STATUS OF THESIS

Title of thesis: Sentiment Classification of Online Customer Reviews and Blogs Using Sentence-level Lexical Based Semantic Orientation Method

I, AURANGZEB KHAN

hereby allow my thesis to be placed at the Information Resource Center (IRC) of Universiti Teknologi PETRONAS (UTP) with the following conditions:

1. The thesis becomes the property of UTP
2. The IRC of UTP may make copies of the thesis for academic purposes only.
3. This thesis is classified as

□ Confidential

☑ Non-confidential

If this thesis is confidential, please state the reason:

________________________________________

________________________________________

________________________________________

The contents of the thesis will remain confidential for _______ years.

Remarks on disclosure:

________________________________________

________________________________________

________________________________________

Endorsed by

_____________________________

Signature of Author

Permanent address:

Village Shahbury P/O Bahrat, Teh &
District, Bannu, KPK, Pakistan

Date: _________________________

_____________________________

Signature of Supervisor

Name of Supervisor

Assoc. Prof. Dr. Baharum Baharudin

Date: _________________________

UNIVERSITI TEKNOLOGI PETRONAS

# SENTIMENT CLASSIFICATION OF ONLINE CUSTOMER REVIEWS AND BLOGS USING SENTENCE-LEVEL LEXICAL BASED SEMANTIC ORIENTATION METHOD

By

AURANGZEB KHAN

The undersigned certify that they have read, and recommend to the Postgraduate Studies Programme for acceptance this thesis for the fullfilment of the requirements for the degree stated.

Signature: ______________________________________

Main Supervisor: Assoc. Prof. Dr. Baharum Baharudin____

Signature: ______________________________________

Head of Department: Dr. Mohd Fadzil Bin Hassan____________

Date: ______________________________________

# SENTIMENT CLASSIFICATION OF ONLINE CUSTOMER REVIEWS AND BLOGS USING SENTENCE-LEVEL LEXICAL BASED SEMANTIC ORIENTATION METHOD

By

AURANGZEB KHAN

A Thesis

Submitted to the Postgraduate Studies Programme

As a requirement for the degree of

DOCTOR OF PHILOSOPHY

INFORMATION TECHNOLOGY

UNIVERSITI TEKNOLOGI PETRONAS

BANDAR SERI ISKANDAR

PERAK

NOVEMBER, 2011

## DECLARATION OF THESIS

Title of thesis: Sentiment Classification of Online Customer Reviews and Blogs Using Sentence-level Lexical Based Semantic Orientation Method

I, AURANGZEB KHAN

hereby declare that the thesis is based on my original work except for quotations and citations, which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTP or other institutions.

Witnessed by

| | |
|---|---|
| Signature of Author | Signature of Supervisor |
| Permanent address: | Name of Supervisor |
| Village Shahbury P/O Bahrat, Teh & District, Bannu, KPK, Pakistan | Assoc. Prof. Dr. Baharum Baharudin |
| Date: ____________ | Date: ____________ |

*To my parents, wife and children*

# ACKNOWLEDGMENT

First of all, I would like to thank the Almighty Allah, for His divine guidance and providence. His support, blessings, goodness, and kindness were always with me. He blessed me with motivation, passion, and hard work. It was his blessings which made me able to plan, visualize, and execute my dreams into the reality. I would like to dedicate this achievement to Him and to my mother and wife, whose immeasurable sacrifices have led to what I am today.

Every PhD student dreams for a visionary supervisor. I am extremely thankful to Allah, for connecting me up with the best possible supervisor Associate Professor Dr. Baharum Baharudin. Prof. Dr. Baharum’s guidance, inspirations and motivations enabled me to bring out the best of myself. I managed to learn a lot from him. The list would become quite long if I mention each and everything, to be short, it is Dr. Baharum’s dedication and perseverance that made me learn scientific writing, guided me to the art of critical thinking, taught me to work smart, gave me freedom to explore different ideas, linked me up with the best researchers, and provided me the best possible working environment. Without his continuous support and appreciation this research work would not have been accomplished. I pay my deep regards to him and his greatness. I am also very grateful to other lecturers and the entire staff of my department for their support during the course of my study. I am very obliged to the Universiti Teknologi PETRONAS (UTP) Malaysia, for awarding me a fully funded scholarship. Without it, I would not be able to pursue my studies.

With my deep heart, I acknowledge my wife, who took care of my children while I was busy in my research activities. Her encouragement, motivations, support and understanding made me able to work on my research activities and to write this thesis. I would like to acknowledge my children: Hina Zeb, Maria Zeb and Ahmed Zeb khan for their patience while I was away for this noble cause.

# ABSTRACT

Sentiment analysis is the process of extracting knowledge from the peoples’ opinions, appraisals and emotions toward entities, events and their attributes. These opinions greatly impact on customers to ease their choices regarding online shopping, choosing events, products and entities. With the rapid growth of online resources, a vast amount of new data in the form of customer reviews and opinions are being generated progressively. Hence, sentiment analysis methods are desirable for developing efficient and effective analyses and classification of customer reviews, blogs and comments.

The main inspiration for this thesis is to develop high performance domain independent sentiment classification method. This study focuses on sentiment analysis at the sentence level using lexical based method for different type data such as reviews and blogs. The proposed method is based on general lexicons i.e. WordNet, SentiWordNet and user defined lexical dictionaries for sentiment orientation. The relations and glosses of these dictionaries provide solution to the domain portability problem.

The experiments are performed on various datasets such as customer reviews and blogs comments. The results show that the proposed method with sentence contextual information is effective for sentiment classification. The proposed method performs better than word and text level corpus based machine learning methods for semantic orientation. The results highlight that the proposed method achieves an average accuracy of 86% at sentence-level and 97% at feedback level for customer reviews. Similarly, it achieves an average accuracy of 83% at sentence level and 86% at feedback level for blog comments.

# ABSTRAK

Analisis sentimen adalah proses untuk mengekstrak pendapat manusia, penilaian dan emosi terhadap entity, situasi dan ciri-ciri mereka. Pendapat ini memberi kesan yang besar terhadap pelanggan untuk memudahkan dalam pilihan mereka untuk membeli-belah melalui talian internet, menentukan situsai, produk dan entiti. Dengan pertumbuhan yang pesat melalui sumber dari internet, jumlah besar data baru dalam bentuk pemerhatian dan pendapat pelanggan dapat dihasilkan dan diperoleh dengan banyak. Oleh itu, kaedah analisis sentimen wajar untuk menghasilkan analisis yang berkesan dan cekap dan klasifikasi pandangan, blog dan pendapat pelanggan.

Inspirasi utama untuk disertasi ini adalah untuk menghasilkan kaedah bebas domain klasifikasi sentimen yang berprestasi tinggi. Kajian ini memberi tumpuan kepada analisi sentimen pada peringkat ayat menggunakan kaedah ‘lexical untuk data yang berbeza-beza seperti ulasan dan blog. Kaedah yang dicadangkan adalah berasakan kepada ‘lexicon’ umum’ i.e. WordNet, SentiWordNet dan takrifan pengguna kamus’ lexical’ untuk orientasi sentimen. Hubungan dan glos kamus ini memberi penyelesaian kepada masalah domain mudah alih.

Eksperimen yang dijalankan berdasarkan pelbagai set data seperti ulasan pelanggan dan ulasan dari blog. Keputusan menunjukkan bahawa kaedah yang dicadangkan dengan menggunakan maklumat perenggan konteks adalah berkesan untuk klasifikasi sentimen. Kaedah yang dicadangkan bertindak dengan lebih baik dari perkataan dan teks peringkat korpus berdasarkan pelajaran kaedah jentera untuk orientasi semantik. Keputusan menunjukkan bahawa kaedah yang dicadangkan mencapai ketepatan purata 86 % pada peringkat ayat dan 97 % di peringkat maklum balasbagi ulasan pelanggan. Begitu juga, ia mencapai purata ketepatan 83% pada peringkat ayat dan 86% pada peringkat maklum balas bagi ulasan blog.

## TABLE OF CONTENTS

# LIST OF TABLES

LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| WWW | World Wide Web |
| NLP | National Language Processing |
| IR | Informational Retrieval |
| BoW | Bag of Words |
| KDT | Knowledge Discovery in Text |
| BOS | Bag of Sentences |
| FS | Feature Selection |
| FE | Feature Extraction |
| DR | Dimension Reduction |
| WSD | Word Sense Disambiguation |
| HTML | Hyper Text Markup Language |
| AI | Artificial Intelligence |
| WCM | Web Content Mining |
| POS | Part of Speech |
| JJ | Adjective |
| JJS | Adjective |
| NN | Noun |
| NNS | Nouns |
| VB | Verb |

| | |
|---|---|
| VBZ | Verbs |
| RB | Adverb |
| SO | Semantic Orientation |
| PMI | Pointwise Mutual Information |
| NB | Naïve Bayes |
| ME | Maximum Entropy |
| SVM | Support Victor Machine |

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

In recent years, the World Wide Web (WWW) has become an important and emerging source of information. This is because of its rapid growth due to the increasing phenomenon of social network contacts, online discussion forums, blogs, digitals libraries and quick streaming news stories.

In order to acquire the specific knowledge needed to complete a certain task, a method for sorting through the vast amount of data available is extremely important. For extraction, retrieval and analysis of data available on the Web, the use of data mining and linguistics techniques are employed. The distributed environment of the Web gives users access to various locations where a large variety of information is kept. Adequate tools which are able to extract only the most pertinent, and hidden, knowledge from the huge Web content are required when considering the tremendous amount of data storage and manipulation (Falinouss, 2007) . Unstructured text data is the type of information most often found on the Web. Several challenges have arisen as a result of the rapid increase in Web data; included in these are the finding of relevant information, extracting patterns and retrieving pertinent knowledge. Efficient and appropriate handling of this Web content is required by utilizing new or modified tools and algorithms as proposed by Web mining (Yao, Y. Yu, Shou, & Li, 2008) (Sebastiani, 2002).

There are three types of Web mining; namely, Web usage mining, Web structure mining and Web content mining. Web usage mining is a process of extracting useful information from server logs i.e. history. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site i.e.

extracting patterns from hyperlinks in the Web (Castellano, Mastronardi, Aprile, & Tarricone, 2007) (B. Liu & Chen-Chuan-Chang, 2004) (Falinouss, 2007) . The scope of this work is Web content mining, which deals with the detection of valuable information from Web content and text data. To deal with unstructured (text) data various kinds of text representation techniques are used as pre-processing steps for information extraction from the Web contents which is described in chapter- 4. However, text is not the only type of Web content. There are also other varieties such as symbolic, audio video, hyperlinked and meta data. Of all these, this research is mostly focussed around text and hypertext content because the reviews and feedback are available in unstructured text format on the Web (B Pang & L Lee, 2008).

Mining Unstructured data is termed as text mining or knowledge discovery in texts (KDT) (Sebastiani, 2002). The text mining studies are gaining more importance because of the availability of an increasing number of electronic documents from a variety of sources. The resources of unstructured and semi-structured information include the WWW, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blog repositories. Therefore, proper classification and knowledge discovery from these resources are very important. Data Mining, Natural Language Processing (NLP), Information Retrieval(IR) and Machine Learning techniques work together to automatically classify and discover knowledge from the Web text (Srivastava, Desikan, & Kumar, 2002) (Gupta & Lehal, 2009) (Raghavan, Amer-Yahia, & Gravano, 2004).

The main objective of text mining is to give users the ability to extract information from textual resources and enable them to deal with the necessary operations like, classification retrieval, and summarization. Most of the information on the web is in an unstructured text form.  So, it has attracted the attention of research communities from data mining, NLP, Artificial Intelligence (AI) and many others for managing the dynamic nature of unstructured text for useful knowledge extraction. The Web content, in the form of unstructured text, like reviews, blogs comments, views and news are useful in decision making. Knowledge extraction from such online text is very important for planning market strategies and decision making process. Particularly sentiment from online reviews has a great influence on others in decision making. Therefore, it is desirable to create an effective sentiment analysis technique or system to support both the customer and manufacturer in their decision making. Hence, there is a real and urgent need in carrying out web mining for sentiment analysis.

## 1.2 Sentiment Analysis and Classification

Before defining and explaining the process of Sentiment Analysis, the concept of sentiment/opinion and its importance is described that becomes the inspiration for this dissertation.

### 1.2.1 Sentiment or Opinion

Reality is perceived when we flux together diverse approaches of people with different experience, wisdom and knowledge regarding any particular issue or phenomena. What other people think has always been an important piece of information for most of us during the decision-making process. Many people come together with diverse point of view, helping an individual to realize what the right choice is and what is not. Opinion is a private state of a person's thinking about some something of interest. The statement of a person is always based on his/her feelings, thoughts, observations, knowledge and expertise. Opinions are channels for humans to make right decisions. Sentiment or Opinions provide a gateway to individuals to take right steps. It helps people to asses and evaluate the process underwent to take decisive steps. We collect, compare and analyze opinions for making a decision. According to the Napoleone Bonaparte "Public opinion is the thermometer a monarch should constantly consult" (Esuli, 2008).

Sentiments or Opinion have a major impact on our daily life as they are used to present our point of view to others with whom we interact. Sentiments are contained in the opinions of those around us and give us knowledge about how reality is perceived by other people around the globe. A comparison is made of the sentiments collected from other people's opinions and we use this information to aid in our decision making. People are usually interested in what other people think about something before they make an important decision. We often read product reviews before making our decision on what product we should buy. We enjoy expressing our opinions and sharing our advice with each other. Today, it is normal for us to participate in online forums, blogs and newsgroups in order to present our opinion (Balahur et al., 2010). Industries and large organizations are also interested in the opinions of customers in regards to their particular services or products. These reviews or feedbacks help them to improve their services and quality of those products and items which have critical or bad opinions (Hu & B. Liu, 2004). A comparison of the results of market surveys carried out by these industries is to gather the opinions of customers about their products as well as competing products; this comparison enables them, based on the customers' reviews, to form new market strategies. Even politicians or political parties keep track of their status in the public eye by reading the public opinion presented in electronic polls and blogs. Fortunately, opinions can be gleaned from an abundant supply of sources which are readily available, such as newspapers, television and the Internet. Considering how widespread its coverage is, and how accessible and liberal it is, the Internet is potentially the most valuable source. The vast discussion outlets available through online forums, blogs, newsgroups and even specialized sites providing information feed in the millions, from which opinions can be gathered, are all a result of Internet's coverage and accessibility. Obviously, a single person or even a group could not handle such an immense amount of input data in any practical way. They would need automatic processing tools capable of filtering and discriminating the irrelevant information from the relevant. A new discipline, Sentiment Analysis and Opinion Mining, gradually emerged in the fields of Information Retrieval and/or Computational Linguistics as a result of the growing interest of researchers who recognized the practical need for opinion analysis tools (B. Liu, 2010a).

### 1.2.2 Sentiment Analysis

Sentiment analysis is the procedure by which information is extracted from the opinions, appraisals and emotions of people in regards to entities, events and their attributes (B. Liu, 2010a). In decision making, the opinions of others have a significant effect on customers ease in making choices regarding online shopping, choosing events, products, entities, etc. When an important decision needs to be made, consumers usually want to know the opinion, sentiment and emotion of others. With rapidly growing online resources such as online discussion groups, forums and blogs, people are communicating more and more today. As a result, a vast amount of new data in the form of customer reviews, comments and opinions about products, events and entities are being generated. The reviews about any entity, e.g. banks, hotels and airlines as well as online shopping items such as books, digital cameras, mobile phones, notebooks, etc. are useful in decision making; both for the customer and the manufacturer. For instance, a customer travelling abroad, has to make decisions on airline selection, hotels and restaurants, shopping, foreign exchange facilities etc. as per his/her needs. The sentiments gleaned from online reviews have an immense influence on how customer would make these decisions.

Before the Web, collecting reviews containing the opinions and sentiments of consumers was relatively very difficult. Moreover, due to lack of opinionated text no computational studies required. At that time, one would make his/her decision based on the opinions collected from friends, families and the people in the surrounding community. On the other hand, an organization would have to conduct surveys in order to gather vital information about their events, services or products from relevant groups of people. Only then they were able to make their necessary decisions. The world is totally different today and with the rapid growth of the social media and its content on the internet over the past few years, decision making has also changed. The Web is not only the easiest way for consumers to give their opinions regarding anything and everything but also the best way for the various industries to collect data related to these opinions. If one wants to buy a product, travel abroad or stay at hotel for instance, one is not limited to only seeking the advice of one's friends and family's, as there are abundant of user reviews available on the Web. As for a company, it is no longer required to conduct manual surveys, with focus groups in order to collect and review the opinions of consumers regarding its products and/or its competitors' products; this is because of the overwhelming abundance of such publicly available information on the Internet (B. Liu, 2010b) (B Pang & L Lee, 2008).

Sentiment analysis allows for a better understanding of customers' feelings regarding various companies, their products and services or the way they handle customer services, as well as the behaviour of their individual agents. It can be used to help in customer relationship management, employees training, identifying and resolving difficult problems as they appear. Therefore, sentiment analysis technique is desirable for developing efficient and effective analysing and classifying of customer reviews and blog comments into positive, negative or neutral opinions. Several researchers have been working on sentiment analysis using a domain dependent framework for feature and feedback level opinion classification. A few are using machine learning techniques for classification at document level (B Pang, L Lee, & Vaithyanathan, 2002) (Hu & B. Liu, 2004) (Balahur et al., 2010) (Andreevskaia & Bergler, 2008) (Nathan, 2009) (B. Liu, 2010b) (Ohana, 2009) (Ding et al., 2009) (B Pang & L Lee, 2008). In this work, a domain independent rule based method is proposed for semantically classifying sentiments from online customer reviews and comments. The method is quite effective, as it takes reviews, checks individual sentences and decides its semantic orientation considering the sentence structure and contextual dependency of each word.

The main purpose of sentiment analysis is the extraction of human perception from users' generated text. This is done by applying concepts obtained from IR, NLP and data mining. The main issue is how to automate the extraction of opinionated, sentimental and emotional expressions from unstructured text, select proper feature and their semantic orientation and finally analyse and summarise the sentiments, as described below. The process of sentiment analysis is shown in Fig. 1.1.

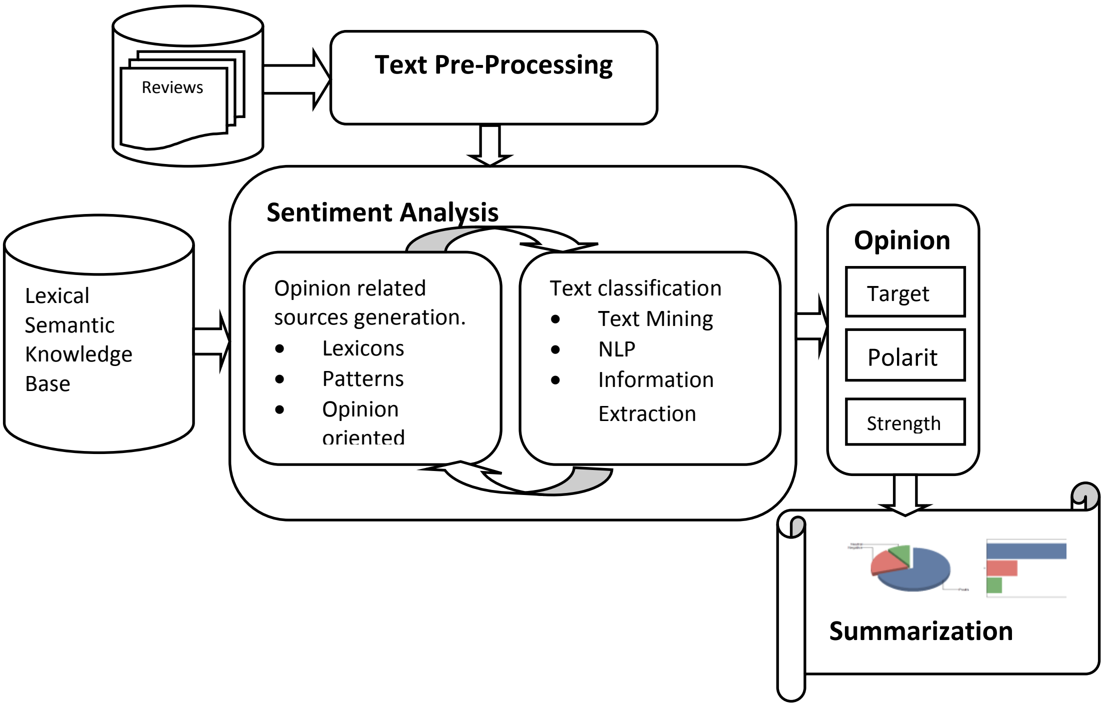


Figure 1.1 The Basic Model of Sentiment Analysis (Gurevych & Oprak, 2010)

### 1.3 Research Trends and Challenges in Text Sentiment Analysis

Exponential growth in the size of the Web has posed several challenges (Hoffman, 2008). One of the biggest challenges is that semantic orientation and classification of the web content is either only semi-structured or not structured at all. This becomes an inherited problem when systems like e-marketing, e-business, e-shopping, e-banking, etc. want to utilize this huge amount of information efficiently. Subsequently, this prevents the development of quality services for users and makes it difficult to provide them with the intended information product (F. Zhu & X. Zhang, 2010).

Researchers have been taking a keen interest in sentiment analysis for the past few years; this is particularly true in regards to the more recent explosion of blogs and other Web 2.0 services. It has attracted a great deal of interest because of the challenging research problems and the wide range of applications for both academia and industry. It needs a computational study for extracting useful knowledge from the peoples' opinions and emotions. In today's global world market with a steep growth in internet usage, a preference for online shopping, banking, ticket reservations, hotel bookings, etc. is rising. Therefore, sentiment analysis from online customer reviews is becoming a major requirement for customers as well as organizations, for effective decision making. During the decision making process, most people depend on the views and emotions of others. It is a natural phenomenon that good decisions can be made on the basis of others' sentiments (Balahur et al., 2010) (Andreevskaia & Bergler, 2008) (Nathan, 2009) (F. Zhu & X. Zhang, 2010) .

The extraction of information and discovery of useful knowledge from the users' views in the form of unstructured text is an essential as well as a challenging area of research. This is because the information stored in the Web is very dynamic in nature. Over 45,000 new blogs are created on a daily basis along with 1.2 million new posts each day. Moreover, 40% of the people in today's society rely on opinions, reviews and recommendations which are gathered from blogs, forums and other related resources. This data is rapidly changing due to round-the-clock updating of information on the Web. For instance, a survey has been done on more than 2000 Americans and the following results were concluded (B Pang & L Lee, 2008) (Andrew Lipsman, 2007):

- 81% of Internet users have, at least once, done online searches about a particular product.
- Between 43% and 84% of internet users who read online reviews of hotels, restaurants and various services like, medical services or travel agencies, report that their purchase choices are significantly influenced by the online comments and reviews.
- 32% using an online ratings system, have rated a product, or person service, and 30% (18% of online senior citizens included) have posted an online review or a comment about a particular service or product (F. Zhu & X. Zhang, 2010) (Hoffman, 2008) (Andrew Lipsman, 2007).

With the vigorous growth and development of internet and its usage, sentiment analysis for online reviews, opinions and comments has become need of the hour. Quite a number of researchers have been working on various aspects of this area to address the current problems (Esuli, 2008) (B. Liu, 2010a) (B Pang, L Lee, & Vaithyanathan, 2002) (Hu & B. Liu, 2004) (Balahur et al., 2010) (Andreevskaia &

Bergler, 2008) (Nathan, 2009) (B. Liu, 2010b) (Ohana, 2009) (Ding et al., 2009). The following general challenges are pointed out in this area.

- Multi-lingual, multi-source sentiment analysis: Online discussion forums and blogs contain multi-lingual texts from multiple sources like Twitter, Facebook and other Web 2.0 media. These are bursting with opinions regarding various issues and topics. They are often capacious, puzzling, multi-lingual and deeply interconnected. These are the new critical sources for marketing planning. There are new varieties of challenges for the researchers in this domain. They necessitate solutions for a more sophisticated data retrieval system, advanced multi-lingual analysis and a suitable infrastructure for managing terabytes of data daily.
- Comments or views (in the form of unstructured text) on the Web in the social networks are mostly noisy. Open forums and blogs are most often created by non-professional writers; therefore, the reviews provided are not in a proper text form. There is no standard rule on how to write comments on the social network forums and blogs. Opinion sources are typically informally written and highly diverse, e.g. Twitter - abbreviations, lack of capitals, poor spellings, poor punctuation, poor grammar etc. Removing the noise and extracting semantics from the symbols and specials characters which are often used, is a challenge for the semantic Web to extract knowledge from the users' comments and views.
- Domain Dependent: Normally opinions are about specific issues, problems or topics. Therefore, the methods are normally domain dependent. However, this leads to the problem of non-generalization.
- Effects of syntax on semantics: Breaking multi-word expressions, mapping of synonyms into different elements, words with multiple meanings used as one single component (polysemous), sentence document complexity, contextual sentiments, heterogeneous documents, reference resolution, and modal operators: might, could and should continue to pose challenging problems in this area .
- Effect of sense on terms, finding subjective terms, and multi-word document analysis.
- The use of ontology for sentiment classification and informational retrieval.

- Mining trends, i.e. marketing, financial (stock exchange trends) and business trends, and from e-documents (Online views, news, stories and events).
- New information management techniques and methods are required for stream texts.
- In order to recover senses from the words used in a specific context, the utilization of a sense-based text classification procedure is necessary.

## 1.4 Applications of Sentiment Analysis

Opinion has equal importance in every field. It can be used in business organization, social work organization, politics, health and education for decision making. Business organizations spend a huge amount of money to find consumer sentiments and opinions through consultants and surveys. Similarly individuals are interested in other’s opinions about products, services, topics and event for finding best choices. Humans intentionally or naturally have the instinct of observation and analysis after that he/she comes up to opine helping in individuals to decide accurately. If the opinion is positive it helps to reform or improve our day to day transactions (Esuli, 2008) (B. Liu, 2010a) . Essential for online services, that exist today, is the abundance of applications for sentiment analysis. Mining customer reviews or gathering feedback from opinions about a given product, event or object (Airline, Hotel. digital camera, car, mobile phone, etc.) can give companies valuable information as to the satisfaction or dissatisfaction of their customers. This information is also immensely valuable for customers in their decisions to purchase a particular product. Furthermore, sentiment analysis enables trend watchers, marketing research teams and recommendation systems to track emotions or opinions over time; tracking of online trends provides interesting as well as valuable data for these groups. In the case of moderating, opinion mining has also proven to be very useful whereby the ability to react quickly and efficiently to messages which have been posted on forums or discussion boards wherein dissatisfied consumers discuss product deficiencies (F. Zhu & X. Zhang, 2010) (B. Liu, 2010a) (B Pang & L Lee, 2008).

- Sentiment Analysis in Products: The opinion of consumers regarding a product might be needed by a company. This information is valuable in helping a company to improve their products. The data they receive could help to identify new marketing strategies, market intelligence, and product and service benchmarking.
- Sentiment Analysis in Company Stocks: Investors are able to identify the humor of the market towards a company’s stocks based on the opinion of analysts who utilize this important data, whereby the trends in their prices are identified.
- Sentiment Analysis on Places: Opinion mining is very helpful during the planning of a travel itinerary. For example, a person who wants to travel might be interested in knowing about places to visit, best and cheap airline flights or restaurants to dine at. Opinion mining would help him here by recommending suitable places and facilities.
- Sentiment Analysis on Elections and Administration Activities: The opinion of voters regarding a particular candidate could be useful to other voters and utilizing sentiment analysis would help them in making their decision. The public opinion is very useful in administration and government sector like get public opinion about government machinery, government intelligence, political issues and events and government regulations proposal and policy making.
- Analysis on Games and Movies: Mining of sentiments regarding the latest games and movies plays an important role for their marketing strategy.
- Sentiment Analysis in Education: Sentiment analysis is very useful and helpful in education sector for ranking of the universities and institutes, as well as, in student and teacher performance evaluation and quality enhancement and assurance.

## 1.5 Motivations, Thesis Objective and Contributions

### 1.5.1 Motivation

The growth of Web information has increased exponentially. It has become a challenge to manage the huge quantity of online information efficiently without having a real implementation of the original semantic-text-analysis-model. As the information objects are not annotated and do not contain meta information, it is very difficult to find context specific information related to user requirements. Although the intended information exists, the users battle the problem of finding context specific information. Sentiment analysis is a burgeoning area of research, one where a cross-disciplinary study could very possibly result in both a theoretical as well as practical gain. It has attracted a great deal of attention because of its challenging research problems and wide range of applications for both academia and industry. Therefore, sentiment analysis from online customer reviews is becoming a major prerequisite for organizations in making their assessments towards the improvement of their products and services as per customers' requirements. It needs a computational study for extracting knowledge from the people's opinions, appraisals and emotions toward entities, events and their attributes. Some of the mutilative perspectives of sentiment analysis are as below.

- People, who want to buy a product, may need comments and reviews of others who have already used that product; they look for comments and reviews of others for their decision making.
- People who have just bought a product can comment on it and can write about the experience they had.
- Manufacturers can get feedback from customers and improve their product and service quality, and can also adjust marketing strategies.
- A Company or Agent/Actor/Individual/Team can get feedbacks from clients or the public for their particular manifest and can plan/scheme their strategies towards exploring new opportunities.

### 1.5.2 Objectives

The main objective of this work is to develop a sentence-level lexical based method for sentiment orientation using general lexical dictionaries for the effective classification and to address domain portability problem. The objectives are defined as follow:

- To develop sentence-level lexical based method for domain independent sentiment classification.
- To remove noise from reviews/comments, identify and extract semantics of short notations and symbols for classification.
- To develop knowledge base from lexical dictionaries, intensifiers, phonetics features and opinion terms.
- To identify opinion sentences and extract sense from opinion expressions considering the sentence structure and contextual information.
- To determine opinion orientation for the polarity classification (positive, negative or natural) of each recognized sentence.

### 1.5.3 Scope of the Thesis

As discussed in section 1.2.2, the thesis presents a technique for sentiment analysis, which is used to create a knowledge base and semantic orientation of opinionated terms at the sentence level. The WWW is most probably the largest digital archive enabling a wide range of different communities to make available large sets of diverse resources and information. This information is further utilized for knowledge discovery which is used in effective planning, decision making, marketing strategies, etc. a method of sentiment classification is proposed at the sentence level by applying rules for all parts of speech to score their semantic strength, contextual valence shifter and expression or sentence structure based on dynamic pattern matching as well as addressing word sense disambiguation. The system identifies opinion type, strength, confidence level and reasons. It deals with the SentiWordNet and WordNet, as the knowledge base, with the additional capability of strengthening the knowledge base with modifiers, contextual valence shifter information and usage of all parts of speech.

### 1.5.4 Research Questions

This section describes research questions which are addressed in this thesis. Wherever applicable, these questions are broken down into more specific ones. The first two research questions are addressed in chapter 4 while the remaining questions are discussed in chapter 3. The process of sentiment analysis about data acquisition, pre-processing and post-processing (summarization), is discussed throughout the thesis. Based on the aforementioned issues, this work is cantered on the following questions.

1. How to remove noise from reviews/comments and to identify and extract semantics of short notations and symbols for semantic orientation?
2. How to develop knowledge base using lexical dictionaries, intensifiers, phonetics features and opinion terms?
3. How to identify opinion sentences and how to extract sense from opinion expressions considering the sentence structure and contextual information?
4. How to determine opinion orientation for the domain independent polarity classification (positive, negative or natural) of each recognized sentence, review or comment?

### 1.5.5 Thesis Contributions

The main contributions of this thesis are as follows:

- Development of sentence-level lexical based method for domain independent sentiment classification.
- Removal of noise from reviews/comments, identification and semantics extraction of short notations and symbols for classification.
- Development of knowledge base from lexical dictionaries, intensifiers, phonetics features and opinion terms.

- Identification of opinion sentences and extraction of sense from opinion expressions considering the sentence structure and contextual information.
- Determination of opinion orientation for the polarity classification (positive, negative or natural) of each recognized sentence, review or comment.

### 1.5.6 Thesis Organization

This chapter serves as an introduction to the thesis and explains the importance of sentiment analysis, research challenges and contributions in this field. The remaining parts of the thesis are categorized as follow. Chapter 2 describes state-of the-art work in the area of semantic analysis by highlighting existing methods, techniques and developed systems/prototypes. Chapter 3 elaborates the proposed framework, methodology and the prototype. Chapter 4 describes the datasets used in this work, their pre-processing and noise removal steps. In Chapter 5, discussion on simulation results, visualization and comparison with other works are presented. The thesis ends with the conclusions and recommendations for future work as highlighted in Chapter6.

# CHAPTER 2

# BACKGROUND AND RELATED RESEARCH

## 2.1 Introduction

This chapter includes the detailed overview of the literature and related research on web mining and sentiment analysis which has been reviewed and consulted during the course of this study. This literature facilitated in understanding and framing ideas in the development of "Domain Independent Lexical Based Method for Sentiment Classification".

## 2.2 Web Mining

Extraction and mining of useful knowledge from the Web data is the main goal of Web mining. The single largest source of data available today is on the Web, as it is well known to any internet user. A better understanding of customer behaviour can be obtained because of this data. Moreover, it is quite useful in evaluating a website's effectiveness or quantifying a marketing campaign's success. The WWW has become a vital source of searching the desired information using automated tools. Therefore, in order to extract the necessary knowledge from these online resources, effective data mining techniques are quite essential (B. Liu, 2010a) (B Pang & L Lee, 2008). Extraction of information from web documents, Web Mining, is done using data mining techniques. Rapid growth of information resources, interest of various communities and steady development of e-commerce have made this area so huge that data mining is one of the hottest issues in information technology research (Etzioni, 1996). The desired information is distributed in a heterogeneous multi-organizational is based on flexible architecture and is implemented by few steps able to examine web content and to extract useful hidden information through mining techniques

(Yao, Y. Yu, Shou, & Li, 2008). Owing to its rapid growth and continuous introduction of modern advanced features as well as its enormous attraction by the public, new challenges have arisen in web mining that must be dealt with. In this regard, researchers in other fields and disciplines, such as IR and NLP have become interested in carrying out related research activities.

Internet availability and popularity in mid 1990s led to the beginning of research into web mining. Web mining, which is used with structured, semi-structured and unstructured data, is the process whereby knowledge is extracted from various types of contents found on the Web (Srivastava et al., 2002) (Falinouss, 2007) . The scope of this thesis is web content mining of unstructured text of online blogs, reviews and comments (Figure 2.1).

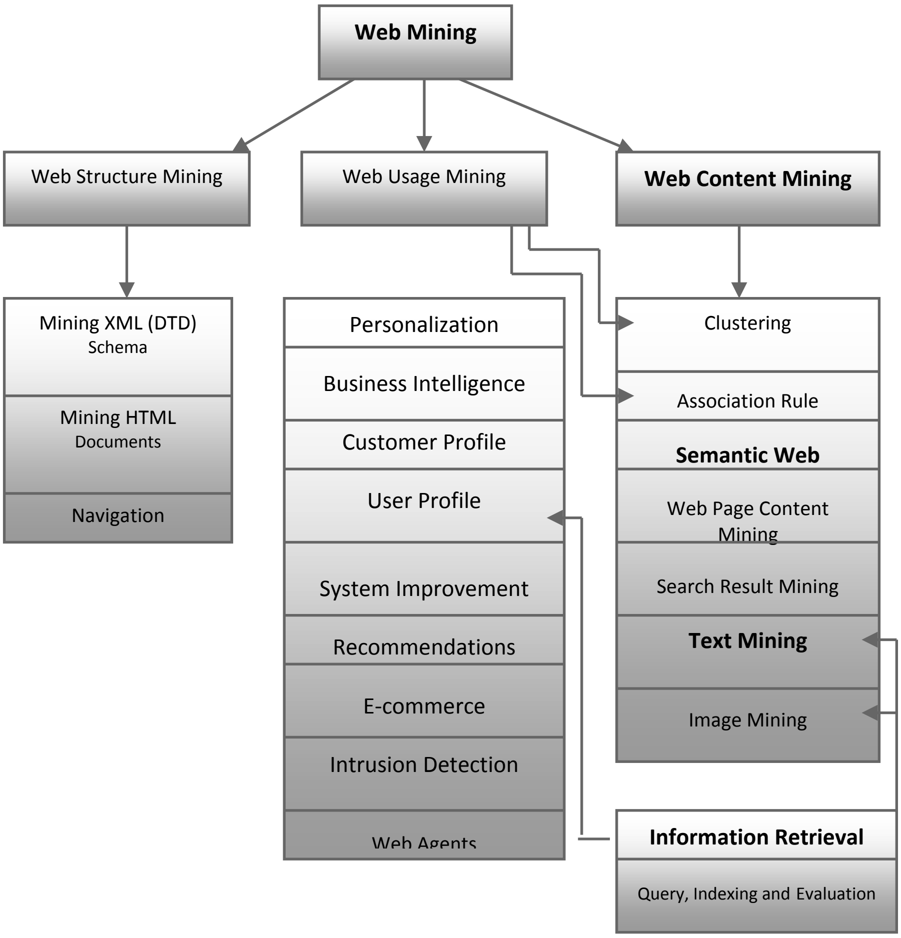


Figure 2.1 Web Mining Taxonomy

### 2.2.1 Web Content Mining

Discovery of important information from the Web is the main area dealt within Web Content Mining (WCM). However, text is not the only type of web content on the internet rather a very wide range of data such as audio visual data, symbolic data as well as hyperlinked data and meta data are also found. Text and hypertext contents are the main focus of our research.

In recent years, an increased amount of attention has been given to Web content mining by research communities. Web content mining is very important for useful information extraction, as evidenced by the research works described below.

A page model for web content mining is described in (Di, Yao, Duan, J. Zhu, & Li, 2008), which gives importance to the named entities in the pages. Examples of these entities are person, location and time. If the proposed technique is utilized to find certain entities or the relationship between certain entities, classify and label the relationship of pairs or calculate weights, it can easily be seen that location and time can be attributed to a person's activities.

Concept based method of knowledge discovery process from Web text content is presented in (Loh, Wives, & de Oliveira, 2000). This method work to extracts concepts rather than analysing words or attribute values, It has been suggested by some approaches that restructuring of the document content into a machine readable format would be more useful as a text document rather than the content that has no machine readable semantics. (Hu & B. Liu, 2004) aimed to introduce a feature-based summary containing numerous customer reviews focusing on products for online sale. In order to accomplish this, a set of approaches was proposed in which product reviews could be mined and summarized utilizing techniques used for data mining and NLP.

This is true for the management of news groups, maintenance of web directories and even emails. Mining from services is also a growing area in the field of information extraction. This is greatly in part due to the staggering number of online services like Usenet, digital libraries, news groups, customer comments and reviews and mailing lists which are popping up all over the Web. A survey has been carried out to demonstrate that Web content mining plays an important part as an efficient tool for the extraction of semi-structured and structured data in order to discover useful knowledge (Srivastava et al., 2002) (Pol, Patil, Patankar, & Das, 2008).

In (Esuli & Sebastiani, 2005) a technique was founded on semi-supervised learning. The method is used to gather term representations which were obtained by utilizing term glossaries from machine-readable dictionaries. That was used to introduce the orientation of subjective terms determination. First, subjectivity analysis is considered as a binary classification in this method; then by applying part-of-speech (POS) information about the terms, opinions are identified.

In (Jindal & B. Liu, 2006), the issue of finding comparative sentences in evaluative texts is described as well as the method used for extracting their comparative relations. Many applications can make use of these types of sentences, e.g., sentiment analysis, e-commerce and marketing intelligence as well as product benchmarking (Jindal & B. Liu, 2006) (B. Liu & Chen-Chuan-Chang, 2004).

While a lot of research work has been done in this area, there still remains a variety of issues needing to be solved. While it is true that an extraordinary opportunity is offered by the Web in the area of web mining, yet it also presents it with difficult challenges to overcome. Some attributes related to Web are; tremendous amount of easily accessible data/information in the form of texts, structured tables, multimedia etc and the probability of finding almost anything because of the Web's wide and diverse coverage of  information. There exist service providing web sites offering a variety of products and service to their users/clients. Since the Web is quite dynamic, yet, keeping up with and monitoring the constantly changing Information on the Web are serious issues to be handled. Above all, Web is a virtual society of communities providing a platform for interactions among people, organizations and automatic systems and not just for data, information and services that it provides.

### 2.2.2 Text Mining

Today Web is the main source of any kind of information. The amount of textual data available to us is consistently increasing, and approximately 80% of the information of an organization is stored in unstructured textual format i.e. in the form of reports, email, views and news etc. Furthermore, of the entire world’s data, approximately 90% is stored in unstructured formats (“www.Oracle.com,” 2008). As a consequence of so much unstructured data, businesses which have information intensive processes require that we surpass the use of simple document retrieval and move forward to the use of knowledge discovery. The necessity for automatic retrieval of pertinent knowledge from the tremendous amount of textual data available in order to aid analyses carried out by humans is in no way unapparent (Raghavan et al., 2004)

The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. The resources of unstructured and semi structured information include the word wide web, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blog repositories. Therefore, proper classification and knowledge discovery from these resources is an important area for research. Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the electronic documents(Sharp, 2001) (Falinouss, 2007).

KDT (knowledge discovery in texts) or text data mining or text mining are terms used for the mining of unstructured or semi-structured data. It is a new sub-discipline of data mining that considers textual data. The fact is that, "text data mining" is an intermediate evolutionary lexical form (M.A. Hearst, 1999).

The majority of the online information about data mining is misleading. Such ambiguous/misleading information implies to the mining metaphor that it is like extracting precious unknown’s hidden patterns/information from the huge data. Data mining has not only directed dealings with the information, but it also attempts to uncover or glean previously unknown, information from the data (text). Three main steps are always involved in the process of text mining; they are (a) acquiring texts which are relevant to the area of concern usually called IR; (b) presenting contents collected from these texts in a format that can be processed, such as statistical modelling, natural language processing, etc.; and (c) actually using the information in the presented format, (Sharp, 2001) (Falinouss, 2007) as shown in Figure 2.2.

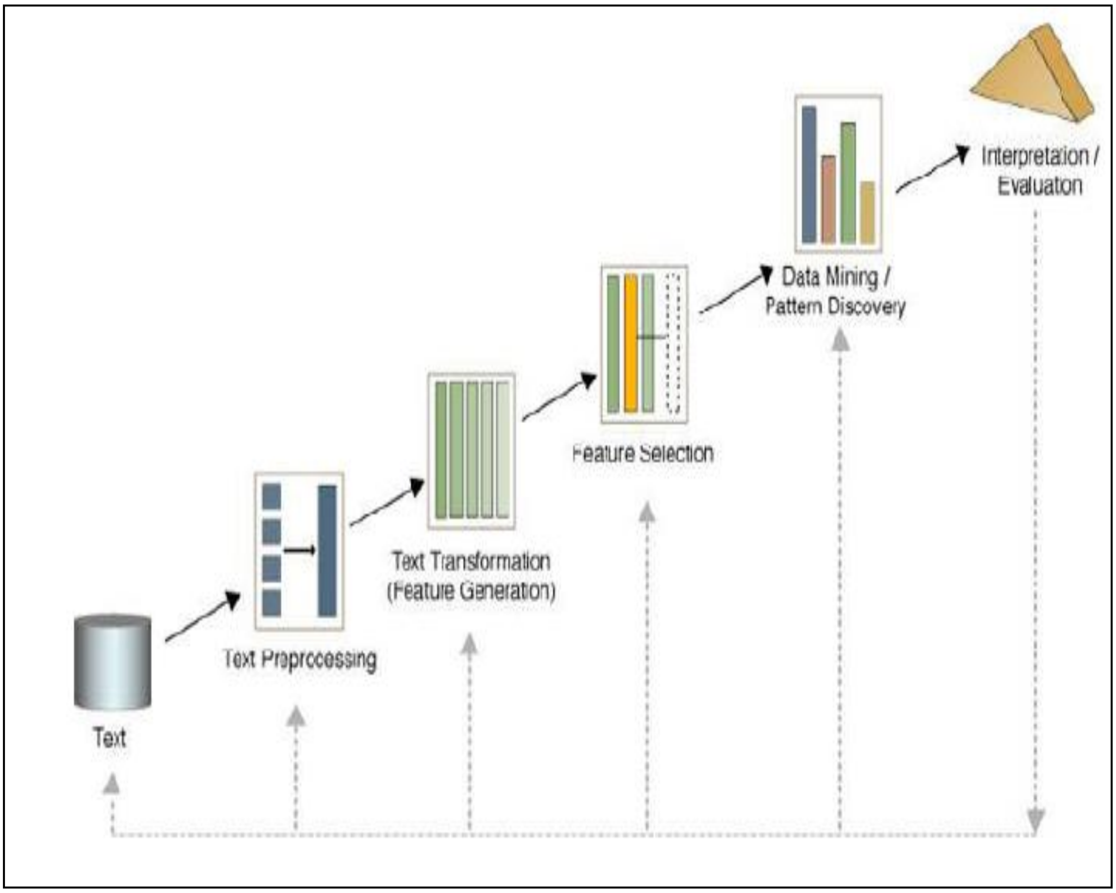

Text
Text Preprocessing
Text Transformation (Feature Generation)
Feature Selection
Data Mining / Pattern Discovery
Interpretation / Evaluation

Figure 2.2  KDT Process (Falinouss, 2007)

Text mining provides a user with the ability to extract the necessary information from textual resources as well as deal with various procedures like, retrieving, classifying (supervised, unsupervised and semi- supervised) and summarizing. However, how to properly classify, annotate and present these documents is an issue (Sebastiani, 2002). Therefore, several challenges exist, such as suitable representation, correct annotation, dimensionality reduction to handle algorithmic issues, and an appropriate classifier function to obtain good generalization and avoid over fitting. Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents are important for the research communities (Hotho, N urnberger, & Paaß, 2005) (Sebastiani, 2002).

Market trends based on the content of the online news articles, sentiments, and events is an emerging area of research in data mining and text mining community (Falinouss, 2007). Related to this, state-of-the-art applications for classifying texts are given in (Sebastiani, 2002) in which three issues are discussed: representation of the text, construction of a classifier and the evaluation of that classifier. As a result, the main points in the classification of texts are as follows; the construction of a data structure which represents the text and construction of a classifier based on that structure in order to establish a very accurate class label for the document.

Syntactic and semantic matters of text, concern for tokenization, domain ontology as well as a variety of NLP and machine learning techniques used for classifying, extracting and retrieving text information are all vital for successful text processing.

The goal of Information Extraction (IE) method is to locate specific information in text documents and extract that information. It is assumed in this first approach, that text mining has a direct relation to the extraction of information. IR, on the other hand, focuses on finding answers contained in the text to specific questions. This goal is accomplished by utilizing statistical measures and other relevant methods so that text data can be automatically processed and compared to the question needing to be answered. In a much broader sense, information retrieval plays a vital part in information processing in its entirety, starting with the retrieval of data and ending with the retrieval of knowledge (Hotho et al., 2005).

Obtaining a better understanding of natural language text is the goal of NLP which uses computers in order to present text semantically, thereby improving the process of information retrieval and classification. To this end, sentences and paragraphs are linguistically parsed into key ideas, nouns, adjectives and verbs, in the process known as semantic analysis.  These words are then compared to the taxonomy by utilizing statistics-backed technology.  Ontology is the explicit and abstract model representation of already defined finite sets of terms and concepts, involved in knowledge management, knowledge engineering and intelligent information integration (Fensel, 2004a).

Text representation must be in a specific format before that text can be classified and the information extracted; it must be represented according to the specific classifier or algorithmic requirements.  This process of text formatting is known as pre-processing which is described in detail in chapter 4.

## 2.3 Sentiment Analysis

Sentiment analysis, a sub category of text mining, is involved with online customer reviews, blog comments, and other related online social network contents (views and news). People recognizing the usefulness, in the immense expansion of the Web, are being drawn more and more towards online services like, shopping, e-banking, e-commerce etc. as well as to the feedback given in the form of reviews and comments about various products and services (B Pang & L Lee, 2008). Online reviews and comments added on a daily basis to various online sites, like epinion.com, cnet.com, amazon.com, facebook.com, and twitter.com are quite helpful for consumers in making decisions and for companies planning market strategies (F. Zhu & X. Zhang, 2010). This has attracted a lot of attention of research communities from industries as well as academia. Consequently, the steady flow of interest towards online resources in recent times has resulted in a tremendous amount of research activity in the field of sentiment analysis and opinion mining (B. Liu, 2010a). This has led to the appearance of Web 2.0 which, combined with the vast social media content has caused quite a bit of excitement as it provide ample opportunities to get a better understanding of what the general public, especially consumers, think about company strategies, product preferences and marketing campaigns as well as social events and political movements. Analysis of the thousands, possibly millions, of reviews, comments and other feedback expressed in various forums (Yahoo Forums), blogs (blogosphere), social network and social media sites (including You Tube, Flikr, and Facebook, Twitter etc) and virtual worlds (like Second Life) can potentially answer the numerous new and interesting research questions regarding social, economical, cultural, geo political and business issues (H. Chen & Zimbra, 2010) (B. Liu, 2010a).

Sentiment analysis is often used in opinion mining (a sub-discipline within data mining) to identify subjectivity, sentiments, affects and other states of emotions within the text found in the above mentioned online resources. Opinion mining is in reference to computational techniques utilized to extract, assess, understand and classify the numerous opinions that are expressed in a variety of online social media comments, news sources and other content created by the user (H. Chen & Zimbra, 2010) .

The early work of sentiment analysis began with subjectivity detection, dating back to the late 1990’s. Later, it shifted its focus towards the interpretation of metaphors, point of views, narrations, affects, evidentiality in text and other related areas. Shown below is the literature describing the early works of subjectivity and detection of affects in the text. With the increase in internet usage, the Web became a source of importance as text repositories. Consequently, a switch was slowly made away from the use of subjectivity analysis and towards the use of sentiment analysis of the Web content. Sentiment analysis has now become the dominant approach used for extracting sentiment and appraisals from online sources. So the early work in subjective and objective analysis and classification the separating of non opinionated, neutral and objective sentences and texts from subjective sentences carrying heavy sentiments is a very difficult job; however, it has been explored earnestly in a closely related yet separate field, (Turney, 2002) (J. M. Wiebe, 1994). It concentrates on making a distinction between 'subjective' and 'objective' words and texts; on one hand, the subjective ones give evaluations and opinions and on the other, the objective ones are used to present information which is factual .This is different than sentiment analysis in regards to the set of categories into which language units are classified by each of these two analyses. Subjectivity analysis focuses on dividing language units into two categories: objective and subjective, whereas sentiment analysis attempts to divide the language units into three categories; negative, positive and neutral. The area of concentration in some of the early works was with subjectivity detection only, which was used for the classification and extraction of subjective terms from the unstructured text using NLP techniques. With the passage of time and a need for better understanding and extraction, momentum slowly increased towards sentiment classification and semantic orientation (J. M. Wiebe, 1990) (J. Wiebe, Wilson, R. Bruce, Bell, & Martin, 2004) (J. Wiebe & Riloff, 2005) (B Pang & L Lee, 2008). Some descriptions of earlier works are as follows

In (J. M. Wiebe, 1990), an algorithm was presented which was able to identify subjective characters in regards to normal patterns. It was able to recognize the way a character's point of view begins, continues and resumes in a text. Rules were given on how to distinguish between the two interpretations of private-state sentences; how the subjective elements appear in the sentence and how the textual situation is. After the characters have been identified, a decision is made as to when the subjective character should be chosen from among them. Thorough examinations of natural narratives were the basis for the results of their experiment.

In (Hearst, 1992), an approach was created for forcing the meanings of sentences into a metaphoric model; this model was the basis for which only semantic interpretation was necessary for determining the sentence direction. The design of this approach to interpreting a sentence enables it to be a component that is easily integrated into a hybrid information access system.

(J. M. Wiebe, 1994) presented an approach which could search for normal patterns in the method they used for point of view manipulation. Their search is accomplished through thorough examinations of naturally occurring narratives. An algorithm was developed which was able to track point of view according to the normal patterns that were found in the text.

In (Sack, 1995), a more realistic story understanding was determined by encoding a way to recognize the *point of view* in the story. The encoding technique was helpful when performing a task for information retrieval which required searching for stories that were credible.

A probabilistic classifier model was introduced by (J. M. Wiebe & R. F. Bruce, 2001) which were utilized to solve issues related to discourse segmentation where segmentation, belief and reference resolutions were all addressed. Subjective sentences, present in a segment or block of text were identified using this technique.

In (J. M. Wiebe, R. F. Bruce, & O'Hara, 1999) a case study was introduced by analysing and improving intercoder reliability in discourse tagging using statistical techniques. Bias corrected tags were formulated and successfully used to guide a revision of the coding manual and develop an automatic classifier. Their focus was on sentences; about private states, such as belief, knowledge, emotions, etc.; and sentences about speech events, such as speaking and writing. Such sentences may be either subjective or objective. From the coding manual: Subjective speech-event (and private-state) sentences are used to communicate the speaker's evaluations, opinions, emotions, and speculations.

A combination of fuzzy logic and NLP approaches were used to form a technique for analysis and management of documents as introduced by (Huettner & Subasic, 2000). This particular technique used semantic typing from NLP and was referred to as fuzzy typing for document management.

Like other developing fields of research today, sentiment analysis terminology is yet to be matured; moreover, just attempting to define a sentiment can be difficult to accomplish. The words sentiment polarity, opinion semantic orientation and valence are used to represent similar if not the same ideas (B Pang & L Lee, 2008). These words are, more often than not, used either to make reference to various aspects of one particular phenomenon, an example being (Hariharan, Srimathi, Sivasubramanian, & Pavithra, 2010) where sentiment is defined as an affective part of opinion, or simply used as synonyms for each other without any true definition of their own. Furthermore, some of these words can be confusing because of their multiple meanings already in linguistic tradition (e.g. polarity, valence) and therefore are confusing (B Pang & L Lee, 2008). This work focus is on capturing expressed sentiment in a text as negative, positive or neutral; therefore, we will refer this domain of research as sentiment analysis.

Early in this century, sentiment analysis became an important subfield for text mining and information management. Moreover, from 2003 to date, sentiment analysis has been recognized as a vital area of research as the term opinion mining appeared in 2003 for the first time in a paper by Dave et al (Dave, Lawrence, & Pennock, 2003), which attracted the attention of various research communities. This interest is due in part to the sudden explosion in the amount of online discussion forums, reviews, blogs and e-commerce etc. as well as the vast range of applications available to academia and industry. Furthermore, data mining and computational linguistics have resulted in challenging research problems which are needed to be solved. Currently, text mining and information retrieval are at the heart of NLP in the area of sentiment analysis.

Classifying documents according to sentiment on the whole rather than just by topic, e.g., a specific review of positive or negative; was presented by (B Pang, L Lee, & Vaithyanathan, 2002). They showed that human-produced baselines are overwhelmingly outperformed by standard machine learning methods as was seen when using information gathered from movie reviews. They utilized Maximum Entropy (ME) Classification, Naive Bayes (NB) and Support Vector Machines (SVM) as their machine learning methods. However, the usual topic-based categorization received better results using these methods than sentiment classification.

A basic unsupervised three step learning algorithm created for rating reviews as either thumbs up or thumbs down was presented by (Turney, 2002). The steps involved are: (1) phrases which possess adverbs or adjectives are removed, (2) the semantic orientation of every phrase is estimated; this is the base of the algorithm and the estimation is calculated with Pointwise Mutual Information-I (PMI-I) (B Pang, L Lee, & Vaithyanathan, 2002), and the final step, (3) the review is classified on the basis of the average phrase semantic orientation.

After introducing a plan for low-level annotation which can represent localized and individual expressions of opinions in the form of opinion-based "template relations", (Cardie, J. Wiebe, Wilson, & Litman, 2003) suggested a method for extracting information from naturally occurring text in order to find and organize opinions as a way to achieve multi-purpose question responses.

A method used to analyse and integrate product reviews was specified by (Dave et al., 2003); it could automate the type of work that aggregation sites or clipping services carry out. To start, they began with reviews for training and testing which were structured. Next, they identified the suitable features, and then they rated approaches for extracting information which could determine the positive or negative state of reviews. The achieved results showed a performance on par with the machine learning techniques traditionally used. Once suitable results have been achieved, identification and classification of the review sentences on the Web are carried out with appropriate classifier, although classification is more difficult here. Moreover, use of a simple technique for identifying the pertinent characteristics of a product achieves a generally useful summary.

There are six emotion categories; they are sad, happy, fearful, angry, surprised and disgusted. In (H. Liu, Lieberman, & Selker, 2003) introduced a method whereby the affect of a text could be classified into these categories. By leveraging a real-world knowledge base called Open Mind with 400,000 pieces of knowledge, they evaluated the affective nature of the underlying semantics of sentences in a robust way. Other methods have been attempted to classify textual affect but have been found to contain limitations; many of these limitations were addressed with their method.

In (Nasukawa & Yi, 2003) a method was described for sentiment analysis where only sentiments in relation to the polarities of positive or negative were extracted for specific areas in a document, rather than categorizing an entire document into either negative or positive.

(Hu & B. Liu, 2004) aimed to introduce a feature-based summary containing numerous customer reviews focusing on products for online sale. In order to accomplish this, a set of approaches was proposed in which product reviews could be mined and summarized utilizing techniques used for data mining and NLP.

In (Esuli & Sebastiani, 2005) a technique was founded on semi-supervised learning. The method is used to gather term representations which were obtained by utilizing term glossaries from machine-readable dictionaries. That was used to introduce the orientation of subjective terms determination. First, subjectivity analysis is considered as a binary classification in this method; then by applying part-of-speech (POS) information about the terms, opinions are identified.

(Choi, Breck, & Cardie, 2006) Utilized a global inference method whereby entities were directly involved in the opinion expressions, used for extraction of the both, opinion holder and expression. This was done in order to investigate the effects of attempting to jointly extract the opinion holders as well as the opinion expressions.

Recently, there has been a change of attitude in the field of sentiment analysis whereby the concentration is now on classification, which has added a third category known as neutrals. Therefore, it is no longer focused on the binary classification of only positive/negative (B Pang & L Lee, 2008) (Turney, 2002). Through empirical observations, there came a realization that it is much easier to separate positive elements from negative ones than it is to differentiate positives or negatives from neutrals. Majority of disagreements amongst human annotators as well as the errors resulting from utilizing automatic systems are associated with attempting to separate neutral words, sentences or texts from those that are either negative or positive (B Pang & L Lee, 2008). In this dissertation the review or comments are called neutral which have equal weight of positive and negative opinion.

Sentiment analysis is given a variety of names in related literature such as sentiment classification, effect analysis, opinion extraction; opinion mining and review mining are some of them. While both sentiment analysis and sentiment classification are terms used in the same sense, their concepts are different (B Pang & L Lee, 2008). The complete process by which sentiment is taken from the text and understood is sentiment analysis; on the other hand, showing semantic orientation is the job of sentiment classification which does this by the assignment of a label to a text or part of a text. The main concern of such analysis is that a sentence or a document may contain a mixture of positive and negative opinions. Sentiment analysis is broken down into three levels, which are word level, sentence level and document level sentiment analysis (Westerski, 2007). The main focus of this work is to discuss sentence and document level sentiment analysis as shown in Figure 2.3.

### 2.3.1 Features Identification and Semantic Orientation of Text

There are three different levels for text feature identification; words, sentences and documents. Existing research works present different methods and ideas for extraction and semantic orientation of sentimental terms from various texts (B Pang & L Lee, 2008). For Feature Extraction (FE) and orientation, statistical and linguistic rules are used. Words and phrases are classified as nouns, verbs, adjectives and adverbs in accordance to the rules of linguistics. For a syntactic or corpus based method, the feature is selected by using the BoW method while the term frequency is used for machine learning classification. Rule based methods are used for selecting features from the lexicon dictionary, and both semantic and statistical methods are used for sentiment classification (Dave et al., 2003). Most of the related works use Part of Speech (POS), stop words removal, stemming, punctuation, fuzzy pattern matching, phrase patterns, semantic orientation, polarity tags, appraisal groups, and link-based patterns in order to extract features and sentiments (J. Wiebe et al., 2003)

Using adjectives and adverbs for subjectivity identification is the main focus of polarity classification (Chesley, Vincent, L. Xu, & Srihari, 2006). WordNet is the tool most often used for adjective identification. WordNet is used by researchers for sentiment analysis in the identification of words as adjectives and for semantic orientation (Breck, Choi, & Cardie, 2007). In most of the existing works, sentiment expressions usually depend on only a few words which are related to subjective sentiment orientation. For example, the word “good” is considered as positive and the word “bad” is considered as a negative sentiment. Subjective words such as these are called adjectives in linguistic terms. The role of identifying verbs is very important when trying to find the relationship between subjective and objective terms. Several researchers have looked into acquiring the verbs meaning and the sub-categorizing of verb frames in particular to aid in natural language processing. An interactive machine learning system has been introduced in (Nedellec, 2000), which has the ability to acquire sub-categorization frames of verbs and taxonomic relations based on syntactic inputs. Nouns, verbs, adjective and adverbs are suggested as all are grammatical categories which have the ability to express subjectivity or emotions (Turney, 2002). Semantic orientation is the classification of sentimental expressions according to their meaning and background knowledge. While syntactic analysis is unable to extract the concept from the text using syntax only, it does play a main role in document classification. Breaking multi-word expressions, mapping of synonyms into various components and words with multiple meanings used as a single component are all problems which semantic analysis can solve. (Turney & Littman, 2003) have used BoW and the semantic concept to improve the depiction of text classification and to extract the concept from a particular text.

Ontology based learning is a novel technique of semantic orientation and concept extraction from text and now the research is focused on using and integrating this method in sentiment analysis process. Ontology integrates the domain knowledge of individual words with terms used for learning and capturing concept from a text. The relationship between terms in a text is helpful in understanding the background knowledge. Ontology has been defined as a formal knowledge representation system consisting of three main elements: classes (concepts or topics), instances (which are individuals belonging to a class) and properties (which link classes and instances allowing for information to be inserted, in regards to the word represented, into the ontology) (DAvanzo, Lieto, & Kuflik, 2008). The combination of semantic information such as ontology and metadata was used by (Kawamura et al., 2008) which retrieved the structured part with the conventional natural language processing such as syntactic parsing from the unstructured part. These ontology working is a specified domain.

In this work, a domain independent sentiment classification method proposed at sentence level by applying rules for all parts of speech to score their semantic strength. Using contextual valence shifter and expression or sentence structure based on dynamic pattern matching as well as addressing word sense disambiguation. The system identifies opinion type, strength, confidence level and reasons. It deals with the SentiWordNet and WordNet as the knowledge base, with the additional capability of strengthening the knowledge base with modifiers, contextual valence shifter information and usage of all parts of speech.

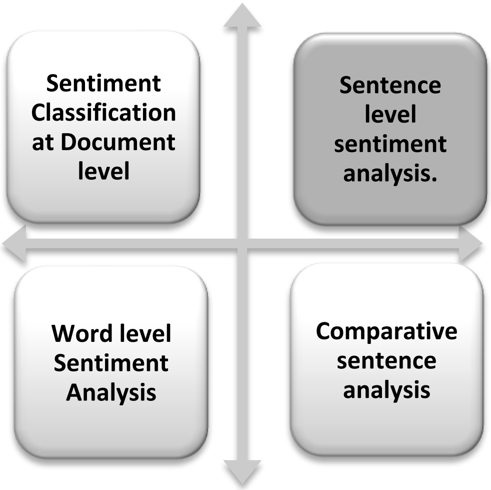

Sentiment Classification at Document level
Sentence level sentiment analysis.
Word level Sentiment Analysis
Comparative sentence analysis

Figure 2. 3 General Trend of Semantic Orientation of Reviews

#### 2.3.1.1 Document level Sentiment Analysis

Document level sentiment analysis is the process of classifying the overall sentiments expressed by the writers in the entire text of the document; the document being positive, negative or neutral about a certain object. Machine learning algorithms and lexical methods are mostly used by the researchers for the document level sentiment classification. Statistical methods provide encouraging results as far as processing speed is concerned, but the accuracy level is low because of the lack of semantic consideration. Document level sentiment analysis deals with a document as a whole and classifies all the sentiments which have been expressed about a certain object by the authors showing whether the overall document is positive, negative or neutral. However, the text documents or reviews are broken down into sentences for sentiment analysis at the sentence level. These sentences are then evaluated by utilizing lexical or statistical methods in order to determine their semantic orientation. This process involves two functions; first is to determine the subjectivity or objectivity of a sentence and the next function is of taking the sentences with an opinion orientation which is subjective. Some existing work involves analysis at different levels. Particularly, the level of semantic orientation involving words regarding opinion as well as the phrase level. Semantic orientation can be accumulated from the words and phrases to find out the overall Semantic Orientation of a particular sentence or review (Hu & B. Liu, 2004) (Leung & Chan, 2008) (Turney, 2002) (B. Liu, 2010b).

#### 2.3.1.2 Word or Feature based Sentiment Analysis

Word or feature level sentiment analysis gains importance by the application of NLP and statistical methods. It is the most detailed study of the text. Several researchers have worked on extracting features and opinion-oriented words by utilizing a predefined seed word list for extracting semantic orientation and opinion classification. The objective is to be able to determine text subjectivity and polarity as well as the author’s likes and dislikes about the object. Typically this objective is split into the following tasks (Westerski, 2007) :

- Extract object, features and their attributes
- Find the orientation of the text for positive, negative and neutral opinions
- Set feature synonyms and create a summary

Sentiment analysis suffers from various challenges, such as, determining which segment of text is opinionated, identifying the opinion holder and determining the positive or negative strength of the opinion. Since, sentiment analysis focuses on people's reviews, emotions and other relevant discussions. It is a challenging task, as every one of us has its own perception and concern about a particular problem, issue or topic. Moreover, opinionated text may be fictitious, irrelevant and/or contain ambiguous information. Opinions are much harder to describe than facts. Sources of opinions are usually informally written and highly diverse in nature (B Pang & L Lee, 2008). Semantic characteristics, like word sentiment, of each word are greatly acknowledged as good indicators of semantic characteristics of a phrase or a text that contains them, e.g. in (Turney, 2002). A sentence or text level sentiment annotation system uses words as indicators (features) of sentiment and therefore, requires the creation of words lists annotated with sentiment markers. The research on word-level sentiment annotation has produced a number of such lists of words that were manually or automatically tagged as sentiment or classified as related to sentiment (Balahur & Montoyo, 2009) (Andreevskaia & Bergler, 2008).

(Bethard, H. Yu, Thornton, Hatzivassiloglou, & Jurafsky, 2004) suggested a method that would use different information occurring at the same time in order to acquire words related to opinion (e.g., disapproval, accuse, commitment, belief) from texts as a way to carry out analysis of subjectivity at the word level. Two different techniques were used. The log-likelihood ratio is computed with the first technique; using data obtained by calculating how often words obtained from one sentence occur with seed words taken from Relative frequencies of words found in documents, either subjective or objective, are computed by using the second technique.

### 2.3.1.3 Sentence level Sentiment Analysis

In sentence level sentiment analysis, the text document or reviews are split into sentences and each sentence is checked for its semantic orientation by using lexical or statistical techniques. It can be associated with two tasks. The first of these two tasks is to identify whether the sentence is subjective or objective. And the second is to subjective sentences for their opinion orientation to classify as positive, negative or neutral. Sentence level semantic orientation is important because it takes each sentence individually for semantic orientation. NLP methods are useful for such types of semantic orientations. Sentence level analysis decides what the primary or comprehensive semantic orientation of a sentence is while the primary or comprehensive semantic orientation of the entire document is, handled by the document level analysis (B Pang & L Lee, 2008) (Hu & B. Liu, 2004). . However, the text documents or reviews are broken down into sentences for sentiment analysis at the sentence level. These sentences are then evaluated by utilizing lexical or statistical methods in order to determine their semantic orientation. This process involves two functions; first is to determine the subjectivity or objectivity of a sentence and the next function is of taking the sentences with an opinion orientation which is subjective. Some existing work involves analysis at different levels. Particularly, the level of semantic orientation involving words regarding opinion as well as the phrase level. Semantic orientation can be accumulated from the words and phrases to find out the overall Semantic Orientation of a particular sentence or review(Leung & Chan, 2008) (Westerski, 2007) (Hu & B. Liu, 2004) (Turney, 2002) (Andreevskaia & Bergler, 2008) (B. Liu, 2010b) .

When NLP and statistical techniques are utilized, much importance is given to sentiment analysis at the word or feature level because it is an analysis of the text with the most detail. The semantic orientation of a phrase or an opinion word is determined by the techniques proposed by (Andreevskaia & Bergler, 2008) (Kim & Hovy, 2005) and (Neviarouskaya, Prendinger, & Ishizuka, 2009). Several researchers used a preset seed word to enable extraction of opinion-oriented words and features (L Dey & S. K. Haque, 2008). (Popescu & Etzioni, 2005) (Hu & B. Liu, 2004) and form a list used for semantic orientation, extraction and classification of opinion. Determining the polarity and subjectivity of a text is not the only aim of sentiment analysis. On the contrary, what the author of the text specifically likes or dislikes regarding an object is also of importance (B. Liu, 2010a).

Three entirely different types of sentence-level sentiment classification are investigated in the present works by researchers. They are reviews of products, reviews of movies and blog comments which have been receiving a lot of attention with text-level sentiment analysis for the past several years, although, blogs have received less attention than the reviews. At the same time, sentence level analysis of comments and reviews has continued to receive relatively little attention (Balahur, Steinberger, Goot, Pouliquen, & Kabadjov, 2009).

Specific challenges are found for sentiment classification with each type and domain presented. A combination of neutral sentences which give a description of the film plot along with sentences which are full of sentiments are often found in movie reviews; whereas, reviews of products tend to be very domain-specific. Moreover, systems focused on one domain cannot be used on another type of domain with the same performance results (Balahur et al., 2009) (Blitzer, Dredze, & Pereira, 2007) (Andreevskaia & Bergler, 2008) .

In contrast to other types, a blog can be an extremely emotional context. Colloquial style, careless presentation manner (e.g. typos, grammatically incorrect sentences), and more often than not, making use of emoticons instead of words are some of the features of blogs. Sometimes they are tagged with the author's mood; however, as there are hundreds to choose from, these mood labels are very diverse. Furthermore, most of these labels are not used consistently (Balahur et al., 2010) (Leshed & Kaye, 2006) (Hariharan et al., 2010) (Andreevskaia & Bergler, 2007).

(Shamma, Kennedy, & Churchill, 2009) investigated the twitter blogs comments for the 2008 American Presidential Electoral debates. They illustrated that the analysis of twitter usage is important and closely yield the semantic structure and contents of the media objects. The twitter can be a predictor of the change in any media event. (Go, Bhayani, & L. Huang, 2009) presented a machine learning method for classifying sentiment of twitter messages and described that pre-processing is more important to remove noisy text in the case of short messages and comments to achieve high accuracy. They have achieved an accuracy of 80% using machine learning algorithms for positive and negative sentiments. So mining blogs comments play an important role that can be leveraged to evaluate and analyse any activity. Newspaper articles and news blogs are more challenging towards sentiment analysis articles as they are more likely to show a "balanced" view of their subject. These articles are often a combination of differing and many times conflicting opinions. They cite views from opposing parties and present not only objective facts but also subjective "points of view"; a variety of news events could even be presented in one text. It is possible to bring about an emotional response from a reader even if the facts are presented in an objective way, "good" vs. "bad" news (Andreevskaia & Bergler, 2008). Miscellaneous sentiments can often be found in a single newspaper text as well as in comments on social network sites and in reviews. That is why, for this dissertation, sentence level sentiment analysis will be carried out rather than analysing the text as a whole. The difficulty in sorting out the opinions of various types that may exist together in a particular text can be avoided, by sentiment classification of small units of language; however, there are issues that still make it more difficult than analysing the sentiment of homogenous texts that contain similar sentiments. The relatively small size of sentences means that the decision about the sentiment of a sentence has to be made based on a small number of sentiment clues and, thus, is more sensitive to system errors and to model sparseness (J. Wiebe et al., 2003) (Andreevskaia & Bergler, 2008) (H. Chen & Zimbra, 2010) (Balahur et al., 2010) (Balahur et al., 2009).

The creation of annotation at the sentence-level as a sub region in the study of sentiment has provided an opportunity for the creation of several applications involving the areas of information retrieval and text mining. Firstly, determining the sentiment that an opinion holder has expressed in regards to a specific topic, issue or event needs a fine-tuned level of annotation and not just extracting the sentiment of the text as a whole. Sentence-level research provides this fine tuning. Secondly, in scientific literature, the study of hedging/fabrication is also a part of the realistic applications of annotation where applications for retrieval of information often require processing, summarizing, and categorizing of the text at the sentence level (Balahur et al., 2010) (Andreevskaia & Bergler, 2008) (Nathan, 2009).

There are algorithms for sentiment analysis which focus on summarizing the opinions which have been expressed regarding a specific product or its features in reviews (Lu, Kong, Quan, W. Liu, & Y. Xu, 2010) (Hu & B. Liu, 2004). It should be noted that this sentiment summarization differs from the traditional type of summarization which attempts to identify the main sentences in a text in order to summarize its major ideas. Classifying opinions according to their semantic orientation is also a subtask of sentiment summarization. So, there are three popular methods of sentiment detection which are: methods based on machine learning, semantic orientation or semantic analysis or lexical based method and the combination of these two (Ding, B. Liu, & P. S. Yu, 2008) (Andreevskaia & Bergler, 2008) (B Pang & L Lee, 2008) (Westerski, 2007) (Taboada, J Brooke, & Stede, 2009).

Lexical methods are utilized for the term semantic orientation which makes use of the so called sentiment lexicons, also known as opinion lexicons in online dictionaries like SentiWordNet, Sentiful, and WordNet etc. (Ding, B. Liu, & P. S. Yu, 2008) (Balahur & Montoyo, 2009) (Nathan, 2009). For machine learning methods, only the lemmas are not enough for detecting sentiment, however, they also make use of features (corpus or seed words) to successfully classify the sentiment. Machine learning and lexical methods are also combined and used as another method in order to extract sentiments. The details about these two approaches are described below in the subsequent sections of this chapter. The common process (Leung & Chan, 2008) (Ohana, 2009) involved for the sentiment analysis in these approaches is depicted in Figure 2.3.

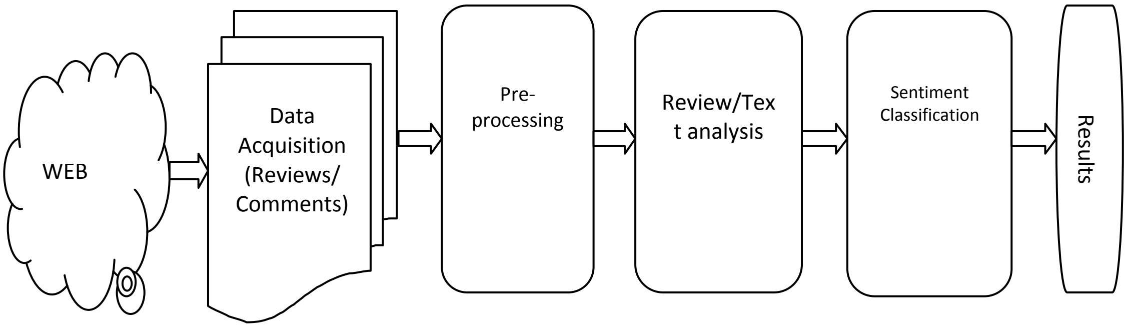


Figure 2.4 Sentiment Analysis Model

Three main steps are involved (Leung & Chan, 2008), namely *Pre-processing*, *Text Analysis* and *Sentiment Classification.* A compilation of specific reviews are taken as input by the model and are then processed according to the above three steps to obtain results. Review classification and evaluation of sentences or expressed opinions in the reviews are the results produced by the model. The following steps are required for sentiment analysis process.

### 2.3.2 Data Acquisition

Data searching, extraction and collection are the steps involved with data acquisition[1], which is a subtask of pre-processing, in sentiment and text analysis process. Crawler is a tool which is usually used to extract the text data from the web. The pre-processing steps are then applied for removal of noise and Feature Selection (FS).

### 2.3.3 Review Analysis and Pre-processing

Text pre-processing involves data preparation and cleaning of the datasets which is essential for the accurate execution of the next step i.e. Text Analysis. Some pre-processing steps typically used in data preparation/cleaning are as follows: removal of content that is non-textual as well as mark-up tags, and removal of non-essential review data, such as dates of the reviews and names of the reviewers, as this type of data is not needed for the sentiment analysis. Taking samples of the reviews in order to build a classifier may also be a part of the data preparation Pre-processing is one of the important steps of text analysis. Text representation is one of the pre-processing techniques which change a document from the full version into a document vector by reducing the complexity of the document; subsequently, the document is easier to deal with. Text representation which is an important aspect in document classification and information extraction signifies the preparation of a document into a concise form. Typically, a text document is presented as a vector of term weights (word features) derived from a set of words (dictionary), where each word is found at least once in a predetermined number of documents (Fensel, 2004a) (H. Liu & Motoda, 1998).

[1] The details of data collection and pre-processing for this work is described in chapter- 4

Document pre-processing or dimensionality reduction (DR) allows for an efficient data manipulation and representation. Irrelevant and redundant features lead to the degradation of how classification algorithms perform by affecting its classification accuracy and speed. Moreover, it tends to reduce over-fitting. This is the reason that DR is a vital step in text classification. As illustrated, DR techniques can be categorized into two approaches, either Feature Extraction (FE) or Feature Selection (FS) (H. Liu & Motoda, 1998).

FE is the pre processing step which is used to present text documents in a clear word format. The documents involved in text classification are represented by a large amount of features, most of them are possibly irrelevant or noisy (Montañés, Fernández, Díaz, Combarro, & Ranilla, 2003). DR is the exclusion of a large number of keywords, based preferably on a statistical process, to create a low dimension vector. DR techniques have received much attention recently because effective DR makes the learning task more efficient and saves more storage space (J. Yan et al., 2005). Commonly, the steps taking place for the FE are:

- Tokenization: A document is treated as a string, and then partitioned into a list of tokens.
- Sentences Boundaries Identification: The sentence boundary identification is important in order to split reviews/comments or text documents into correct sentences.
- Remove Noisy Text: In online web forums, social networks, blogs etc., people mostly write short forms of words and use symbols in comments to express their views. How these symbols and short forms are made is useful in extracting their semantics from such sentences. Noise removal from a text improves the efficiency of the semantic orientation and classification process.
- Removing Stop Words: Stop words such as “the”, “a”, “and” etc. occur frequently so the insignificant words need to be removed.
- Lemmatization: Applying the stemming algorithm that converts different word forms into similar canonical forms. This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute etc.

- Part of Speech Tagging (POS): POS is used for assigning a tag to each word in a sentence, tracing the position of a word in the sentence and extracting the structure of the sentence for semantic orientation. The sentence patterns extracted are also used in WSD where the tag is assigned to each word, like, JJ, JJS, VB, VBS, RB, NN, NNS, and DT etc.

- Word Sense Disambiguation: Determining in which sense a word having a number of distinct senses is used in a given sentence.

In order to obtain an accurate classification, extraction of the appropriate keywords from the text is important. A feature vector, e.g. F = (f1, f2 ... fn) is the form that keywords related to the original data are usually kept in. A different word, also known as a feature, of the original text is represented by each coordinate of a feature vector. The value for each feature may be either an integer or a binary value. The intensity of the feature in the original text could be expressed further by an integer while the presence or absence of the feature is indicated by the binary value. The machine learning process is strongly influenced by these features; therefore, it is vital to have a good selection of features for learning to be achieved. Capturing the desired characteristics of the original text in relation to the sentiment analysis at hand is the reason behind selecting the best possible features. Unfortunately, an application which is able to easily find these best features does not exist yet. For now, the user must totally depend on his/her intuition, domain knowledge and a lot of experimentation in order to choose the best set of features. Therefore, NLP techniques are important for FS in sentiment analysis (Sebastiani, 2002)(Clark, 2003).

The proposed approach includes the use of a knowledge base as FS for rule-based lexical sentiment analysis as well as a popular Bag-of-words model used as Bag-of-Sentence which takes individual words in a sentence as a feature. This results in the creation of a vector consisting of an unorganized collection of words representing the entire text. Moreover, one word is represented by each feature in the vector. The major challenge with this approach is choosing words that are suitable to become features. It is obvious that for any real use, a comparison of this vector with a feature vector containing a large number of words, a dictionary of the language in fact, would have to be made. However, this particular model would overfit, which in turn would lead to bad performance when presented with a new dataset causing it to be quite inefficient.  To reduce the size of the features and to solve the problem of overfitting features, the selection method is used as described in the following section.

The second most important step in the pre-processing of text classification after FE is FS. A vector space is constructed using FS for improvement in efficiency, scalability and accuracy of the text classifier. Basically, the properties of the domain and algorithm are considered by a good FS technique (Z. Q. Wang, Sun, D. X. Zhang, & Li, 2006) (Sebastiani, 2002).

Filters and wrappers are the two main kinds of FS techniques involved in machine learning. Filters do not apply FS based on the specific learning algorithms that make use of the selected features; rather, an evaluation metric is used in order to evaluate a feature. The filters use the matrix vector to measure how well the feature can differentiate each class. On the other hand, wrappers do utilize some learning algorithms as their evaluation function which relates to the particular algorithm's classification accuracy. However, for text classification, wrappers are usually not suitable. This unsuitability is in relation to their general time consumption. When the number of features is large, wrappers may take more time because of the need to train a classifier for each feature subset to be evaluated.  A text document could possibly fit partially into a variety of categories which poses a challenge to text classification in finding the category that the text document best fits in. The term (word) frequency/inverse document frequency (TF/IDF) approach is typically used to capture the relevancy among words, text documents and particular categories by weighing each word in the text document in regards to its uniqueness (H. Liu & Motoda, 1998) (Sebastiani, 2002).

Of the various feature evaluation matrices which have been evaluated, those worthy of mention  are Gini index, information gain (IG), expected cross entropy, term frequency, the weight of evidence of text, Term frequency and Document frequency (TF/DF), mutual information, Odds Ratio and Chi-square.  A good FS matrix will consider problem domains as well as algorithm properties (Z. Q. Wang, Sun, D. X. Zhang, & Li, 2006) (Sebastiani, 2002).

In the literature, a typical approach is to choose the most important keywords or features manually, for example, a good gauge of the author's opinion would be a word used as an adjective. The most important keywords are those which express the polarity of a sentence such as 'fabulous', 'excellent' and 'horrible'; words like these would be selected as features. However, it was shown by (B Pang, L Lee, & Vaithyanathan, 2002) that statistical models outperform manual keyword models. Statistical models have a good set of words which represent features which are selected according to their occurrence in the existing training compilation. Therefore, the size of the compilation and the similarity of domains of the training and test data have a bearing on the quality of selection.

For FS using corpus based method, a unigram FS technique is utilized and in the beginning a large number of features are retrieved making the model of a higher variance. As a consequence, a lot more training data will be needed to avoid overfitting. While the training set does contain hundreds of thousands of sentences, it is still a considerably large number of features for the training data set; hence, it is better for us to remove as many irrelevant features as possible. Therefore, an attempt is made to accomplish this task by using different FS algorithms, few are define as follows.

- Frequency-based Feature Selection

This is the easiest method to use for FS. Features (unigram words in this work) for a particular class are to be chosen which are occurring most frequently in this class. In practice, a good feature for a particular class is one which occurs more often than a predetermined threshold.

- Mutual Information(MI)

The idea behind mutual information is that, for each feature F and each class C, there is a score that is used to measure how much contribution could be made by F towards

correct decisions about class C. After the MI score is calculated, choose only the top k (threshold value) features possessing the highest scores for the feature set to be used for testing. It is observed that when k is small, the data is underfitting because the model is too simple. However, if k is large, the data is overfitting because the model is too complex.

By learning the characteristic categories from a set of classified texts, machine learning algorithms can construct a classifier automatically. This classifier can then be used to classify a particular text into preset categories. However, machine learning techniques have some disadvantages: (1) a large number of training text words must be collected by humans in order to train a classifier which is an extremely laborious process. If changes occur in the predefined categories, a new set of training text words must be collected using the same techniques. (2) The semantic relations between words are not taken into consideration by many of these traditional techniques; therefore, it is quite difficult for the accuracy of these classification techniques to be improved on (Sebastiani, 2002). (3) The issue of translatability, between one natural language into another natural language is another disadvantage. These types of issues prove that machine understanding systems are facing problems. Some of these may be addressed if we have machine readable ontology (Song, Lim, Kang, & S. J. Lee, 2005), and that’s why the ontology based text representation is very important for knowledge extraction. During the text mining process, ontology can be used to provide expert background knowledge about a domain. Recent research shows the importance of the domain ontology in the text and sentiment classification process (Fensel, 2004a).

Hence Feature Extraction(FE) and Feature selection are the pre-processes used to represent the text before the text is to be re-orientated in some structured form, the noisy text should be removed, i.e. removal of unnecessary irrelevant words and symbols. This is because online discussion forums, blogs and customer reviews may contain a lot of noisy text and the opinion sources are typically informally written and are highly diverse. For sentiment analysis, we should also remove all facts that do not express opinions. The subjective phrases are considered and the objective portion of the text is removed because the focus of this work is only on the users' opinions (L Dey & S. K. Haque, 2008) (Turney, 2002). Text analysis and pre-processing is one of the steps of sentiment analysis model. This is where the linguistic features of the reviews are analysed so that identification of opinions and/or product features as well as other interesting data takes place. Extractions of opinions and product features from the processed reviews often take place after applying some computational linguistics applications such as POS tagging.

### 2.3.4 Sentiment Classification

Data mining, machine learning or NLP methods are used for the sentiment classification and knowledge extraction after the data has been pre-processed. Generally, sentiment polarity is classified into positive, negative or neutral opinions, which is explain in details in the below subsequent section.

### 2.3.5 Presentation and Summarization

In this step, a summarization of the sentiment extraction from several opinions and the resulting classification, expressed in graphical, table or text form, must be completed in order to be presented to the user. The goal is to give a general comprehension and to facilitate understanding about what is being said in the opinion.

## 2.4 Sentiment Analysis

Basically there are two types of approaches used for sentiment analysis: supervised and unsupervised sentiment classification methods.

1. Supervised Classification Methods are Naïve Bayes (NB), Maximum Entropy (ME) Classifier, and Support Vector Machine (SVM) (Machine Learning Techniques)

which are mostly used in sentiment classification using predefined feature set or corpus.

2. Unsupervised methods are used in the form of lexical and rule based approaches for sentiment classification. These types of methods use dictionaries or user defined corpus as knowledge base.

### 2.4.1 Corpus Based Machine Learning Approaches

The machine learning method and topic classification are similar in the sense that topics are classes of sentiment such as Negative and Positive. This is how it works: a review is broken down into phrases or words, the review is then presented as a document vector (bag-of-words), and finally, the review is classified on the basis of the document vectors (Leung & Chan, 2008). The majority of existing approaches today for classification of sentiment at the document level are on the basis of supervised learning; however, a few unsupervised methods are also available, which are introduced in the next section.

It is apparent that classifying a sentiment can easily be formulated as a supervised learning problem which has two class labels, negative and positive. In regards to the assumption above, it is not a surprise that the reviews utilized in existing research regarding data for training and testing are mostly product based. Data for training and testing is easily available due to any typical review site having already assigned a reviewer rating (e.g. 1-5 stars) to each review (Sarvabhotla, Pingali, & Varma, 2009). Commonly, a thumbs-up or positive review will be assigned 4-5 stars while a negative or thumbs-down review is assigned with only 1-2 stars. Studies present to date have taken unlabeled data from the domain of interest with labelled data from another domain as well as general opinion words and made use of them as features for adaptation (Zhao, K. Liu, & G. Wang, 2008) (Leung & Chan, 2008). A method was proposed in (Balahur & Montoyo, 2009) for web reviews to haul out, categorize and summarize opinions on products. It was based on the prior building of product features arrangement and on the semantic relatedness given by the Normalized Google Distance and SVM learning. For features and attributes extraction SVM classifier was used with WordNet and Concept.

Unlike the earlier works which mainly focused on the lexical feature at word level, (Ding, B. Liu, & P. S. Yu, 2008) (Nathan, 2009) utilized the supervised learning approach to present a method for sentence level classification considering the contextual information. In their works they proposed training of two classifiers, one using word level lexical features, which can be used to label each sentence in the document and another one based on the labelled sentences or possibly combined with subjectivity Summary. These classifiers can then be used to classify accordingly. The problem of attributing a numerical score (one to five stars) to a review is presented in (Sarvabhotla et al., 2009). Their main focus was on feature representations of widely used reviews, problems related to them and solutions to address these problems. They presented it as a multi label classification (supervised learning) problem and proposed two approaches, using NB and SVM. A set of tools and experiments were anticipated in (Saggion & Funk, 2010) (Taboada, J Brooke, & Stede, 2009) for the text based opinion classification. This set of tools can compute word based and sentence based sentiment features utilizing SentiWordNet as a lexical resource and by classifying short English texts in accordance with its rating (the positive or negative value of the opinions) using machine-learning based semantic and linguistic analysis.

A combination of machine learning and polarity feature improvement was proposed by (Waltinger, 2009) for identification of sentiment polarity. Detecting sentiment polarity of colloquial language was presented using the dataset of the Urban Dictionary project. (Baccianella, Esuli, & Sebastiani, 2009) used BoW, sentence position and Part-Of-Speech (POS) information as features for analysing reviews. The reviews were then represented as feature vectors for a learning device, such as NB and SVM. However, approaches of FE also require tools such as POS tagger since there is no consideration for contextual information. (Zhao et al., 2008) suggested a method for sentiment classification on the basis of conditional random fields (CRFs). This was in response to the two specific characteristics of “label redundancy” and “contextual dependency” in sentence level sentiment classification. Contextual constraints in sentence sentiments are captured by CRFs. A hierarchical framework is used for introducing redundant labels and capturing label redundancy from among various classes of sentiments. However, it is apparent that the hierarchical structure in a large scale data set is not only very costly but also quite ineffective.

The issue with assigning a numerical value (e.g. 1-5 stars) to a review is introduced in (Sarvabhotla et al., 2009). The feature representations of reviews were utilized and it was described as a multi-label classification (supervised learning) problem while utilizing NB and SVM. (H. Yu & Hatzivassiloglou, 2003) proposed a system which would check the subjectivity of sentences after it classifies a document. The method of machine learning with integration of compositional semantics with sentiment classification is described in (Choi & Cardie, 2008). (Whitelaw, Garg, & Argamon, 2005) presenting the SVM algorithm with BoW which is used for classification of movie reviews. However, this method does have its limitations i.e. it only considers adjectives and their modifiers which indicate assessment/judgment. Polarity of phrases is extracted by the method in (Turney, 2002) which utilizes the PMI between seed words and phrases. The flat feature vector BoW method is utilized in the majority of the above mentioned applications in order to represent the documents. On the other hand, methods which are based on statistics depend on subject, language style and domain as well as huge amounts of significant statistical data, while neglecting syntactical structure and contextual information. As a consequence, in small textual composition levels, the accuracy of sentiment classification is affected. In turn, data that might be extracted at the sentence level could possibly be inaccurately represented by these methods. Some methods have been proposed to help solve this issue. Simple methods for combining individual sentiments (Esuli & Sebastiani, 2005) and supervised (Alm, Roth, & Sproat, 2005) statistical techniques were proposed which can measure sentiment on the phrase or sentence level using opinion oriented words. Another method, proposed by (Wilson, J. Wiebe, & Hoffmann, 2005), makes use of both lexical and syntactic features for sentiment analysis and is a machine learning approach. This method, however, missed pertinent contextual information which indicates that the individual sentence itself is vital when extracting semantic orientation.

Corpus based machine learning method or methods based on compilations are able to compile lists of negative and positive words with a high accuracy. However, in order to reach their full potential, most of these approaches need immense annotated training datasets. Lexical-based methods can overcome some of these limitations by utilizing dictionary-based approaches since these approaches depend on existing lexicographical resources (such as WordNet) to provide semantic data in regards to individual senses and words (Andreevskaia & Bergler, 2008) (Hu & B. Liu, 2004).

### 2.4.2 Lexical Based Semantic Orientation Approaches

Lexical methods are utilized for the term semantic orientation which makes use of the so called sentiment lexicons, also known as opinion lexicons, in online dictionaries like SentiWordNet, Sentiful, and WordNet etc. A compilation of recognized sentiment terms along with their semantic values are contained in these lexicon dictionaries. In most cases these semantic values are in numerical form ranging from -1 to 1 (Leung & Chan, 2008) (Nathan, 2009) (Andreevskaia & Bergler, 2007).

The complete process by which sentiment is taken from the text and understood is sentiment analysis; on the other hand, showing semantic orientation is the job of sentiment classification which does this by the assignment of a label to a text or part of a text. This process involves two functions, first is to determine the subjectivity or objectivity of a sentence and the next function is of taking the sentences with an opinion orientation which is subjective. Some existing work involves analysis at different levels. Particularly, the level of semantic orientation involving words regarding opinion as well as the phrase level. Semantic orientation can be accumulated from the words and phrases to find out the overall semantic orientation of a particular sentence  or review(Andreevskaia & Bergler, 2008) (Hu & B. Liu, 2004) (Ohana, 2009) (Turney, 2002) (B. Liu, 2010b).

Dictionary based techniques make use of the data found in references and lexicographical resources, such as WordNet and the thesaurus which can be used for assigning sentiments to a large number of words. Majority of these methods utilize various relationships between words (synonymy, antonymy, hyponymy /hyperonymy) in order to find the seed words and other entries as described earlier. The data existing in dictionary definitions is made use in word-level sentiment orientation in some of the recent methods. For semantic orientation lexical based semantic terms are extracted using dictionaries like SentiWordNet, ConceptNet etc. for the sentence level classification (Ohana, 2009) (Ding, B. Liu, & P. S. Yu, 2008) (Andreevskaia &

Bergler, 2007) (Westerski, 2007) (Leung & Chan, 2008) (Taboada, J Brooke, & Stede, 2009).

In this work, sentiment analysis is to extract polarity from the text, and semantic orientation refers to the polarity and strength of words, phrases, or texts. The concern of this work is primarily with the semantic orientation of sentence and texts, but the sentiment of words and phrases were extracted towards that goal.

Two sub tasks are involved in the semantic orientation approach. The first sub-task determines the semantic orientation of the opinions which were taken from reviews in the Review Analysis step, while the second sub-task determines the overall semantic orientation of a review or sentence, or is based on the semantic orientation of the opinions it contains. The following tasks are performed in the analysis of the sentiment at the sentence level (B. Liu, 2010a).

- *Subjectivity classification*: A sentence is to be determined either a subjective or an objective sentence.
- *Sentence-level sentiment classification*: In the case of the sentence being subjective, it must be determined whether a positive opinion is expressed or a negative opinion is expressed.

"Classification at the sentence level is more often than not an intermediate step. In the majority of applications, the object or features of the object for which the opinions are given is what is required to be known. All the same, the two sub tasks of classification at the sentence level remain vital because (1) they weed out those sentences which have no opinion, and (2) after gaining knowledge of what particular objects and features of the objects are mentioned in a sentence, this step helps in determining if the opinions on the objects and their features are negative or positive. While the majority of today's researchers study both problems, there are some of them who devote their attention to only one. Since, the issues are regarding classification, typical supervised learning methods are again appropriate (B. Liu, 2010a) . The manual effort needed to annotate a large number of training examples is one of the bottlenecks in utilizing supervised learning" (B. Liu, 2010b).

"The sentence asserts a single opinion from a single opinion holder. This assumption is only suitable when applied to simple sentences with a single opinion, e.g., "*The picture quality of this camera is amazing*." However, in the case of compound sentences, more than one opinion may be expressed in a single sentence, considering the sentence, "*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*", and both negative and positive opinions are expressed. The sentence is positive as far as "picture quality" and "battery life" is concerned, but for "viewfinder", it is negative" (B. Liu, 2010a).

(Q Ye, Z Zhang, & R Law, 2009), use machine learning approach using datasets in the travelling domain and perform different experiments on different number of reviews documents using the feature selected in that domain. Their method is applicable that domain dependent on the important n- gram feature from the travel blogs.

In (H. Yu & Hatzivassiloglou, 2003) an attempt has been made to classify subjective sentences while at the same time determining their opinion orientations. Supervised learning is applied for identification of subjective or opinion sentences. Sentence similarity, naïve Bayesian classification, and multiple Naïve Bayesian classifiers were the three learning approaches evaluated. In (Hariharan et al., 2010) a technique is proposed to extract the opinion words from reviews. This proposed extraction algorithm assigns scores to each of the words in the review. The recommendation of the product to a user by the opinion miner is based on the cumulative weight of the scores. This method is domain independent and can be applied to any review formats, provided the reviews are structured and formatted. However, low accuracy in results may be experienced due to unfairness in this type of scoring in some contexts.

Natural Language Processor Linguistic Parser has been introduced by (Balahur & Montoyo, 2009), which can be used for parsing of reviews,  splitting text into sentences and producing tags for each word's part of speech, i.e. verb, noun, adjective etc. Very few authors have considered word sense disambiguation rather an assumption was made about various senses of a solitary word which in turn can provide different opinions. Synset from WordNet is utilized for various senses of the same word (Andreevskaia & Bergler, 2008) (Ohana, 2009).

Applications based on the dictionary are often utilized with, not only two way, positive vs. negative classification, as with machine learning corpus based methods, but also with positive vs. negative vs. neutral in three way classification. Sentiment can be assigned by dictionary based approaches to not only words, but also their senses. (Ding, B. Liu, & P. S. Yu, 2008) (Andreevskaia & Bergler, 2007) This is because the labels are based on sense level definitions. Thus, one of the main advantages of these methods is their appropriateness with sentiment classification, not only at the sense level but also at the sentence level. These types of approaches used the lexicon dictionaries as described below.

#### 2.4.2.1 Analysis of Linguistic and Lexical Resources for Sentiment Analysis

For sentiment extraction, knowledge of various linguistic terms and acquisition of the sense of the opinion terms as well as their semantic orientation are necessary. The classification of the contents of documents into positive or negative, and subjective or objective terms is the prime issue of sentiment analysis. The terms are identified by either their syntactic features or the lexical semantics. According to (Polanyi & Zaenen, 2006), “The most salient clues about attitude are provided by the lexical choice of the writer, but the organization of the text also contributes information relevant to assessing attitude”. Subjectivity detection is the area of importance for sentiment analysis. It is a general term used to mean opinions, evaluations, beliefs, perceptions, emotions, speculations, etc. is private state (Jindal & B. Liu, 2006). Subjectivity is used to express these private states in regards to a text or conversation. An objective statement presents information in accordance with the author’s intention. If the feedback of the user has no judgment or opinion on the source content, it is considered objective. The lexical resources are used for semantic orientation and function as a knowledge base for sentiment analysis. These lexical resources contains WordNet, SentiWordNet, SentiFul, ConceptNet etc are described below in details (B Pang & L Lee, 2008) (Ohana, 2009).

##### *2.4.2.1.1 General Inquirer*

General inquire is a computer-assisted approach for content analysis. It allows for access-points to various resources containing textual data associated with the specific General Inquirer (Stone, Dunphy, & Smith, 1966). This includes manually-classified terms which are labelled as positive or negative semantic orientation in a variety of types, and words that are related to agreement or disagreement.

#### *2.4.2.2 Opinion Finders*

Opinion finder is available for download and is a list of subjectivity clues which make up a Subjectivity Lexicon. These clues, which were used in (Wilson et al., 2005) were compiled over a period of several years with a great effort using a variety of sources.

#### *2.4.2.3 WordNet*

WordNet is made up of English words put into a large lexical database which is most often used in classification of text, semantic orientation, computational linguistics and natural language processing. It contains sets of cognitive synonyms (synsets) made up of various parts of speech such as nouns, adjectives, verbs and adverbs which express their own distinct concept. These synsets are interlinked by conceptual-semantic and lexical relations. WordNet is also freely and publicly available for download (Fellbaum, 1998).

#### *2.4.2.4 VerbNet*

VerbNet (VN) is currently the largest verb lexicon available on-line for the English language. It is a hierarchical domain-independent broad-coverage verb lexicon. It also contains mappings to other lexical resources like WordNet. VerbNet is arranged into classes of verbs which extend Levin (1993) classes by the refinement and the addition of subclasses in order to achieve syntactic and semantic coherence among members of a specific verb class (Kipper-Schuler 2006).

#### *2.4.2.5 SentiFul*

SentiFul is a lexicon-based system for sentiment analysis which is strongly dependent upon the availability of sentiment-conveying terms in its database. In order to solve the issue of lexicon coverage, an original method was introduced for building and expanding the sentiment lexicon (SentiFul). It is represented by words that convey sentiments and are annotated by sentiment polarity, polarity scores and weights.

#### *2.4.2.6 ConceptNet*

ConceptNet is common-sense knowledge based natural-language-processing toolkit which is freely available for download. It supports a variety of practical textual-reasoning tasks related to real-world documents right out-of-the-box; there is no additional statistical training required. ConceptNet is a resource which is rather unique as it captures a wide range of common sense concepts and relations, like the ones available in the Cyc-knowledgebase; however, this knowledge is not arranged as a complex and intricate logical framework, but instead as a simple, easy-to-use semantic network, similar to WordNet.

#### *2.4.2.7 SentiWordNet*

SentiWordNet (Esuli & Sebastiani, 2006) is an opinion mining lexical resource made up of synset from WordNet, a thesaurus-like resource, which allocates a sentiment score of positive, negative or neutral. These scores are automatically generated using the semi-supervised method which is described in (Esuli & Sebastiani, 2006). It is also free to the public.

#### *2.4.2.8 Turney adjective list*

This is a list of 1400 adjectives with their semantic-orientation values which were rated by using the method proposed by (Turney, 2002); the list is available through the Yahoo! Sentiment AI group.

According to (B Pang & L Lee, 2008), the first try at employing WordNet relations in word sentiment annotation was made by (Kim & Hovy, 2004)(Kim & Hovy, 2005). They made the suggestion about an extension to lists of manually tagged positive and negative words by adding to the list the synonyms for those words. They began with just 54 verbs and 34 adjectives. The method was applied in two occurrences and acquired 6079 verbs and 12113 adjectives. Then, on the basis of the strength of sentiment polarity which had been assigned to each word, the words which had been acquired were ranked. This strength-of-sentiment score or rank for each word was calculated by maximizing the probability of the category of the word's sentiment in regards to its synonyms.  An alternative method was suggested by (Kamps, Marx, Mokken, & De Rijke, 2004) for utilizing WordNet's synonymy relations for tagging words with Osgood's three semantic dimensions. The shortest path joining a particular word to the words 'good' and 'bad' was calculated through WordNet relations in order to assign values of positive or negative to the word.

In (Lu et al., 2010), the authors proposed an approach to evaluating the sentiment strength of reviews. They first extract the opinion phrases which consist of the opinion noun word and the modified opinion sentiment features from reviews, and then calculate the sentiment strength of review based on the extracted feature phrases. The strength of the opinion phrase is determined by the strength of the adjective word along with the adverb that modifies it. They mark the strength of adverbs manually and employ the link analysis method for calculation of adjective strength based on a progressive relation between adjective words.

Dictionary-based methods for sentiment classification at the word-level have no need of large corpora, or search engines having special functionalities. Rather, they depend on readily available lexical resources existing today such as WordNet. They are able to compile comprehensive, accurate and domain-independent word lists containing their sentiment and subjectivity annotated senses. Such lists provide a vital resource for sentence or text sentiment classification and because of early compilation they are able to increase efficiency of sentiment classification at text and sentence level. In contrast to the other works, this work presents sentence level lexical/dictionary knowledge base method to tackle the domain adaptability problem for different type's data (Andreevskaia & Bergler, 2008).

A domain dependent rule-based method was introduced by (L Dey & S. K. Haque, 2008) for analysing word dependency and structure in contextual information for sentiment classification. Extraction of opinion from noisy text data with granularity at multiple levels was introduced. Domain knowledge was utilized for contextual structure and WordNet was used for semantic orientation. These techniques, however, do have limitations. They are domain-dependent lexicons which are developed manually, and are unable to handle long complex sentences. A lexical system for analysis of sentiment at various grammatical levels is presented in (Neviarouskaya et al., 2009). This method made use of a wide-coverage lexicon, accurate parsing and sentiment sense disambiguation semantic orientation. So, contextual information of all the parts of speech is vital for the semantic orientation. Structure of the sense in sentences and all content parts of speech play an imperative role in analysis of sentiments. There are several limitations of the methods available today. These approaches focused on one domain and cannot be used on another type of domain and genre; reviews and blogs have a different genre and domains. Moreover, concentration on the structure of the sentence and the contextual valence shifter is low, word sense disambiguation is ignored, the system is based on lexicons suffering from a lexical coverage limitation, less attention is given to attenuate, the rule of term weighting and polarity score is too generalized, the imperial expression or confidence level sentiment orientation in the expression is ignored and there is no proper rule for handling the noisy text with photonic symbols or special characters given.

In this work a technique for domain independent sentence level classification of sentiment is introduced. Rules for all parts of speech are applied so that they can be scored on the strength of their semantics, contextual valence shifter, and sentence structure or expression on the basis of dynamic pattern matching. Moreover, word sense disambiguation to extract accurate sense of the sentence has also been addressed. Opinion type, confidence level, strength and reasons are all can be identified using this system. SentiWordNet and WordNet are utilized as the primary knowledge base which has the further capability of being strengthened by using modifiers, information in the contextual valence shifter and all parts of speech.

## 2.5 Summary

This chapter conclude with the solution of the few limitations found during the thorough study of the existing literature. The task of sentiment and subjectivity analysis has attracted considerable interest since 1990s when the first automatic system for these tasks was developed. It has become a major research stream within NLP. After the growth of online sources and web, sentiment analysis has gained much of its popularity and significance due to tremendous attention given by research communities to make it a major area of research in integrating other disciplines like text mining, IR and knowledge management etc. In this chapter the background and literature related to the development of this research field is studied (Starting from 1990's to 2000; subjectivity and emotion extraction using NLP; and then from 2000 to 2003; the development in internet, e-commerce and increasing number of communication via internet; and now its a major area of research in integrating other disciplines like text mining, IR and knowledge management in the recent work of last few years). Based on the methods and approaches presented in this chapter enable to frame a new idea which is capable to overcome few limitations like domain adoptability, word sense extraction and taking the contextual information of all part of speech at the sentence level.

# CHAPTER 3

# LEXICAL BASED SENTENCE LEVEL SEMANTIC ORIENTATION

## 3.1 Introduction

This chapter describes the proposed sentence level sentiment classification method. The proposed method is proficient for semantic orientation of online customer reviews, blogs and social network comments. Sentence-level lexical contextual information and Word Sense Disambiguation (WSD) is proposed for accurate sense extraction from each individual sentence. SentiWordNet, WordNet, and the information of other lexical contents (such as intensifier, enhancer and reducer dictionaries) are used as knowledge base for semantic score of each term in the sentence. The chapter starts by identifying the methodology and the key components of the method that presents the foundation, which structures the basis for building this method. The foundation components include some set of definitions that are related to pre-processing, source and object finding, knowledge base and WSD for integration and formulation of the method. This chapter then gives explanation of the building block, which is about the assembly of the identified foundation components to form new approach, the available data models and tools for editing and exporting the semantic information that can be implemented in the method components. The chapter thereafter depicts the process of integrating different components for building new sentence level sentiment analysis method. Finally, evaluation of this approach as to how this method meets the semantic orientation and sense extraction using the lexical knowledge base is described. The subsequent chapter illustrates the substantiation of the proposed method.

## 3.2 Sentence level Semantic Orientation

The proposed method describes different components of the sentiment analysis process that is used to identify, aggregate and evaluate the web text sources i.e; customer reviews, blogs, comments, discussion forums, about events and entities available on the web. This method is the combination of different components to extract the semantic score and to summarize the unstructured web contents. The steps are interlinked and each step is the precursor for the next. The basic elements of sentimental reviews which are useful for semantic orientation and classification are(B. Liu, 2010a):

- **Sentiment holder:** The holder of an opinion is the person or organization that holds or expresses a particular opinion about something, e.g. *I love playing cricket.*
- **Object (Target):** "An object '*o*' is an entity which can be a product, person, event, organization, or topic about which a specific opinion is expressed. The expressed opinion is usually about object or specific features '*f*' of the object". e.g. " *I don't like this phone" or " the battery life of this phone is not long".* The object can be represented as follow:

"An object *o* is represented with a finite set of features, $F = \{f1, f2, \ldots, fn\}$, which includes the object itself as a special feature. Each feature $fi \in F$ can be expressed with any one of a finite set of words or phrases $Wi = \{wi1, wi2, \ldots, wim\}$, which are *synonyms* of the feature, or indicated by any one of a finite set of feature indicators $Ii = \{ii1, ii2, \ldots, iiq\}$ of the feature." (B. Liu, 2010a)

- **Opinion:** This is a view, attitude or appraisal regarding an object from an opinion holder. *E.g. "I like Nano. However, I don't like the steering system of Nano"*

Based on the above details the semantic orientation can formally define as follows. The semantic orientation on an object *O* or its feature *f* specifies whether the opinion is positive '*Pos'*, negative '*Neg'* or neutral '*Nut'*. Semantic orientation is also known as polarity of sentiment, opinion orientation, or sentiment orientation.

The major part of this work is the lexical base semantic orientation of reviews at sentence level for domain independent sentiment classification for different genre.

Two sub tasks are performed in the analysis of the sentiment at the sentence level. They are as follows:

1. *Subjectivity classification*: Sentences is determined to be either a subjective sentence or an objective sentence.
2. *Sentence-level sentiment classification*: In the case of being subjective, it must be determined whether a positive opinion is expressed or a negative opinion is expressed.

These tasks are involved in the semantic orientation approach. The first task determines the semantic orientation of the opinions which were taken from reviews in the Review Analysis step, while the second sub task determines the overall semantic orientation of a review or sentence, or as based on the semantic orientation of the opinions contains. The final opinion strength is then decided to check all the part of speech in sentences with contextual information (B. Liu, 2010b).

Classification at the sentence level is more often than not an intermediate step. In the majority of applications, the object or features of the object which the opinions are on, is what is required to be known. All the same, the two sub tasks of classification at the sentence level remain vital because (1) they weed out those sentences which have no opinion, and (2) after gaining knowledge of what particular objects and features of the objects are mentioned in a sentence, this step helps in determining if the opinions on the objects and their features are negative or positive. While the majority of today’s researchers study both problems, there are some of them who devote their attention to only one. Since, the issues are regarding classification, typical supervised learning methods were used.

In this work, a rule based module is used to extract those sentences which contain opinions and subjective expressions or terms using SentiWordNet, WordNet or the subjectivity lexicon knowledge base. This work proposes a few steps for rule based lexicon method to determine the subjectivity of the sentences for the semantic orientation of the opinionated text for classification into positive, negative or neutral. From subjective sentences the opinion expressions are extracted and checked for their semantic scores using the SentiWordNet directory. The final weight of each individual sentence is calculated after considering the whole sentence structure, contextual information and word sense disambiguation. The steps are described as follows; Fig. 3.1 shows the broad view of sentiment analysis of the proposed method and all steps involved are briefly explained as follow.

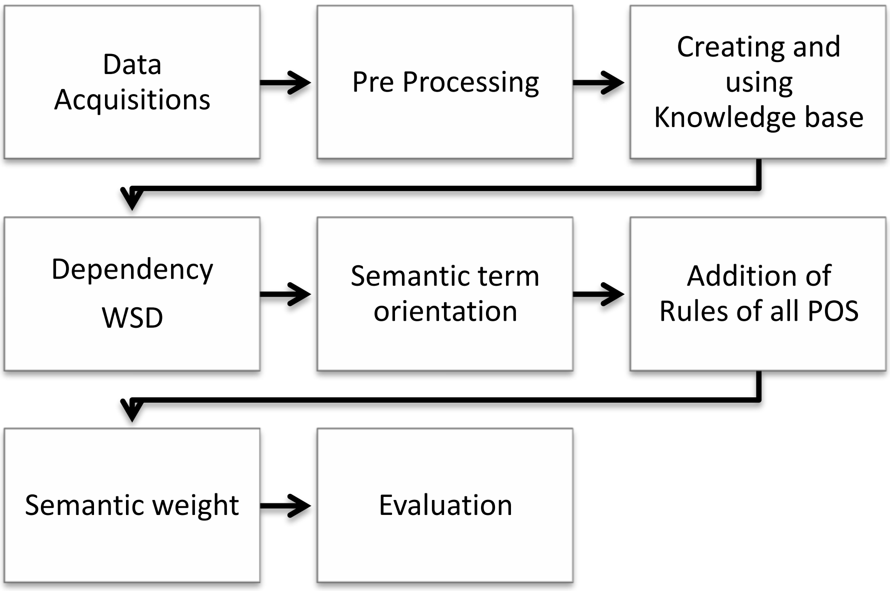

Data Acquisitions
Pre Processing
Creating and using Knowledge base
Dependency WSD
Semantic term orientation
Addition of Rules of all POS
Semantic weight
Evaluation

Figure 3.1 Steps of Proposed Method

**Data Acquisitions:** The first step is to collect the data which are available in the form of unstructured text (customer's reviews and comments) on web. There are many sites that contain feedback and reviews such as cnet.com, amazoon.com, skytrax.com, twitter.com.

- **Pre Processing:** Reviews/comments are split into sentences to form a Bag of Sentences (BOS). Noise is removed from sentences using spelling correction, convert special characters and symbols (phonetics) to their text expression. POS is used for tagging each word of the sentence and the position of each word is stored.
- **Creating and using Knowledge base:** A comprehensive dictionary (feature vector) of the important features with its position in the sentence is developed.

Sentences are classified into objective and subjective sentences using lexical feature.

- **Dependency (WSD):** The correct sense of the sentiment word is extracted using WordNet.
- **Semantic Term Orientation:** A lexical dictionary is used as a knowledge base and the polarity of the subjective sentence is checked as positive, negative or neutral.
- **Addition of Rules of all POS and Semantic Weight:** Polarity is updated using the sentence structure and contextual feature of each term in the sentence.
- **Evaluation:** Finally, results are evaluated summarized.

Sentence-level sentiment classification supposed to be the sentence expresses a single opinion from a single opinion holder. This supposition is only suitable for simple sentences with a single opinion, e.g., “The hotel was at good location” However, for complex sentences, a single sentence may express more than one opinion. For example, the sentence, “The hotel was nice and at good location but the room was so small that it felt like a prison cell”, expresses both positive and negative opinions. For “hotel” and “location”, the sentence is positive, but for “room”, it is negative (B. Liu, 2010a).

The sentiment bearing document can be represented as “A general opinionated document d which contains opinions on a set of objects {O1, O2… On} from a set of opinion holders {h1, h2, …, hm}. The opinions on each object obj are expressed on a subset Fj of features of Obj.”

The detailed architecture of the proposed model is shown in Fig. 3.2. The steps illustrated in Fig. 3.2 describe the overall process for semantic orientation for different genre and domains using lexical dictionaries. It has four major components: 1) collecting data (text), processing and removal of noise form text data. 2) Developing and using knowledge base which is the collection of lexical dictionaries.(3) processing of text data at sentence level using WSD for extraction of sentence sense (4) checking the polarity of each sentence according to sentence structure and deciding about opinion to positive, negative or neutral. The detailed description of each component is given in the subsequent sections.

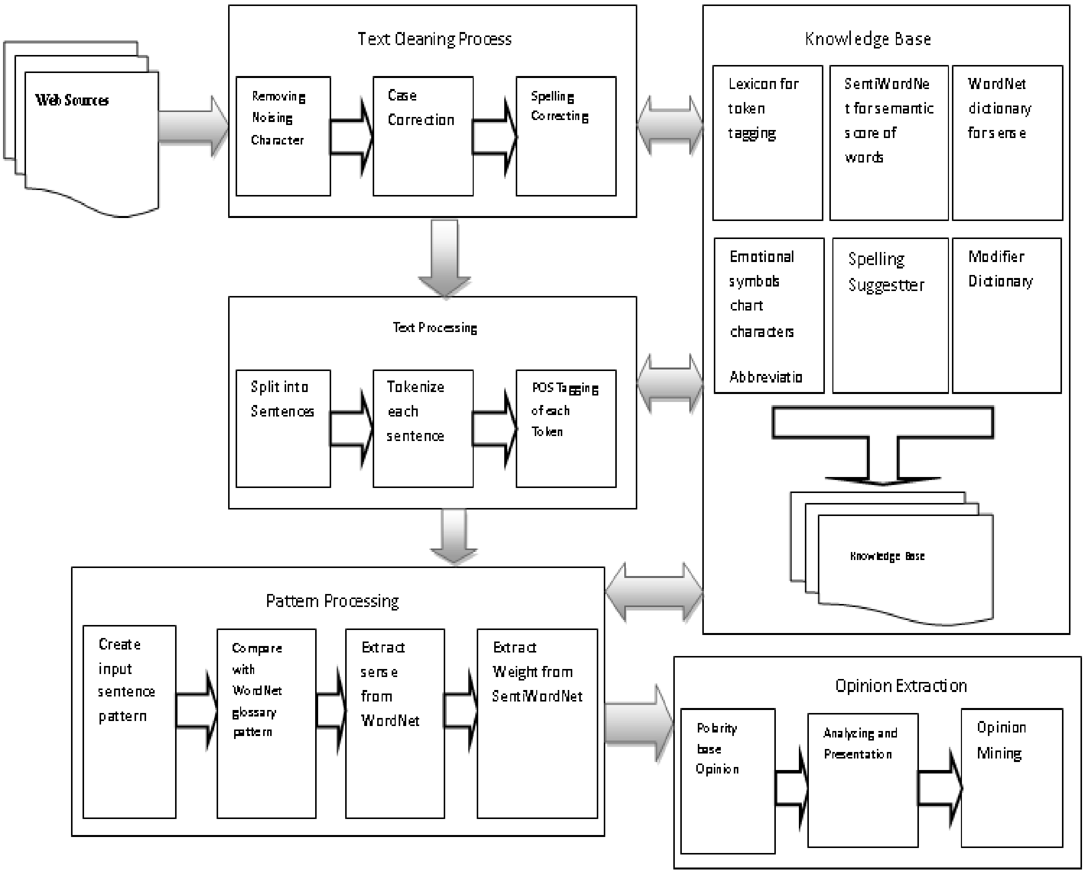

Web Sources
Text Cleaning Process
Removing Noising Character
Case Correction
Spelling Correcting
Knowledge Base
Lexicon for token tagging
SentiWordNet for semantic score of words
WordNet dictionary for sense
Emotional symbols chart characters
Abbreviatio
Spelling Suggestter
Modifier Dictionary
Text Processing
Split into Sentences
Tokenize each sentence
POS Tagging of each Token
Knowledge Base
Pattern Processing
Create input sentence pattern
Compare with WordNet glossary pattern
Extract sense from WordNet
Extract Weight from SentiWordNet
Opinion Extraction
Polarity base Opinion
Analyzing and Presentation
Opinion Mining

Figure 3.2 Details Architecture of Proposed Method

### 3.3 Data Acquisition and Pre-processing

The data collection and pre-processing is a first important step in online reviews and comments mining for sentiments classification. After data collection the pre-processing is used to reduce the complexity of the text and to select that features which are important for the classification process. Because of the lack of any regulations on reviews and blog sites, users do not often use formal structure of language when generating contents. There is a lot of diversity in such text like spelling mistakes, use of symbols and short abbreviated words, and the homonyms for similar sounding words. It becomes very difficult to extract important features in the presence of these errors for effective semantic orientation.

So the NLP techniques like parsing, part of speech tagging and WSD fail to perform with a high accuracy on noisy text. This work describes the pre-processing steps[2] needed in order to achieve high accuracy. The data pre-processing, and cleaning of the dataset which are essential for the resulting analysis are performed in the data preparation and pre-processing steps described in next chapter (L Dey & S. K. Haque, 2008) (Turney, 2002). One of the contributions of this work is to collect new datasets and process them to extract semantics of emotions, symbols and short abbreviated words for efficient sentiment classification and orientation.

### 3.4 Creating and using Knowledge base for Domain Independent Sentiment Classification

Effective sentiment orientation of text is dependent upon annotated words list with lexical semantic features. Dictionaries and corpora can be used to produce these lists of annotated words. Dictionaries are referred as lexicon based approach while corpora are called corpus based approach. Dictionaries or lexicon based approach is independent of domain and uses sentiment bearing words to classify text (Andreevskaia & Bergler, 2006). It takes advantage of totality and general nature of dictionaries like SentiWordNet and WordNet. In contrast, corpus based method is domain dependent to acquire these features. It is more sensitive towards domain changes and needs more time and efforts to obtain the desired training for a particular domain.

On the other hand lexicon based approach is advantageous over corpus based machine learning approach because it relies on the existing resources and does not depend on specific search facilities and it is domain independent. Corpus based methods need immense annotated training datasets. Some of these limitations can be overcome by utilizing dictionary based approaches; these approaches depend on existing lexicographical resources (such as WordNet) to provide semantic data in regards to individual senses and words (Esuli, 2008). Extraction of word sentiment information from dictionaries and lexical resources is important for feature acquisition, subsequent annotation, semantics and lexicography in the development of

[2] Next chapter describes data collection and pre-progressing steps, used to improve the performance of sentiment classification.

automatic extraction systems. These systems are able to automatically assign a semantic tag to each term or feature to classify sentiments into positive, negative or neutral category (Hu & B. Liu, 2004) (Leung & Chan, 2008) (Nathan, 2009) (Andreevskaia & Bergler, 2007) (Taboada, J Brooke, & Stede, 2009).

Applications based on the dictionary are often utilized with, not only two ways positive and negative classifications, as with corpus based methods, but also with three way classifications as positive , negative and. neutral. Sentiment can be assigned by dictionary based approaches to not only words, but also their senses. Thus, one of the main advantages of these methods is their appropriateness with sentiment classification not only at the sentence level but also check the sense in the text (J. Wiebe & Mihalcea, 2006) (Taboada, J Brooke, & Stede, 2009). Dictionary based methods for sentiment classification at the sentence level have no need of large corpora, or search engines having special functionalities. Rather, they depend on readily available lexical resources existing today such as ConceptNet, SentiWordNet and WordNet. They are able to compile comprehensive, accurate and domain independent word lists containing their sentiment and subjectivity annotated senses. Such lists provide a vital resource for sentence or text sentiment orientation and, because of early compilation; they are able to increase efficiency of sentiment classification at text and sentence level.

This work creates a knowledge base which conations SentiWordNet, WordNet and predefined intensifier dictionaries for domain independent polarity classification for positive, negative and neutral opinions. Sentiment words are usually classified into positive and negative categories. For this purpose, the semantic score of each opinion word is extracted using the SentiWordNet dictionary containing the semantic score of more than 117662 words. Then, the structure and associated words (which affect the weight of the opinion word) in the sentence is checked and the polarity updated accordingly. The main aspect of this work is a knowledge base for the contextual information of each part of speech in a sentence which really modifies the strength of the opinion. The knowledge base (calculates semantic strength for each sentence) contains negation words, enhancers, reducers, model nouns, context shifters and other intensifiers with their semantic scores.

This work combines and interlinks the lexical dictionaries (WordNet, SentiWordNet, intensifiers etc) to make a knowledge base and used to extract the sense of terms, and semantic score, as described in the next section. Below are the description of different dictionaries from the literature and their usage in this work.

### 3.4.1 WordNet

The research efforts of the Department of Linguistics and Psychology at Princeton University for better understanding of English language and semantics resulted in WordNet lexicon. It is a complete lexicon where English language terms and semantics can be searched and retrieved as per their conception and semantic affiliations. At its third version, WordNet is available as a database, searchable via web interface or via a variety of software APIs, providing a comprehensive database of over 150,000 unique terms organised into more than 117,000 different meanings (WORDNET, 2006). WordNet also grew with extensions of its structure applied to a number of other languages (WORDNET, 2009) (B Pang & L Lee, 2008) (Esuli, 2008) (Fellbaum, 1998).

WordNet is an electronic lexical dictionary. It is made up of English words put into a large lexical database which is most often used in classification of text, semantic orientation, computational linguistics and natural language processing. It contains sets of cognitive synonyms called synset made up of various parts of speech such as nouns, adjectives, verbs and adverbs which express their own distinct concept. These synset are interlinked by conceptual-semantic and lexical relations. WordNet is freely and publicly available on internet (Fellbaum, 1998). Due to its effectiveness in semantic information extraction and its reach semantic and syntactical information about words, an important measure and source is adopted in this work as to extract the sense and also used in the FS process. Psycholinguistics and computational theories of human lexical memory are the main driving factors in the design process of WordNet. Part of speech such as verbs, adverbs, nouns and adjectives are sorted out into sets of synonyms called as synset. Every synset represents one lexical concept where each word used in different sense has its respective sense and concepts in glossary. The above mentioned parts of speech are further organized by WordNet into sets of lexicographer source file as per their syntactic class (Ohana, 2009). Nouns and verbs are classified according to their semantic sense while adverbs are stored in a separate file. Adjectives are stored in a different file depending upon their descriptive and relational behaviour. A list of synset for every part of speech is contained in a source file, which comprises synonymous word form, related pointers and other information. These related pointers include hyponymy /homonymy, antonyms, entailment, and meronymy / holonymy (Esuli, 2008) (Ohana, 2009).

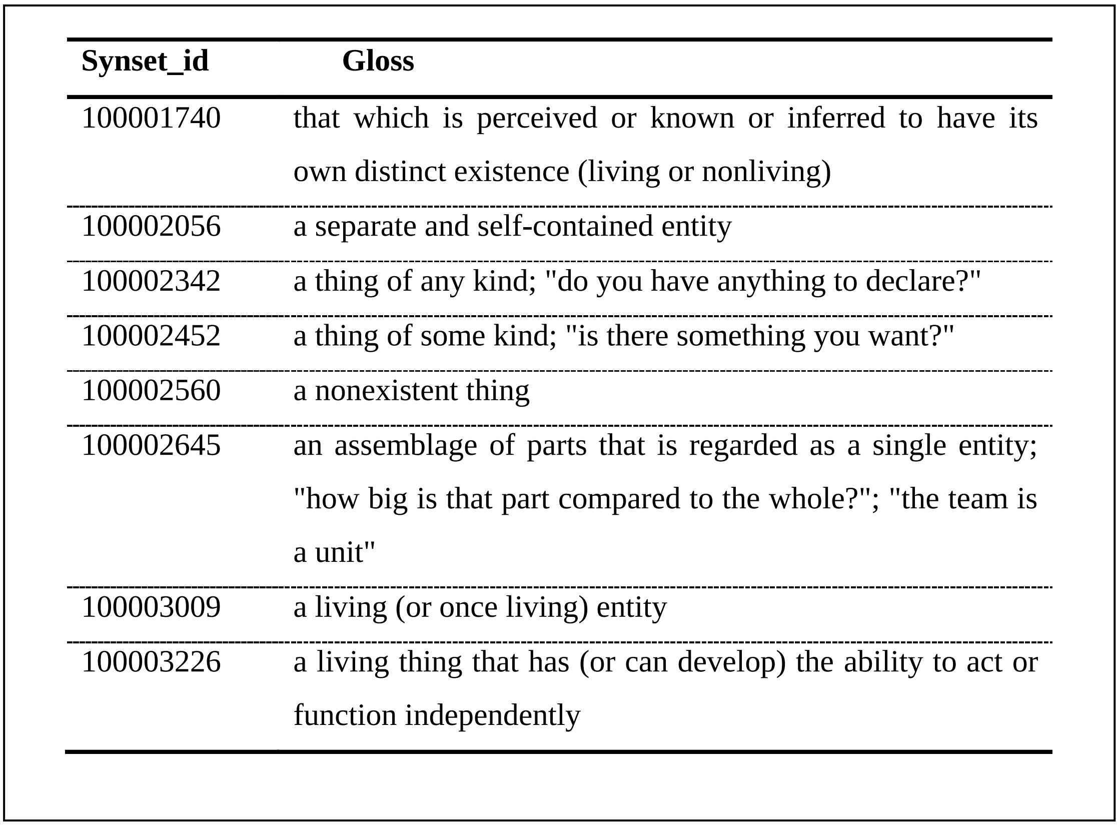
Polysemous word forms, are those having or characterized by many meanings, appear in more than one synset. A textual glossary for a synset is usually maintained by a lexicographer who helps in interpreting the true semantics for synonymous words and their usage. In this work a textual gloss as a dictionary of concepts is included to use for correct sense extraction of the words used in the sentences using pattern of that sentence with the close match of the WordNet gloss concept patterns. Table 3.1 shows the sample of synset information of WordNet for all parts of speech.

Table 3.1 WordNet Synset Information

| **Synset_id** | **W_num** | **Word** | **Ss_type** | **Sense_number** | **Tag_count** |
|---|---|---|---|---|---|
| 100001740 | 1 | entity | n | 1 | 11 |
| 100002056 | 1 | thing | n | 12 | 0 |
| 100002342 | 1 | anything | n | 1 | 0 |
| 100002452 | 1 | something | n | 1 | 0 |
| 100002560 | 1 | nothing | n | 2 | 0 |
| 100002560 | 2 | nonentity | n | 3 | 0 |
| 100002645 | 1 | whole | n | 2 | 0 |
| 100002645 | 2 | whole_thing | n | 1 | 0 |
| 100002645 | 3 | unit | n | 6 | 0 |
| 100003009 | 1 | living_thing | n | 1 | 1 |
| 100003009 | 2 | animate_thing | n | 1 | 0 |
| 100003226 | 1 | organism | n | 1 | 9 |

Table 3.2 contains the sample information about each synset with their corresponding gloss details.

Table 3.2 WordNet Gloss Information



| Synset_id | Gloss |
|---|---|
| 100001740 | that which is perceived or known or inferred to have its own distinct existence (living or nonliving) |
| 100002056 | a separate and self-contained entity |
| 100002342 | a thing of any kind; "do you have anything to declare?" |
| 100002452 | a thing of some kind; "is there something you want?" |
| 100002560 | a nonexistent thing |
| 100002645 | an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit" |
| 100003009 | a living (or once living) entity |
| 100003226 | a living thing that has (or can develop) the ability to act or function independently |

In Table 3.3 all the information about each word was combined i.e. their sense no, synset id extracted from Table3.1 and their meaning or details about the term sense from glossary Table 3.2. All the information of are sorted and linked with SentiWordNet to extract the correct sense for each part of speech. The sense-no and synset-id is use to link SentiWordNet for the information about sense of each word score.

Table 3.3 WordNet Sense Gloss Information

| Synset_id | Word | Sense_number | Gloss |
|---|---|---|---|
| 100001740 | entity | 1 | that which is perceived or known or inferred to have its own distinct existence (living or nonliving) |
| 100002056 | thing | 12 | a separate and self-contained entity |
| 100002342 | anything | 1 | a thing of any kind; "do you have anything to declare?" |
| 100002452 | something | 1 | a thing of some kind; "is there something you want?" |
| 100002560 | nothing | 2 | a nonexistent thing |
| 100002560 | nonentity | 3 | a nonexistent thing |
| 100002645 | whole | 2 | an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit" |
| 100002645 | whole_thing | 1 | an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit" |
| 100002645 | unit | 6 | an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit" |
| 100003009 | living_thing | 1 | a living (or once living) entity |
| 100003009 | animate_thing | 1 | a living (or once living) entity |
| 100003226 | organism | 1 | a living thing that has (or can develop) the ability to act or function independently |
| 100003226 | being | 2 | a living thing that has (or can develop) the ability to act or function independently |

### 3.4.2 SentiWordNet

SentiWordNet is sentiment analysis lexical resource made up of synset from WordNet, a thesaurus-like resource; they are allocated a sentiment score of positive, negative or objective. These scores are automatically generated using the semi-supervised method which is described in (Esuli & Sebastiani, 2006). It is also available freely for research purpose on web. The possible POS which having their score used in SentiWordNet is given in Table 3.4 with observations.

Table 3.4 Inside POS Information of SentiWordNet

| SentiWordNet_Abrv | POS_Abbrivation | POS_Name |
|---|---|---|
| n | NN | Noun |
| a | JJ | Adjective |
| v | VB | Verb |
| r | RB | Adverb |

SentiWordNet is one of the sources of sentiment analyses. It is a semi-automatic way of providing word/term level information on sentiment polarity by utilizing WordNet database of English terms and relations. Each term in WordNet database is assigned a score of 0 to 1 in SentiWordNet which indicates its polarity. Strong partiality information terms are assigned with higher scores whereas less subjective terms carry low scores. How opinion information appears in SentiWordNet, is shown in table 3.5.

Table 3.5 SentiWordNet Dictionary information

| POS | ID/Offset | PosScore | NegScore | SynsetTerms |
|---|---|---|---|---|
| a | 10073761 | 0.125 | 0.625 | Strained, forced constrained |
| n | 10036762 | 0.375 | 0.125 | Feat, exploit, effort |
| v | 311113 | 0.25 | 0.25 | Slur, dim, blur |
| r | 139759 | 0.125 | 0.125 | Unsuitably, inappropriately |

In SentiWordNet each set of synonymous terms is assigned with three numerical scores ranging from 0 to 1 which indicates its objectiveness i.e. positive and negative polarity. One of the key features of SentiWordNet is that it assigns both positive and negative scores for a given term according to the following rule (Esuli & Sebastiani, 2006). The dictionary database information is described as follows.

- The pair (POS, offset) uniquely identifies a WordNet synset. Numeric ID called offset associated with POS uniquely identified a synset in a database.
- The values PosScore and NegScore are the positivity and negativity score assigned by SentiWordNet to the synset
- The objectivity score can be calculated as

  Pos(s) + Neg(s) + Obj(s) = 1 (Eq. 3.1)

  ObjScore = 1 - (PosScore + NegScore) (Eq. 3.2)
- Last column reports the terms, with POS and sense number, belonging to the synset (separated by spaces).

(Where NegScore= negative Score, PosScore=Positive Score, Pos(s) = Positive score of synset s., Neg(s) = Negative score of synset s., Obj(s) = Objectiveness score of synset s.)

As described in the above section, SentiWordNet terms are sorted according to their meaning, expression or the part of speech the term is used in a given sentence. According to (Ohana, 2009) the opinion score presented by (Esuli & Sebastiani, 2006) is illustrated in Table 3.6 that shows how it is affected by the parts of speech.

Table 3.6 POS Score Information adopted from (Ohana, 2009)

| **Part of Speech** | **% Synsets with Objectiveness= 1** | **Average Objective Score** | **Average Pos. Score** | **Average Neg. Score** |
|---|---|---|---|---|
| Noun | 83.50 % | 0.944 | 0.022 | 0.034 |
| Verb | 81.05 % | 0.940 | 0.026 | 0.034 |
| Adverb | 32.97% | 0.698 | 0.235 | 0.067 |
| Adjective | 44.71% | 0.743 | 0.106 | 0.151 |

From the (Table 3.6), it can be seen that nouns and verbs are mainly objective in nature with little or no polarity. Weaker association of nouns and verbs with other terms in WordNet, carrying positive or negative bias, has been realized in the building process of SentiWordNet. Adverbs and adjectives are such part of speech which possesses the highest percentage of terms with positive subjective score. Adjectives or adverbs (modifiers) are more common in expressing subjective opinion than verbs and nouns, which are more frequently used in objective scenarios. One more observation about adverbs is that although they own substantial polarity weight (only 32.97% of terms contain no subjective bias) yet their average score is significantly positive (Ohana, 2009).

After analysing the database structure of SentiWordNet, this section explores key aspects that need to be taken into consideration when designing features to be used in sentiment classification. As illustrated in Table 3.6 the data in SentiWordNet is grouped in terms of part of speech and synset, depending upon their objectiveness in which they are grammatically used. Source documents are classified for extracting the information on POS so that accurate SentiWordNet scores can be applied. Part of speech tagging algorithm is utilized on the source document to automatically sort the words into groups as per their part of speech. A relevant tag is assigned to each term, such as verb, noun, adjective etc, which specifies its role in the sentence. POS taggers and their use within opinion mining and sentiment analysis research are discussed in next chapter. The details about the dictionaries are shown in Appendix C.

### 3.4.3 Other Dictionaries

#### *3.4.3.1 Intensifiers*

Intensifiers can be categorized into two major types, depending on their polarity: amplifiers (e.g., *very*) and downtoners (e.g., *slightly*) (Quirk, Greenbaum, Leech, & Svartvik, 1985). The amplifiers increase the semantic strength of contiguous lexical items, whereas the downtoners decrease the semantic score of the sibling term in a sentence. (Polanyi & Zaenen, 2006) uses contextual shifter with intensifiers by simple addition and subtraction of fixed values. This method is limited to a small range of intensifiers within the same category, which can be considered as one of the drawbacks of this method. (Juilen Brooke, 2009) formulated a dictionary for intensifiers which has been used to calculate semantic orientation.

In this work the intensifiers are used with modifiers with their percentage score which alter the semantic weightage of the associated opinion terms. Due to which an obvious and valiant decision can be made to classify the terms into positive, negative or neutral.

### *3.4.3.2 Modifiers (Enhancer and Reducer)*

Modifiers can be defined as such words which enhance or reduce the strength of polarity of a sentiment term or expression in a sentence or document. If there is a modifier word in a sentence (e.g. *Slightly, Somewhat, Pretty, Really, Very, Extremely, (the) most),* closer to the sentiment term, then its polarity will be recalculated by referring to its weightage dictionary. The score of the opinion word will be affected in the sentence by checking its position in the sentence. e.g., in the sentence *"The staff at the reception was very nice and good",* the modifier "*very*" is enhancing the weight of the nearest opinion word, i.e. nice. The uniqueness of this module is that, if the modifier is an adverb and its semantic score exists in the SentiWordNet dictionary then it will extract that score and it will be added/subtracted with the weight of the sentiment term. Otherwise it will refer to the enhancer and reducer score dictionaries for extraction of the respective score.

In the example, *"The staff at the reception was very nice and good",* score calculated by the module is given in equation 3, which for the particular sentence is equal to *1.75.* In this sentence the sentiments about the *staff (employee/people)* at the *reception (location)* are *"NICE" and "GOOD" (sentiment/opinion terms)* while "VERY" is the modifier which is used to enhance the semantic strength of the adjacent opinion term "Nice" as shown in Eq. 3.3.

Semantic weight = Nice + very and Good = (0.875 +0.25) + 0.625 = 1.75 (Eq. 3.3)

Each word in the sentences is stored with their POS tag, respective position in the sentences and WordNet and SentiWordNet referral tag to extract the semantic score as shown in Table 3.7. The description of the above sentence semantic score extracted from the lexicon dictionary is shown in Table 3.8.

Table 3.7 POS-Types and Their Abbreviations Used in SentiWordNet

| POS_ID | POS_Name | POS_Abbrivation | SentiWordNet_Abrv |
|---|---|---|---|
| 1 | Noun | NN | n |
| 2 | Adjective | JJ | a |
| 3 | Verb | VB | v |
| 4 | Adverb | RB | r |
| 5 | Nouns | NNS | nns |
| 6 | Adjectives | JJS | a |
| 7 | Verbs | VBS | v |

Table 3.8 shows the details of the sentence *"The staff at the reception was very nice and good".*

Table 3.8 Extraction of Words Score Using SentiWordNet

| Word | POS_ID | POS-Score | NEG-Score | Position |
|---|---|---|---|---|
| staff | 1 | | | 2 |
| reception | 1 | | | 5 |
| was | 3 | | | 6 |
| very | 4 | 0.25 | | 7 |
| nice | 2 | 0.875 | 0 | 8 |
| good | 2 | 0.625 | 0 | 10 |

Table 3.10 shows the semantic scores extracted from knowledge base which decide about the final opinion strength. The examples show the semantic score of different words with their part of speech tag information and position in the sentence.

In the sentence *"The stay was great and the meal service was very good*" the adverb very enhance the strength of the opinion word good shown in Table 3.9.

Table 3.9 Extraction of Words Score Using SentiWordNet

| Word | POS_ID | POSScore | NEGScore | Position |
|---|---|---|---|---|
| stay | 1 | | | 2 |
| was | 3 | | | 3 |
| great | 2 | 0.25 | 0.125 | 4 |
| meal | 1 | | | 7 |
| service | 1 | | | 8 |
| was | 3 | | | 9 |
| very | 4 | 0.25 | 0 | 10 |
| good | 2 | 0.625 | 0 | 11 |

For further details see the modifier list with their corresponding semantic weights in Appendix C.

#### *3.4.3.3 Modifiers of Certain Nouns*

Some nouns can be used as modifiers in a sentence effecting its opinion expression and polarity. For example words like (*a (little) bit of, a few, Minor, Some, a lot, Deep, Great, a ton of*) are nouns but they effect the sentence polarity, hence can be considered as modifiers. If such words occur in a sentence, recalculation of its polarity is recommended. This recalculation can be done by using dictionary of weights of words/terms by assigning weights to each term accordingly. This work implements the intensifier of modifier using semantic weighted score which reflects its enhancing or reducing nature and modifies the semantic strength of the adjacent term. This can be applied to adverbs, adjectives and verbs for the final semantic weightage calculation.

There are some types of adjectives like total, huge etc that have no semantic orientation at their own, but they contribute to the word following them. e.g. *"the management at the entrance was a total failure."* In this sentence the word "total" is used with word "failure", which is emphasizing on the extent of failure. The word "total" itself is an adjective but when it is used with the following word "failure" it is increasing its polarity. There are different types of such adjectives added to our intensifier dictionary, which have semantic scores and affect the nouns if they appear before them. These are treated as intensifiers e.g. the total failure is worse than just failure.

### *3.4.3.4 Negation*

Negation words are obviously of some importance as when they appear in a sentence they often change opinion orientation. One example would be in the sentence "*I don't like this airline*" which is definitely negative. On the other hand, not all appearances of negation words result in a negative opinion; this is why it is vital that these words are handled with care. Take the case of term "not" in the phrase "*not only ... but also*"; in this context 'not', though it is negative, does not change the direction of the orientation.

Negation words reverse the polarity of opinion words by checking their position in a sentence. The words like Not*, Never, N't, Doesn't, Can't, Nor, Don't, Wouldn't, No, etc* are usually used in a negative scenario. If these terms are not accurately recognized by the system in a sentence then the result will be opposite.

So, for the recognition of semantic expression in a sentence, the WSD is used to extract the exact or nearest semantic score of the opinion expression.

### *3.4.3.5 Contact Shifter*

The term "valence shifters" is the most widely used for this category of words and expressions. Occasionally they are also called polarity modifiers and polarity shifters.

There are a few types of context shifters to populate the knowledge base with semantic scores; they are followed by some specific rules for semantic weight extraction from sentences and are shown below.

- The contact shifter (but*, except, however, only, although, though, while, whereas, etc*.)
- Contradictory nature contact shifter (*Although, Despite, While*)
- Mobilizing or modal contact shifter (*Would, Should*)

- Pre-Supposition contact shifter (*Miss, forget, refused, assumed, hard, harder, less, etc*.)

If sentences have any such type of words, then the polarity will be recalculated by checking their position in respect to the opinion expression because these words affect the polarity of the opinion word. The negation words reduce its effect to nothing. The examples of such sentences are as follows.

a. "Only sampled the breakfast but that was very nice with quite a lot of variety"

b. "The outside of the building did look scruffy but the lobby was really nice"

c. "Therefore I would not recommend the jolly hotel to anybody"

d. "Despite all these minor and trivial problems details this is a well kept hotel"

e. "If you are just looking for a basic but comfortable stay this should be sufficient"

Table-3.10 shows the overall example of intensifier dictionary with their corresponding semantic weights which is used in the decision of the semantic orientation of the opinion term and the overall semantic orientation of the sentence.

Table 3.10 Sample of Intensifiers Semantic Weight Dictionary

| **Adverb** | | **Nouns and Verb** | | **Modifier** | | **Modifier Assured Nouns** | | **Negation** | | **Contact Shifter** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Word | Weightage | Word | Weightage | Word | Weightage | Word | Weightage | Word | Weightage | Word | Weightage |
| Foolishly | -2 | shame | -3 | very | 3 | few | 2 | not | -1 | but | 5 |
| absolutely | 4 | inspire | 2 | highly | 4 | some | 2 | no | -1 | accept | 4 |
| Purposely | 2 | hate | -4 | totally | 3 | deep | 3 | didn't | -1 | although | 3 |
| obviously | 3 | fabricate | -2 | fully | 3 | great | 4 | Don't | -1 | while | 4 |

## 3.5 Feature and Opinion Word Position Extraction

The algorithm for sentiment classification uses opinion terms or expressions to determine polarity of sentences based on contextual information and sentence structure. The position of each word in a sentence is important for the semantic orientation and correct pattern extraction for word sense disambiguation. Product features and opinion words are also extracted from tagged sentences using the word position. This work selects features from the list at run time after suggesting the most frequent features extracted from the opinionated sentences. To extract opinion words from sentences, the first focus is on finding features that emerge explicitly as nouns or noun phrases in reviews. The following steps are used.

- Use POS tagger to tag every word of the sentence and store each word position with its assigned tag.
- Collect the nouns, noun phrases and adjectives with their positions.
- Noun phrases are observed as product features.
- For each sentence in the review, if it contains any feature word, extract any nearby adjective and consider such adjectives as opinion words.
- Adjectives and/or adjective preceded by adverbs are observed as opinion words.
- Frequent product features are selected from key noun phrases.

## 3.6 Word Sense Disambiguation (WSD)

WSD is an important step in semantic orientation to extract the correct sense of a term or expression in a sentence. Sentiment analysis, in most cases, relies on lexicons of words that may be used to express prejudice or subjectivity. These works do not address the peculiarity of different senses of a word in a way that its true sense is not categorized. Moreover, subjective lexicons are not accumulated as word meanings; rather they are compiled as lists of keywords. In most cases, these keywords have both opinionated and factual senses. Depending upon the contextual appearance, some degree of positive or negative polarity can be experienced even with the purely subjective sense (Esuli, 2008) (Ohana, 2009).

The contribution of this work is to check the WSD using unsupervised approach using the existing public resources. The proposed method extracts the semantic pattern of the desired sentence using the opinion expression position in the sentence. Then, all possible patterns for that opinion expression for all possible senses are extracted based on the WordNet glossaries; the system locates an exact pattern match of the desired sentence and extracts the sense number from the WordNet synset. The semantic score for that sense number is extracted from SentiWordNet, which gives efficient results. If patterns are not exactly matched, then it checks for the nearest pattern and the score of that nearest pattern is extracted from SentiWordNet. The results of proposed process are described in Tables- 3.13, 3.14, 3.15, 3.16. In Table-

3.13 reviews are split into sentences and only subjective sentences are selected for semantic orientation.

There comes an issue while evaluating a particular term in SentiWordNet as to what specific WordNet synset this terms belongs and what would be its score. Consider the example for the term “Prosperous”, with four synsets in WordNet.

Table 3.11 Single Term with multiple score in SentiWordNet

| POS | ID/Offset | PosScore | NegScore | SynsetTerms | Gloss |
|---|---|---|---|---|---|
| a | 163948 | 0.25 | 0.125 | wrapped#2 intent#1 enwrapped#1 engrossed#1 captive#2 absorbed#1 | giving or marked by complete attention to; "that engrossed look or rapt delight"; "then wrapped in dreams"; "so intent on this fantastic...narrative that she hardly stirred"- Walter de la Mare; "rapt with wonder"; "wrapped in thought" |
| a | 177547 | 0.75 | 0 | prosperous#4 lucky#3 golden#6 favourable#4 favorable#3 | presaging or likely to bring good luck; "a favourable time to ask for a raise"; "lucky stars"; "a prosperous moment to make a decision" |
| v | 2641463 | 0 | 0.25 | wait#2 hold_off#2 hold_back#4 | wait before acting; "the scientists held off announcing their results until they repeated the experiment" |
| a | 2386612 | 0.125 | 0.75 | short#3 little#6 | low in stature; not tall; "he was short and stocky"; "short in stature"; "a short smokestack"; "a little man" |

In the above example, (Table.3.12), four meanings can possibly be referred to the adjectives “wrapped″ and “Prosperous”. The question here is as to what meaning this word is referring in a particular sentence and what particular score, positive or negative, should be assigned to it in SentiWordNet. Determining which synset needs to be applied on a specific context is analogous to the problem of WSD. So term sense extraction according to the structure and contents of the sentence is challenging task. A technique is proposed here for the extraction of term sense extraction according to sentence structure for effective sentiment classification.

Table 3.12 Semantic Weight Assigned to Sentences

| SEN_ID | Sentence | Weight |
|---|---|---|
| 1 | KUL-BKK A320 pretty modern cabin crew okay need to polish on their smiles and social skills | -0.3 |
| 2 | Nice flight cheap price | 1.75 |
| 3 | For the price I paid no complaints | 0.625 |
| 7 | The one on the way in was really dirty | -0.5 |
| 8 | On the way out of Bali the plane seemed brand new it was clean too | 0.875 |
| 9 | AirAsia service was bad on the way in but great on the way out | -0.525 |
| 10 | The flight attendants seemed to ignore us on the way in but were kinder on the way out | 0.1 |
| 11 | One thing I've noticed though is the lack of safety cards along with the magazine and Buy-on-board list in every seat are we supposed to share safety cards | 0.25 |
| 14 | AirAsia offers good value for money considering the ticket prices but is definitely not my carrier of choice even for short flights | 0.225 |
| 15 | But their cheap tickets allowed us to stay at a better hotel than we would have if wed flown a full-fare airline KUL-TWU SDK-KKI and KKI-SIN | 0.475 |
| 16 | Overall a good experience | 0.325 |
| 17 | Only downside was not receiving the meals we had prepaid for 3 months in advance when booking the tickets | 0.125 |
| 18 | This is a major inconvenience for vegetarians who have nearly no other choice to get a meal on-board because the meal selection in general is very poor on Air Asia they are usually out of stock on most items you ask about | 0.625 |
| 20 | There was no way to reassign your seat using online check-in two days before the flight not even if you are willing to pay for it | 1.25 |
| 23 | Check in was fine and boarding not a problem either Seats were more than adequate and the cabin staff were as helpful as they needed to be | -0.625 |
| **25** | **To be honest for a low cost airline this was actually a fantastic flight** | **0.625** |

In this work, experiments with different data sets have been performed for semantic orientation using WSD technique and its impact on the complexity of the data sets have been addressed. As a first step, part of speech tagging is used to obtain some level of disambiguation for extracting semantic scores from SentiWordNet. However if there occurs a multiple sense within the same part of speech a simpler approach can be used to assign scores as to evaluate WSD, this approach is used to extract the sentence contextual pattern and refer this to WordNet glossy for the correct sense extraction and the same sense score is selected from the SentiWordNet as described in Table 3.12, 3.13 and 3.14. It is significant to the overall performance of this method, however future developments of the SentiWordNet model taking into account more sophisticated techniques of WSD could yield positive results.

From the Table 3.12 sentence number 25 is taken with its semantic weight. Table-3.14 shows the semantic scores of each term in the sentence. The matching algorithm is applied to this sentence to extract the sense of the semantic term "fantastic" from WordNet. The proposed method extracts the pattern for the sentiment term and matches it with WordNet synset terms; there are four possible senses of the word "fantastic" with both negative and positive scores, but here the sense with the positive score is to be extracted. So, the system exactly extracts the positive score for the term "fantastic" as 0.375 from SentiWordNet, as shown in Table 3.13 and 3.1 4. The process is described in Algorithm 3.1.

Table 3.13 Description of Terms Weight

| **25** | **To be honest for a low cost airline this was actually a fantastic flight** | | | **0.625** |
|---|---|---|---|---|
| Word | POS_ID | POS-Score | NEG-Score | Position |
| To | 1 | | | 1 |
| be | 3 | 0.25 | 0.125 | 2 |
| honest | 2 | 0.75 | 0 | 3 |
| low | 2 | 0 | 0.25 | 6 |
| cost | 1 | | | 7 |
| airline | 1 | | | 8 |
| was | 3 | | | 10 |
| fantastic | 2 | 0.375 | 0.375 | 13 |
| flight | 1 | 0.25 | 0 | 14 |

The tag sentence is ("[To/NN be/VB honest/JJ for/IN a/DT low/JJ cost/NN airline/NN this/DT was/VBD actually/RB a/DT fantastic/JJ flight/NN ./. ") and the pattern extracted is VBD//WRD-//NN, which matches the sense number 5 in WordNet, and the semantic score of sense number 5 for the term "fantastic" is 0.375 as shown in Table 3.14.

So, it extracts the positive score of the term “fantastic” which is an accurate semantic score according to the sentence structure.

Table 3.14 SentiWordNet Semantic Score for Term Fantastic

| ID1 | POS | ID | Pos-Score | Neg-Score | Synset-Terms | Gloss |
|---|---|---|---|---|---|---|
| **9869** | **a** | **1796452** | **0.375** | **0.375** | **fantastic#5** | **extravagantly fanciful in design, construction, appearance; "Gaudi's fantastic architecture"** |
| 10611 | a | 1936778 | 0 | 0.625 | fantastical#1<br>fantastic#4 | existing in fancy only; "fantastic figures with bulbous heads the circumference of a bushel"- Nathaniel Hawthorne |

There are still problems with semantic scores as seen in Table-3.13. e.g. the word “low” has a negative score when it is used alone, but in sentence number 25 Table-3.12, its sense appears to be positive; “Low-cost”. To tackle this problem bigram word or term extraction method is proposed for future step.

Table 3.15 WordNet Sense Patterns

| Word | Sense No. | Pattern |
|---|---|---|
| fantastic | 2 | /IN-/NNS-/DT-WRD-/NN-/NN-/DT |
| **fantastic** | **5** | **/NN-/VBP-/VBD-WRD-/NN-/JJ-/NNS** |
| fantastic | 4 | /IN-/JJ-RB-WRD-/NNS-/IN-/NN |
| **fantastic** | **3** | **/JJ-/JJ-/DT-WRD-/NN-/IN-/PRP$** |

```
Step 1:
Function:  DES_PATTERN
INPUT:
        TAG_SENT – POS Tagged Sentence
OUTPUT:
DES_PATTERN – Desired pattern of tagged sentence consisting NN JJ RB VB
PROCESS:
        SELECT only NN JJ RB VB from TAG_SENT
        Place WRD at NN JJ RB VB place
        CONCATINATE tags of k+3 and k-3 with WRD
        RETURN DES_PATTERN
Step 2:
Function:  SELECT_PATTERN
INPUT: SENT_SENTIM_WORD – Sentiment word which has different senses in
WordNet
OUTPUT: SLT_PATTERN – Extract pattern from WordNet glossary of INPUT
        USING:
        WORDNET – Dictionary for extracting pattern from WordNet glossary of
        INPUT
PROCESS:
        SELECT glossary of INPUT
        CREATE pattern of INPUT using k+3 and k-3 with WRD
        RETURN SLT_PATTERN
Step 3:
IF the DES_PATTERN is similar   SLT_PATTERN THEN
        Function:  EXTRACT_SENSE
                INPUT:
                SLT_PATTERN – Extracted pattern from WordNet glossary
                OUTPUT:
                SENSE_NO – Extract sense number of INPUT
                USING:
                WORDNET - For sense extraction
        PROCESS:
        SELECT SENSE_NO of SLT_PATTERN from WORDNET
ELSE the DES_PATTERN is not similar   SLT_PATTERN THEN
        Function:  NEAREST_PATTERN
        INPUT:
        SLT_PATTERN – Extracted pattern from WordNet glossary
        DES_PATTERN – Desired pattern of tagged sentence consisting NN JJ RB VB
        OUTPUT:
        SENSE_NO – Extract sense number of INPUT
Step 4:
        SENTIM_WORD_SCORE:= Extract positive negative score from the
        SentiWordNet according to SENSE_NO
        IF the POSITIVE_SCORE is greater than NEGATIVE_SCORE THEN
                SENTIM_WORD_SCORE:= POSITIVE_SCORE
        ELSE the POSITIVE_SCORE is less than NEGATIVE_SCORE THEN
                SENTIM_WORD_SCORE:= NEGATIVE_SCORE
```

Figure 3.3 Algorithm for POS and Word Sense Disambiguation

## 3.7 Contextual Semantic Orientation of Sentences

In this section, the process of assigning weight to each sentence is described, which decides whether the review is positive, negative or neutral. Rule based method is used to check the polarity of sentences and the contextual information at the sentence level. The process is used to extract the contextual information from the sentence and calculate its semantic orientation using SentiWordNet, WordNet and predefined intensifier semantic score dictionaries. From the results, it is clear that contextual information and consideration of sentence structure for correct sense extraction is very important for useful sentiment classification. The main contributions of this work are sentence level semantic pattern extraction for WSD, by considering all POS of the sentence for semantic orientation and generic sentiment polarity classification (domain independent). However, the limitations of this work include the dependency on a lexical dictionary and limited WSD. The system is evaluated on several datasets and online comments and its results are outperformed. The following process shows the overall polarity calculation of the proposed method to split the sentence structure.

**Step1-** Split the reviews into sentences; a Bag of Sentences is created (BOS). Assign each review and each sentence an –id.

**Step-2.**Clear noise from text and apply POS.

**Step-3.**Check each sentence and finds the required word (WRD), if it exists in the sentence, then extract its position in the sentence. X= Pos_WRD. Check the opinion word (OW) in the sentence by calculating its position as (X-5) and (X+5) in the sentence. If found, then mark it as an opinion sentence and assign the word to N. (N=OW)

**Step-4.**Classify sentences into subjective and objective on the basis of the opinion expression extracted in the previous step.

**Step-5.**Calculate its word semantic orientation and assign a weight to this word from the SentiWordNet dictionary. (OW←SEM_SCOR).

**INPUT:**

***CORPUS*** - Input corpus

**OUTPUT:**

*SENT_SENTIM_WORDS [ ]* - Stores the sentiment words list extracted from SENT and it's Positive Negative Score accordingly

*SENT_SENTIM_NONSENTIM [ ]* - Stores sentiment and non sentiment SENT

*SENT_TSCORE [ ]* - Stores the total strengths of positive and negative identified in SENT

*SENTIM_WORD_SCORE* – Store the sentiment word score extracted from SentiWordNet

*REVIEW_SCORE* - Store polarity of positive negative and neutral reviews

**METHOD:**

**Step 1:**

```
    REVIEWS: = Split Corpus
    SENT: = Split Reviews
    REW_ID:= Assign ID to each Review
    SENT_ID:= Assign ID to each sentence
    WORD_LIST:= list of words in sentence
    WORD_POSITION: = Position of each word in a sentence
```

**Step 2:**

***CALL***

```
    Function: NOIS_CLR
    Function:  POS_TAG
```

**Step-3**

***Extract Opinion Sentences***

```
    OW=Find opinion word and position in sentences
    OW-P= Position of OW
```

**Step-4:**

```
    Sub-sent = opinion express sentences
    Subj –sentences = non –opinion sentence
```

**Step-5:**

```
    Function:  DES_PATTERN
    Function:  SELECT_PATTERN
    Function:  EXTRACT_SENSE
    Function:  NEAREST  PATTERN
```

```
Step 6:
    SENTIM_WORD_SCORE:= extract positive negative score from the SentiWordNet
    according to SENSE_NO
IF the POSITIVE_SCORE is greater than NEGATIVE_SCORE THEN
SENTIM_WORD_SCORE:= POSITIVE_SCORE
ELSE the POSITIVE_SCORE is less than NEGATIVE_SCORE THEN
SENTIM_WORD_SCORE:= NEGATIVE_SCORE
Step 7:
MODIFIER_WEIGHT:= weight of SENT_SENTIM_WORD in MODIFIER_DICT
    MODIFIER_DICT: = list of Modifier which affects the score of positive and
    negative polarity
    IF SENT_SENTIM_WORD is similar JJ OR SENT_SENTIM_WORD is similar RB
     THEN CHECK (SENT_SENTIM_WORD + 3) and (SENT_SENTIM_WORD - 3)
    for Modifier from MODIFIER_DICT
IF WORD found as MODIFIER
THEN Repeat Step 7
Step 8:
IF the MODIFIER is a negation modifier THEN
SENTIM_WORD_SCORE:= Reverse the polarity of SENT_SENTIM_WORD
Step 9:
IF the MODIFIER is a intensifier or contact shifter THEN
    SENTIM_WORD_SCORE:= intensifying MODIFIER_WEIGHT obtained from
    MODIFIER_DICT
SENTIM_WORD_SCORE:= SENTIM_WORD_SCORE + MODIFIER_WEIGHT
Step 10:
    IF the MODIFIER is a decelerator OR IF the MODIFIER is enhancer OR IF the
    MODIFIER is context shifter THEN
    SENTIM_WORD_SCORE:= intensifying MODIFIER_WEIGHT obtained from
    MODIFIER_DICT
    SENTIM_WORD_SCORE:= SENTIM_WORD_SCORE + MODIFIER_WEIGHT
Step 11:
For Each SENTIM_WORD_SCORE in SENT
SENT_TSCORE:= SENT_TSCORE + SENTIM_WORD_SCORE
RETURN SENT_TSCORE
Step 12:
For Each SENT in REVIEW
REVIEW_SCORE:= REVIEW_SCORE + SENT_TSCORE
```

Figure 3.4  Algorithm for Sentiment Analysis

**Step-6.**For the correct sense, extract the sense-id from WordNet using the semantic pattern of the desired sentence, refer to SentiWordNet, the semantic score of the WRD is extracted on the basis of that sentence structure. The sentence level polarity is calculated considering the weight of each term in the sentence.

**Step-7.**If there is a negation word (Not, Never, N’t, Doesn’t, Can’t, Nor, Don’t, Wouldn’t, No) near the N, Check (N+3) and (N-3) then reverse its polarity, e.g. (OW=+0.8 →OM= -0.8).

**Step-8.**If there is any type of context shifter in the sentence or enhancer/reducer, then the polarity will be recalculated, because these words affect the polarity. The position of the contact shifter is checked in the sentence and then the nearest opinion word is checked; this may be JJ, JJS, noun NN, NNS or VB, VBS. If its score is negative, then it will be changed after recalculating its weights and vice versa. The negation words reduce its effect to nothing.

**Step-9.**Check the modifier word in the sentence, if it exists, then recalculates the polarity referring to the weightage dictionary. The same process will be repeated until the score is same as that of which the opinion word will be affected. There are a few types of certain nouns which affect the sentence polarity, so recalculate the polarity if such types of words occur, assign weights to each sentence accordingly from the dictionary of weights of words/terms.

**Step-10 & 11.**Calculate the final weights of each sentence and each review to decide if it is positive, negative or neutral. So, opinion strength for both sentence and feedback is calculated by assigning the combined opinion weight to the sentence and review using Equations 3.4, 3.5, 3.6 and 3.7.

$$SentenceScore(Sen) = \sum_{i=1}^{n} Score(w) \quad \text{(Eq.3.4)}$$

$$SentenceScore(Sen) = \frac{\sum_{i=1}^{n} Score(w)}{n} \quad \text{(Eq.3.5)}$$

Where, Score (Sen.), is the positive or negative score of the word w, i is the positive or negative score of the ith word in sentence S and n is the total number of words in Sen.

$$\mathrm{Re}\,viewScore(\mathrm{Re}\,w) = \sum_{i=1}^{n} Score(Sen) \qquad \text{(Eq.3.6)}$$

$$\mathrm{Re}\,viewScore(\mathrm{Re}\,w) = \frac{\sum_{i=1}^{n} Score(Sen)}{n} \qquad \text{(Eq.7)}$$

Where, Score (Review), is the positive or negative score of the sentence Sen, i is the positive, or negative score of the ith sentence in the review and n is the total number of sentences in the review.

## 3.8 Summary

This chapter introduces a new method for sentiment classification using WordNet and SentiWordNet as a knowledge base at sentence level. This approach uses WordNet relations with learning from WordNet glosses and lexical relations for word sense extraction and SentiWordNet for the sentiment orientation. The approach is applicable to the acquisition of sentiment-bearing words as well as of words with some other semantic categories, such as words with increasing/decreasing semantics and other valence shifters, which are also relevant for sentiment analysis. The resulting wordlists can then be used as an input for sentence and text-level sentiment analysis. This chapter, thus, first describes the WordNet-based approach to sentiment orientation at the word and sense level, and then evaluates the obtained wordlists as a part of the sentence and text-level sentiment classification system. The development of portable (domain and genre independent) sentiment determination system poses a substantial challenge for researchers in text mining, NLP and knowledge management.

# CHAPTER 4

# DATA ACQUISITION AND PRE-PROCESSING

## 4.1 Introduction

The first step of text mining is data collection and pre-processing. Data pre-processing is necessary to enable further processing for information/ knowledge extraction using various algorithms. In this chapter the data acquisition, collection, pre-processes steps were described and highlight its importance in the view of some current literature. One of the contributions of this dissertation is the preparation of new datasets and application of a new technique for pre-processing in the sentiment analysis process. Pre-processing is an essential step in the data mining process. Most of the data mining algorithms depend on pre-processing for selection of the appropriate subsets of the data, like FS and FE, in order to format the data according to the requirements of the algorithms. Data pre-processing is characterized by any kind of processing which is applied to raw data to make it ready for further processing using other applications. Data pre-processing, which is usually a technique utilized for preliminary data mining, changes the data into a format that is more readily and effectively processed in regards to the user’s requirements. Data pre-processing consists of removing noise from the data, extracting specific data that is pertinent in some particular context, and organizing data for more efficient accessibility. It might be desirable or even necessary to carry out some form of data pre-processing before beginning with the analysis; this, of course, depends on the type of analysis to be performed. Moreover, algorithmic constraints may require pre-processing. The KDT process also requires the pre-processing of data before it can be used. A key issue with data mining is quality; consequently, 80% of mining experts more often than not spend their time on data quality. Therefore, the pre-processing steps play a major role in data mining process (Wong, W. Liu, & Bennamoun, 2006) (H. Liu & Motoda, 1998). There are two major types of data that require pre-processing, structured data and unstructured or textual data. As data mining is the process of extracting knowledge from huge amounts of data, the size of the data is almost always large; therefore, selection of the appropriate subset of the data for efficient processing is necessary. Hence, there are three pre-processing steps for data mining which need to be considered: data collection, FE and FS. As the domain of this work is web content (unstructured data), the pre-processing steps for textual data are described as shown in Figure 4.1.

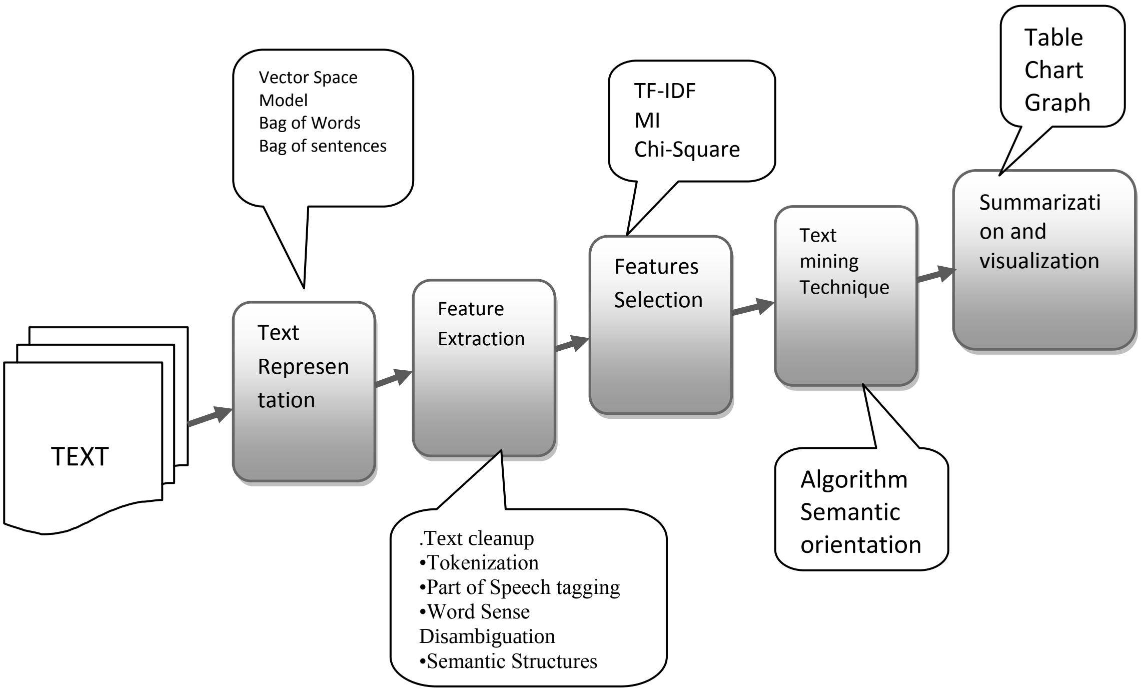


Figure 4.1 Steps of Pre-processing

## 4.2 Text Representation

Text representation is one of the pre-processing techniques which change a document from the full version into a document vector by reducing the complexity of the document; subsequently, the document is easier to deal with. Text representation which is an important aspect in document classification and information extraction signifies the preparation of a document into a concise form. Typically, a text document is presented as a vector of term weights (word features) derived from a set of words (dictionary), where each word is found at least once in a predetermined number of documents. The immensely high dimensionality of text data is a major characteristic of the challenge involved with classification of a text. The number of training documents is often exceeded by the number of potential features. A document is defined as a joint partnership of words having various patterns of occurrences. An important element in many applications involving management of information is the classification of the text.  Therefore, algorithms which are able to improve efficiency as well as maintain accuracy during the classification process are highly desirable as a result of the sudden growth of web data (J. Yan et al., 2005) (Shang et al., 2007). As illustrated, DR techniques can be categorized into two approaches, either Feature Extraction (FE) or Feature Selection (FS) (H. Liu & Motoda, 1998).

### 4.2.1 Feature Extraction

The aim of pre-processing is to make the border of each language structure clear and to eliminate, as much as possible, language dependent factors, tokenization, stop word removal, sentence boundary identification, spelling corrections, noise removal and lemmatization (L Dey & S. K. Haque, 2008), (Y. Wang & X. J. Wang, 2005).  FE is the pre processing step which is used to present text documents in a clear word format. The documents involved in text classification are represented by a large amount of features, most of them are possibly irrelevant or noisy (Montañés et al., 2003).

### 4.2.2 Feature Selection

The most important step in the pre-processing of text classification after FE is FS. A vector space is constructed using FS for improvement in efficiency, scalability and accuracy of the text classifier. Basically, the properties of the domain and algorithm are considered by a good FS technique (Z. Q. Wang, Sun, D. X. Zhang, & Li, 2006). The main idea behind the FS is that it takes the original documents and selects a subset of features from them. FS is carried out by considering the predetermined measure of importance of particular words and then storing the words with the highest predetermined scores (Montañés et al., 2003). The original physical meanings of the features which have been selected are kept for a better understanding of the data for the learning process (H. Liu & Motoda, 1998). The high dimensionality of the feature space is definitely a major issue for text classification. It is a fact that, a tremendous number of features are contained in almost all text domains where most of these features are irrelevant for the function of text classification and provide no benefit at all. Some of them may even cause the accuracy of the classification to be reduced drastically, e.g. noise features (J. Chen, H. Huang, Tian, & Qu, 2009). Hence, FS is typically used in text classification for improvement of the accuracy and efficiency of the classifiers while reducing the dimensionality of the feature space.

Inaccurate results can very well be the consequence of utilizing a compilation of words from various domains which have dissimilar properties to the domain of the text being processed for classification. One such case would be if the analysis is carried out using a set of tweets regarding a particular product but it is trained using a set based on movie reviews; this would lead to the misclassification of most of the sentences (Go et al., 2009) (Shamma et al., 2009). Furthermore, creation of a dictionary that can extract the important keywords or features to classify previously unseen sentences is essential in order to achieve the most accurate analysis possible (Lipika Dey & S. M. Haque, 2009). In addition, some of the irrelevant words like articles, pronouns, prepositions etc. (List of stop words) can be removed. Not surprisingly, the models presented in literature are very basic, and has several limitations, including being unable to capture the polarity relation between words and distinguishing between the various meanings which might be given to one word. Utilizing regular expressions for dealing with negation and parts of speech for a syntax analysis of a word could overcome other limitations.

### 4.2.3 Semantic and Ontology Based Text Representation

This section focuses on semantic and ontology techniques, language, associated issues for FS and text classification. According to (Yeh, Hirschman, & Morgan, 2003) statistical techniques are not sufficient for text mining; better classification will be achieved when considering the semantics. Ontology is a data model representing a set of concepts in a specific domain and the relationships these concepts have with each other. It is used to speculate about the objects in that particular domain. Ontology is the explicit and abstract model representation of already defined finite sets of terms and concepts, involved in knowledge management, knowledge engineering and intelligent information integration (Fensel, 2004b) .The characteristics of objects and entities (individuals and instances) are  real things and association (relations) with their attributes are used for the titles of the two concepts or entities. Ontology has been proposed for handling heterogeneity semantically when extracting information from various text sources such as the internet (Tenenboim, Shapira, & Shoval, 2008).

Ontology based text representation can also be called as semantic representation used in specified domain. The sentences and paragraphs are linguistically parsed into key concepts, verbs and proper nouns in a procedure called Semantic analysis. Statistics-backed technology is then utilized to compare these words to taxonomy (categories) and categorize them in relation to their relevance (Yeh et al., 2003). Better classification will be performed when taking the semantics under consideration; hence, the semantical representation of a text and web document is the key challenge for the sentiment classification and knowledge management (Kawamura et al., 2008).

Table 4.1 Common Challenges in Text Pre-processing for Sentiment Analysis and Classification

| Challenge | Description |
|---|---|
| Sentence Splitting | How we Identify sentence boundaries in a document? |
| Tokenization | How the documents are tokenized and tokens are recorded or annotated, by word or phrase. This is important because many down stream components need the tokens to be clearly identified for analysis? |
| Part-of-Speech(POS) Tagging | In regards to the part of speech characteristics and the data annotation, how such components are assigned a pos tag to token pos information? |
| Stop word list | How stop word lists will be taken, and which words are to be considered as stop words as well as in which domain? |
| Stemming | If we reduce the words to their stems, how it will affect the meaning of the documents? |
| Noisy Data | Which steps are required for the document to be clear from noisy data? |
| Word Sense | How we clarify the meaning of the word in the text/ ambiguity problem? |
| Collocations | What about the compound and technical terms? |
| Syntax | How should make a syntactic or grammar analysis? What about data dependency/anaphoric problems? |
| Text Representation | Which will be more important for representation of the documents: phrases, words or concepts, and noun or adjective? And, for this, which techniques will be feasible to use? |
| Domain and data understanding for Ontology | How to define the area, data availability and its relation to ontology construction? |

In this work, customer reviews, blog and social network comments in the form of unstructured text is extracted from Web. The text is processed to extract the important feature for semantic orientation and sentiment classification. A lexicon based method is used for semantic orientation and classification of sentiments in a text into positive, negative or neutral opinions. Moreover, the increasing volume of user sentiments in the form of unstructured text needs IR and NLP techniques for knowledge discovery.

Some of the common challenges in text pre-processing are shown in Table 4.1 context.

## 4.3 Data Collection and Pre-processing

In this section, the data collection and the pre-processing steps are defined, which are used in this work. Pre-processing is one of the important steps of text analysis as described earlier. The 1st step is data collection and acquisition. The data collection is related to information retrieval. Specific text are retrieved the from the web contents which we need, depending on some product, event, or person. Once the text set is retrieved and collected, it needs to be pre-processed. Pre-processing involves re-orientation of the text in some structured form; where we should remove the noisy text, i.e. removal of unnecessary irrelevant words and symbols and to extract important features for classification and semantic orientation. This is because online discussion forums, blogs and customer reviews may contain a lot of noisy text and the opinion sources are typically informally written and are highly diverse. In this work, two types of datasets are employed for the proposed method's evaluation; one is the own collected and processed datasets and the other one is acquired from already processed datasets, freely available on the internet (benchmark datasets) for research purposes.

Basically, there are three types or formats of reviews available on the Web (B Pang & L Lee, 2008).

- Format I – Pros, cons and the detailed review: Pros and cons are described by the reviewers using short phrases where details of the reviews are written separately.
- Format II - Pros and cons: In this format pros and cons are described by the reviewers separately and are written in full sentence form.
- Format III - Free format: The reviews are written by the reviewers in free text form, usually consisting of short phrases and incomplete sentences with no separation of pros and cons; this is followed by a detailed review.

Following review has been taken from www.cnet.com as an example to explain the above mentioned formats. The reviewer is describing the pros and cons of Review as below.

Pros: Blazingly fast, incredible handling, excellent media interface and sound. It is the first vehicle I have ever had that exceeded my expectations.

Cons: needs backup camera, limited rear visibility. I would say poor mileage.

Format II and III usually consist of long sentences and complete sentence reviews. For example, “The larger lens of the g3 gives better picture quality in low light, and the 4-times optical zoom gets you just that much closer”. However, the product features extraction from reviews of format II and III is more challenging because the complete sentences are more complex and contain a large amount of irrelevant information.

In this work, format III type reviews are collected from Skytrax, airline reviews and blog comments from Cricinfo. The other datasets used in this work are movie reviews, hotel reviews and twitter comments acquired from Tripadvisor, Twitter and (Bo Pang & Lillian Lee, 2005) respectfully as shown in Figures 4.2 and 4.3. The data is classified mainly as positive and negative sets for testing purposes. One of the contributions of this work is collecting and processing two types of datasets, as mentioned above, for sentiment analysis. The second type of data set is already freely available on the Web for research purposes.

1000 comments are collected from the twitter datasets, publicly available for research purposes (Shamma et al., 2009) and extracted 500 short comments from Cricinfo, about the performance of the Pakistani team in the Cricket Word Cup 2011. Table 4.2 shows the blog comments dataset information.

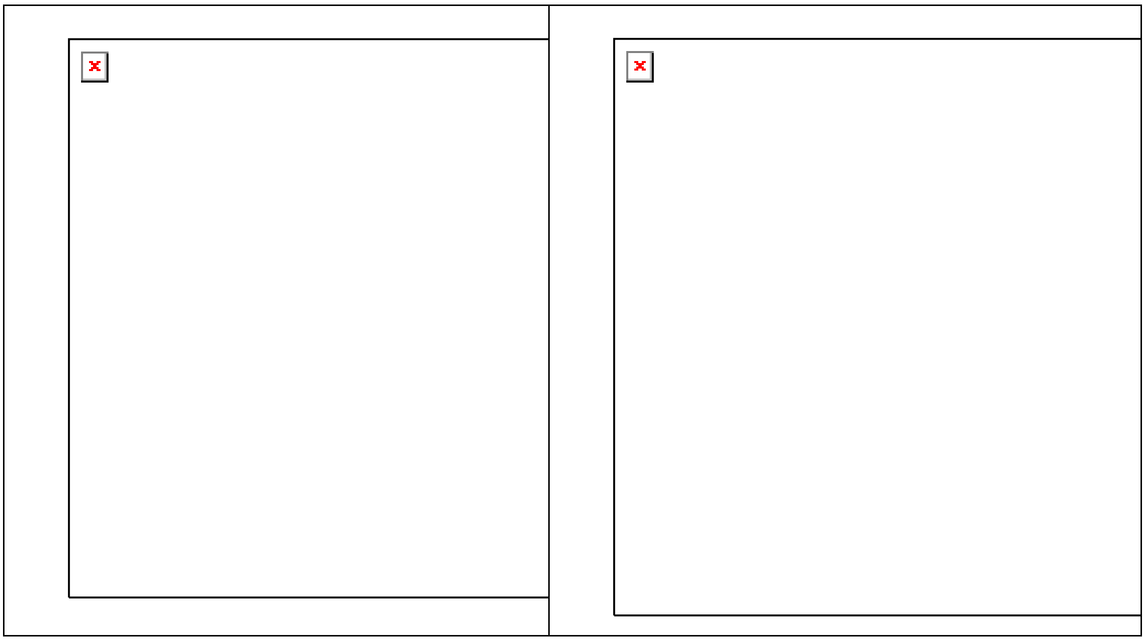
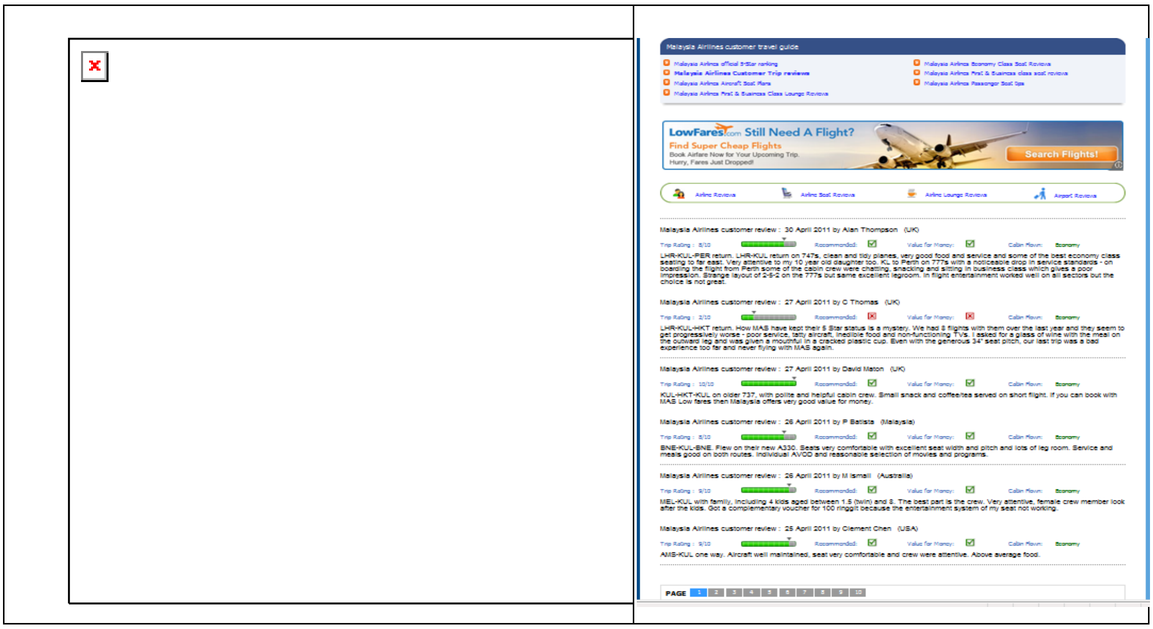


Figure 4.2 Cricket Blog Reviews

For review collection, three types of online customer review datasets were acquired for the proposed method's performance evaluation. The types of reviews and their details are shortly described in the bullets below:


Malaysia Airlines customer travel guide
Malaysia Airlines official 5-Star ranking
Malaysia Airlines Customer Trip reviews
Malaysia Airlines Aircraft Seat Plans
Malaysia Airlines First & Business Class Lounge Reviews
Malaysia Airlines Economy Class Seat Reviews
Malaysia Airlines First & Business class seat reviews
Malaysia Airlines Passenger Seat tips
Airline Reviews
Airline Seat Reviews
Airline Lounge Reviews
Airport Reviews

Malaysia Airlines customer review : 30 April 2011 by Alan Thompson (UK)

Trip Rating : 8/10 Recommended: ☑ Value for Money: ☑ Cabin Flown: Economy

LHR-KUL-PER return. LHR-KUL return on 747s, clean and tidy planes, very good food and service and some of the best economy class seating to far east. Very attentive to my 10 year old daughter too. KL to Perth on 777s with a noticeable drop in service standards - on boarding the flight from Perth some of the cabin crew were chatting, snacking and sitting in business class which gives a poor impression. Strange layout of 2-5-2 on the 777s but same excellent legroom. In flight entertainment worked well on all sectors but the choice is not great.

Malaysia Airlines customer review : 27 April 2011 by C Thomas (UK)

Trip Rating : 2/10 Recommended: ☒ Value for Money: ☒ Cabin Flown: Economy

LHR-KUL-HKT return. How MAS have kept their 5 Star status is a mystery. We had 8 flights with them over the last year and they seem to get progressively worse - poor service, tatty aircraft, inedible food and non-functioning TVs. I asked for a glass of wine with the meal on the outward leg and was given a mouthful in a cracked plastic cup. Even with the generous 34" seat pitch, our last trip was a bad experience too far and never flying with MAS again.

Malaysia Airlines customer review : 27 April 2011 by David Maton (UK)

Trip Rating : 10/10 Recommended: ☑ Value for Money: ☑ Cabin Flown: Economy

KUL-HKT-KUL on older 737, with polite and helpful cabin crew. Small snack and coffee/tea served on short flight. If you can book with MAS Low fares then Malaysia offers very good value for money.

Malaysia Airlines customer review : 26 April 2011 by P Batista (Malaysia)

Trip Rating : 8/10 Recommended: ☑ Value for Money: ☑ Cabin Flown: Economy

BNE-KUL-BNE. Flew on their new A330. Seats very comfortable with excellent seat width and pitch and lots of leg room. Service and meals good on both routes. Individual AVOD and reasonable selection of movies and programs.

Malaysia Airlines customer review : 26 April 2011 by M Ismail (Australia)

Trip Rating : 9/10 Recommended: ☑ Value for Money: ☑ Cabin Flown: Economy

MEL-KUL with family, including 4 kids aged between 1.5 (twin) and 8. The best part is the crew. Very attentive, female crew member look after the kids. Got a complementary voucher for 100 ringgit because the entertainment system of my seat not working.

Malaysia Airlines customer review : 25 April 2011 by Clement Chen (USA)

Trip Rating : 9/10 Recommended: ☑ Value for Money: ☑ Cabin Flown: Economy

AMS-KUL one way. Aircraft well maintained, seat very comfortable and crew were attentive. Above average food.

Figure 4.3 Airline Reviews

- Popular publicly available corpus from movie-review polarity dataset i.e. v2.0 IMDB movie reviews . The data set consists of 1000 positive and 1000 negative reviews in individual text files; also, the sentences polarity dataset (includes 5331 positive and 5331 negative processed sentences / snippets (Bo Pang & Lillian Lee, 2005). Positive and negative sentences have been taken to check the performance of the proposed method.
- 1000 reviews have been extracted from Skytrax, where there are more than 2.5 million independent reviews for over 670 airlines and 700 airports. After splitting the reviews into sentences, an average of 8 sentences per review is found. The subjective lexicons and semantic orientation were extracted from all the positive and negative sentences.
- 2600 hotel reviews have been downloaded as a data set for the experimentation, which are collected from TripAdvisor, one of the popular review sites about hotels and travelling. Only the texts from these reviews using text files were extracted.

Table 4.2 Processed Datasets

| Datasets | Comments | Sentences | Sentences/Comments(Average) |
|---|---|---|---|
| Twitter | 1000 | 2045 | 2 |
| Cricket World Cup | 500 | 1630 | 3 |
| Movie Reviews | ---- | 10662 | 10 |
| Airline Reviews | 1000 | 7730 | 8 |
| Hotel Reviews | 2600 | 25663 | 10 |

After data collection, the major step is to clean the data from noise, and represent it in a specific form according to the requirements of the algorithms. Text data has more challenges as compared to numeric data in pre-processing because of its unstructured diverse nature. All the datasets are processed to remove noise, cleaning up the special characters and symbols and also checked them for spelling mistakes. Furthermore, the POS tagger is applied and classifies the sentences into subjective and objective sentences as described in previous chapter. The movie reviews data has already been processed for positive and negative sentences. Subjective sentences were hauled out only for further processing to find the semantic orientation at the individual sentence level. The pre-processing steps taken in this work are as follows.

### 4.3.1 Sentence Splitting and Processing of Noisy Text

In this section, the pre-processing steps used in this work are described. After removing the noise, the reviews/comments are split into sentences to extract the feature level sentiment score from SentiWordNet. A BOS is made from the split sentences, and each sentence is stored with a Review-ID and Sentence-ID. After applying the POS, the position of each word in the sentence is also stored for further processing. The noisy text degrades the performance of the classifier and is a main hurdle in semantic orientation. Machine based learning methodologies are often used, where pre-processing for noise removal is done using a generative model and noisy channel method. Unrestrained vocabulary, spelling mistakes, casual capitalization of words, white spaces etc are assumed to be possibly contained in the text. Sentence boundary detection for speech transcripts is yet another well researched issue. Majority of these systems make use of Hidden Markov Models (HMM) as well as a set of lexical and prosodic features which have been learnt from a manually tagged training set (L Dey & S. K. Haque, 2008). In this work the idea of (L Dey & S. K. Haque, 2008) was followed for noise removal with some modification and implementation of new technique for text cleaning and their possible semantic orientation especially in case of blogs as shown in Figure 4.4.

```
INPUT:
CORPUS - NOISY_REVIEW (noisy text)
OUTPUT:
REVIEW_CLEAN (clean text)
METHOD:
    REVIEWS: = Split Corpus
    SENT: = Split Reviews
    REW_ID:= Assign ID to each Review
    SENT_ID:= Assign ID to each sentence
    WORD_LIST:= list of words in sentence
    WORD_POSITION: = Position of each word in a sentence
For Rewiew1 to n
Identify Sentence boundary
Check for "." Exclude the predefined words like [Prof. Org. Pvt. Gov. Ltd. etc ]
Merge two sentences
    IF new line start with lower case non dictionary word fragment
    For Sentence1 to n
        For Word1 to word n
    Case correction;
    Spelling correction;
            Check special characters and symbols;
IF character or symbol = predefine word or symbols then
Replace the word with the dictionary word
Else remove all
End For
    End For

End For
```

Figure 4. 4 Algorithm for Noise Removal and Symbols/Short Words Processing

### 4.3.2 Sentence Boundary Identification

For sorting of reviews/comments into correct sentences, sentence boundary identification is very important. A rule based module has been implemented for this purpose. In this method, "." is considered as the sentence boundary, if it is not preceded by a predefined word: i.e., Pvt., Ltd., etc. The "." is also ignored after an abbreviation list (defined in the dictionary) and immediately after digits which do not follow a space character. Sentences commonly begin with a capital letter which is the most identifiable marker for sentence breaks. It is rational to consider two lines as merging together if it starts with a small letter and the line before it does not have any recognizable punctuation symbol. However, if a new line begins with a small letter with a non dictionary term, then the first word the last sentence is checked and merged with it, if it become a dictionary word and then a sentence is made by joining the contents of the two lines (L Dey & S. K. Haque, 2008) (Lipika Dey & S. M. Haque, 2009) (Wong et al., 2006).

### 4.3.3 Sentence Cleanliness

To remove noise from a text, an algorithm is applied to remove symbols, check spellings and correct those words which are incorrectly written. The semantic score of those symbols were extracted, from which the reviewer wants to express something meaningful. In online Web forums, social networks, blogs etc., people frequently write short forms of words and use symbols in comments to express their views. How these symbols and short words are made is useful in extracting their semantics from such sentences (L Dey & S. K. Haque, 2008). Up till now, no such tool has been available to extract and calculate the semantic score of such words and symbols because there is no standard rule available for writing comments, reviews on online forums, blogs etc. Such symbols or shortened words are B4, Gr8, bcz, :), ##123, @@@, >> etc (Go et al., 2009). Here, it is attempted to overcome this problem by collecting such symbols and words that are most often used in conveying special messages instead of writing a full sentence. (Wong et al., 2006).The same algorithm is applied to online customer reviews and comments and results are encouraging as well as improve the sentiment analysis process. However, there should be proper rules for writing reviews and comments to express our views on online forums and blogs. Table-4.3 shows a collection of such types of symbols with their meanings. When viewing these little things, which are called "emoticons", often the idea is to turn the head sideways so a picture is made on a lot of the smiley faces [;-)]for example, where the [ ; ] semi-colon are the eyes, the [ - ] hyphen is the nose, and the [ ) ] parenthesis is the mouth. Also, some people use the hyphen [-] to show the nose, while others will show the same expression without the nose, e.g.: [;-) and ;)] represent the same thing. These symbols show emotions and expressions which are very important for sentiment analysis (Wong et al., 2006).

Table 4.3 Symbols and Characters used in Blogs

The reviews/comments which are taken for pre-processing to remove noise and to split them into sentences are shown in Table 4.4.

Table 4.4 Pre-processing Noisy Text

| Types of Cleaning | Noisy Text | Clean Text |
|---|---|---|
| Symbols cleaning<br>Semantic extraction | ????You are rite kamran but idk why they dont listen to US what we all thinking , anyways GAME ON HAY :)<br>Wish you GOOD Luck Pakistan team please play ++++++) not ------- :)<br>I request all ppl please support them ll find out after WC hope everything going good for US ( inshalla) | You are rite kamran but I dont know why they dont listen to US what we all thinking anyways GAME ON HAY smile. Wish you GOOD Luck Pakistan team please play positive not negative smile. I request all ppl please support them ll find out after WC hope everything going good for US inshalla |
| Merging sentences | 1. Why Pakistan's NRR shows 1.747 on 05 March 2011…<br>2. they did not play or played against 150 and 150 overs respectively | Why Pakistan's NRR shows 1. 747 on 05 March 2011. They did not play or played against 150 and 150 overs respectively. |
| Symbols and special character cleaning | Why Pakistan's NRR shows 1.747 on 05 March 2011… >>> they did not play or played against 150 and 150 overs respectively..!?! should it not be 143 and 125.6 overs respectively..!?!? if i m not wrong... | Why Pakistan's NRR shows 1. 747 on 05 March 2011. They did not play or played against 150 and 150 overs respectively. Should it not be 143 and 125.6 overs respectively. If i am not wrong. |

The rule-based system can get rid of such short abbreviated words/character and symbols. Remove the repetitious characters and symbols, then refer the symbol to dictionary and perform search to check, if immediate both sides of the symbols are words in the dictionary, or not. If not, another check is done to see if together they form a recognizable word. If they do, then the fragments are joined; if not, a blank space replaces the symbols. If a symbol represents a valid punctuation mark like "?", ":", etc. A dictionary of symbols based on inputs from the site can be compiled by the users for their own use (L Dey & S. K. Haque, 2008). For example, the sentence presented in Table 4.4 shows various symbols like "+++++)" which have made their way in because of encoding problems while crawling this particular site.

There are special words and symbols like "Idk= I don't know and :) = smile" as shown in Table-4.4, such symbols are useful to complete the sentence for semantic orientation. Mostly, people want to pass a message through these short words and symbols; so, the meanings for such symbols and short abbreviated words were extracted, to make a dictionary. Furthermore, these symbols with their respective meanings were replaced, referring to the dictionary for sentence semantic extraction.

### 4.3.4 Part of Speech (POS) Tagger

For assigning a tag to each word in a sentence, POS tagger is used, by adopting the Stanford trigger lexical database as the knowledge base. The tagger is connected with the proposed method with some changes for efficient and effective tagging. A tag is assigned to each word, like, JJ, JJS, VB, VBS, RB, NN, NNS, DT etc. as described in Table 4.5.

Table 4.5 POS Types with Abbreviations

| Pos-id | POS_Name | POS_Abbrivation | SentiWordNet_Abrv |
|---|---|---|---|
| 1 | Noun | NN | n |
| 2 | Adjective | JJ | a |
| 3 | Verb | VB | v |
| 4 | Adverb | RB | r |
| 5 | Nouns | NNS | n |
| 6 | Adjectives | JJS | a |
| 7 | Verbs | VBZ | v |

This work focuses on free text format (Format-III) reviews/comments as illustrated in Figure-4.4. The system extracts the reviews and comments from the Web using a crawler, and then cleans it and applies the POS for tagging. Airline reviews are selected which have been taken to process for tagging as described in Figure-4.5.

Figure 4.5 Free format review

From Figure 4.5, it is clear that the system assigns a tag to each word using a lexicon dictionary. The Stanford lexicon dictionary is used for effective part of speech tagging.

A plain text document is used as an input to POS tagger and returns an output; a document, in the form of tagged words and punctuation marks that indicate the part of speech the terms is used as. For example, the input text of Figure 4.4 gives the result as shown in Figure 4.5.

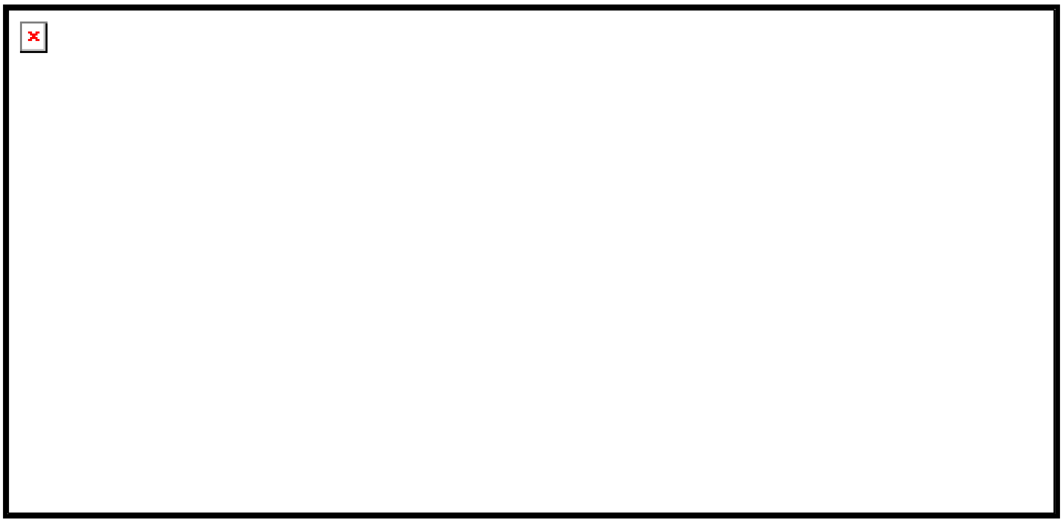


Figure 4.6 Tagged review using POS

The sentences are stored with Sentence-ID and Review-ID to make a BOS for further processing and semantic orientation as shown on Figure 4.6.

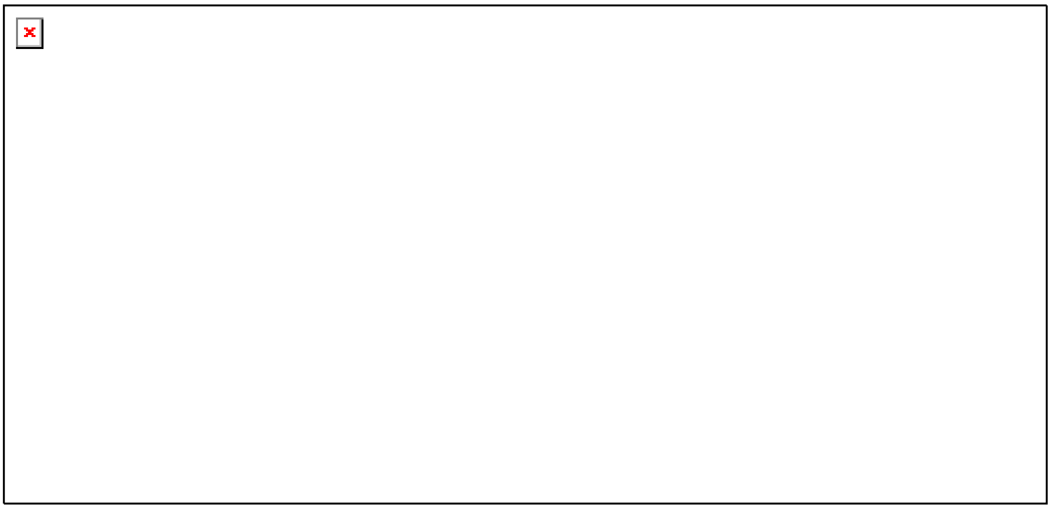


Figure 4.7  Bag of Sentences (BOS)

Another review of semi-structured data is taken from hotel reviews dataset (Figure 4.7); the data is in XML format which contains different tags. Only the text files of reviews are extracted, remove noise and process for POS to select the needed features as described in Figures 4.8 and 4.9.

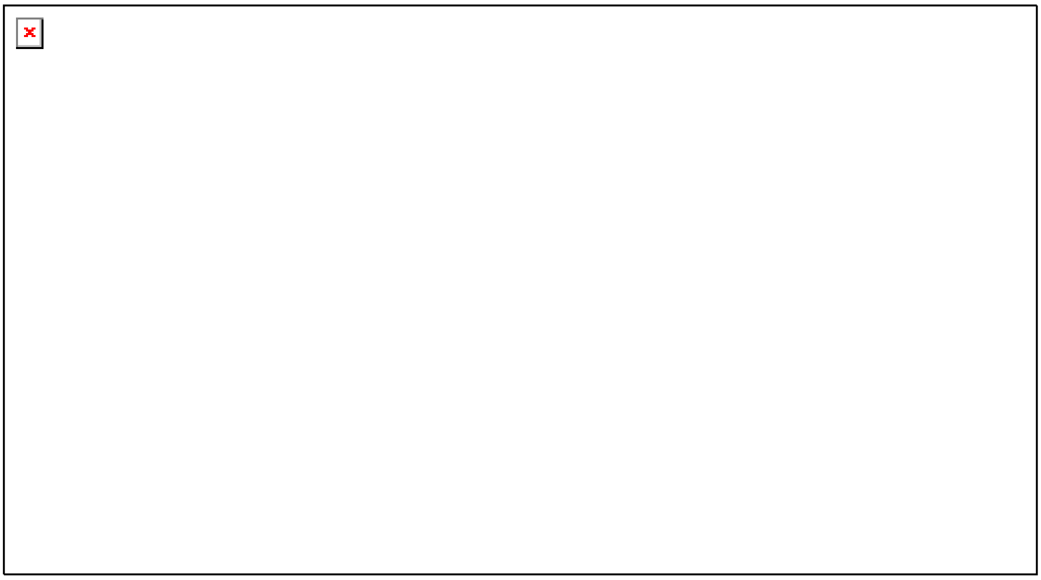


Figure 4.8  Hotel Review Data

After extracting the text file from the reviews, the text is processed to remove noise and split them into sentences.  Each sentence is tagged using POS tagger and stored with their Sentence-ID. For example, the sentences "*this small hotel is in a* *fabulous location* “and “*the people at the reception are friendly and helpful*” for the tagging process and the important FS as described in Figure 4.8 and 4.9.

The clean sentences with their Id’s are stored for further processing, Figure 4.9.

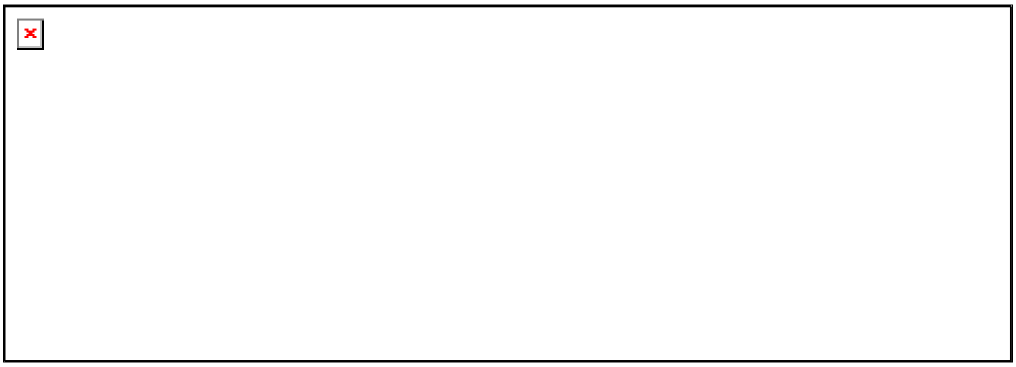


Figure 4.9  Sentences with Their IDs

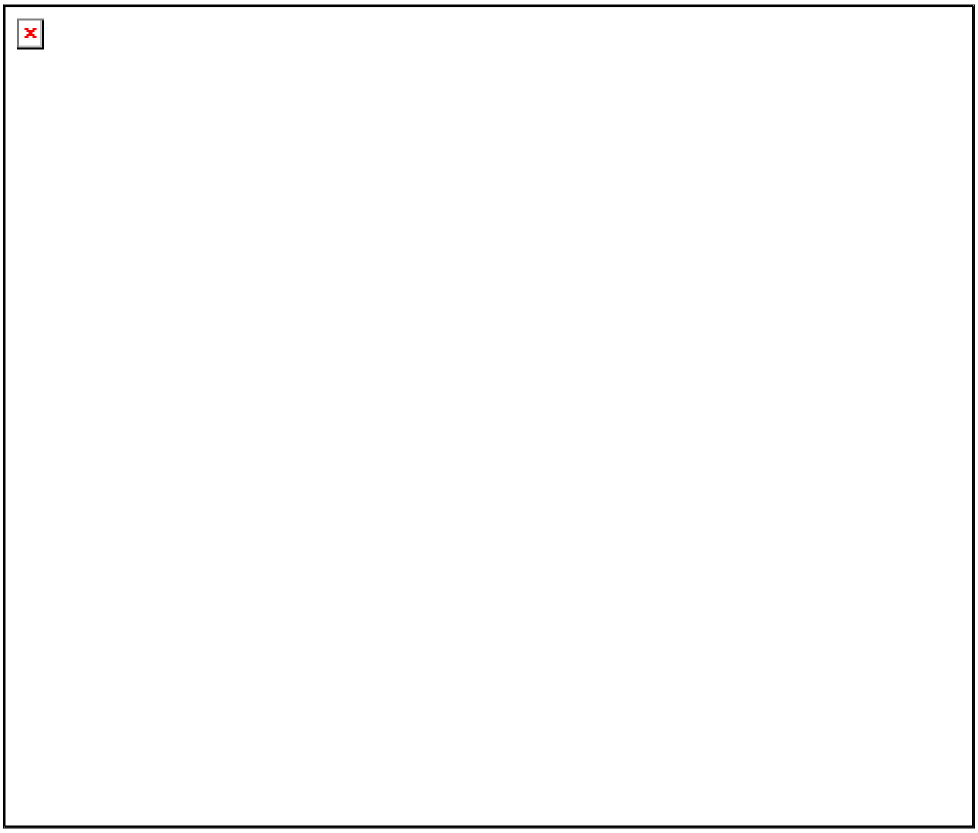


Figure 4.10  POS Tagging Process and Selecting the Important Features

## 4.4 Summary

Pre-processing is an important step in the data mining process, particularly for unstructured data due to its diverse nature. Web content contains a great deal of noisy text, especially social network sites and customer reviews. There is no standard rule for writing comments and reviews to express views on such forums. When extracting information and knowledge from such sources, there is a need to remove noise and process it according to the requirements of the mining tool. This chapter describes the collection and acquisition of data from such sources and the pre-processing steps used for sentiment extraction from user comments and reviews. This work contributes two new datasets for research purposes, namely, customer review for airlines and sports blog comments. Furthermore, it proposes a rule for semantic information extraction from short abbreviations and symbols.

# CHAPTER 5

# RESULTS AND DISCUSSIONS

## 5.1 Introduction

This chapter presents the evaluation of the proposed method with respect to the overall performance. The aim is to examine the performance of the method and highlight the performance increase that can be attained by utilizing this method. The chapter begins with the methodology that has been adopted to conduct the evaluation of the proposed method and about the simulation description. Thereafter it identifies the experiment settings, reports and discusses the achieved results.

## 5.2 Evaluation methods

To ensure the reliability and consistency of the evaluation procedure, a set of principles has been introduced from the evaluation of different systems throughout this study. Corpus-based methods have been taken as benchmarks for the evaluation of performance of the Lexical Based sentence level method which enables to assess the comparative improvement in the obtained results.

In the study presented here, sentences are classified into positive, negative, and neutral sentences depending upon the sentiment content. In other words a sentiment is understood by this method as a ternary category i.e. positive, negative or neutral. Accuracy, precision, recall, and F-measure are the common performance measures which have been utilized for performance measurement of the approaches presented here (Andreevskaia & Bergler, 2008) (Go et al., 2009). The accuracy calculation is dependent on the component and dataset for binary (positive vs. negative) and ternary (positive vs. negative vs. neutral) classification (L Dey & S. K. Haque, 2008).

The accuracy for ternary classification is measured as a percentage of correct labels for all three categories out of the entire size of test set, as follow (Ding, B. Liu, & P. S. Yu, 2008).

The performance of binary classification is evaluated by its accuracy, precision, recall, and F-measure. Binary accuracy is computed as the percentage of correctly assigned positive and negative labels over the number of all sentences with positive and negative labels in the test standard dataset.

For the performance evaluation using precision and recall, this is the standard evaluation criterion for the classification. Precision of binary, positive/negative classification is defined here as a proportion of correct positive and negative labels given by the system over the number of all positive and negative labels assigned by the system. Sentences that were not tagged as positive or negative are ignored, the precision for positive and negative is calculated as follow.

Recall is the percentage of correct positive and negative labels assigned by the system over the sum of positives and negatives in the standard dataset as follow.

Finally, F1-measure is computed based on precision and recall as follow.

The following table confusion matrix, which is a de-facto standard in NLP and is widely used in the comparison of algorithm performance shown in Table-5.1. The same is used for the performance of the proposed method using the above mention equations.

Table 5.1 Confusion Matrix

| | Machine Says Yes | Machine Says No |
|---|---|---|
| Human Says Yes | TP | FN |
| Human Says No | FP | TN |

Where TP= True Positive, FP = False positive, TN, = True Negative and FN= False Negative. According to Table 5.1 the Precision and Recall can be calculated as follows (Ye et al., 2009) (L Dey & S. K. Haque, 2008).

### 5.3 Experimental Setup

The experimental setup of this work is divided into three components namely noise removal, WSD and semantic orientation and classification. Initially the reviews/comments from the web are extracted and process them for noise removal using noise removal module (see chapter-4). Short abbreviated words and symbols are also taken into consideration during noise removal process in this module. POS tagging module is applied in parallel for tagging words according to their respective parts of speech. Then the semantic score of the opinion word is extracted using WSD module. At last the final semantic score is calculated considering all the parts of speech and the contextual information in a sentence. Experiments are set for a test run on different datasets to evaluate the performance of the proposed method. The results are collected in two ways differing only in the scoring methods used. The first test run determines whether a sentence is positive, negative or neutral depending on all part of speech, semantic score and the contextual information contained. The second run calculates the average score of the sentence in a review of feedback and determines whether the review is positive, negative or neutral. These results are in binary and turnary form which are easy to compare with other methods because earlier research in this area used binary way (positive or negative) for sentiment classification.

For simulation, C# program is implemented by using .net framework (Microsoft visual studio 2008) that performs the experiments and handles the results. The code is available in Appendix A with screen shots in Appendix C. Windows-Vista 2008 and Windows 7 is utilized for the experiments on standalone systems. The datasets used in the experiments (reviews and comments) are discussed in detail in chapter 4.

### 5.3.1 Noise removal and spelling correction

Reviews and blogs comments extracted from web needs to be pre-processed. Pre-processing involves re-orientation of the text in some structured form; where the noisy text should be removed, i.e. removal of unnecessary irrelevant words and symbols and to extract important features for classification and semantic orientation. This is because online discussion forums, blogs and customer reviews may contain a lot of noisy text and the opinion sources are typically informally written and are highly diverse (L Dey & S. K. Haque, 2008) (Lipika Dey & S. M. Haque, 2009). During the process of noise removal for comments and reviews, it is observed that comments possess more noise than reviews (Go et al., 2009). To validate this, the performance of the proposed method is test out on Airline reviews and blogs comments. A large number of unwanted symbols, text, special characters and digits were detected and were removed by noise removal module. At the same time spelling correction module was also activated for spelling correction. The proposed spelling correction module consists of the built in dictionary of MS Word 2007, which is capable of giving spelling suggestions as per the requirement of the user. It can also merge sentences and clean out symbols. The evaluation has been done on 1630 sentences from blogs, 2045 comments from twitter and 7730 sentences from reviews. The system initially identified 470 words from blogs, 674 from twitter comments and 1020 words from the reviews which were possibly erroneous. These sentences were merged and cleaned for symbols and the spellings were corrected in accordance with the dictionary. The erroneous words were replaced by the topmost suggestion made by the system for correction. This increases the performance of the sentiment orientation especially in sentences. Table 5.2 summarizes these results.

Table 5.2 Evaluation of Spelling Correction and Noise Removal Module

| Reviews/ Comments | Sentences | Words Extracted as Incorrect | Corrected Successfully | Accuracy |
|---|---|---|---|---|
| Airline Reviews | 7730 | 1020 | 938 | 0.91 |
| Blog comments | 1630 | 470 | 422 | 0.90 |
| Twitter comments | 2045 | 674 | 568 | 0.84 |

### 5.3.2 Sentiment Expression Detection & Word Sense Disambiguation

Word Sense Disambiguation is important for semantic orientation for actual sense extraction from sentence. In this work the main focus is to extract sentence and document level sentiment analysis. Sentence level analysis decides what the primary or comprehensive semantic orientation of a sentence is while the primary or comprehensive semantic orientation of the entire document is handled by the document level analysis. The text documents or reviews are broken down into sentences for sentiment analysis at the sentence level. These sentences are then evaluated by utilizing lexical methods in order to determine their semantic orientation. This process involves two functions; first is to determine the subjectivity or objectivity of a sentence and the next function is of taking the sentences with an opinion orientation which is subjective. Semantic orientation can be accumulated from the words and expression to find out the overall Semantic Orientation of a particular sentence. Hence, contextual information of all the parts of speech is vital for the semantic orientation. Structure of the sense in sentences and all content parts of speech play an imperative role in analysis of sentiments.

After noise removal and spelling correction the sentences are tagged using POS tagger module (described in chapter 4). The tagged texts in each sentence are then checked to see if it contains sentiment words or not. Those sentences which have sentiment words are then sorted as subjective sentences and the opinion word sense is extracted considering the sentence structure, while those sentences which don’t have sentiment expression are discarded. The results of sentiment expression detection as subjective sentences are described in Table 5.3.

Table 5.3 Sum of Opinion Sentences

| Dataset | Reviews | Sentences | Subjective | Objective | Percentage (Sub/Obj) |
|---|---|---|---|---|---|
| Movie Reviews | --- | 10662 | 8530 | 2132 | 80/20 |
| Airline Reviews | 1000 | 7730 | 5405 | 2325 | 70/30 |
| Hotel Reviews | 2600 | 25663 | 17704 | 7969 | 68/32 |
| Twitter | 1000 | 2045 | 1636 | 409 | 80/20 |
| Cricket Blog | 500 | 1630 | 1238 | 392 | 76/24 |

Sentence structure plays an important role for extraction of sense of the sentiment word. WordNet gloss is utilized to extract the semantic scores for that sense from SentiWordNet. As a first step, part of speech tagging is used to obtain some level of disambiguation for extracting semantic scores from SentiWordNet. However if there occurs a multiple sense within the same part of speech a simpler approach can be used to assign scores based on the evaluated scores for each synset for a given term. If there are conflicting scores e.g. positive and negative scores exist for the same term then check the sense of the sentence using their contextual lexical pattern and return the SentiWordNet positive or negative score according to the sense of the sentence.

The WSD module is used to extract the sentence contextual pattern and referring it to WordNet glossary for the correct sense extraction. Correct score selection for these senses from SentiWordNet has been described in chapter-3.

Existing experimental literature on binary (positive/negative) sentiment classification reported that non-statistical approaches (lexical methods) give better accuracy as compared to statistical approaches (corpus-based or machine learning methods). This is probably due to the lack of annotated training data for statistical methods (Andreevskaia & Bergler, 2008) (Go et al., 2009).

## 5.4 Lexical Based Sentence Level Semantic Orientation

Lexical methods are utilized for the term semantic orientation which makes use of the so called sentiment lexicons, also known as opinion lexicons in online dictionaries like SentiWordNet, Sentiful, and WordNet etc. For machine learning methods, only the lemmas are not enough for detecting sentiment, however, they also make use of features (corpus or seed words) to successfully classify the sentiment. A lexical based method is proposed in order to extract sentiments with out using seed word. After extracting the sense of the opinion word and their semantic scores the system processes the rules for the sentence level contextual structure and checking the valence shifter (described in chapter-3) for different domains. The method is evaluated by dividing the datasets into two type's i.e. long reviews blog comments. For reviews dataset movie reviews, airline reviews and hotel reviews (detail definition of these datasets are in chapter-4) and for blogs cricket and twitter comments were processed.

### 5.4.1 Evaluation and Performance Measure on Blog Datasets

Both types of datasets are processed for lexical based semantic orientation at sentence level and feedback level. Results of blog comments and twitter datasets for performance evaluation of the proposed method considering both sentence level and feedback level are described in Table 5.4, 5.5, 5.6 and 5.7.

For this evaluation 210 twitter feedbacks are taken from twitter dataset which are split into 540 sentences and manually evaluated for positive, negative and neutral. From the human evaluation 106 were judged as positive, 77 as negative and 27 as neutral feedbacks as described in Table-5. The objective is to evaluate the capability of the proposed method to correctly classify the semantic orientation of the sentiments of these sentences and also to access the positive, negative or neutral sentences from the dataset.

Table 5.4 Sentiment Orientation of Twitter Comments at Sentence Level.

| | | Actual Orientation | | | |
|---|---|---|---|---|---|
| | | Positive | Negative | Neutral | Total |
| **System Assigned** | Positive | 200 | 27 | 14 | 241 |
| | Negative | 36 | 155 | 2 | 193 |
| | Neutral | 11 | 5 | 90 | 106 |
| | Total | 247 | 187 | 106 | 540 |
| **Accuracy at Sentence level** | | **0.824** | | | |

Table-5.4 and 5.5 presents the confusion matrix of the sentiment orientation at sentence and feedback level respectively.

Table 5.5 Sentiment Orientation of Twitter Comments at Feedback Level

| | | Actual Orientation | | | |
|---|---|---|---|---|---|
| | | Positive | Negative | Neutral | Total |
| **System Assigned** | Positive | 95 | 13 | 3 | 111 |
| | Negative | 9 | 62 | 2 | 73 |
| | Neutral | 2 | 2 | 22 | 26 |
| | Total | 106 | 77 | 27 | 210 |
| **Accuracy at Feedback Level** | | **0.85** | | | |

It is observed that the accuracy of sentiment orientation of comments of blog at feedback level (overall) achieved better results compared to at sentence level. Achieved results from the blog comments are 82% at sentence level and 85% at blog level respectively. After a closer review it is observed that the method has performed well in recognizing positive and negative sentences in blogs. Most of the errors involve are in detection of neutral sentences. Since all blogs contain more sentimental sentences than neutral sentences, therefore sentiment classification is comparatively easier at blog level. However, it is noted that the blogs comments contain more noisy text in the form of short abbreviated words, symbols and special characters etc as compared to reviews text, which degrades the performance of sentiment classification. This problem is tackled in this work to some extent which shows improvement in achieved results. Details regarding noisy text can be found in details in chapter 4.

Since in the previous results Twitter comments were evaluated, which is a popular social network blog having public opinion on different day to day scenarios and topics (Shamma et al., 2009) (Go et al., 2009). To evaluate of the proposed method at a different domain selected sports blog comments (cricinfo blog comments) as a dataset to access the performance of the system. It is observed that change of domain has little effect on the performance and variation in output results is quite less, which shows the domain adoptability of the proposed method. Results for sports domain dataset are shown in Table 5.6.

Table 5.6 Sentiment Orientation Cricket Blog Comments at Sentence Level

| | | **Actual Orientation** | | | |
|---|---|---|---|---|---|
| | | Positive | Negative | Neutral | Total |
| **System Assigned** | Positive | 222 | 26 | 12 | 260 |
| | Negative | 30 | 170 | 8 | 208 |
| | Neutral | 14 | 10 | 100 | 124 |
| | Total | 266 | 206 | 120 | 592 |
| **Overall Accuracy** | **0.831** | | | | |

157 Cricket blog feedbacks are taken from www.cricinfo.com as a dataset. This dataset is split into 592 sentences which are manually evaluated for positive, negative and neutral sentiments. Out of these manually evaluated sentences, 266 are labelled as positive, 206 as negative and 120 as neutral sentences. When the proposed method is evaluated on this dataset with others approaches for sentiment orientation, an accuracy of 83% is achieved at sentence level with respect to manual evaluation for this new domain of blogs comments. It can be observed that at sentence level, the accuracy of sentiment orientation for twitter is 82% (Table-5.4) while for Cricinfo it is 83%., therefore it can be deduced that changing domain has little effect on the achieved results.

Table 5.7 Sentiment Orientation of Cricket Blog Comments at Feedback Level

| | | Actual Orientation | | | |
|---|---|---|---|---|---|
| | | Positive | Negative | Neutral | Total |
| **System Assigned** | Positive | 80 | 7 | 3 | 90 |
| | Negative | 5 | 44 | 2 | 51 |
| | Neutral | 1 | 2 | 13 | 16 |
| | Total | 86 | 53 | 18 | 157 |
| **Overall Accuracy : 0.87** | | | | | |

The blog comments of the above dataset are manually evaluated for performance checking at the feedback level. Among 157 feedbacks, 86 comments as whole feedback are judged as positive, 53 as negative and 18 as neutral feedbacks as described in Table-5.7. The objective of this work is to evaluate the capability of the proposed method to correctly classify the semantic orientation of sentiments of the sentences and also access the positive, negative or neutral sentiments from the dataset. The proposed method achieved 87% results at the feedback level from the sports blog, which is 2% higher than was achieved at sentence level for same dataset. From the results in Table 5.5 (Twitter) and Table 5.7 (Cricinfo), it is observed that number of sentences in blogs can affect the accuracy at feedback level. If feedback or blog contains more sentences, its accuracy could be higher compared to those having less number of sentences.

#### *5.4.1.1 Blogs Evaluation using Precision Recall and F-Test*

There are two standard criteria for classification of text which are Accuracy (precision) and Recall as defined in Eq. 5.3 and Eq. 5.4. Accuracy and recall reflects the quality of classification by using F1 test which is commonly used in literature for text classification. F1 test is computed on the bases of precision and recall values as given in Eq5.5, as shown in Tables 5.8 and 5.9.

Table 5.8 Evaluation of Twitter Comments using Precision Recall and F-Test

| | | Total | Positive | Negative | Accuracy | Recall | F1 Value |
|---|---|---|---|---|---|---|---|
| **Sentences** | Positive | 250 | 212 | 38 | 0.848 | 0.819 | 0.833 |
| | Negative | 250 | 47 | 203 | 0.812 | 0.842 | 0.827 |
| **Feedback** | Positive | 75 | 64 | 11 | 0.853 | 0.831 | 0.842 |
| | Negative | 75 | 13 | 62 | 0.827 | 0.849 | 0.838 |

Table 5.9 Evaluation of Cricket Blog Comments using Precision Recall and F-Test

| | | Total | Positive | Negative | Accuracy | Recall | F1 Value |
|---|---|---|---|---|---|---|---|
| **Sentences** | Positive | 250 | 216 | 34 | 0.864 | 0.837 | 0.850 |
| | Negative | 250 | 42 | 208 | 0.832 | 0.860 | 0.846 |
| **Feedback** | Positive | 75 | 67 | 8 | 0.893 | 0.848 | 0.870 |
| | Negative | 75 | 12 | 63 | 0.840 | 0.887 | 0.863 |

250 sentences and 75 feedbacks have taken containing positive and negative opinion respectively, from twitter and cricinfo shown in Tables 5.8 and 5.9. These sentences and feedbacks are then evaluated to test the performance the proposed method using accuracy (precision), recall and F1 measures. It is clear from the results that average accuracy is 83% at sentence level and 87% at feedback level. Hence it can be concluded that the proposed lexical based method’s performance is better and is adoptive with different domains datasets.

Both statistical (machine learning or corpus-based) and non-statistical (lexicon-based) methods have certain advantages, as reported in results from the recent literature, in terms of sentiment and subjectivity classification at the sentence level. Research conducted on sentence-level subjectivity and sentiment is quite less, and due to the diverse nature of datasets and approaches the results reported in these studies are not directly comparable with each other. Therefore an extensive research is necessary to investigate the benefits and short comings of these approaches. Some of these issues are tried to address in this study.

#### *5.4.1.2 Comparison with Related Work on Blogs Results*

The results of proposed method were compared with corpus based machine learning methods on same datasets from the recent research work. (Go et al., 2009) presented a machine learning method for classifying sentiment of twitter messages and described that pre-processing is more important to remove noisy text in the case of short messages and comments to achieve high accuracy. They have achieved an accuracy of 80% using machine learning algorithms for positive and negative sentiments. (Shamma et al., 2009) investigated the twitter blogs comments for the 2008 American Presidential Electoral debates. They illustrated that the analysis of twitter usage is important and closely yield the semantic structure and contents of the media objects. The twitter can be a predictor of the change in any media event. So mining blogs comments play an important role that can be leveraged to evaluate and analyse any activity.

The proposed method is also compared with (Andreevskaia & Bergler, 2008); presented machine learning based lexical method for different dataset with accuracy of 71% in blogs dataset and 82% for movie reviews and news datasets. The proposed method achieved better results than this approach as shown in Table 5.10. Most corpus based techniques use flat feature vector or BoW methods to represent the documents. However, statistical based techniques rely on subject, domain and language style to gather large amounts of significant data with statistics, while neglecting contextual information and syntactical structure, which in turn affects the accuracy of the sentiment classification at small textual composition levels. So the techniques may not accurately represent the information that can be extracted at sentence level. Therefore for an individual sentence it is imperative for extracting semantic orientation.

The main limitations of the corpus-based approaches are the low attention towards sentence structure and the lower level of contextual valence shifter. On the other hand lexicon based systems suffer from limitations in lexical coverage, WSD, rule of term weighting and a generalized polarity score. Moreover, less attention is given to attenuation, imperial expression or the confidence level of the sentiment orientation in the expression, and there is no proper rule for handling the noisy text with photonic symbols and special characters.

Table 5.10 shows the overall performance of the proposed method in comparison with the machine learning corpus based methods proposed by (Andreevskaia & Bergler, 2008), (Go et al., 2009) using same blog datasets. The main contribution of the proposed method is the extraction of sentence level semantic orientations taking into account all parts of speech and sentence contextual structure.

Table 5.10 Compression with Other Related Works on Blog Datasets

| | | Andreevskaia Bergler,(2008) | Go, Bhayani, Huang,(2009) | Proposed Method |
|---|---|---|---|---|
| **Sentiment Orientation at** | Sentence | 71 | 80 | 83 |
| | Feedback | 82 | 82 | 87 |

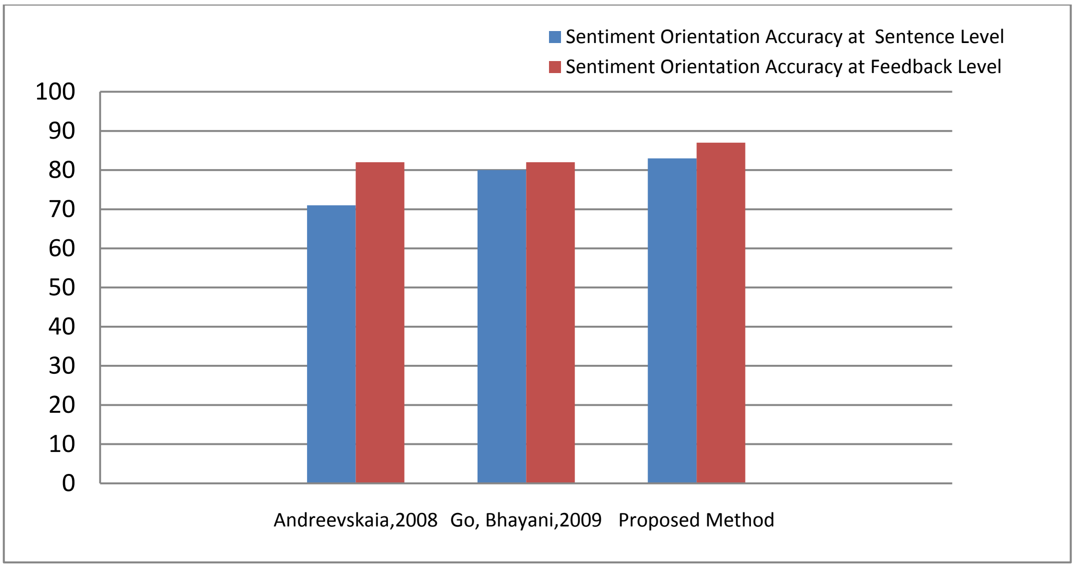


Figure 5.1 Comparison with Other Methods using Blogs Datasets

### 5.4.2 Evaluation and Performance Measure on Customer Reviews Data

In the evaluation of online customer reviews datasets, distinguished between positive and negative reviews is as, a review is positive if it consists of more positive sentences than negative ones and vice versa. Different domain datasets are considered for evaluation and performance of the proposed method.

### *5.4.2.1 Evaluation on Movie Reviews*

Movie review data is collected which has already been processed (Details in chapter-4). There have been 5331 positive and 5331 negative sentences available out of which 1470 sentences are selected. These sentences were then sorted for positive, negative and neutral sentiments, out of which 816 were positive, 446 negative and 208 were neutral. For the feedback/text level classification the same movie reviews dataset is considered which consisting of 1000 positive and 1000 negative reviews in individual text files. For performance checking the proposed method only 185 reviews were taken into consideration, as shown in Table 5.11. These reviews were then manually arranged into 108 positive, 58 negative and 19 neutral reviews. These reviews are processed using the proposed method, based on this study the proposed method achieved results of 86% for sentence level semantic orientation and 97% at feedback level sentiment classification.

Table 5.11 Sentiment Orientation for Movie Reviews

| | | **Actual Orientation** | | | | |
|---|---|---|---|---|---|---|
| **System Assigned at** | | Positive | Negative | Neutral | Total | Accuracy |
| **Sentence Level** | Positive | 750 | 82 | 32 | 864 | 0.868 |
| | Negative | 48 | 350 | 18 | 416 | 0.841 |
| | Neutral | 18 | 14 | 158 | 190 | 0.831 |
| | Total | 816 | 446 | 208 | 1470 | |
| | Overall accuracy | **0.86** | | | | |
| **Feedback Level** | Positive | 105 | 3 | 0 | 108 | 0.972 |
| | Negative | 2 | 55 | 1 | 58 | 0.948 |
| | Neutral | 1 | 0 | 18 | 19 | 0.947 |
| | Total | 108 | 58 | 19 | 185 | |
| | Overall accuracy | **0.97** | | | | |

The proposed method show significant improvement for sentiment classification as compared to other approaches. The baseline system of (Hu & B. Liu, 2004) which uses corpus based method achieved results of 84.4% on same dataset. While (Lipika Dey & S. M. Haque, 2009), achieved 85 % results at sentence level and 97% percent at feedback level. They incorporated the corpus based method and noise remeovel process using seed lists for sementic orientation and also checked the contextual structure of sentences using dictionaries. (Andreevskaia & Bergler, 2008), presented machine learning based lexical method for movie reviews dataset achieved 81% at sentence level and 84% at feedback level. As compared to these approaches when the proposed method was used for semantic classification for the same data set, an accuracy of 86% at sentence level and 97 % at feedback level was achieved as shown in Table 5.12 & Figure 5.2.

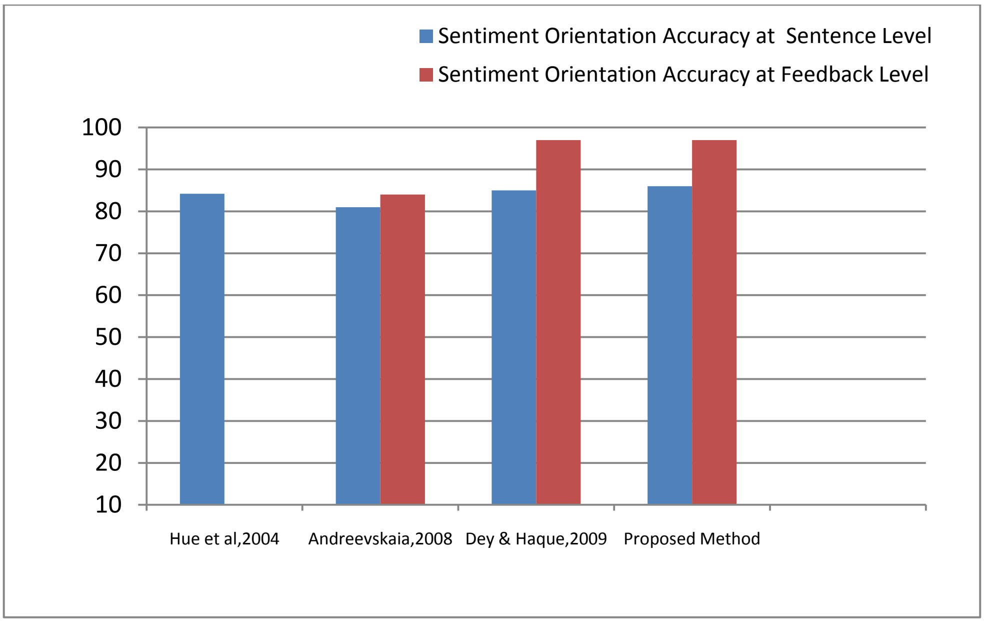


Figure 5.2 Comparison of Proposed Method with Other Approaches for Movie Reviews Dataset

Table 5.12 Comparison with Other Related Approaches Using Movie Reviews

| | | Hue et al ,2004 | Andreevskaia & Bergler,2008 | Dey & Haque, 2009 | Proposed Method |
|---|---|---|---|---|---|
| **Sentiment Orientation Accuracy at** | Sentence | 84.2 | 81 | 85 | 86 |
| | Feedback | --- | 84 | 97 | 97 |

There are several limitations of the methods available today. These approaches focused on one domain and cannot be used on another type of domain and data types; reviews and blogs have a different types and domains. Moreover, concentration on the structure of the sentence and the contextual valence shifter is low, WSD is ignored, the system is based on lexicons suffering from a lexical coverage limitation, less attention is given to attenuate, the rule of term weighting and polarity score is too generalized. From the results it is observed that contextual information in a sentence as well as the sentiment term according to the sentence semantic structure plays an important role in sentence level sentiment classification.

### *5.4.2.2 Evaluation on Hotel Reviews Dataset*

To evaluate the performance of the proposed method on hotel reviews, 2600 hotel reviews have been selected from TripAdvisor, which is one of the popular review sites about hotels and travelling. The dataset is available in XML format and the proposed system is compatible with text only, therefore only the texts were extracted from these reviews using text files. For experimental purpose 120 reviews are chosen which are manually tagged; 56 as positive, 37 as negative and 27 as neutral reviews. A total of 392 sentences were selected out of 120 reviews. These sentences were then manually marked for subject polarity such as 211 were marked as positive, 128 as negative and 53 as neutral sentences. After processing for noise removal and WSD this data was then processed with the proposed method for semantic orientation for positive, negative and neutral sentiments. Results of 81% at sentence level and 84% at feedback level were achieved as compared to manual sentiment tagging. Table 5.13 shows the details of results obtained for hotel reviews dataset. It has been observed during the processing of the hotel reviews dataset that if the number of sentences in a review is large it can have adverse affects on the performance and results of the proposed method. Moreover noise content in the text also degrades the performance and can significantly affect the results.

Table 5.13 Sentiment Orientation of Hotel Reviews

| System Assigned at | | Actual Orientation | | | | |
|---|---|---|---|---|---|---|
| | | Positive | Negative | Neutral | Total | Accuracy |
| **Sentence level** | Positive | 183 | 31 | 9 | 223 | 0.82 |
| | Negative | 23 | 95 | 7 | 125 | 0.76 |
| | Neutral | 5 | 2 | 37 | 44 | 0.84 |
| | Total | 211 | 128 | 53 | 392 | |
| | Overall accuracy **0.81** | | | | | |
| **Feedback level** | Positive | 50 | 5 | 4 | 59 | 0.84 |
| | Negative | 4 | 30 | 3 | 37 | 0.81 |
| | Neutral | 2 | 2 | 20 | 24 | 0.83 |
| | Total | 56 | 37 | 27 | 120 | |
| | Overall accuracy **0.84** | | | | | |

#### *5.4.2.3 Comparison With Other Related Approaches*

The proposed method was compared with other related work in the travelling and hotel domain. The method show good results for sentiment classification as compared to other approaches. The baseline system of (Hu & B. Liu, 2004), which uses corpus based method achieved results of 84.4% as described earlier. (Ye et al., 2009), use machine learning approach using datasets in the same domain they perform different experiments which show an average of 80% results. (Andreevskaia & Bergler, 2008)**,** presented machine learning based lexical method for different domains dataset, which achieved accuracy of 81% at sentence level and 84% at feedback level. As compared to these approaches when proposed method was used for semantic classification for the same hotel reviews dataset, an accuracy of 81% at sentence level and 84% at feedback level was achieved as shown in Table-5.14 & Figure-5.3.

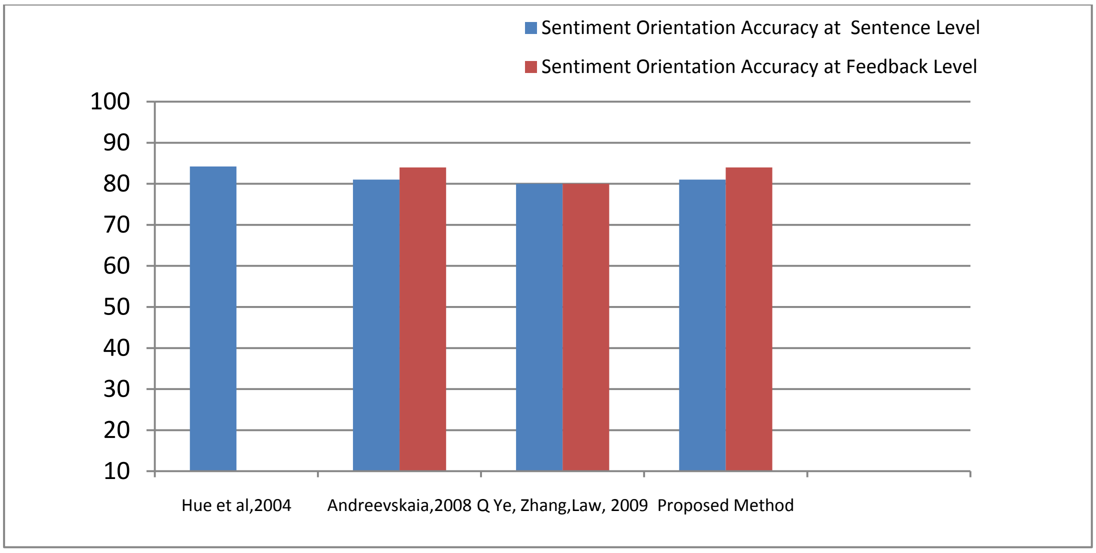


Figure 5.3 Comparison of Proposed Method with other Approaches for Hotel Reviews Dataset

Table 5.14 Comparison with Other Approaches for Hotel Reviews Dataset

| | | Hue et al, 2004 | Andreevskaia & Bergler,2008 | Q Ye, Z Zhang, & R Law, 2009 | Proposed Method |
|---|---|---|---|---|---|
| **Sentiment Orientation Accuracy at** | Sentence | 84.2 | 81 | 80 | 81 |
| | Feedback | --- | 84 | 80 | 84 |

Corpus based machine learning method or methods based on compilations are able to compile lists of negative and positive words dependent on a list of pre defined word list. Most of these approaches need immense annotated training datasets. Lexical based methods can overcome some of these limitations by utilizing dictionary-based approaches since these approaches depend on existing lexicographical resources (such as WordNet) to provide semantic data in regards to individual senses and words.

### *5.4.2.4 Evaluation on Airline Reviews Dataset*

1000 reviews were extracted from Skytrax, a popular airline reviews site; as a dataset to evaluate the proposed method. After splitting these reviews into sentences, an average of 8 sentences per review was found. Subjective lexicons for semantic orientation were extracted from all the positive and negative sentences considering sentence structure for accurate sense extraction. In this dataset a large number of noisy texts was detected and removed during processing. For experiment and evaluation, the dataset is again manually processed for positive, negative and neutral reviews and sentences. 170 reviews out of 1000 are selected and split them into 1296 sentences. Among 170 reviews 103 were marked as positive, 43 as negative and 21 as neutral during manual processing. Out of 1296 sentences, 687 were tagged as positive, 411 as negative and 198 as neutral sentences. From experimental results as shown in Table-5.15, 87% accuracy at sentence level and 96% accuracy at feedback level were achieved.

Table 5.15 Sentiment Orientation for Airline Reviews

| **System Assigned at** | | **Actual Orientation** | | | | |
|---|---|---|---|---|---|---|
| | | Positive | Negative | Neutral | Total | Accuracy |
| **Sentence Level** | Positive | 630 | 52 | 17 | 699 | 0.90 |
| | Negative | 44 | 350 | 21 | 415 | 0.84 |
| | Neutral | 13 | 9 | 160 | 182 | 0.87 |
| | Total | 687 | 411 | 198 | 1296 | |
| | Overall accuracy **0.87** | | | | | |
| **Feedback Level** | Positive | 100 | 2 | 0 | 102 | 0.98 |
| | Negative | 2 | 40 | 0 | 42 | 0.95 |
| | Neutral | 1 | 1 | 21 | 21 | 0.91 |
| | Total | 103 | 43 | 21 | 167 | |
| | Overall accuracy **0.96** | | | | | |

*5.4.2.5 Compassion With Other Relevant Methods*

The proposed method is compared with other related work in this area. However to the best our knowledge no work has been done in the domain of such data types. The relevant work in this area is (Lipika Dey & S. M. Haque, 2009), extracting revies from Web using rule based method by extracting feature list and seed word lists for sementic orientation as corpus based method. Dictionaries were useed for noise remeovel process and for the contextual structure of sentences; which achieved 85 % results at sentence level and 97% percent at feedback level. The methos is also compared with the baseline system of (Hu & B. Liu, 2004), which uses corpus based method achieved results of 84.4% as described earlier.

Others relevant compressions include (Andreevskaia & Bergler, 2008)**,** presented machine learning based lexical method for different domains dataset, which achieved accuracy of 81% at sentence level and 84% at feedback level. The proposed method show good results for sentiment classification as compared to the above approaches, and achieved accuracy, an accuracy of 87% at sentence level and 96% at feedback level using semantic classification for airline reviews dataset, shown in Table 5.16 & Figure-5.4.

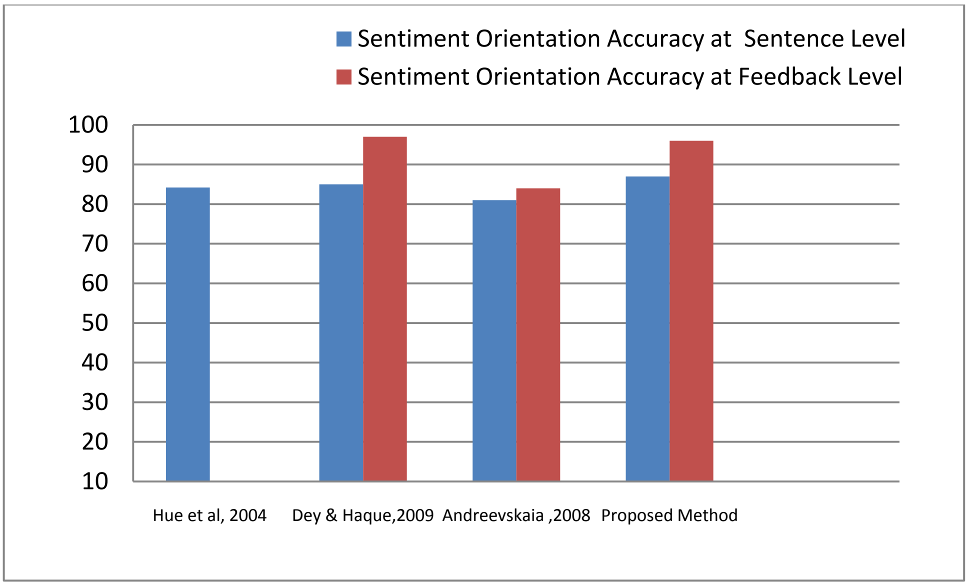


Figure 5.4 Comparison of Proposed Method with other Approaches for Airline Reviews Dataset

Table 5.166 Comparison with Other Approaches for Hotel Reviews Dataset

| | | Hue et al. (2004) | Dey & Haque, 2009 | Andreevskaia & Bergler,2008 | Proposed Method |
|---|---|---|---|---|---|
| **Sentiment Orientation at** | Sentence Level | 84.2 | 85 | 81 | 87 |
| | Feedback Level | | 97 | 84 | 96 |

In this work a technique for domain independent sentence level classification of sentiment is introduced. Rules for all parts of speech are applied so that they can be scored on the strength of their semantics, contextual valence shifter, and sentence structure or expression on the basis of dynamic pattern matching. Moreover, WSD is also addressed to extract accurate sense of the sentence. Opinion type, confidence level, strength and reasons are all can be identified using this system. SentiWordNet and WordNet are utilized as the primary knowledge base which has the further capability of being strengthened by using modifiers, information in the contextual valence shifter and all parts of speech. From the results mentioned above the proposed method perform better as compared to other method in different domains using different datasets.

### *5.4.2.6 Evaluation on Customer Reviews Datasets in Different Domains using Precision Recall and F-Test*

To evaluate the performance of the proposed method on binary classification precision, recall and F-measure were used, shown in Table 5.17.

47 negative and 47 positive feedbacks were taken form movie reviews and 500 positive and 500 negative sentences.  87% precision with 85% recall and 84 % precision with 87% recall is recorded for positive and negative sentiments at sentence level respectively. For the same dataset at feedback level, system achieve 95% precision with 91% recall and 93% precision with 93% recall values for positive and negative sentiments respectively.

Similarly for the hotel reviews each 26 positive negative feedbacks and 300 positive and negative sentence were taken evaluation, The system achieved 85% precision and 82% recall for positive reviews with 83% F1 value and 81% precision and 84% recall for negative reviews with 82% F1 test values at feedback level. For sentence level evaluation 82% precision and 78% recall for positive reviews with 80% F1 values and 76% precision and 81% recall is recorded with 78% F1 value for negative reviews.

Table 5.17 Overall Accuracy of Customer Reviews using Precision Recall and F-Measure

| | | | Total | Positive | Negative | Accuracy | Recall | F1 Value |
|---|---|---|---|---|---|---|---|---|
| **Movie** | Sentences | Positive | 500 | 436 | 64 | 0.872 | 0.850 | 0.861 |
| | | Negative | 500 | 77 | 423 | 0.846 | 0.869 | 0.857 |
| | Feedback | Positive | 47 | 45 | 3 | 0.957 | 0.918 | 0.938 |
| | | Negative | 47 | 4 | 44 | 0.936 | 0.936 | 0.936 |
| **Hotel** | Sentences | Positive | 300 | 245 | 55 | 0.817 | 0.775 | 0.795 |
| | | Negative | 300 | 71 | 229 | 0.763 | 0.806 | 0.784 |
| | Feedback | Positive | 26 | 22 | 4 | 0.846 | 0.815 | 0.830 |
| | | Negative | 26 | 5 | 21 | 0.808 | 0.840 | 0.824 |
| **Airline** | Sentences | Positive | 600 | 527 | 73 | 0.878 | 0.847 | 0.863 |
| | | Negative | 600 | 95 | 505 | 0.842 | 0.874 | 0.857 |
| | Feedback | Positive | 73 | 70 | 3 | 0.959 | 0.886 | 0.921 |
| | | Negative | 73 | 9 | 66 | 0.904 | 0.957 | 0.930 |

For airline reviews better results were achieved in terms of person and recall for positive or negative sentiments orientation both at sentence and feedback levels shown in Table 5.17.

For sentence level evaluation the method achieved 87.8% precision, 84.7% recall and 86.3% F1 measure value for positive text, similarly for negative reviews 84.2% precision and 87.4% recall value with 85.7 % F1 test value is achieved. For feedback level performance on the same dataset, the proposed method achieved 95.9% precision, 88.6% recall and 92.1% F1 value for positive reviews, similarly for negative reviews at sentence level evaluation 90.4% precision, 95.7% recall and 93% F1 valued is achieved as shown in Table 5.17. Hence it can be concluded that the proposed lexical based method's performance is better and is adaptive with different domains datasets in customer reviews and blogs comments.

In this study different factors of the lexical based sentiment orientation approach is examine. This includes those aspects that can result in the enhancement of a lexical based classifier's performance. These factors involves are the acquisition of knowledge-rich lexicon, Noise removal from the text and word or expression actual sense extraction. The study of knowledge-rich lexicon-based methods to achieve sentiment orientation has received relatively little attention in the literature as compared to corpus-based methods. Thus it is clear from the mentioned results that the contribution of this work is a sentence level lexical based method for domain independent sentiment classification.

## 5.5 Summary

This chapter summarizes the obtained results based on different simulation experiments to evaluate the performance of proposed method. The results highlight that the proposed method achieves an average accuracy of 86% at the sentence level and 97% at the feedback level for different customer review datasets. The results also indicate 83% accuracy (on average) at sentence level and 87% accuracy at feedback level for blogs and comments. Hence it can be concluded that the proposed method's performance is better and is adoptive with different domains datasets.

# CHAPTER 6

# CONCLUSIONS AND FUTURE WORK

## 6.1 Introduction

This chapter elaborates the overview of the work, discusses the key results and concludes the thesis. The future possible extensions to this work are also pointed out.

## 6.2 Conclusions

This thesis explores different possibilities of lexical based semantic orientation from online customer reviews and blogs. It includes the thorough study of corpus based machine learning approaches and lexical based sentiment classification to address the issue of domain portability. The sentence level lexical based method is proposed that attempts to enhance the existing lexicon based method with additional benefits of noise removal, improved WSD and knowledge base. The analysis is based on extensive simulation work that confirms the effectiveness of the proposed method.

The work shows its importance for emerging information available online in the form of reviews and comments. The proposed method integrates different components and other lexical resources like POS tagger, WSD, NLP dictionaries, semantic of phonetics and symbols for sentence level classification to address domain portability problem. The method outperforms the existing techniques and is able to classify reviews and blog comments into positive, negative or neutral opinions. The work done in this thesis to address the research questions is concluded as follows.

A module is developed and evaluated, which considers short notations and symbols for their semantics extraction as well as noise removal from the text. For this purpose, the most frequently used symbols and phonetics from web blogs are collected to develop a dictionary to integrate with knowledge base. The symbols and phonetics are referred to the knowledge base for extraction of their semantics. In case of non existence in dictionary, symbols are considered as noise and removed accordingly.

A knowledge base is developed based on the combination of lexical dictionaries that include WordNet, SentiWordNet, predefined intensifiers, POS lexicons, spelling suggester and symbol/phonetics. It is used to assign tag to each word, extract the sense of terms and semantic score of positive, negative or neutral opinions. The semantic score of each opinion word is extracted using the SentiWordNet dictionary that contains the semantic scores of more than 117662 words. Then, the structure and associated words (which affect the weight of the opinion word) in the sentence are checked and the polarity is updated accordingly. The knowledge base calculates semantic strength for each sentence considering the term dependency at sentence level. It contains negation words, enhancers, reducers, model nouns, context shifters and other intensifiers with their semantic scores.

A module is developed to check the opinion terms and to extract their sense based on sentence structure for the removal of WSD. The identification of opinion sentences is performed by checking opinion expressions/terms in sentences using the knowledge base. POS tagging is used to obtain some level of disambiguation for extracting semantic scores from SentiWordNet. However if a multiple sense occurs within the same part of speech then the proposed approach can be used to assign scores based on the predefined rules.

A rule based module is developed to check the polarity of sentences and the contextual information at sentence level. The module extracts the contextual information from the sentence and calculates its semantic orientation using lexical dictionaries. It is used for sentence level semantic pattern extraction by considering all POS of the sentence. The final weight of each sentence and review is calculated to decide about its positive, negative or neutral polarity. The proposed method is compared with other related works for the validation. For blog comments, it is compared with corpus based machine learning methods (Go et al., 2009) (Andreevskaia & Bergler, 2008) and it achieves an average accuracy of 83% at sentence level and 86% at feedback level. For customer reviews, the method achieves an average accuracy of 86% at sentence level and 97% at feedback level which shows good improvement for sentiment classification as compared to other approaches that include (Hu & B. Liu, 2004), (Lipika Dey & S. M. Haque, 2009) and (Andreevskaia & Bergler, 2008). Hence, the method solves domain portability issues as it is validated by comparing with the other related works for different domains.

## 6.3 Contributions

The thesis addresses the issue of semantic orientation from online customer reviews and blogs. The thesis contributions can be structured into four areas to develop lexical based sentence level semantic orientation method to address the issues of domain portability: 1) Noise removal 2) Knowledge base 3) WSD considering sentence structure 4) Polarity at sentence level using contextual sentence structure.

The main contribution of this thesis is the development of lexical based sentence level semantic orientation method for online reviews and blog comments to address the issue of domain portability. The method is used to check the polarity of sentences and contextual information at sentence level. It extracts the contextual information from the sentence and calculates its semantic orientation using lexical dictionaries. From the results, it is evident that contextual information and consideration of sentence structure is effective in sentiment classification used in different domains.

Another contribution is the development of module for noise removal, identification and semantics extraction of short notations/symbols from reviews/comments and also helps to improve the performance of classifier for semantic orientation. The module is used to consider the short notations and symbols for their semantics extraction as well as noise removal from the text.

The de**v**elopment of knowledge base from lexical dictionaries (WordNet, SentiWordNet, POS lexicons, spelling suggester, intensifiers, phonetics features and opinion terms) is another major contribution of this thesis. It deals with text tagging, identifying sentence structure and contextual dependency, contains different senses of opinion terms and semantic scores for each term.

Another major contribution of this thesis is to develop a module for sense extraction at sentence level. The module is developed and evaluated to provide more solid term sense extraction using the sentence structure. WSD has great impact on sentiment classification and helps to determine the actual polarity of opinions. All parts of speech are utilized (nouns, adverbs, verbs and adjectives) to validate the use of polarity words for sentence level sentiment classification.

The final contribution of the thesis is development of a module which defines rules for determining opinion orientation of each recognized sentence, review or comment. It is also used to extract contextual information from the sentence and calculates its semantic orientation using lexical dictionaries.

### 6.4 Study Limitations

As the nature of knowledge, every work has to have some limitations to ensure the future research connections in that field. Similarly, this work also has some limitation as follow.

The limitation of this work includes the dependency on lexical dictionaries. Different lexicons dictionaries are interconnected and used for the semantic orientation and polarity of text. Hence the proposed method is dependent on these dictionaries for semantic orientations.

Another limitation of this work is the lack of word sense disambiguation. WSD is natural language processing topic, which needs more solid methods for extraction of sense of term according to the contextual sentence pattern.

Another limitation is the extraction of semantic and processing of short abbreviated words and symbols. Web contents contain a great deal of noisy text particularly in case of social networking sites and customer reviews. There is a need of more sophisticated methods to remove noise for such sources. Likewise, methods are required to extract information and knowledge, semantics of these symbols and short abbreviated terms for the effective classification.

## 6.5 Future Work

The limitations of any research work open new possibilities for future research. Hence, the limitations of this work may be the foundation for future research. One major possible direction for future research is the combination of lexical based method and corpus based approach with improved and well-rich knowledge base for optimized sentiment classification.

The lexical based method may perform better at sentence level with discourse modifiers, improved sentence contextual information, addition of valence shifter handling and semantic score of all parts of speech in a sentence.

Using WSD to extract the acute sense of sentiment words according to the sentence structure may improve the performance of the method. One opinion term /expression have few senses. Hence the extraction of accurate sense according to the contextual information in sentences can enhance the performance of the semantic orientation.

Removal of noise with accurate semantic extraction from short abbreviated words and symbols may also improve the sentiment orientation especially in blogs. In online Web forums, social networks, blogs etc., people frequently write short abbreviated words and use symbols to express their views. These symbols show emotions and expressions which are very important for sentiment analysis. How these symbols and short words are made useful in extracting their semantics is one of the directions for future research.

# REFERENCES

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 579-586).

Andreevskaia, A., & Bergler, S. (2006). Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. *Proceedings of EACL* (Vol. 6, pp. 209-216).

Andreevskaia, A., & Bergler, S. (2007). CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 119-120).

Andreevskaia, A., & Bergler, S. (2008). When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. *Proceedings of ACL-08: HLT*, 290-298. Citeseer.

Andrew Lipsman. (2007). Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior. *comScore, Inc. Industry Analysis*, 2-28.

Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Multi-facet rating of product reviews. *Advances in Information Retrieval , LNCS 5478* (pp. 461-472). Springer.

Balahur, A., & Montoyo, A. (2009). A Semantic Relatedness Approach to Classifying Opinion from Web Reviews. *Procesamiento del lenguaje natural*, *42*, 47-54.

Balahur, A., Steinberger, R., Goot, E., Pouliquen, B., & Kabadjov, M. (2009). Opinion mining on newspaper quotations. *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03* (pp. 523-526).

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., et al. (2010). Sentiment analysis in the news. *Proceedings of LREC* (Vol. 10).

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text* (pp. 2224-2236).

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Proceeding of the*

*45th Annual Meeting-Association For Computational Linguistics* (Vol. 45, pp. 440-447). Prague, Crech Repunlic.

Breck, E., Choi, Y., & Cardie, C. (2007). Identifying expressions of opinion in context. *Proceedings of IJCAI-2007*, 2683-2688.

Brooke, Juilen. (2009). *A Semantic Approach for Automatic Text Sentiment Analysis*. Department Of Liguistics, SIMON FRASER University.

Cardie, C., Wiebe, J., Wilson, T., & Litman, D. (2003). Combining low-level and summary representations of opinions for multi-perspective question answering. *Working Notes-New Directions in Question Answering (AAAI Spring Symposium Series)* (pp. 1-13).

Castellano, M., Mastronardi, G., Aprile, A., & Tarricone, G. (2007). A Web Text Mining Flexible Architecture. *International Journal of Computer Science and Engineering, 1*(4), 78-85. Citeseer.

Chen, H., & Zimbra, D. (2010). AI and Opinion Mining. *IEEE Intelligent Systems, IEE computer Society*, 74-80. Published by the IEEE Computer Society.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naive Bayes. *Expert Systems with Applications*, *36*(3), 5432-5435. Elsevier.

Chesley, P., Vincent, B., Xu, L., & Srihari, R. K. (2006). Using verbs and adjectives to automatically classify blog sentiment. *Training*, *580*(263), 233-235.

Choi, Y., Breck, E., & Cardie, C. (2006). Joint extraction of entities and relations for opinion recognition. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 431-439).

Choi, Y., & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 793-801).

Clark, A. (2003). Pre-processing very noisy text. *Proc. of Workshop on Shallow Processing of Large Corpora* (pp. 12-22).

D Avanzo, E., Lieto, A., & Kuflik, T. (2008). *Manually vs semiautomatic domain specific ontology building*. Universita Dedli Studi Di Salerno.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web* (pp. 519-528).

Dey, L, & Haque, S. K. (2008). Opinion mining from noisy text data. *Proceedings of the second workshop on Analytics for noisy unstructured text data* (pp. 83-90). Singapore: ACM.

Dey, Lipika, & Haque, S. M. (2009). Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition (IJDAR), 12*(3), 205-226. doi: 10.1007/s10032-009-0090-z.

Di, N., Yao, C., Duan, M., Zhu, J., & Li, X. (2008). Representing a web page as sets of named entities of multiple types: a model and some preliminary applications. *Proceeding of the 17th international conference on World Wide Web* (pp. 1099-1100).

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. *Proceedings of the international conference on Web search and web data mining* (pp. 231-240).

Esuli, A. (2008). Automatic generation of lexical resources for opinion mining: models, algorithms and applications. *ACM SIGIR Forum* (Vol. 42, pp. 105-106).

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 617-624).

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of LREC* (Vol. 6, pp. 417-422).

Etzioni, O. (1996). The World-Wide Web: quagmire or gold mine? *Communications of the ACM, 39*(11), 65-68. ACM.

Falinouss, P. (2007). *Stock trend prediction using news articles: a text mining approach*. Lule{å} tekniska universitet.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press.

Fensel, D. (2004a). *Ontologies: a silver bullet for knowledge management and electronic commerce* (pp. 1-99). Springer Verlag, pp 1-99.

Fensel, D. (2004b). Ontologies: a silver bullet for knowledge management and electronic commerce (pp. 1-119). Springer Verlag.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* (pp. 1-12). Stanford University.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1), 60-76.

Gurevych, P. I., & Oprak, D.-I. C. (2010). Sentiment Analysis for User Generated Discourse in eLearning 2.0 (SENTAL). *Technische Vniversitat Darmstadt*. Retrieved from http://www.ukp.tu-darmstadt.de/?id=2664.

Hariharan, S., Srimathi, R., Sivasubramanian, M., & Pavithra, S. (2010). Opinion mining and summarization of reviews in web forums. *Proceedings of the Third Annual ACM Bangalore Conference* (pp. 1-4).

Hearst, M. (1992). Direction-based text interpretation as an information access refinement. *Text-Based Intelligent Systems. Lawrence Erlbaum Associates*, 257-274. Citeseer.

Hoffman, T. (2008). Online reputation management is hot but is it ethical. *Computerworld, February*, 1-4.

Hotho, A., N urnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum-GLDV Journal for Computational Linguistics and Language Technology* (Vol. 20, pp. 19-62).

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

Huettner, A., & Subasic, P. (2000). Fuzzy typing for document management. *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, 26-27.

Jindal, N., & Liu, B. (2006). Identifying comparative sentences in text documents. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 244-251).

Jindal, N., & Liu, B. (2006). Mining comparative sentences and relations. *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, pp. 1331-1336).

Kamps, J., Marx, M. J., Mokken, R. J., & De Rijke, M. (2004). *Using wordnet to measure semantic orientations of adjectives*. European Language Resources Association (ELRA).

Kawamura, T., Nagano, S., Mizoguchi, Y., Inaba, M., Yamasaki, T., & Okamoto, M. (2008). Ontology-based WOM extraction service from weblogs. *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 2231-2236).

Khan, A., Baharudin, B., & Khan, K. (2011a). Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews, *6*(10), 1141-1157.

Khan, A., Baharudin, B., & Khan, K. (2011b). Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure.

*Communications in Computer and Information Science, Software Engineering and Computer Systems*, (175th ed., pp. 317-331). Springer Verlag.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology, 1*(1), 4-20. Academy Publisher, PO Box 40 Oulu 90571 Finland.

Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. *Proceedings of the 20th international conference on Computational Linguistics* (pp. 1367-1374).

Kim, S. M., & Hovy, E. (2005). Automatic detection of opinion bearing words and sentences. *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)* (pp. 61-66).

Leshed, G., & Kaye, J. J. (2006). Understanding how bloggers feel: recognizing affect in blog posts. *CHI'06 extended abstracts on Human factors in computing systems* (pp. 1019-1024).

Leung, C. W. K., & Chan, S. C. F. (2008). Sentiment Analysis of Product Reviews. *Encyclopedia of Data Warehousing and Mining Information Science Reference,,* 1794-1799. Citeseer.

Liu, B. (2010a). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,,* 978-1420. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN.

Liu, B. (2010b). Sentiment analysis: a multifaceted problem. *IEEE Intelligent System. v25*, 76-80.

Liu, B., & Chen-Chuan-Chang, K. (2004). Editorial: special issue on web content mining. *ACM SIGKDD Explorations Newsletter, 6*(2), 1-4. ACM.

Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. *Proceedings of the 8th international conference on Intelligent user interfaces* (pp. 125-132).

Liu, H., & Motoda, H. (1998). Feature extraction, construction and selection: A data mining perspective (pp. 1-411). Kluwer Academic Publishers.

Loh, S., Wives, L. K., & Oliveira, J. P. M. de. (2000). Concept-based knowledge discovery in texts extracted from the Web. *ACM SIGKDD Explorations Newsletter, 2*(1), 29-39. ACM.

Lu, Y., Kong, X., Quan, X., Liu, W., & Xu, Y. (2010). Exploring the sentiment strength of user reviews. *Web-Age Information Management* (pp. 471-482). Springer.

Montañés, E., Fernández, J., Díaz, I., Combarro, E., & Ranilla, J. (2003). Measures of rule quality for feature selection in text Categorization. *Advances in Intelligent Data Analysis V*, 589-598. Springer.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70-77).

Nathan, S. (2009). *Improving Sentimental Classifications Using Contextual Sentences Lexical Base* (pp. 1-8). Stanford University: Department of Computer, Stanford University.

Nedellec, C. (2000). Corpus-based learning of semantic relations by the ILP system, Asium. *Learning language in logic*, 461-491. Springer.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009). Semantically distinct verb classes involved in sentiment analysis. *Proc. IADIS Int. Conf. Applied Computing, 2009 IADIS AC (1)* (pp. 27-35).

Ohana, B. (2009). *Opinion mining with the SentWordNet lexical resource*. Dublin Institute of Technology.

Pang, B, & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135. Now Publishers Inc.

Pang, B, Lee, L, & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86).

Pang, Bo, & Lee, Lillian. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the ACL* (pp. 115-124).

Pol, K., Patil, N., Patankar, S., & Das, C. (2008). A Survey on Web Content Mining and extraction of Structured and Semistructured data. *Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on* (pp. 543-546).

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, 1-10. Springer.

Popescu, A. M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 339-346).

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Pearson Education India.

Raghavan, P., Amer-Yahia, S., & Gravano, L. (2004). *Structure in Text: Extraction and Exploitation*. *Proceeding of the 7th international Workshop on the Web and Databases (WebDB), ACM SIGMOD/PODS*. © 2002 Verity, Inc.

Sack, W. (1995). On the computation of point of view. *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE* (pp. 1488-1492).

Saggion, H., & Funk, A. (2010). Interpreting SentiWordNet for Opinion Classification. *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10* (pp. 1129-1133).

Sarvabhotla, K., Pingali, P., & Varma, V. (2009). Supervised Learning Approaches for Rating Customer Reviews. *Journal of Intelligent Systems, 19*(1), 79-94.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR), 34*(1), 1-47. ACM.

Shamma, D. A., Kennedy, L., & Churchill, E. F. (2009). Tweet the debates: understanding community annotation of uncollected sources. *Proceedings of the first SIGMM Workshop on Social Media* (pp. 3-10).

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, *33*(1), 1-5. Elsevier.

Sharp, M. (2001). Text mining. *Rutgers University, School of Communication, Information and Library Studies* (pp. 1-12). Allen, J. Natural Language Understanding, Second Edition. Redwood City, CA: Benjamin/Cummings.

Song, M. H., Lim, S. Y., Kang, D. J., & Lee, S. J. (2005). Automatic classification of web pages based on the concept of domain ontology. *Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05)* (pp. 114-120). IEEE Computer Society.

Srivastava, J., Desikan, P., & Kumar, V. (2002). Web Mining: Accomplishments & Future Directions. *National Science Foundation Workshop on Next Generation Data Mining (NGDM'02)* (pp. 51-69).

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Taboada, M., Brooke, J, & Stede, M. (2009). Genre-based paragraph classification for sentiment analysis. *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 62-70).

Tenenboim, L., Shapira, B., & Shoval, P. (2008). Ontology-based classification of news in an electronic newspaper. *International Book Series "Information Science and Computing"* (pp. 98-97). Institute of Information Theories and Applications FOI ITHEA.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)* (pp. 417-424).

Turney, P., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* (pp. 211-219).

Waltinger, U. (2009). Polarity reinforcement: Sentiment polarity identification by means of social semantics. *IEEE AFRICON'09., AFRICON,* (pp. 1-6). September 2009, Nairobi, Kenya.

Wang, Y., & Wang, X. J. (2005). A New Approach to feature selection in Text Classification. *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on* (Vol. 6, pp. 3814-3819).

Wang, Z. Q., Sun, X., Zhang, D. X., & Li, X. (2006). An Optimal Svm-Based Text Classification Algorithm. *Machine Learning and Cybernetics, 2006 International Conference on* (pp. 1378-1381).

Westerski, A. (2007). Sentiment Analysis : Introduction and the State of the Art overview. *Universidad Politecnica de Madrid, Spain* (pp. 211-218).

Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 625-631).

Wiebe, J. M. (1990). Identifying subjective characters in narrative. *Proceedings of the 13th conference on Computational linguistics-Volume 2* (pp. 401-406).

Wiebe, J. M. (1994). Tracking point of view in narrative. *Internation Journal of Computational Linguistics*, *20*(2), 233-287. MIT Press.

Wiebe, J. M., & Bruce, R. F. (2001). Probabilistic classifiers for tracking point of view. *PROGRESS IN COMMUNICATION SCIENCES*, 125-142. ABLEX PUBLISHING CORPORATION.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 246-253).

Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., et al. (2003). *Recognizing and organizing opinions expressed in the world press. Working Notes-New Directions in Question Answering (AAAI Spring Symposium Series)*. Working Notes-New Directions in Question Answering -AAAI Spring Symposium Series.

Wiebe, J., & Mihalcea, R. (2006). Word sense and subjectivity. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 1065-1072).

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. *Computational Linguistics and Intelligent Text Processing*, 486-497. Springer.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Internation Journal of Computational linguistics*, *30*(3), 277-308. MIT Press.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347-354).

Wong, W., Liu, W., & Bennamoun, M. (2006). Integrated scoring for spelling error correction, abbreviation expansion and case restoration in dirty text. *Proceedings of the fifth Australasian conference on Data mining and analystics-Volume 61* (pp. 83-89).

www.Oracle.com. (2008). *ORACLE*. Retrieved November 2008, from www. oracle .com.

Yan, J., Liu, N., Zhang, B., Yan, S., Chen, Z., Cheng, Q., et al. (2005). OCFS: optimal orthogonal centroid feature selection for text categorization. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 122-129).

Yao, C., Yu, Y., Shou, S., & Li, X. (2008). Towards a global schema for web entities. *Proceeding of the 17th international conference on World Wide Web* (pp. 999-1008).

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*(3), 6527-6535. Elsevier.

Yeh, A. S., Hirschman, L., & Morgan, A. A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *International Journal of Bioinformatics, 19*(suppl 1), i331-i339. Oxford Univ Press.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of the 2003 conference on Empirical methods in natural language processing-Volume 10* (pp. 129-136).

Zhao, J., Liu, K., & Wang, G. (2008). Adding redundant features for CRFs-based sentence sentiment classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 117-126).

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, *74*(2), 133-148. Am Marketing Assoc.

# LIST OF PUBLICATIONS

## CONFERENCES

1. **Aurangzeb Khan**, Baharum Baharudin, Khairullah Khan, “An Overview of E-Documents Classification." International Conference on Machine Learning and Computing, , Perth, Australia 2009, IPCSIT (2011), pp.544-552

2. **Aurangzeb Khan**, Baharum B Baharudin, Khairullah Khan, "Frequent Patterns Mining of Stock Data Using Hybrid Clustering Association Algorithm," IEEE-ICIME, 2009, pp.667-671.

3. Khairullah Khan, Baharum Baharudin, & **Aurangzeb Khan**, “Mining opinion from text documents: A survey”. 3rd IEEE International Conference on Digital Ecosystems and Technologies (DEST-09), 2009, pp. 217-222.

4. **Aurangzeb Khan**, Baharum Baharudin, Khairullah Khan, “Efficient Feature Selection and Domain Relevance Term Weighting Method for Text Classification”, 2nd IEEE- ICCEA- Indonesia – 2010, pp. 98-403.

5. **Aurangzeb Khan**, Baharum Baharudin, Khairullah Khan, “Semantic Based Features Selection and Weighting Method for Text Classification”, IEEE, ITSIM, Kula Lumpure-2010, pp 850-855.

6. **Aurangzeb Khan**, Baharum Baharudin, Khairullah Khan, “Subjectivity Based Feature and Opinion Sentences Extraction from Online Reviews." International Conference on Intelligence and Information Technology (ICIIT’10) IEEE 2010. pp. 547-551.

7. **Aurangzeb Khan,** Baharum Baharudin, Khairullah Khan, “Sentence Based Sentiment Classification from Online Customer Reviews." Frontiers of information Technology, ACM, FIT '10, -2010, pp.25

8. **Aurangzeb Khan,** Baharum Baharudin, Khairullah Khan; “Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence

Structure" ICSECS 2011: 2nd International Conference on Software Engineering and Computer Systems, Springer Pahang, Malaysia, 2011, pp. 317-331.

9. **Aurangzeb Khan,** Baharum Baharudin, Khairullah Khan; "Sentiment classification using sentence-level semantic orientation of opinion terms from blogs and online social network forums", Accepted NPC-IEEE, UTP, Malaysia, 2011.

**JOURNALS**

10. **Aurangzeb Khan**, Baharum B Baharudin, Lam Hong Lee, Khair ullah Khan, and "A Review of Machine Learning Algorithms for Text-Documents Classification" Journal of Advances in Information Technology JAIT, Volume-1, 2010, pp.4-20.

**11.** **Aurangzeb Khan**, Baharum Baharudin, Khairullah Khan' "Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews" International Journal of Trends in Applied Sciences**.** ISI- SCOPUS. Volume-6, 2011, pp. 1141-1157.

12. **Aurangzeb Khan**, Baharum Baharudin, Khairullah khan; "Mining Customer Data for Decision Making Using New Hybrid Classification Algorithm", Journal of Theoretical and Applied Information Technology, SCOPUS, Vol, 27. No.1, 2011, pp.54-61,

**13.** **Aurangzeb Khan**, Baharum Baharudin, Khairullah Khan; "Sentence Level Semantic Orientation of Online Reviews and Blogs using SentiWordNet for Effective Sentiment Classification", International Journal of Computer Applications in Technology, Under-review (2011).

14. **Aurangzeb Khan**, Baharum Baharudin, Khairullah Khan; "Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs" Int. J Comp Sci. Emerging Tech. Vol-2, No 4, 2011.

## BOOK CHAPTER

15. **Aurangzeb khan**, Baharum Baharudin and Khairullah Khan “Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure” Communications in Computer and Information Science, 1, Volume 179, Software Engineering and Computer Systems, Part 5, Pages 317-331.

# APPENDIX A

## Main Simulation File

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Data.OleDb;
using System.Data;
using System.Windows.Forms;

namespace PrjOpeningMining
{
    public  class ClsConnection
   {
     OleDbConnection Cn = new
OleDbConnection("Provider=Microsoft.ACE.OLEDB.12.0;Data Source=F:\\Project
Material\\Sentiment Analysis\\Data Base\\sentiwordnet.accdb");
     OleDbDataAdapter Adp = new OleDbDataAdapter();
      OleDbCommand Cmd = new OleDbCommand();

      private void Openconnection()
      {
         if (Cn.State == ConnectionState.Closed)
            Cn.Open();
      }
      public void Executecommand(string Sqlstatements)
      {
         Openconnection();
         Cmd.Connection = Cn;
      Cmd.CommandType = CommandType.Text;
      Cmd.CommandText = Sqlstatements;
      Cmd.ExecuteNonQuery();
    }
        public int ExecuteScaler(string Sqlstatements)
    {
      int k;
     Openconnection();
      Cmd.Connection = Cn;
      Cmd.CommandType = CommandType.Text;
      Cmd.CommandText = Sqlstatements;
      k = Convert.ToInt16(Cmd.ExecuteScalar());
      return k;
    }

public string  ExecuteScalerStrin(string Sqlstatements)
    {
      string  k;
     Openconnection();
      Cmd.Connection = Cn;
      Cmd.CommandType = CommandType.Text;
      Cmd.CommandText = Sqlstatements;
      k = Cmd.ExecuteScalar().ToString();
      return k;
    }

    public void FillDset(ref DataSet Dset, string Statements)
    {
      Openconnection();
      Adp.SelectCommand = new OleDbCommand ();
      Adp.SelectCommand.Connection = Cn;
      Adp.SelectCommand.CommandType = CommandType.Text;
```

```
        Adp.SelectCommand.CommandText = Statements;
        Adp.SelectCommand.ExecuteNonQuery();
        Dset.Clear();
        Adp.Fill(Dset, "Dtable");
    }
    public void FillCombo(ref DataSet Dset, string Statements, ref ComboBox CBO,
string DisplayMember, string ValueMember)
    {
        Openconnection();
        Adp.SelectCommand = new OleDbCommand ();
        Adp.SelectCommand.Connection = Cn;
        Adp.SelectCommand.CommandType = CommandType.Text;
        Adp.SelectCommand.CommandText = Statements;
        Adp.SelectCommand.ExecuteNonQuery();
        Adp.Fill(Dset, "Dtable");
        CBO.DataSource = Dset.Tables[0].DefaultView;
        CBO.DisplayMember = DisplayMember;
        CBO.ValueMember = ValueMember;
    }
    public void FillListBox(ref DataSet Dset, string Statements, ref ListBox LST,
string DisplayMember, string ValueMember)
    {
        Openconnection();
        Adp.SelectCommand = new OleDbCommand ();
        Adp.SelectCommand.Connection = Cn;
        Adp.SelectCommand.CommandType = CommandType.Text;
        Adp.SelectCommand.CommandText = Statements;
        Adp.SelectCommand.ExecuteNonQuery();
        Adp.Fill(Dset, "Dtable");
        LST.DataSource = Dset.Tables[0].DefaultView;
        LST.DisplayMember = DisplayMember;
        LST.ValueMember = ValueMember;
    }

    public void FillLGrid(ref DataSet Dset, string Statements, ref DataGridView
GRD)
    {
      Openconnection();
      Adp.SelectCommand = new OleDbCommand ();
      Adp.SelectCommand.Connection = Cn;
      Adp.SelectCommand.CommandType = CommandType.Text;
      Adp.SelectCommand.CommandText = Statements;
      Adp.SelectCommand.ExecuteNonQuery();
      Adp.Fill(Dset, "Dtable");
      GRD.DataSource = Dset.Tables[0].DefaultView;
    }

    public int  Generate_Maximum(string Table_Name, string Field_Name)
    {
      int k;
      Openconnection();
      Cmd.Connection = Cn;
      Cmd.CommandType = CommandType.Text;
      Cmd.CommandText = "select iif(isnull(max(" + Field_Name + ")),0,max(" +
Field_Name + "))+1 from " + Table_Name + "";
      k = Convert.ToInt32(Cmd.ExecuteScalar  ());
      return k;
    }

  }
}
//++++++++++++++++++++++++++++++++++++++++++++++++++++

using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.Linq;
using System.Text;
using System.Windows.Forms;
using System.Collections;
using  System.Runtime.InteropServices ;
using System.Reflection;
using Microsoft.Office.Core ;
using Word;

namespace PrjOpeningMining
{
    public partial class FrmOpenion : Form
    {
        public FrmOpenion()
        {
            InitializeComponent();
        }
        // cls is a instance of clsconnection class
        ClsConnection cls = new ClsConnection();
        string str;
        bool opinion_sentence;

        int ssno;
        int sent=1;
        int pos;
        string Pattern ;
        string patternselected;
        int sense_number;
        string pptrn;
        string sentencesensepattern;

        //comparative section
        int totalsent;
        decimal totalwe;
```

```
int compsent;
int totalpos;
decimal poswei;
int totalneg;
decimal negpos;
//end of comparative section
//browse button start
private void btnBrowse_Click(object sender, EventArgs e)
{
    OpenFileDialog de = new OpenFileDialog();
    de.CheckFileExists = true;
    de.Title = "Choose Text File";
    de.Filter = "Text Format File|*.txt";
    de.ShowDialog();
    if (de.FileName.Length > 0)
    {
        txtReadFile.LoadFile(de.FileName, RichTextBoxStreamType.PlainText);
    }
}
// end of browse button


// start of tagging
private void btnTag_Click(object sender, EventArgs e)
{
    // here jjscore is used for the score of jj ,nn and nns
    decimal jjscore = 0;
    opinion_sentence = false;

    // text cleaning start
    this.txtReadFile.Text = this.txtReadFile.Text.Replace(")", " ");
    this.txtReadFile.Text = this.txtReadFile.Text.Replace("]", " ");
    this.txtReadFile.Text = this.txtReadFile.Text.Replace("[", " ");
    this.txtReadFile.Text = this.txtReadFile.Text.Replace("    ", " ");
    this.txtReadFile.Text = this.txtReadFile.Text.Replace("   ", " ");
    this.txtReadFile.Text = this.txtReadFile.Text.Replace("  ", " ");
    this.txtReadFile.Text = this.txtReadFile.Text.Replace("\n", " ");
```

```
            this.txtReadFile.Text = this.txtReadFile.Text.Replace("\"", " ");
          //Call module noise removal and semantic extractuion from symbols and short
words
            this.txtReadFile.Text = this.txtReadFile.Text.Replace("*k*", "kiss");
            this.txtReadFile.Text = this.txtReadFile.Text.Replace("*K*", "kiss");
            this.txtReadFile.Text = this.txtReadFile.Text.Replace(";-)~~~~~~~~~", "giving
someone the raspberries.");
            this.txtReadFile.Text    =    this.txtReadFile.Text.Replace("(((((person)))))",
"giving them a virtual hug.");
            this.txtReadFile.Text = this.txtReadFile.Text.Replace("\\~//", "glass with a
drink. (usually booze)");
            this.txtReadFile.Text = this.txtReadFile.Text.Replace("^5", "high five");
            this.txtReadFile.Text = this.txtReadFile.Text.Replace("?^", "Whats Up?");


            //end of replace simbols
            // delete from sentences , words and pattern from database tables
            cls.Executecommand("delete from Sentences");
            cls.Executecommand("delete from Words");
            cls.Executecommand("delete from pattern");
            // end delete from sentences and words from database



            // tokenization and tagging
            NLPlib tagger = new NLPlib();
            string s = this.txtReadFile.Text;
            ArrayList v = tagger.tokenize(s);
            ArrayList t = tagger.tag(v);
            // end tokenization and tagging

            // select negation words from database

            DataSet dsetnegation = new DataSet();

            // end select negation words from database

            // main loop for tagging each words
            // PB. for progress bar
            this.PB.Maximum = v.Count;
            for (int i = 0; i < v.Count; i++)
```

```
        {


             //display in textbox in specific format
            str = str + v[i] + t[i] + " ";

            //set words position
            pos = pos + 1;



            // check noun and then insert into database words(table)
            if (t[i].ToString () == "/NN")
                {
                  //variable for sentence whis is selected from database
                  string selectedsentence;
                  selectedsentence   =   cls.ExecuteScalerStrin("select   sentence   from
sentences where sentence_id= " + sent + "").ToString();

               // sensefind is a function to which we pass word and selected sentence
pattern

                  ssno= SenseFind(v[i].ToString(), sentencesensepattern);
               //dataset for nn
               DataSet dsetnn = new DataSet();
               // if sense is not found select pos neg score without sense
               if (ssno == 0)
               {
                  cls.FillDset(ref  dsetnn,  "select  posscore,negscore  from  sentiword
where  synsetterms like '" + v[i].ToString() + "#_' and (posscore > 0 or negscore > 0)
and pos = 'n' ");
               }
                  //if sense is found select pos neg score of the selected sense
               else
               {
                cls.FillDset(ref dsetnn, "select posscore,negscore from sentiword where
synsetterms like '" + v[i].ToString() + "#" + ssno + "' and (posscore > 0 or negscore >
0) and pos = 'n' ");

               }

                     //  if  records  found  means  word  have  score  then  consider  the
sentences as a opinon
               if (dsetnn.Tables[0].Rows.Count > 0)
               {
                  opinion_sentence = true;
                  // if positive score is > negative score
Convert.ToDecimal(dsetnn.Tables[0].Rows[0].ItemArray[1]) * -1;
```

```
            }
          }

          //insert word into words table without score bcoz we will update the
score at update polrity button
          cls.Executecommand("insert into Words values ('" + v[i].ToString() +
"',1,Null,Null," + sent + "," + pos + ")");
           }
        // end of noun checking

        //check adjective and insert into database
        else if (t[i].ToString() == "/JJ" || t[i].ToString() == "/JJS" || t[i].ToString()
== "/JJR" )
          {
            string selectedsentence;
            selectedsentence = cls.ExecuteScalerStrin("select sentence from
sentences where sentence_id= " + sent + "").ToString();
            sentencesensepattern = WordSentence(v[i].ToString(),
selectedsentence);

            ssno= SenseFind(v[i].ToString(), sentencesensepattern);
            //find sense number


            ////return sense number


             cls.Executecommand("insert into Words values ('" + v[i].ToString() +
"',2,Null,Null," + sent + "," + pos + ")");

          // select positive and negative score from sentiword net of the the
particular adjective
             DataSet dsetjj = new DataSet();

             if (ssno == 0)
             {
               cls.FillDset(ref dsetjj, "select posscore,negscore from sentiword
where  synsetterms like '" + v[i].ToString() + "#_' and (posscore > 0 or negscore > 0)
and pos = 'a' ");

             }
               if ( dsetjj.Tables[0].Rows.Count == 0)
               {
                 cls.FillDset(ref dsetjj, "select posscore,negscore from sentiword
where  synsetterms like '" + v[i].ToString() + "#_' and (posscore > 0 or negscore > 0)
and pos = 'a' ");
```

```
                            }

                        }
                            // if records found
                            if (dsetjj.Tables[0].Rows.Count > 0)
                                {
                                    opinion_sentence = true;
                                    // if positive score is > negative score
                                    if (Convert.ToDecimal(dsetjj
.Tables[0].Rows[0].ItemArray[0]) >
Convert.ToDecimal(dsetjj.Tables[0].Rows[0].ItemArray[1]))
                                    else
                                    {
                                    jjscore = jjscore +
Convert.ToDecimal(dsetjj.Tables[0].Rows[0].ItemArray[1]) * -1;
                                    }
                                }

                        // check adverb before adjective
                    if (i>0)
                    {
                        if (t[i - 1].ToString() == "/RB" || t[i].ToString() == "/RBS")
                            {
                                // select positive and negative score from sentiword net of the the
particular adverb

                                DataSet dsetrb = new DataSet();

                                cls.FillDset(ref dsetrb, "select posscore,negscore from sentiword
where synsetterms like '"+ v[i - 1].ToString()+"#_' and (posscore > 0 or negscore > 0)
and pos = 'r'" );
                                    // record found
                                if (dsetrb.Tables[0].Rows.Count > 0)
                                {

                                    // if positive score is > negative score
                                    if
(Convert.ToDecimal(dsetrb.Tables[0].Rows[0].ItemArray[0]) >
Convert.ToDecimal(dsetrb.Tables[0].Rows[0].ItemArray[1]))
                                    {

                                        if (jjscore > 0)
                                        {
                                            // adjective score + adverb score
                                            jjscore =
Convert.ToDecimal(dsetrb.Tables[0].Rows[0].ItemArray[0]) + jjscore;
```

```
                    }
                    else
                    {
                      jjscore = -1 *
Convert.ToDecimal(dsetrb.Tables[0].Rows[0].ItemArray[0]) + jjscore;
                    }
                  }

                  else

                  // if Negative score is > Positive score
                  {
                    jjscore = -1 *
Convert.ToDecimal(dsetrb.Tables[0].Rows[0].ItemArray[1]) + jjscore;
                  }

                }
                else
                {
                // select modifier word list from database to compare with k-1

                      if                              (Convert.ToInt16
(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 1)
                      {
                       jjscore = jjscore * -1;
                      }

                      // 3 for enhancer modifier
                      if
(Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 3)
                      {
                       jjscore    =    jjscore    +    Convert.ToDecimal(
dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                      }

               }

              }
        }


          if (i > 0)
          {
            cls.FillDset(ref  dsetnegation,  "select  *  from  Enhancers  where
Enhancer_Name= '" + v[i - 1].ToString().ToUpper() + "'");

            if (dsetnegation.Tables[0].Rows.Count > 0)
            {
```

```
          if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 1)
                {
                  jjscore = jjscore * -1;
                }


          if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 3)
                {
                  jjscore = jjscore +
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                }

                if  (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1])
== 4 || Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 5)
                {
                  jjscore = jjscore -
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                }
             }
             // for k+1
             cls.FillDset(ref  dsetnegation,  "select  *  from  Enhancers  where
Enhancer_Name= '" + v[i + 1].ToString().ToUpper() + "'");

             if (dsetnegation.Tables[0].Rows.Count > 0)
             {

          if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 1)
                {
                  jjscore = jjscore * -1;
                }


          if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 3)
                {
                  jjscore = jjscore +
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                }

                if  (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1])
== 4 || Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 5)
                {
                  jjscore = jjscore -
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                }
             }
           }
```

```
                    if (i > 1)
                    {
                        //for k-2
                        cls.FillDset(ref dsetnegation, "select * from Enhancers where
Enhancer_Name= '" + v[i - 2].ToString().ToUpper() + "'");

                        if (dsetnegation.Tables[0].Rows.Count > 0)
                        {

                  if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 1)
                            {
                                jjscore = jjscore * -1;
                            }


                   if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 3)
                            {
                                jjscore = jjscore +
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                            }

                   if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 4)
                            {
                                jjscore = jjscore -
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                            }
                        }
                    }
                   //*************
                    // for k - 3
                    if (i > 2)
                    {
                        //for k-2
                        cls.FillDset(ref dsetnegation, "select * from Enhancers where
Enhancer_Name= '" + v[i - 3].ToString().ToUpper() + "'");

                        if (dsetnegation.Tables[0].Rows.Count > 0)
                        {

                  if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 1)
                            {
                                jjscore = jjscore * -1;
                            }


                   if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 3)
                            {
```

```
                            jjscore = jjscore +
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                        }

                  if (Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 4)
                        {
                            jjscore = jjscore -
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                        }
                    }
                }
                //*************


                // for k+2
                if (i < v.Count - 2)
                {
                        cls.FillDset(ref dsetnegation, "select * from Enhancers where
Enhancer_Name= '" + v[i + 2].ToString().ToUpper() + "'");

                        if (dsetnegation.Tables[0].Rows.Count > 0)
                        {

                            if
(Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 1)
                            {
                                jjscore = jjscore * -1;
                            }



                            if
(Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 3)
                            {
                                jjscore = jjscore +
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                            }

                            if
(Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 4)
                            {
                                jjscore = jjscore -
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                            }
                        }
                }

                //****** k+3
if (i < v.Count - 3)
```

```
          {
              cls.FillDset(ref dsetnegation, "select * from Enhancers where
Enhancer_Name= '" + v[i + 3].ToString().ToUpper() + "'");

              if (dsetnegation.Tables[0].Rows.Count > 0)
              {

                if
(Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 1)
                {
                  jjscore = jjscore * -1;
                }



                if
(Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 3)
                {
                  jjscore = jjscore +
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                }

                if
(Convert.ToInt16(dsetnegation.Tables[0].Rows[0].ItemArray[1]) == 4)
                {
                  jjscore = jjscore -
Convert.ToDecimal(dsetnegation.Tables[0].Rows[0].ItemArray[3]);
                }
              }
          }

          //*****

         }

        // checking Verb
        else if (t[i].ToString() == "/VB" || t[i].ToString() == "/VBD" ||
t[i].ToString() == "/VBG")
          {
            string selectedsentence;
            selectedsentence = cls.ExecuteScalerStrin("select sentence from
sentences where sentence_id= " + sent + "").ToString();
            sentencesensepattern = WordSentence(v[i].ToString(),
selectedsentence);

            ssno = SenseFind(v[i].ToString(), sentencesensepattern);
            DataSet dsetvb = new DataSet();

            if (ssno == 0)
            {
            }
            else
            {
              cls.FillDset(ref dsetvb, "select posscore,negscore from sentiword
where  synsetterms like '" + v[i].ToString() + "#" + ssno + "' and (posscore > 0 or
negscore > 0) and pos = 'v' ");
            }
            // if records found
            if (dsetvb.Tables[0].Rows.Count > 0)
            {
              opinion_sentence = true;
              // if positive score is > negative score
              if (Convert.ToDecimal(dsetvb.Tables[0].Rows[0].ItemArray[0]) >
Convert.ToDecimal(dsetvb.Tables[0].Rows[0].ItemArray[1]))
              {
                jjscore = jjscore +
Convert.ToDecimal(dsetvb.Tables[0].Rows[0].ItemArray[0]);
              }
              // if negative score > positive score
              else
              {
                jjscore = jjscore +
Convert.ToDecimal(dsetvb.Tables[0].Rows[0].ItemArray[1]) * -1;
              }
            }

          cls.Executecommand("insert into Words values ('" + v[i].ToString() +
"',3,Null,Null," + sent + "," + pos + ")");
          }
          //same process as for NN


        else if (t[i].ToString() == "/NNS")
         {
           string selectedsentence;
           selectedsentence = cls.ExecuteScalerStrin("select sentence from
sentences where sentence_id= " + sent + "").ToString();
           sentencesensepattern = WordSentence(v[i].ToString(),
selectedsentence);

           ssno = SenseFind(v[i].ToString(), sentencesensepattern);
           DataSet dsetnns = new DataSet();

           if (ssno == 0)
           {
                   cls.FillDset(ref dsetnns, "select posscore,negscore from sentiword
where synsetterms like '" + v[i].ToString() + "#_' and (posscore > 0 or negscore > 0)
and pos = 'n' ");

                }
                else
                {
                   cls.FillDset(ref dsetnns, "select posscore,negscore from sentiword
where synsetterms like '" + v[i].ToString() + "#" + ssno + "' and (posscore > 0 or
negscore > 0) and pos = 'n' ");

                }
                // if records found
                if (dsetnns.Tables[0].Rows.Count > 0)
                {
                   opinion_sentence = true;
                   // if positive score is > negative score
                   if (Convert.ToDecimal(dsetnns.Tables[0].Rows[0].ItemArray[0]) >
Convert.ToDecimal(dsetnns.Tables[0].Rows[0].ItemArray[1]))
                   {
                      jjscore = jjscore +
Convert.ToDecimal(dsetnns.Tables[0].Rows[0].ItemArray[0]);
                   }
                   // if negative score > positive score
                   else
                   {
                      jjscore = jjscore +
Convert.ToDecimal(dsetnns.Tables[0].Rows[0].ItemArray[1]) * -1;
                   }
                }

               cls.Executecommand("insert into Words values ('" + v[i].ToString() +
"',5,Null,Null," + sent + "," + pos + ")");
            }
            // checking Adverb
            else if (t[i].ToString() == "/RB" ||t[i].ToString() == "/RBS" )
               {
                  cls.Executecommand("insert into Words values ('" + v[i].ToString() +
"',4,Null,Null," + sent + "," + pos + ")");
               }
               //checking /./ for seperation sentences
            else if (t[i].ToString() == "/.")
               {
               //check if sentence is opinion then update it's weights
               if (opinion_sentence == true)
                  {
```

```
            cls.Executecommand("update sentences set weight = " + jjscore + "
where sentence_id= " + sent + "");
            DataSet dsetcompsent = new DataSet();
            string compsent;
            cls.FillDset(ref dsetcompsent, "select comp_word from
comp_words");
            compsent = cls.ExecuteScalerStrin("select sentence from sentences
where sentence_id = " + sent + "");
            for (int c = 0; c <= dsetcompsent.Tables[0].Rows.Count - 1; c++)
            {
              if
(compsent.Contains(dsetcompsent.Tables[0].Rows[c].ItemArray[0].ToString()))
              {
               cls.Executecommand("update sentences set sentence_type ='C'
where sentence_id= " + sent + "");
              }
            }
            // increase sentence
            }
          //for new sentence assign false to opinion sentence
          opinion_sentence = false;
          //increase sentence
              sent = sent + 1;

          // assign 0 to jjscore for new sentencce at start
              jjscore = 0;
          //also asign 0 to position
              pos = 0;

          }
        // increse the bar of progress bar
        this.PB.Value = i+1;
      }
      // end of main loop

      //assign tagged text to textbox
      this.txtTagged.Text = str;
    }
    // end of tagging



    private void button1_Click(object sender, EventArgs e)
    {
      DataSet dsetsentiwords = new DataSet();
      DataSet dsetword = new DataSet();
```

```
          cls.FillLGrid(ref dsetsentiwords, "select * from sentiword where synsetterms
like '"+ this.textBox1.Text   +"#_' and (posscore > 0 OR negscore > 0)", ref
this.dataGridView1);

        }

//*************************************************
        // start of update polarity
        private void btnUpdatePoliarity_Click(object sender, EventArgs e)
        {
              //datasets for words of input sentence and for sentword of sentiword
dictionary
                DataSet dsetword = new DataSet();
                DataSet dsetsentiword = new DataSet();
            //fill dataset from the counting view
                cls.FillDset(ref dsetword, "select Word,SentiWordNet_Abrv from
Counting_View");
                for (int j = 0; j < dsetword.Tables[0].Rows.Count - 1; j++)
                {
                    // assign words and it's abbriaviation to variable

                    string wwd = dsetword.Tables[0].Rows[j].ItemArray[0].ToString();

                    string abr = dsetword.Tables[0].Rows[j].ItemArray[1].ToString();

                    string ssql = "select * from sentiword where synsetterms like '" + wwd

+ "#_" + "' and (posscore > 0 OR negscore > 0) and POS = '"+ abr  +"'";


                    cls.FillDset(ref dsetsentiword,ssql  );
                if (dsetsentiword.Tables[0].Rows.Count > 0)
                {
                   //assign positive polarity and negative polarity word to separte vairable
as below
                    string poswd;
                    string negwd;
                 poswd = dsetsentiword.Tables[0].Rows[0].ItemArray[3].ToString();

                negwd = dsetsentiword.Tables[0].Rows[0].ItemArray[4].ToString();

                    //update it's score in word datable

                cls.Executecommand("update words set POSScore = " +

Convert.ToDecimal (poswd) + ", NEGScore= " + Convert.ToDecimal(negwd ) + "

where word= '" + wwd + "'");


                }

                }
                MessageBox.Show("Saved");
            }
        // end of update polarity
        //**********************************************

        //**********************************************
        // start of sentence level feature/term/word sementic orientaion, considering contextual structure and term dependency.

        private void BtnPlot_Click(object sender, EventArgs e)
        {
            DataSet DsetReport = new DataSet();

            FrmReports FR = new FrmReports();

            OM CR = new OM();

            cls.FillDset(ref DsetReport, "select * from Total");

            CR.SetDataSource (DsetReport.Tables[0].DefaultView);

            FR.CRV.ReportSource = CR;

            FR.Show();


        }


        private void btnOpinionSentences_Click(object sender, EventArgs e)
        {
            DataSet DsetReport = new DataSet();
            FrmReports FR = new FrmReports();
            CR_Sentences CR = new CR_Sentences();
            cls.FillDset(ref DsetReport, "select * from Total_Sentences_View");
            CR.SetDataSource(DsetReport.Tables[0].DefaultView);
            FR.CRV.ReportSource = CR;
            FR.Show();
        }

        private void btnPositiveNegative_Click(object sender, EventArgs e)
        {
            DataSet dsetposneg = new DataSet();

            decimal  positivemid;

            decimal negativemid;

            int strongpositivecount;

            int weakpositivecount;

            int strongnegativecount;

            int weaknegativecount;
```

```
      int neutralcount;
      cls.FillDset(ref dsetposneg, "select * from Total");
      if (dsetposneg.Tables[0].Rows.Count > 0)
      {
        positivemid =
Convert.ToDecimal(dsetposneg.Tables[0].Rows[0].ItemArray[3]) /
Convert.ToDecimal ( dsetposneg.Tables[0].Rows[0].ItemArray[0]);
        positivemid= Math.Round(positivemid, 3);
        negativemid =
Convert.ToDecimal(dsetposneg.Tables[0].Rows[0].ItemArray[4]) /
Convert.ToDecimal(dsetposneg.Tables[0].Rows[0].ItemArray[1]);
        negativemid = Math.Round(negativemid , 3)* -1;

        neutralcount = cls.ExecuteScaler("select neutral from total");
        strongpositivecount = cls.ExecuteScaler("select count(sentence_id) from
Positive_Weights_View where Weight >= " + positivemid + "");
        weakpositivecount = cls.ExecuteScaler("select count(sentence_id) from
Positive_Weights_View where Weight < " + positivemid + "");

        strongnegativecount = cls.ExecuteScaler("select count(sentence_id) from
Negative_Weights_View where Weight >= " + negativemid + "");
        weaknegativecount = cls.ExecuteScaler("select count(sentence_id) from
Negative_Weights_View where Weight < " + negativemid + "");

        cls.Executecommand("insert into Positive_Negative_Ranges values (" +
strongpositivecount + ", " + weakpositivecount + ", "+ neutralcount +" ," +
strongnegativecount + ", " + weaknegativecount + ")");

        DataSet DsetReport = new DataSet();
        FrmReports FR = new FrmReports();
        StrongGraph CR = new StrongGraph();
        cls.FillDset(ref DsetReport, "select * from Positive_Negative_Ranges");
        CR.SetDataSource(DsetReport.Tables[0].DefaultView);
        FR.CRV.ReportSource = CR;
        FR.Show();
      }

    }
    // function for finding sense of given pattern
    private int SenseFind(string wrd, string Patternparam)
    {
     //dataset for wordnet sentence to create pattern
      DataSet dsetpattern = new DataSet();
      //dataset for wordnet selected pattern
      DataSet dsetselectedpattern = new DataSet();

      // string varible for wordnet glossary
      string gloss;

      string dwrd;
      // select sense number,glossry etc from sense_view of particular word
      cls.FillDset(ref dsetpattern, "select * from sense_View where word = '" + wrd
+ "'");
      for (int j = 0; j <= dsetpattern.Tables[0].Rows.Count - 1; j++)
      {
        //assign wordnet glossary of given word to variable
        gloss = dsetpattern.Tables[0].Rows[j].ItemArray[3].ToString();
        //clean glossary
        gloss  = gloss.Replace ("\"", "");
        gloss = gloss.Replace(";", "");

        //tokenizeing and tagging glossary
        NLPlib tagger = new NLPlib();
        string s = gloss;

        ArrayList v = tagger.tokenize(s);
        ArrayList t = tagger.tag(v);

        // creating pattern for K+3 and K-3
        for (int k = 0; k < v.Count ; k++)
        {
          dwrd = v[k].ToString();
          if  (dwrd.ToUpper()      ==  wrd.ToUpper()  ||    dwrd.ToUpper()     ==
wrd.ToUpper()+"S")
          {
            patternselected = "WRD";
            // creating patterns
            if (t.Count - 1 >= k + 3)
            {
    patternselected = patternselected + "-" + t[k + 1] + "-" + t[k + 2] + "-" + t[k + 3];
            }
            else
            {
              if (t.Count - 1 == k + 2)
              {
```

```
                                patternselected = patternselected + "-" + t[k + 1] + "-" + t[k + 2];
                            }
                            else
                            {
                                if (t.Count - 1 == k + 1)
                                {
                                    patternselected = patternselected + "-" + t[k + 1];
                                }

                            }

                        }

                    if (k >= 4)
                        {
                            patternselected = t[k - 3] + "-" + t[k - 2] + "-" + t[k - 1] + "-" +
patternselected;
                        }
                        else
                        {
                            if (k == 3)
                            {
                                patternselected = t[k - 2] + "-" + t[k - 1] + "-" + patternselected;
                            }
                            else
                            {
                                if (k == 2)
                                {
                                    patternselected = t[k - 1] + "-" + patternselected;
                                }
                            }
                            //end of creating pattern

                        }

                       }
                     }
                                //store pattern in a database for future processing
                cls.Executecommand("insert into pattern values ('" + wrd + "'," +
Convert.ToInt16(dsetpattern.Tables[0].Rows[j].ItemArray[2]) + ", '" +
patternselected + "')");


            }
            // now compare the input pattern and wordnet glossary pattern if matched then
return sense number
```

```
    cls.FillDset(ref  dsetselectedpattern,  "select  *  from  pattern  where  pattern  like
'%"+ Patternparam  +"%' and Word = '"+ wrd  +"'");
    sense_number = 0;
        if (dsetselectedpattern.Tables[0].Rows.Count > 0)
        {

          sense_number = Convert.ToInt32 (
dsetselectedpattern.Tables[0].Rows[0].ItemArray[1]);

        }
      return sense_number;
    }

    // report to find strong positive and storng negative, weak positive and weak
negative and neutral
    private void btnstrongneg_Click(object sender, EventArgs e)
    {
      //first delete existing data
      cls.Executecommand("delete from strong_weak_neg_pos");
      DataSet dsetposneg = new DataSet();
      //variable for each one
      decimal positivemid;
      decimal negativemid;
      int strongpositivecount;
      int weakpositivecount;
      int strongnegativecount;
      int weaknegativecount;
      int neutralcount;
      //select these counts from total view
      cls.FillDset(ref dsetposneg, "select * from Total");
      if (dsetposneg.Tables[0].Rows.Count > 0)
      {
        // to find postive mid divide postive weight by positve count
        positivemid =
Convert.ToDecimal(dsetposneg.Tables[0].Rows[0].ItemArray[3]) /
Convert.ToDecimal(dsetposneg.Tables[0].Rows[0].ItemArray[0]);
        //round decimial point upto 3 places
```

```
            positivemid = Math.Round(positivemid, 3);
            // to find negative mid divide negative weight by negative count
            negativemid =
Convert.ToDecimal(dsetposneg.Tables[0].Rows[0].ItemArray[4]) /
Convert.ToDecimal(dsetposneg.Tables[0].Rows[0].ItemArray[1]);
            //round decimial point upto 3 places
            negativemid = Math.Round(negativemid, 3) * -1;
            // select nutral
            neutralcount = cls.ExecuteScaler("select neutral from total");
            //now devide strong pos strong negative according to the positive and
negative mid
            strongnegativecount = cls.ExecuteScaler("select count(sentence_id) from
Negative_Weights_View where Weight >= " + negativemid + "");
            weaknegativecount = cls.ExecuteScaler("select count(sentence_id) from
Negative_Weights_View where Weight < " + negativemid + "");
            //insert into table for report
            cls.Executecommand("insert into strong_weak_neg_pos values ('Strong
Positive', " + strongnegativecount + ")");
            cls.Executecommand("insert into strong_weak_neg_pos values ('Weak
Positive', " + weakpositivecount + ")");
            cls.Executecommand("insert into strong_weak_neg_pos values ('Strong
Negative', " + strongnegativecount + ")");
            cls.Executecommand("insert into strong_weak_neg_pos values ('Weak
Negative', " + weaknegativecount + ")");
            cls.Executecommand("insert into strong_weak_neg_pos values ('Neutral', "
+ neutralcount + ")");

            //crete report from the above
            DataSet DsetReport = new DataSet();
            FrmReports FR = new FrmReports();
            CRStrongPOSNEg CR = new CRStrongPOSNEg();
            cls.FillDset(ref DsetReport, "select * from strong_weak_neg_pos");
```

```
        CR.SetDataSource(DsetReport.Tables[0].DefaultView);

        FR.CRV.ReportSource = CR;

        FR.Show();


      }
    }

    // end
    //***************************************************
    //function to return a pattern of a given sentence for a specifiec word for k+3 and
k-3
    private string  WordSentence(string wrd, string Sentence)
    {

      string gloss ;
       string dwrd;
        gloss = Sentence;
        gloss = gloss.Replace(",", "");
        gloss = gloss.Replace("'", "");
        gloss = gloss.Replace("\"", "");
        gloss = gloss.Replace(";", "");

        NLPlib tagger = new NLPlib();
        string s = gloss;

        ArrayList v = tagger.tokenize(s);
        ArrayList t = tagger.tag(v);


        for (int k = 0; k < v.Count; k++)
        {
          dwrd = v[k].ToString();
          if   (dwrd.ToUpper()   ==   wrd.ToUpper()   ||   dwrd.ToUpper()   ==
wrd.ToUpper() + "S")
          {
            pptrn = "WRD";
            // creating patterns
            if (t.Count - 1 >= k + 3)
            {
              pptrn = pptrn + "-" + t[k + 1] + "-" + t[k + 2] + "-" + t[k + 3];
            }
            else
            {
              if (t.Count - 1 == k + 2)
              {
                pptrn = pptrn + "-" + t[k + 1] + "-" + t[k + 2];
              }
```

```
                            else
                            {
                                if (t.Count - 1 == k + 1)
                                {
                                    pptrn = pptrn + "-" + t[k + 1];
                                }
                            }
                        }

                        if (k >= 4)
                        {
                            pptrn = t[k - 3] + "-" + t[k - 2] + "-" + t[k - 1] + "-" + pptrn;
                        }
                        else
                        {
                            if (k == 3)
                            {
                                pptrn = t[k - 2] + "-" + t[k - 1] + "-" + pptrn;
                            }
                            else
                            {
                                if (k == 2)
                                {
                                    pptrn = t[k - 1] + "-" + pptrn;
                                }
                            }
                            //end of creating pattern
                        }

                    }


                }


            return pptrn ;
        }


        private void FrmOpenion_Load(object sender, EventArgs e)
        {

        }

        private void BtnGarbage_Click(object sender, EventArgs e)
        {
            cls.Executecommand("delete from sentences where sentence_ID in (select
sentence_ID from word_counts where  wordcount < 2 and Weight is Null)");
```

```
      cls.Executecommand("delete from sentences where sentence_ID not in (select
sentence_ID from word_counts)");
      MessageBox.Show("Cleared");
    }

    private void btnFeaturesList_Click(object sender, EventArgs e)
    {
      FrmFeatures frm = new FrmFeatures();
      frm.Show();
    }


    private void btnFeatureWeight_Click(object sender, EventArgs e)
    {
      DataSet DsetReport = new DataSet();

      FrmReports FR = new FrmReports();

      CRF  CR = new CRF ();

      cls.FillDset(ref DsetReport, "select * from Total_Features_View");

      CR.SetDataSource(DsetReport.Tables[0].DefaultView);

      FR.CRV.ReportSource = CR;

      FR.Show();

    }

    private void btnFeatureCount_Click(object sender, EventArgs e)

    {

      DataSet DsetReport = new DataSet();

      FrmReports FR = new FrmReports();

      CRFC   CR = new CRFC  ();

      cls.FillDset(ref DsetReport, "select * from Total_Features_View");

      CR.SetDataSource(DsetReport.Tables[0].DefaultView);

      FR.CRV.ReportSource = CR;

      FR.Show();

    }

    private void btnComparative_Click(object sender, EventArgs e)
    {
      DataSet DsetReport = new DataSet();
      FrmReports FR = new FrmReports();
      CRComparative CR = new CRComparative();
      cls.FillDset(ref DsetReport, "select * from comparative_Report_View");
      CR.SetDataSource(DsetReport.Tables[0].DefaultView);
```

```
            FR.CRV.ReportSource = CR;
            FR.Show();

        }

        private void button2_Click(object sender, EventArgs e)
        {


        Word.Application app = new Word.Application();

            int errors = 0;
            if (txtReadFile.Text.Length > 0)
            {
                app.Visible = false;


                // Setting these variables is comparable to passing null to the function.
                // This is necessary because the C# null cannot be passed by reference.

                object template = Missing.Value;

                object newTemplate = Missing.Value;
                object documentType = Missing.Value;
                object visible = true;


                Word._Document doc1 = app.Documents.Add(ref template, ref
newTemplate, ref documentType, ref visible);
                doc1.Words.First.InsertBefore(txtReadFile.Text);
                Word.ProofreadingErrors spellErrorsColl = doc1.SpellingErrors;
                errors = spellErrorsColl.Count;


                object optional = Missing.Value;

                doc1.CheckSpelling(

                    ref optional, ref optional, ref optional, ref optional, ref optional, ref
optional,
                    ref optional, ref optional, ref optional, ref optional, ref optional, ref
optional);


                lblcheck.Text = errors + " errors corrected ";
                object first = 0;

                object last = doc1.Characters.Count - 1;
            txtReadFile.Text = doc1.Range(ref first, ref last).Text;
        }

        object saveChanges = false;
        object originalFormat = Missing.Value;
        object routeDocument = Missing.Value;

        app.Quit(ref saveChanges, ref originalFormat, ref routeDocument);
    }
    }


     }
```

## APPENDIX B

### Part of Speech (POS) Lixicons Dictionery with  Implementaion Module Using C#

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Windows.Forms;
namespace POS
{
   }
// File:      NLPlib.cs
 // Summary:   part of speech tagger

using System;

using System.Text;

using System.IO;

using System.Runtime.Serialization;

using System.Runtime.Serialization.Formatters.Binary;

using System.Collections;

using System.Text.RegularExpressions;
public class NLPlib {

  private static Hashtable lexHash = null;

  public NLPlib() {
        if (lexHash != null) return; // singleton pattern
      lexHash = new Hashtable();

        Stream file = File.Open("lex.dat", FileMode.Open);

        IFormatter formatter = (IFormatter)new BinaryFormatter();

        lexHash = formatter.Deserialize(file) as Hashtable;

        file.Close();
        Console.WriteLine("Initialized lexHash from serialized data.");
  }

  public ArrayList tokenize(string s) {
        ArrayList v = new ArrayList();
        Regex reg = new Regex(@"(\S+)\s");

        MatchCollection m = reg.Matches(s);
        foreach (Match m2 in m) {
```

```
                    if (m2.Length != 0) {
                                string z = m2.ToString().Trim();
                                if (z.EndsWith(";") || z.EndsWith(",") ||
                                    z.EndsWith("?") || z.EndsWith(")") ||
                                    z.EndsWith(":") || z.EndsWith(".")) {
                                            z = z.Substring(0, z.Length - 1);
                                }
                                v.Add(z);
                    }
        }
        return v;
}
public ArrayList tag(ArrayList words) {
    ArrayList ret = new ArrayList();
    for (int i = 0, size = words.Count; i < size; i++) {
        ret.Add("NN");  // default
        string s = (string)lexHash[words[i]];
        // 1/22/2002 mod (from Lisp code): if not in hash, try lower case:
        if (s == null)
            s = (string) lexHash[((string)words[i]).ToLower()];
        if (s != null) {
            int index = s.IndexOf(" ");
            if (index > -1) ret[i] = s.Substring(0, index).Trim();
            else                  ret[i] = s;
        }
    }
    /**
     * Apply transformational rules
     **/
    for (int i = 0; i < words.Count; i++) {
        //  rule 1: DT, {VBD | VBP} --> DT, NN
        if (i > 0 && ret[i - 1].Equals("DT")) {
            if (ret[i].Equals("VBD")
                || ret[i].Equals("VBP")
                || ret[i].Equals("VB")) {
                ret[i] = "NN";
            }
        }
        // rule 2: convert a noun to a number (CD) if "." appears in the word
        if (((string)ret[i]).StartsWith("N")) {
            if (((string)words[i]).IndexOf(".") > -1)
```

```
                ret[i] = "CD";
        }
        // rule 3: convert a noun to a past participle if ((string)words[i]) ends with "ed"
        if (((string)ret[i]).StartsWith("N") && ((string)words[i]).EndsWith("ed"))
            ret[i] = "VBN";
        // rule 4: convert any type to adverb if it ends in "ly";
        if (((string)words[i]).EndsWith("ly"))
            ret[i] = "RB";
     // rule 5: convert a common noun (NN or NNS) to a adjective if it ends with "al"
        if (((string)ret[i]).StartsWith("NN") && ((string)words[i]).EndsWith("al"))
            ret[i] = "JJ";
        // rule 6: convert a noun to a verb if the preceeding work is "would"
        if (i > 0
            && ((string)ret[i]).StartsWith("NN")
            && ((string)words[i - 1]).ToLower().Equals("would"))
            ret[i] = "VB";
      // rule 7: if a word has been categorized as a common noun and it ends with "s",
        //         then set its type to plural common noun (NNS)
        if (((string)ret[i]).Equals("NN") && ((string)words[i]).EndsWith("s"))
            ret[i] = "NNS";
        // rule 8: convert a common noun to a present prticiple verb (i.e., a gerand)
        if (((string)ret[i]).StartsWith("NN") && ((string)words[i]).EndsWith("ing"))
            ret[i] = "VBG";
    }
    return ret;
 }

public static void Main(String[] args) {
        NLPlib tagger = new NLPlib();
        string s = "The dog's paw was bit. We blame the cat; is that fair? ";
        ArrayList v = tagger.tokenize(s);
        ArrayList t = tagger.tag(v);
        for (int i=0; i<v.Count; i++) {
            Console.WriteLine((string)v[i] + "/" + (string)t[i]);
    }
}
```

```
}
//File:      MakeLex.cs
  //Summary:   create a binary lexicon file.

using System;
using System.Text;
using System.IO;
using System.Runtime.Serialization;

using System.Runtime.Serialization.Formatters.Binary;
using System.Collections;

public class App {

   public static void Main(String[] args) {
      try{
            Hashtable hash = new Hashtable();
            //int count = 0;
            FileInfo lexFile = new FileInfo("LEXICON");
            StreamReader reader = lexFile.OpenText();
            string line;
            do {
                        line= reader.ReadLine();
                        if (line == null)  break;
                        int index = line.IndexOf(" ");
                        //Console.WriteLine("line: " + line + " index: " + index);

                        string word = line.Substring(0, index).Trim();

                        string tags = line.Substring(index).Trim();

                        //Console.WriteLine("word: " + word + ", tags: " + tags);
                        //count++;
                        if (hash[word] == null)  hash.Add(word, tags);

            } while (line != null);

            reader.Close();

            Stream file = File.Open("lex.dat", FileMode.Create);

            IFormatter formatter = (IFormatter)new BinaryFormatter();

            // Serialize the object hashto stream

            formatter.Serialize(file, hash);

            file.Close();

      }catch(Exception e2){
         Console.WriteLine("Error: " + e2);
      }
   }
}
```

Sample of POS Lexicons Dictionary Used

" "
-RRB- )
-LCB- (
-LRB- (
-RCB- )
# #
$ $
Prizm /NNP
) SYM /CD
shakeup /NN
. /.
Laurance /NNP
mg /NN /JJ
expressing /VBG
citybred /JJ
Brestowe /NNP
STARS /NNP /NNS
negative /JJ /NN
investors /NNS /NNPS
mountain /NN
mavens /NNS
performing-arts /NNS
car-care /JJ
Athabascan /NNP
founding /NN /VBG /JJ
oversold /VBN /JJ /VB
Sepulveda /NNP
competency /NN
'82 /CD
largely-silent /JJ
ICL-GE /NNP
cf. /NN /FW
stretch /NN /VBP /JJ /VB
Lehder /NNP
scavenger /NN
Lebanese /JJ /NNPS /NNP
sinkt /FW
chorus /NN
common-carrier /NN
Bowles /NNP
Cabbage /NNP /NN
Bremner /NNP
IC /NNP
fleetest /JJS
studio-quality /JJ /NN

Carnegey /NNP
rigueur /FW
self-deprecation /NN
Reeve /NNP
Conn.based /JJ
ill-mannered /JJ
uncompensated /JJ
HIRING /NN /VBG
logistics /NNS
propsed /VBN
glass-like /JJ
interactive /JJ
port-shopping /NN
knuckle-duster /NN
glass-making /NN /JJ
casually /RB
Champs /NNP /NNS
Beatles /NNPS /NNP
Tator /NNP
Branching /NNP
sterility /NN
gate-post /NN
introspection /NN
probation /NN
Takashi /NNP
Kirin /NNP
bank-teller /NN
Tonal /JJ
Pale /NNP /RB
ex-brother-/IN-law /NN
unnavigable /JJ
abstraction /NN
union-owned /JJ
air-traffic /NN
S.D. /NNP
Partecipazioni /NNP
bullies /VBZ /NNS
evinced /VBN /VBD
Copernican /JJ /NNP
debtholders /NNS
start /VB /VBP /NN RP
MLR /NNP
Secondly /RB
Alumina /NNP
Forte /NNP

addict /NN
tempering /VBG
gizmos /NNS
Ham /NNP
Debating /NNP
Aldrin /NNP
generalization /NN
bad-smelling /JJ
motions /NNS /VBZ
sacked /VBD /VBN
weirs /NNS
Teagan /NNP
sketchy /JJ
traffic /NN
suspensor /NN
slows /VBZ /NNS
playable /JJ
Denis /NNP
leafy /JJ
Plummer /NNP
elegy /NN
happily /RB
torments /VBZ /NNS
jingles /NNS
lucidity /NN
preliminary /JJ /NN
throngs /NNS
boat-rocker... :
NT&SA-run /JJ
nonelectrical /JJ
respected /VBN /JJ /VBD
Mondrian /NNP
Casanova /NNP
cross-cultural /JJ /NN
Trifari /NNP
firing /VBG /JJ /NN /NN|/VBG
jelled /VBD
Koenig /NNP
wearing /VBG
owe /VBP /VB
stimulators /NNS
Face /NNP /VBP
midafternoon /NN
Liqueur /NNP

# APPENDIX C

## Sample of Dictionaries Used

| SENTIWORD | | | | | | |
|---|---|---|---|---|---|---|
| **ID1** | **POS** | **ID** | **PosScore** | **NegScore** | **SynsetTerms** | **Gloss** |
| 1 | a | 1740 | 0.125 | 0 | able#1 | (usually followed by `to') having the necessary means or skill or know-how or authority to do something; "able to swim"; "she was able to program her computer"; "we were at last able to buy a car"; "able to get a grant for the project" |
| 2 | a | 2098 | 0 | 0.75 | unable#1 | (usually followed by `to') not having the necessary means or skill or know-how; "unable to get to town without a car"; "unable to obtain funds" |
| 3 | a | 2312 | 0 | 0 | dorsal#2<br>abaxial#1 | facing away from the axis of an organ or organism; "the abaxial surface of a leaf is the underside or side facing away from the stem" |
| 4 | a | 2527 | 0 | 0 | ventral#2<br>adaxial#1 | nearest to or facing toward the axis of an organ or organism; "the upper side of a leaf is known as the adaxial surface" |
| 5 | a | 2730 | 0 | 0 | acroscopic#1 | facing or on the side toward the apex |
| 6 | a | 2843 | 0 | 0 | basiscopic#1 | facing or on the side toward the base |
| 7 | a | 2956 | 0 | 0 | abducting#1<br>abducent#1 | especially of muscles; drawing away from the midline of the body or from an adjacent part |
| 8 | a | 3131 | 0 | 0 | adductive#1<br>adducting#1<br>adducent#1 | especially of muscles; bringing together or drawing toward the midline of the body or toward an adjacent part |
| 9 | a | 3356 | 0 | 0 | nascent#1 | being born or beginning; "the nascent chicks"; "a nascent insurgency" |
| 10 | a | 3553 | 0 | 0 | emerging#2<br>emergent#2 | coming into existence; "an emergent republic" |
| 11 | a | 3700 | 0.25 | 0 | dissilient#1 | bursting open with force, as do some ripe seed vessels |
| 12 | a | 3829 | 0.25 | 0 | parturient#2 | giving birth; "a parturient heifer" |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| 13 | a | 3939 | 0 | 0 | dying#1 | in or associated with the process of passing from life or ceasing to be; "a dying man"; "his dying wish"; "a dying fire"; "a dying civilization" |
| 14 | a | 4171 | 0 | 0 | moribund#2 | being on the point of death; breathing your last; "a moribund patient" |
| 15 | a | 4296 | 0 | 0 | last#5 | occurring at the time of death; "his last words"; "the last rites" |
| 16 | a | 4413 | 0 | 0 | abridged#1 | (used of texts) shortened by condensing or rewriting; "an abridged version" |
| 17 | a | 4615 | 0 | 0 | shortened#4 cut#3 | with parts removed; "the drastically cut film" |
| 18 | a | 4723 | 0 | 0 | half-length#2 | abridged to half its original length |
| 19 | a | 4817 | 0 | 0 | potted#3 | (British informal) summarized or abridged; "a potted version of a novel" |
| 20 | a | 4980 | 0 | 0 | unabridged#1 | (used of texts) not shortened; "an unabridged novel" |
| 21 | a | 5107 | 0.5 | 0 | uncut#7 full-length#2 | complete; "the full-length play" |
| 22 | a | 5205 | 0.5 | 0 | absolute#1 | perfect or complete or pure; "absolute loyalty"; "absolute silence"; "absolute truth"; "absolute alcohol" |
| 23 | a | 5473 | 0.75 | 0 | direct#10 | lacking compromising or mitigating elements; exact; "the direct opposite" |
| 24 | a | 5599 | 0.5 | 0.5 | unquestioning#2 implicit#2 | being without doubt or reserve; "implicit trust" |
| 25 | a | 5718 | 0.125 | 0 | infinite#4 | total and all-embracing; "God's infinite wisdom" |
| 26 | a | 5839 | 0.5 | 0.125 | living#3 | (informal) absolute; "she is a living doll"; "scared the living daylights out of them"; "beat the living hell out of him" |
| 27 | a | 6032 | 0.25 | 0.5 | relative#1 comparative#2 | estimated by comparison; not absolute or complete; "a relative stranger" |
| 28 | a | 6245 | 0 | 0 | relational#1 | having a relation or being related |
| 29 | a | 6336 | 0 | 0 | absorptive#1 absorbent#1 | having power or capacity or tendency to absorb or soak up something (liquids or energy etc.); "as absorbent as a sponge" |
| 30 | a | 6777 | 0.375 | 0 | sorbefacient#1 absorbefacient#1 | inducing or promoting absorption |

| Word Senses | | | |
|---|---|---|---|
| **synset_id** | **word** | **sense_number** | **gloss** |
| 100001740 | entity | 1 | that which is perceived or known or inferred to have its own distinct existence (living or nonliving) |
| 100002056 | thing | 12 | a separate and self-contained entity |
| 100002342 | anything | 1 | a thing of any kind; "do you have anything to declare?" |
| 100002452 | something | 1 | a thing of some kind; "is there something you want?" |
| 100002560 | nothing | 2 | a nonexistent thing |
| 100002560 | nonentity | 3 | a nonexistent thing |
| 100002645 | whole | 2 | an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit" |
| 100002645 | whole_thing | 1 | an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit" |
| 100002645 | unit | 6 | an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit" |
| 100003009 | living_thing | 1 | a living (or once living) entity |
| 100003009 | animate_thing | 1 | a living (or once living) entity |
| 100003226 | organism | 1 | a living thing that has (or can develop) the ability to act or function independently |
| 100003226 | being | 2 | a living thing that has (or can develop) the ability to act or function independently |
| 100004358 | benthos | 2 | organisms (plants and animals) that live at or near the bottom of a sea |
| 100004483 | heterotroph | 1 | an organism that depends on complex organic substances for nutrition |
| 100004609 | life | 11 | living things collectively; "the oceans are teeming with life" |
| 100004740 | biont | 1 | a discrete unit of living matter |
| 100004824 | cell | 2 | (biology) the basic structural and functional unit of all organisms; cells may exist as independent units of life (as in monads) or may form colonies or tissues as in higher plants and animals |
| 100005598 | causal_agent | 1 | any entity that causes events to happen |
| 100005598 | cause | 4 | any entity that causes events to happen |
| 100005598 | causal_agency | 1 | any entity that causes events to happen |
| 100006026 | person | 1 | a human being; "there was too much for one person to do" |
| 100006026 | individual | 1 | a human being; "there was too much for one person to do" |

| | | | |
|---|---|---|---|
| 100006026 | someone | 1 | a human being; "there was too much for one person to do" |
| 100006026 | somebody | 1 | a human being; "there was too much for one person to do" |
| 100006026 | mortal | 1 | a human being; "there was too much for one person to do" |
| 100006026 | human | 1 | a human being; "there was too much for one person to do" |
| 100006026 | soul | 2 | a human being; "there was too much for one person to do" |
| 100012748 | animal | 1 | a living organism characterized by voluntary movement |
| 100012748 | animate_being | 1 | a living organism characterized by voluntary movement |
| 100012748 | beast | 1 | a living organism characterized by voluntary movement |
| 100012748 | brute | 2 | a living organism characterized by voluntary movement |
| 100012748 | creature | 1 | a living organism characterized by voluntary movement |
| 100012748 | fauna | 2 | a living organism characterized by voluntary movement |
| 100014510 | plant | 2 | a living organism lacking the power of locomotion |
| 100014510 | flora | 2 | a living organism lacking the power of locomotion |

**WordNet Synset**

| Synset_id | W_num | Word | Ss_type | Sense_number | Tag_count |
|---|---|---|---|---|---|
| 100001740 | 1 | entity | n | 1 | 11 |
| 100002056 | 1 | thing | n | 12 | 0 |
| 100002342 | 1 | anything | n | 1 | 0 |
| 100002452 | 1 | something | n | 1 | 0 |
| 100002560 | 1 | nothing | n | 2 | 0 |
| 100002560 | 2 | nonentity | n | 3 | 0 |
| 100002645 | 1 | whole | n | 2 | 0 |
| 100002645 | 2 | whole_thing | n | 1 | 0 |
| 100002645 | 3 | unit | n | 6 | 0 |
| 100003009 | 1 | living_thing | n | 1 | 1 |
| 100003009 | 2 | animate_thing | n | 1 | 0 |
| 100003226 | 1 | organism | n | 1 | 9 |
| 100003226 | 2 | being | n | 2 | 7 |
| 100004358 | 1 | benthos | n | 2 | 0 |
| 100004483 | 1 | heterotroph | n | 1 | 0 |
| 100004609 | 1 | life | n | 11 | 31 |
| 100004740 | 1 | biont | n | 1 | 0 |
| 100004824 | 1 | cell | n | 2 | 44 |
| 100005598 | 1 | causal_agent | n | 1 | 0 |
| 100005598 | 2 | cause | n | 4 | 4 |
| 100005598 | 3 | causal_agency | n | 1 | 0 |
| 100006026 | 1 | person | n | 1 | 7229 |
| 100006026 | 2 | individual | n | 1 | 51 |
| 100006026 | 3 | someone | n | 1 | 17 |
| 100006026 | 4 | somebody | n | 1 | 0 |
| 100006026 | 5 | mortal | n | 1 | 2 |
| 100006026 | 6 | human | n | 1 | 7 |
| 100006026 | 7 | soul | n | 2 | 6 |
| 100012748 | 1 | animal | n | 1 | 67 |
| 100012748 | 2 | animate_being | n | 1 | 0 |
| 100012748 | 3 | beast | n | 1 | 4 |
| 100012748 | 4 | brute | n | 2 | 0 |
| 100012748 | 5 | creature | n | 1 | 16 |
| 100012748 | 6 | fauna | n | 2 | 0 |
| 100014510 | 1 | plant | n | 2 | 207 |
| 100014510 | 2 | flora | n | 2 | 0 |
| 100014510 | 3 | plant_life | n | 1 | 0 |
| 100016236 | 1 | object | n | 1 | 64 |
| 100016236 | 2 | physical_object | n | 1 | 0 |
| 100017087 | 1 | natural_object | n | 1 | 0 |
| 100017572 | 1 | substance | n | 1 | 68 |
| 100017572 | 2 | matter | n | 1 | 41 |
| 100018827 | 1 | food | n | 1 | 34 |
| 100018827 | 2 | nutrient | n | 1 | 1 |
| 100019244 | 1 | artifact | n | 1 | 1 |
| 100019244 | 2 | artefact | n | 1 | 0 |
| 100020136 | 1 | article | n | 2 | 6 |
| 100020333 | 1 | psychological_feature | n | 1 | 0 |
| 100020486 | 1 | abstraction | n | 6 | 0 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 100020729 | 1 | cognition | n | 1 | 0 |
| 100020729 | 2 | knowledge | n | 1 | 46 |
| 100020729 | 3 | noesis | n | 1 | 0 |
| 100021213 | 1 | motivation | n | 1 | 5 |
| 100021213 | 2 | motive | n | 1 | 18 |
| 100021213 | 3 | need | n | 3 | 12 |
| 100021668 | 1 | feeling | n | 1 | 50 |
| 100022625 | 1 | location | n | 1 | 992 |
| 100023103 | 1 | shape | n | 2 | 9 |
| 100023103 | 2 | form | n | 6 | 8 |
| 100023548 | 1 | time | n | 5 | 96 |
| 100023929 | 1 | space | n | 1 | 33 |
| 100024197 | 1 | absolute_space | n | 1 | 0 |
| 100024304 | 1 | phase_space | n | 1 | 0 |
| 100024568 | 1 | state | n | 4 | 142 |
| 100025950 | 1 | event | n | 1 | 62 |
| 100026194 | 1 | act | n | 2 | 26 |
| 100026194 | 2 | human_action | n | 1 | 1 |
| 100026194 | 3 | human_activity | n | 1 | 0 |
| 100026769 | 1 | group | n | 1 | 2350 |

| Enhancer Types |  |
|---|---|
| **Enhancer_Type_ID** | **Enhancer_Name** |
| 1 | Negation Words |
| 2 | Context Shifter |
| 3 | Modifiers Enhancer |
| 4 | Modifiers Reducer |
| 5 | Modifiers of Noun |

Negation Words

| **Enhancers** |  |  |
|---|---|---|
| **Enhancer_ID** | **Enhancer_Name** | **Weight** |
| 2 | NOT |  |
| 3 | ISNT |  |
| 4 | DIDNT |  |
| 5 | WOULDNT |  |
| 7 | NOTHING |  |
| 8 | NOR |  |
| 20 | DONT |  |
| 37 | WASNT |  |

Contact shifter

| Enhancers | | |
|---|---|---|
| **Enhancer_ID** | **Enhancer_Name** | **Weight** |
| 21 | BUT | 0.3 |
| 22 | EXCEPT | 0.3 |
| 24 | ALTHOUGH | 0.3 |
| 25 | WHILE | 0.3 |
| 26 | WHEREAS | 0.3 |
| 27 | WOULD | 0.2 |
| 28 | SHOULD | 0.3 |
| 29 | COULD | 0.2 |
| 30 | FORGOT | 0.2 |
| 31 | REFUSED | 0.3 |
| 32 | FORGET | 0.3 |
| 33 | ASSUMED | 0.2 |
| 35 | HARDER | 0.2 |

Modifier Enhancer

| Enhancer | | |
|---|---|---|
| **Enhancer_ID** | **Enhancer_Name** | **Weight** |
| 11 | PRETTY | 0.1 |
| 12 | EXTREMELY | 0.3 |
| 14 | MOST | 0.2 |
| 57 | MORE | 0.1 |
| 58 | FAIRLY | 0.2 |
| 59 | IMMEDIATELY | 0.1 |
| 62 | PERFECTLY | 0.2 |
| 113 | DEEPLY | 0.2 |
| 126 | TOTAL | 0.2 |
| 127 | HUGE | 0.2 |
| 128 | TREMENDOUS | 0.3 |
| 129 | MASSIVE | 0.1 |
| 131 | CLEAREST | 0.1 |
| 132 | BIGGER | 0.1 |
| 133 | BIGGEST | 0.2 |
| 134 | ABVIOUS | 0.1 |
| 135 | SERIOUSLY | 0.1 |
| 136 | DIPPER | 0.1 |
| 137 | DIPPEST | 0.2 |
| 138 | CONSIDERABLE | 0.1 |
| 139 | HIGHEST | 0.2 |

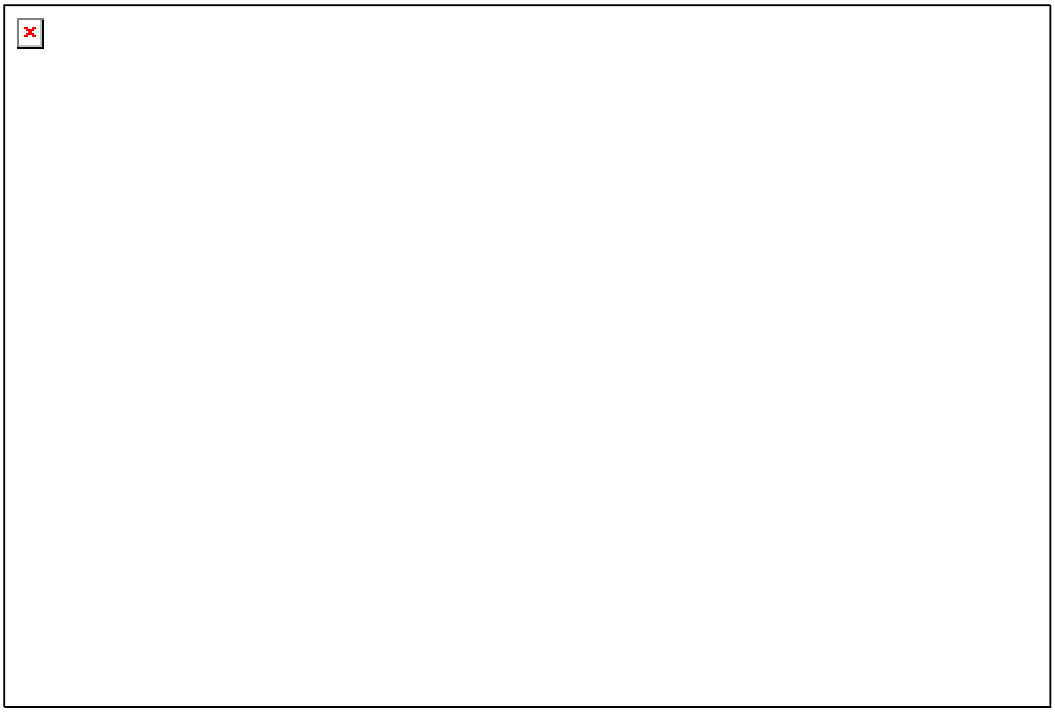
## Modifier Reducers

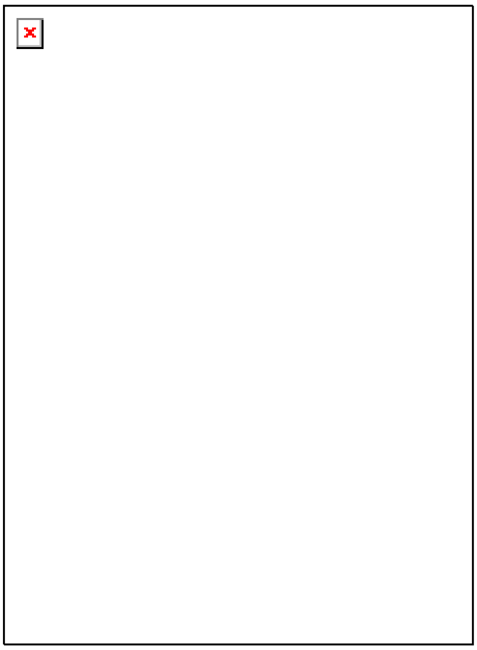
| Enhancers | | |
|---|---|---|
| **Enhancer_ID** | **Enhancer_Name** | **Weight** |
| 149 | hate | 0.5 |
| 1 | LEAST | 0.3 |
| 9 | SLIGHTLY | 0.3 |
| 10 | SOMEWHAT | 0.3 |
| 39 | LESS | 0.3 |
| 40 | HARDLY | 0.4 |
| 41 | ONLY | 0.3 |
| 42 | ALMOST | 0.2 |
| 43 | BARELY | 0.3 |
| 44 | NOT-TOO | 0.3 |
| 45 | A-LITTLE | 0.3 |
| 46 | A-LITTLE-BIT | 0.3 |
| 48 | SLIGHTLY | 0.3 |
| 49 | MARGINALLY | 0.3 |
| 50 | RELATIVELY | 0.3 |
| 51 | MIDLY | 0.2 |
| 52 | MODERATELY | 0.2 |
| 53 | SOMEWHAT | 0.3 |
| 54 | PARTIALLY | 0.3 |
| 55 | RATHER | 0.2 |
| 56 | T0-SOME-EXTENT | 0.3 |
| 130 | INCREDIBLE | 0.4 |
| 140 | SMALLER | 0.3 |
| 141 | SMALLEST | 0.4 |
| 142 | HIGHEST | 0.2 |
| 143 | LOWER | 0.2 |
| 144 | LOWEST | 0.4 |
| 146 | FEWER | 0.3 |
| 147 | FEWEST | 0.4 |
| 148 | few | 0.3 |

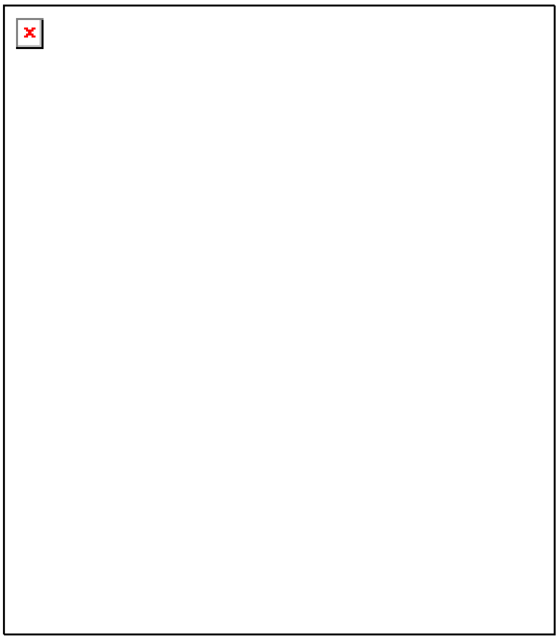
## Certain nouns

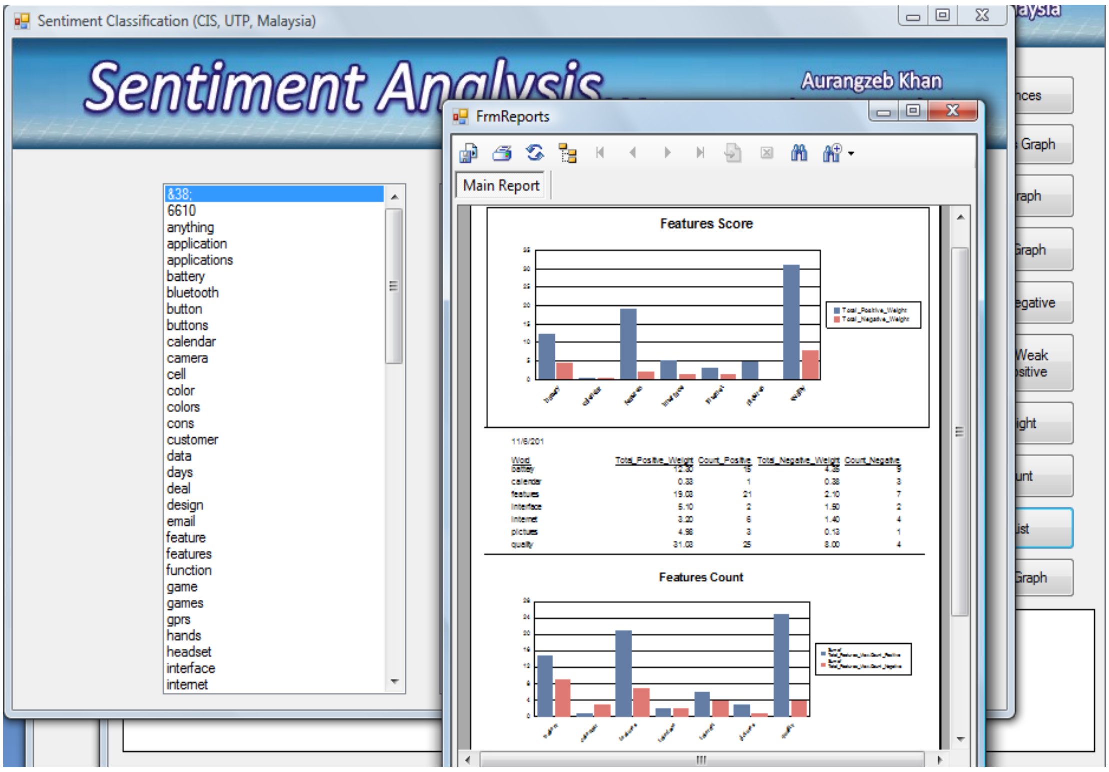
| | | |
|---|---|---|
| **Enhancer_ID** | **Enhancer_Name** | **Weight** |
| 150 | problem | 0.4 |
| 19 | FEW | 0.2 |
| 38 | HATE | 0.5 |

# APPENDIX D
## Screen shots





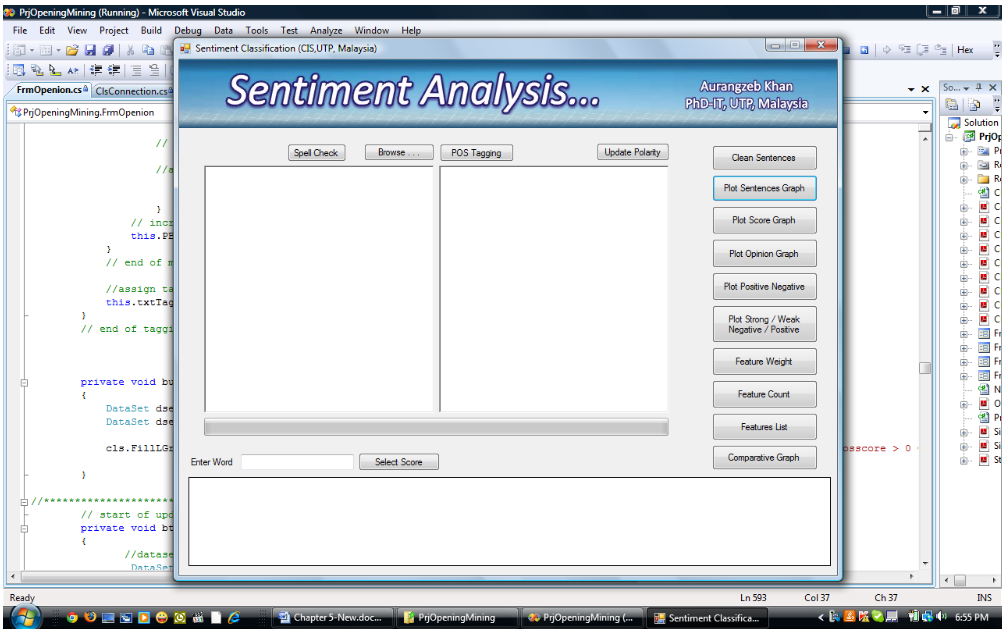




&38;
6610
anything
application
applications
battery
bluetooth
button
buttons
calendar
camera
cell
color
colors
cons
customer
data
days
deal
design
email
feature
features
function
game
games
gprs
hands
headset
interface
internet

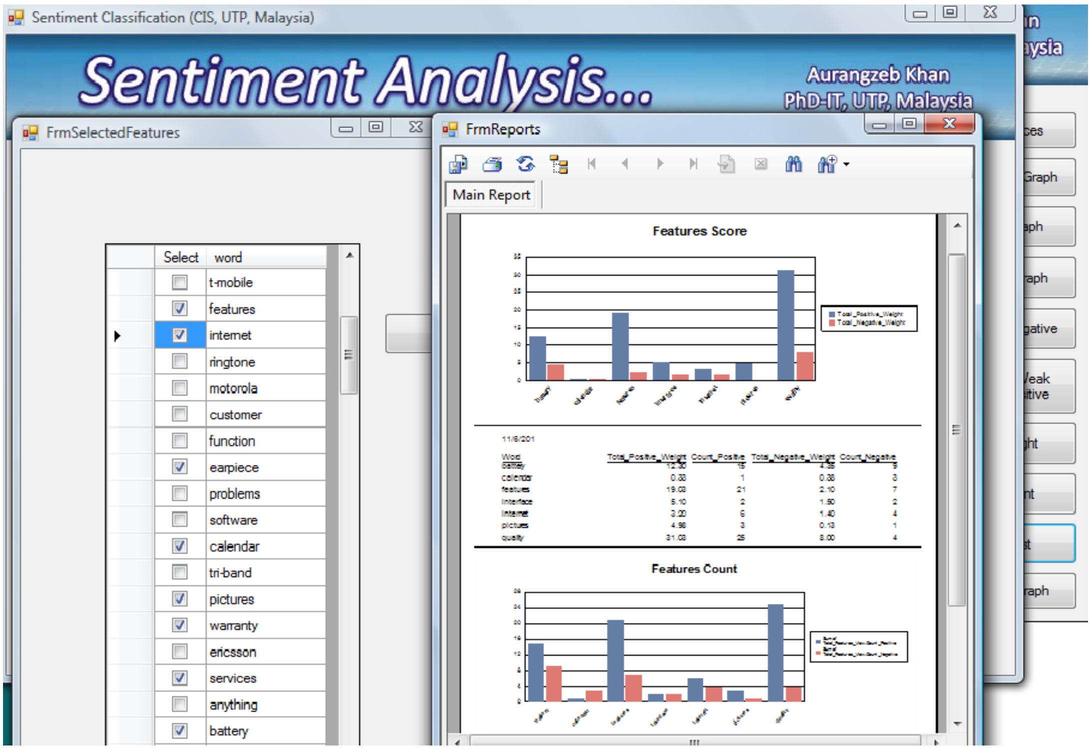
11/6/201

| Word | Total_Positive_Weight | Count_Positive | Total_Negative_Weight | Count_Negative |
|---|---|---|---|---|
| battery | 12.30 | 15 | 4.35 | 9 |
| calendar | 0.33 | 1 | 0.38 | 3 |
| features | 19.03 | 21 | 2.10 | 7 |
| interface | 5.10 | 2 | 1.50 | 2 |
| internet | 3.20 | 6 | 1.40 | 4 |
| pictures | 4.98 | 3 | 0.13 | 1 |
| quality | 31.03 | 25 | 8.00 | 4 |



```
                //
                //a

            }
            // incr
            this.PE
        }
        // end of m

        //assign ta
        this.txtTag
    }
    // end of taggi


    private void bu
    {
        DataSet dse
        DataSet dse

        cls.FillGr

    }

//*********************
    // start of upd
    private void bt
    {
        //datase
        DataSet
```

```
psscore > 0
```



| Select | word |
|---|---|
| □ | t-mobile |
| ☑ | features |
| ☑ | internet |
| □ | ringtone |
| □ | motorola |
| □ | customer |
| □ | function |
| ☑ | earpiece |
| □ | problems |
| □ | software |
| ☑ | calendar |
| □ | tri-band |
| ☑ | pictures |
| ☑ | warranty |
| □ | ericsson |
| ☑ | services |
| □ | anything |
| ☑ | battery |

11/6/201

| Word | Total_Positive_Weight | Count_Positive | Total_Negative_Weight | Count_Negative |
|---|---|---|---|---|
| battey | 12.30 | 15 | 4.35 | 5 |
| calendar | 0.33 | 1 | 0.38 | 3 |
| features | 19.03 | 21 | 2.10 | 7 |
| Interface | 5.10 | 2 | 1.50 | 2 |
| Internet | 3.20 | 6 | 1.40 | 4 |
| pictures | 4.98 | 3 | 0.13 | 1 |
| quality | 31.03 | 25 | 8.00 | 4 |