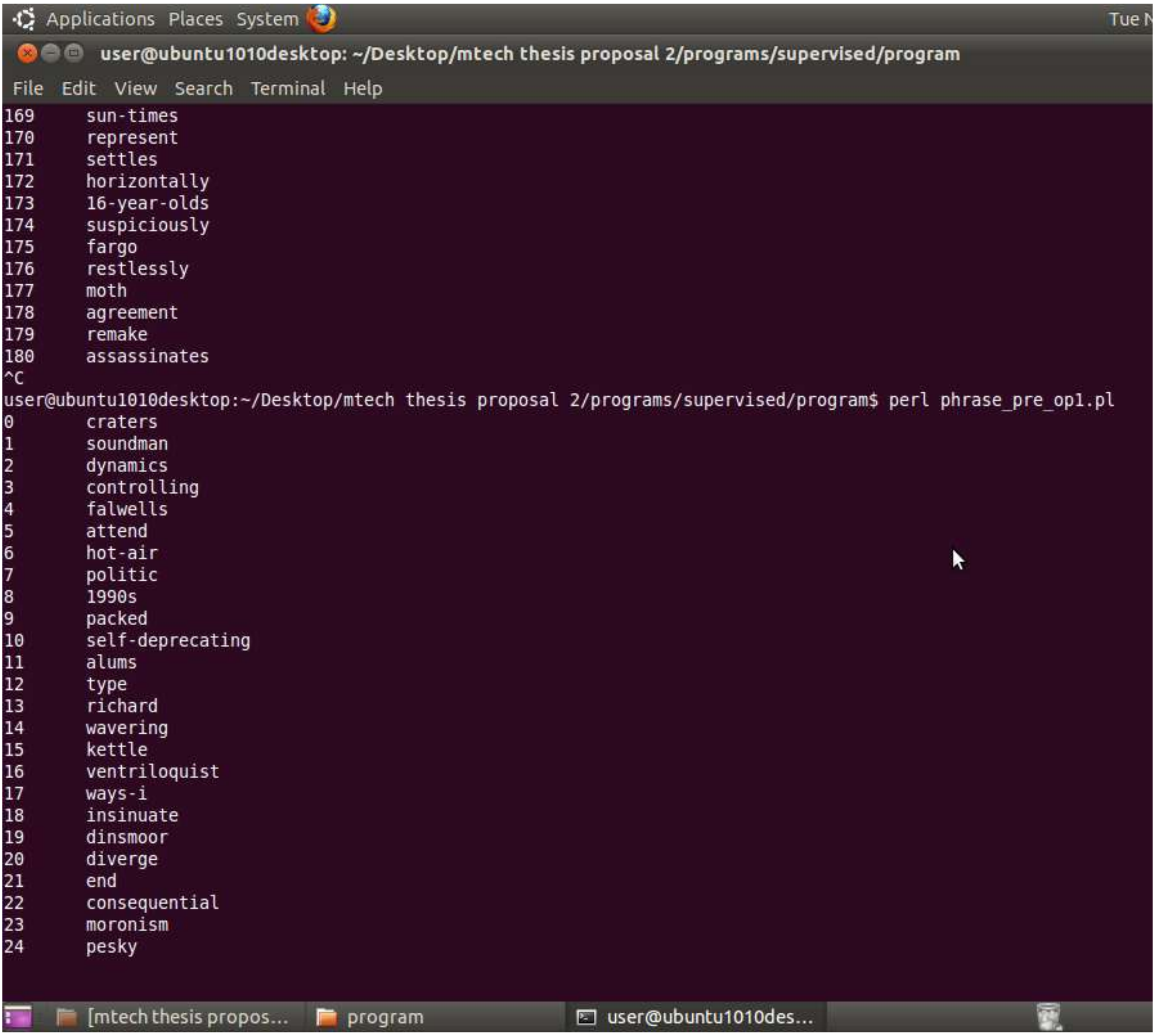
3. We run phrase_count1.pl to calculate negative and positive frequency of unique words in pos and neg documents.


Applications Places System
Tue Nov 1
user@ubuntu1010desktop: ~/Desktop/mtech thesis proposal 2/programs/supervised/program
File Edit View Search Terminal Help
[user@ubuntu...
[mtech thesis ...
program
user@ubuntu1...
[gen_phrase.t...

```
1368    training/pos/pos_134.txt
1369    training/pos/pos_683.txt
1370    training/pos/pos_432.txt
1371    training/pos/pos_532.txt
1372    training/pos/pos_580.txt
1373    training/pos/pos_686.txt
1374    training/pos/pos_396.txt
1375    training/pos/pos_578.txt
1376    training/pos/pos_289.txt
1377    training/pos/pos_675.txt
1378    training/pos/pos_22.txt
1379    training/pos/pos_476.txt
1380    training/pos/pos_530.txt
1381    training/pos/pos_665.txt
1382    training/pos/pos_523.txt
1383    training/pos/pos_211.txt
1384    training/pos/pos_440.txt
1385    training/pos/pos_490.txt
1386    training/pos/pos_568.txt
1387    training/pos/pos_629.txt
1388    training/pos/pos_651.txt
1389    training/pos/pos_132.txt
1390    training/pos/pos_269.txt
1391    training/pos/pos_233.txt
1392    training/pos/pos_186.txt
1393    training/pos/pos_119.txt
1394    training/pos/pos_182.txt
1395    training/pos/pos_236.txt
1396    training/pos/pos_474.txt
1397    training/pos/pos_166.txt
1398    training/pos/pos_446.txt
1399    training/pos/pos_645.txt
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised/program$ perl phrase_count1.pl
0       pos pos_512.txt
1       pos pos_439.txt
2       pos pos_140.txt
3       pos pos_26.txt
4       pos pos_170.txt
5       pos pos_374.txt
```

The statement "perl unique_phrase.pl will generate gen_phrase1.txt having the output.

4. then to remove tags like _jj etc we run “perl phrase_pre_op1.pl” which will create gen_unigram_pmi.txt

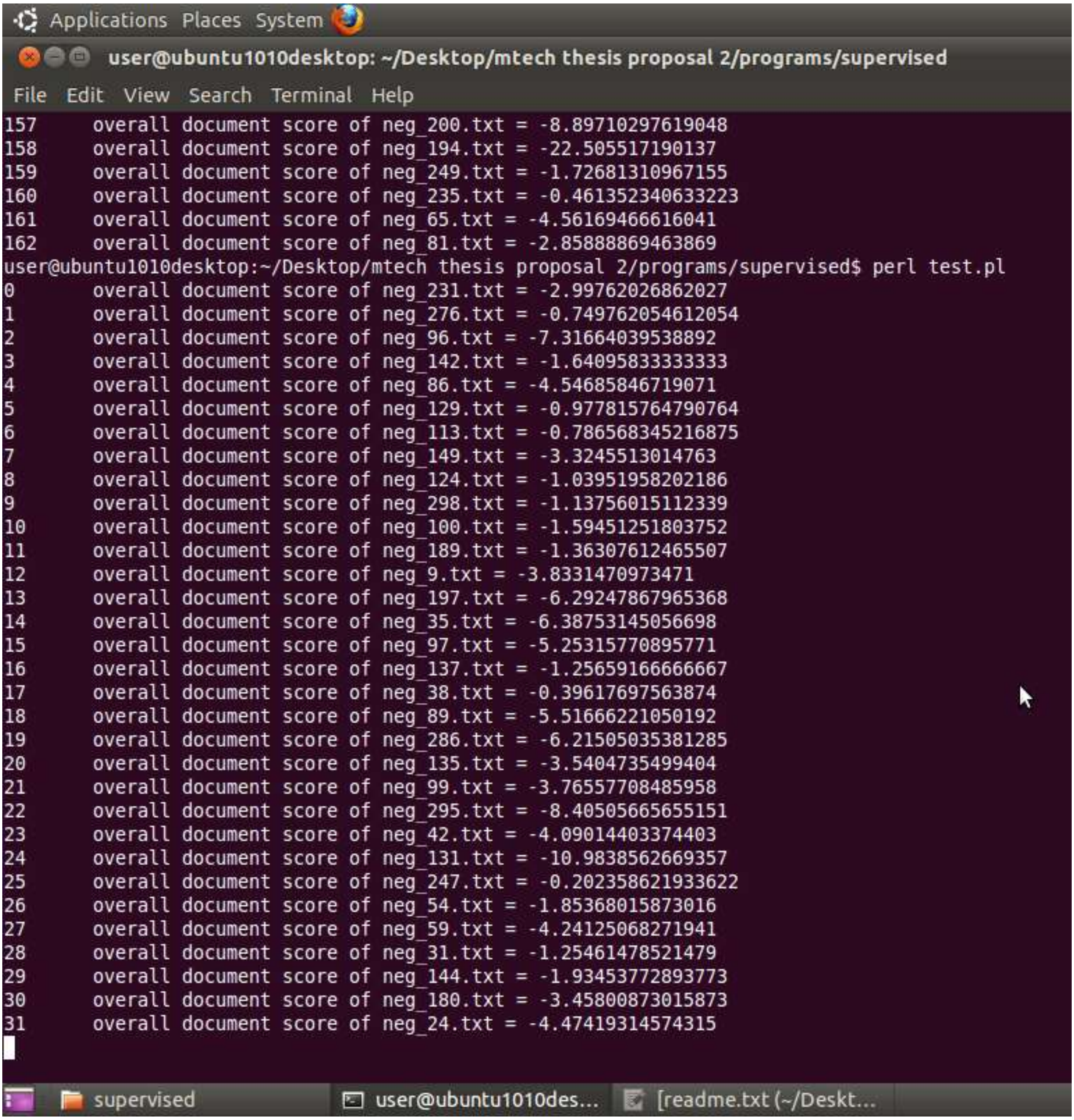


Applications Places System
Tue N
user@ubuntu1010desktop: ~/Desktop/mtech thesis proposal 2/programs/supervised/program
File Edit View Search Terminal Help
[mtech thesis propos...
program
user@ubuntu1010des...

```
169     sun-times
170     represent
171     settles
172     horizontally
173     16-year-olds
174     suspiciously
175     fargo
176     restlessly
177     moth
178     agreement
179     remake
180     assassinates
^C
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised/program$ perl phrase_pre_op1.pl
0       craters
1       soundman
2       dynamics
3       controlling
4       falwells
5       attend
6       hot-air
7       politic
8       1990s
9       packed
10      self-deprecating
11      alums
12      type
13      richard
14      wavering
15      kettle
16      ventriloquist
17      ways-i
18      insinuate
19      dinsmoor
20      diverge
21      end
22      consequential
23      moronism
24      pesky
```

5. at last we will calcute semantic score by running statement phrase_pmi1.pl which create text file pos_phrase.txt and neg_phrase.txt

6. At last we run test.pl to classify negative and positive docs


Applications Places System
user@ubuntu1010desktop: ~/Desktop/mtech thesis proposal 2/programs/supervised
File Edit View Search Terminal Help
supervised
user@ubuntu1010des...
[readme.txt (~/Deskt...

```
157     overall document score of neg_200.txt = -8.89710297619048
158     overall document score of neg_194.txt = -22.505517190137
159     overall document score of neg_249.txt = -1.72681310967155
160     overall document score of neg_235.txt = -0.461352340633223
161     overall document score of neg_65.txt = -4.56169466616041
162     overall document score of neg_81.txt = -2.85888869463869
user@ubuntu1010desktop:~/Desktop/mtech thesis proposal 2/programs/supervised$ perl test.pl
0       overall document score of neg_231.txt = -2.99762026862027
1       overall document score of neg_276.txt = -0.749762054612054
2       overall document score of neg_96.txt = -7.31664039538892
3       overall document score of neg_142.txt = -1.64095833333333
4       overall document score of neg_86.txt = -4.54685846719071
5       overall document score of neg_129.txt = -0.977815764790764
6       overall document score of neg_113.txt = -0.786568345216875
7       overall document score of neg_149.txt = -3.3245513014763
8       overall document score of neg_124.txt = -1.03951958202186
9       overall document score of neg_298.txt = -1.13756015112339
10      overall document score of neg_100.txt = -1.59451251803752
11      overall document score of neg_189.txt = -1.36307612465507
12      overall document score of neg_9.txt = -3.8331470973471
13      overall document score of neg_197.txt = -6.29247867965368
14      overall document score of neg_35.txt = -6.38753145056698
15      overall document score of neg_97.txt = -5.25315770895771
16      overall document score of neg_137.txt = -1.25659166666667
17      overall document score of neg_38.txt = -0.39617697563874
18      overall document score of neg_89.txt = -5.51666221050192
19      overall document score of neg_286.txt = -6.21505035381285
20      overall document score of neg_135.txt = -3.5404735499404
21      overall document score of neg_99.txt = -3.76557708485958
22      overall document score of neg_295.txt = -8.40505665655151
23      overall document score of neg_42.txt = -4.09014403374403
24      overall document score of neg_131.txt = -10.9838562669357
25      overall document score of neg_247.txt = -0.202358621933622
26      overall document score of neg_54.txt = -1.85368015873016
27      overall document score of neg_59.txt = -4.24125068271941
28      overall document score of neg_31.txt = -1.25461478521479
29      overall document score of neg_144.txt = -1.93453772893773
30      overall document score of neg_180.txt = -3.45800873015873
31      overall document score of neg_24.txt = -4.47419314574315
```

Similarly we can have for phrases also.