

**STUDY AND ANALYSIS OF APPROACHES FOR  
FINDING SEMANTIC ORIENTATION IN OPINION  
MINING**

*Dissertation submitted to the Jawaharlal Nehru University  
in partial fulfillment of the requirements  
for the award of the degree of*

*MASTER OF TECHNOLOGY  
in  
COMPUTER SCIENCE AND TECHNOLOGY*

SUBMITTED BY

**SYED SHADAB**



**SCHOOL OF COMPUTER AND SYSTEMS SCIENCES  
JAWAHARLAL NEHRU UNIVERSITY  
NEW DELHI-110067, INDIA  
2009-2010**



# जवाहरलाल नॅहरू विश्वविद्यालय

SCHOOL OF COMPUTER & SYSTEMS SCIENCES

Jawaharlal Nehru University

New Delhi- 110067

## CERTIFICATE

This is to certify that this dissertation entitled “**STUDY AND ANALYSIS OF APPROACHES FOR FINDING SEMANTIC ORIENTATION IN OPINION MINING**” submitted by Syed Shadab to the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, for the award of the degree of **MASTER OF TECHNOLOGY in COMPUTER SCIENCE AND TECHNOLOGY**, is a record of bonafide work carried out by him under the supervision of **Dr. Aditi Sharan**.

This work has not been submitted in part or full to any university or institution for the award of any degree or diploma.

**Dr. Aditi Sharan**

(Supervisor)

Assistant Professor, SC&SS,

Jawaharlal Nehru University,

New Delhi-110067

**Dean,**

SC&SS,

Jawaharlal Nehru University

New Delhi-110067



# जवाहरलाल नॅहरू विश्वविद्यालय

JAWAHARLAL NEHRU UNIVERSITY

School of Computer & Systems Sciences

NEW DELHI- 110067, INDIA

## DECLARATION

I hereby declare that this dissertation entitled “**STUDY AND ANALYSIS OF APPROACHES FOR FINDING SEMANTIC ORIENTATION IN OPINION MINING**” submitted by me to the **School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi**, for the award of the degree of **MASTER OF TECHNOLOGY in COMPUTER SCIENCE AND TECHNOLOGY**, is a record of bonafide work carried out by me under the supervision of **Dr. Aditi Sharan**.

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

**Syed Shadab**

School of Computer & Systems Sciences,

Jawaharlal Nehru University,

New Delhi-110067

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

*To My Loving Family & Friends.....*

## Acknowledgement

*At the very outset I thank Almighty Allah for all the favors He Showered upon me throughout my life.*

*I deem it a sacred duty to express my innermost feeling of gratitude for my supervisor Dr. Aditi Sharan, Asst. Prof., SC&SS, Jawaharlal Nehru University, New Delhi, whose able guidance had always inspired and elevated me to explore the realm of knowledge in all sphere of Opinion Mining. Her sincerity and dedication towards her work has been a motivating factor for me to work with ease and follow her footprints. Her continuous and unconditional co-operation and valuable suggestions have helped me to complete this work successfully.*

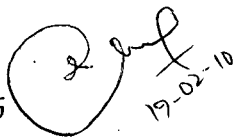
*Without mentioning names, I would, in particular, like to thank all the faculty members of SC&SS for their immense co-operation, help and encouragement. I also want to remember and thank to all the all non-teaching and administrative staff of SC&SS, especially to Mrs. Usha Kaushik, for their generous behavior and co-operation throughout my course of M. Tech.*

*I wish to extend my thanks to my lab buddies Mrs. Manju Lata Joshi, Mrs. Sonia, Mr. Amit Kumar Singh, Ms. Anupma Pandey and Ms. Vajenti Mala who helped me all the way to my project work wherever I needed.*

*I am grateful to all my seniors especially Asif bhai, Tanveer bhai and Amit bhai and all my juniors especially Vinti Agrawal and Liu who provided valuable suggestions and co-operation. Thanks to all my classmates, hostelmates and friends especially Uma, Mary, Sohan, Khalid, Faraz, Sifat and Sandeep who stood by me thick and thin. Words fail me to thank them all.*

*I am deeply indebted to all my family members for supporting me to achieve my goals and showing faith in my abilities even in my hard times I know of and I don't know of and for everything they have done to cherish my dream.*

Thanks  
Syed Shadab



19-02-10

## **Abstract**

---

In recent years World Wide Web has become largest and most widely used repository to obtain information in various fields. However it is being said about web that a person exploring the web is drowning in the information but starving for knowledge. This rapid growth of online data due to this World Wide Web and widespread use of databases have created an immense need for Knowledge Discovery methodologies. The amount of data available today far exceeds our abilities to reduce and analyze data without use of automated analysis techniques.

This dissertation work is an attempt to explore role of opinion mining techniques in the field of Web Data Mining. The process of determining appropriate orientation of any document (review) is full of uncertainty about product, subjectivity, imprecision in ratings etc and cannot be handled by traditional text mining methods. We have studied different approaches for finding product's features and semantic orientation and pointed out the areas where we need to work upon more.

The dissertation is divided in five chapters. First chapter describes present scenario of knowledge discovery from unstructured data and its challenges. In chapter 2, we discuss about opinion mining, its classification and applications. Chapter 3 talks about different existing methods for finding product's features and semantic orientation. It also presents types of review format and features that we encounter in our daily life. Proposed algorithm and experiment works is in chapter 4. Finally, chapter five concludes the work.

# Table of Contents

	Page No.
<b>Declaration</b>	i
<b>Certificate</b>	ii
<b>Acknowledgements</b>	v
<b>Abstract</b>	vi
<b>Chapter 1 Introduction</b>	<b>1-3</b>
<b>Chapter 2 Opinion Mining</b>	<b>4-8</b>
2.1 Opinion Mining Task Classification	4
2.2 General Framework Opinion Mining	5
2.3 Applications of Opinion Mining	7
<b>Chapter 3 Approaches for Finding Product Features and Semantic Orientation</b>	<b>9-23</b>
3.1 Different review Formats and Feature's types	9
3.2 Different Approaches to find Sentiment of Reviews at document level	11
3.2.1 Unsupervised Approach for Finding Semantic Orientation	11
3.2.2 Semi-supervised Approach for Finding Semantic Orientation	14
3.3 Approaches to find Semantic Orientation of Reviews at sentence level (feature based classification)	17
3.3.1 Approach for Finding Semantic Orientation for Format 1 & 2	17
3.3.2 Approach for Finding Semantic Orientation for Format 3	18



<b>Chapter 4 Proposed Work and Experiments</b>	<b>23-30</b>
4.1 Comparative study of different methods for Finding Semantic Orientation	23
4.2 Proposed Algorithm for Finding Semantic Orientation	25
4.3 Experiments for Finding Semantic Orientation	27
<b>Chapter 5 Conclusion</b>	<b>31-32</b>
<b>References</b>	<b>33-34</b>

With rapid expansion of e-commerce, more and more products are sold on the web and more and more people are buying products online. In order to enhance customer satisfaction and shopping experience, it has become a common practice for online merchants to enable their customers to review or express opinions on the products that they have purchased. With larger number of common users becoming comfortable with Web, opinion writing has grown rapidly. It has been seen that online opinions are getting popular day by day and these opinions represent wealth of information which can be beneficial for the industry as well as consumers. Opinions can be very helpful in enabling a person to decide whether to purchase a product or not. However when a huge amount of such reviews and opinions exist in the form of unstructured data, identifying overall opinion about a product becomes difficult. This is where opinion mining comes into picture. Opinion mining provides a way to determine overall sentiment of a review by analyzing documents and sentences in the document. It is preferable to have information in a format that is user friendly, so automating this process is very useful.

Opinion Mining aims to extract opinions from information sources (user generated content or user generated media) such as reviews, and present them to the users in a user friendly manner (graphically for example).

Opinion Mining can also be defined in a more general way as:

“Opinion Mining is a recent discipline at the crossroads of Information Retrieval and of Computational linguistics, which is, concerned not with the topic a document, is about, but with the opinion it expresses [2].”

Opinion Mining is very useful for having a brief as well as detailed idea of any product or service but it has a number of challenges to overcome:

- **Unstructured-ness:** While dealing with opinions, the main problem is its unstructured-ness. For extracting knowledge from structured data, a number of standard techniques [5] and query format exist but for unstructured data, such standard methods are generally not available.
- **Heterogeneity and Voluminous:** When we talk about online reviews, we have very large number of data available on any type of product or service. But these available information are heterogeneous i.e., in the form of text or picture or video etc. So for dealing with such scattered and voluminous data, we need to have more general methods which can handle all these types of data correctly and efficiently.
- **Subjectivity:** Since most of the reviews are available in free format (free text) expressed in natural language, there is a lot of inherent subjectivity in the reviews. It is very difficult to know what people are talking about. Most of the text mining methods are based on keyword based analysis, which may be able to capture general aspect of a document (search as topic discovery), but capturing sentiment of document is a very tedious task.
- **Dealing with semantic aspects:** In traditional text mining methods, bag of words approach is used. Bag of word approach does not consider overall association of words based on their position in the document. But in opinion mining, it is very important to keep track of positions of the words in order to determine their semantic association. For example, a noun preceded by a positive adjective represents a positive sentiment whereas a noun preceded by negative adjective represents a negative sentiment. Need of determining such associations makes opinion mining a challenging task.
- **Extracting appropriate knowledge:** Any unstructured data is enriched with vast amount of knowledge, which is generally hidden. Extracting the knowledge of our use is of great importance [16]. But in opinion mining, there are no pre-specified requirements which can be used to extract the knowledge of our relevance.

Researchers from different areas of AI have been working on automatic identification of opinions and reviews, their classification at document level or sentence level etc [17]. More researches are to explore many other aspects of opinion mining. There are many different areas, which are closely related to opinion mining. IR (Information Retrieval) Systems, IE (Information Extraction), NLP (Natural Language Processing), Machine Learning, Web Data Mining etc are different areas which are closely related and helps the opinion mining to achieve its objectives.

The objective of this work is to study and analyze the different approaches for finding semantic orientation of any review document. This dissertation is organized as follows: Chapter 2 is about opinion mining in general, its categorization, and applications etc. Chapter 3 deals with detailed study of different approaches for extracting product features and finding semantic orientation at document as well as sentence level. Chapter 4 is all about our proposed work and experiments. In last chapter 5, we present conclusion of our work.

Opinion Mining is a recent discipline at the crossroads of Information Retrieval and Computational linguistics, which is, not concerned with the topic of a document but with the opinion it expresses [2].

Opinion Mining identifies and extracts opinions and emotions from different review portals, presents it in a form which is more informative and friendly to user and enables us to predict, to take corrective measure and helps in other important business related issues.

### **2.1 Opinion Mining Task Classification:**

Opinion mining task can be broadly divided in three categories:

**Sentiment classification:** In this category, Opinion mining is treated as text classification problem. It classifies an evaluative text as being positive or negative. For example, given the review, the system determines whether review expresses a positive or negative sentiment of reviewer. It is done at **document level** and no details of document are covered.

This method has some limitations: In a review, a consumer presents its views about different features as well as final opinion about the product "*recommended or not recommended*". It is quite possible that say in a review, a consumer may have shown its dislike for some features, but at last overall review tends to be positive (product is

recommended). This method does not allow to analyze the individual opinions of the consumer about various features of the product.

**Feature based Opinion Mining and summarization:** It classifies the review text at **sentence level** to cover details of items i.e. what features of the item people liked or disliked. For example, in a product review, this task identifies product features that have been commented and allows to comment on quality of individual features.

For example, Camera A has *great zooming power* and *larger battery life*. Here *zooming power* and *battery life* are the product features about which author has talked and opinions about them are great and larger which are positive in nature respectively.

**Comparative Sentence and relation mining:** This method basically deals with comparing one or more similar types of products based on their features and evaluates them accordingly [4].

For example, car X is *cheaper* than car Y. Here we have a sort of comparison on some basis which is price in this case.

## 2.2 General Framework for Opinion Mining:

In this section, we provide a general framework for finding semantic orientation of reviews. To find semantic orientation at sentence level (feature-based), two main tasks are apparent:

1. **Identifying and extracting features of the product:** In this task, we identify features of the product on which the reviewers have expressed their opinion. Mostly, part-of-speech tags provide a good clue for identifying product's features. In simplest case, noun/noun phrases may be identified as product features. For example, in the sentence of any given review:

“the picture quality of this camera is amazing”

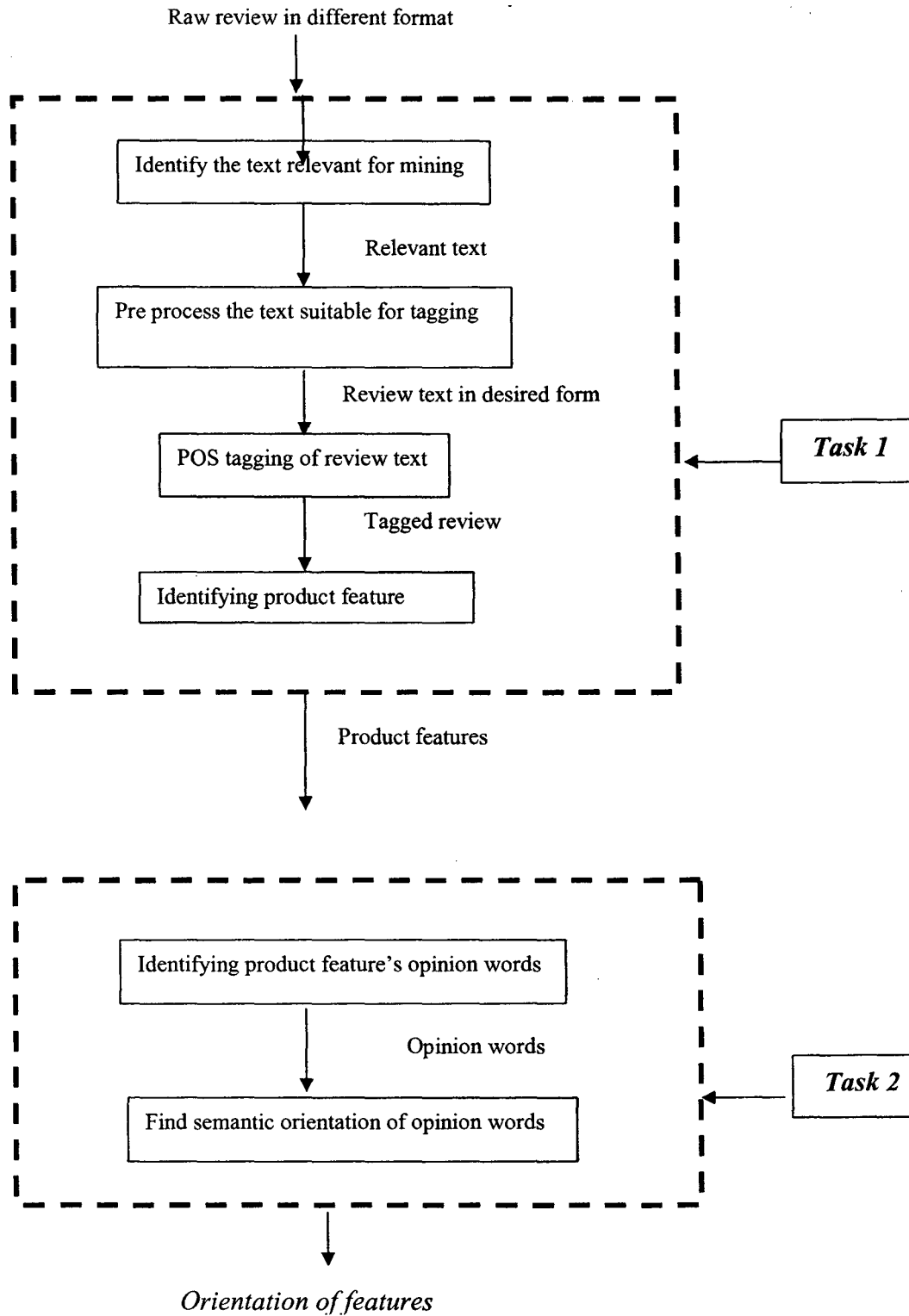
POS tags for the given sentence are:

“the\_DT picture\_NN quality\_NN of\_IN this\_DT camera\_NN is\_VBZ amazing\_JJ”

Then as per rule, product features can be identified i.e., “picture quality”

2. **Determining orientations of the features:** This task finds the opinion words in the review and allows us to determine the orientation of opinions (positive/negative). For example, in above sentence, the opinion word for the feature “picture quality” is

“amazing” and its orientation turns out to be positive. By summarizing individual opinions, one can determine about the overall sentiment of the document.



**Fig. 2.1 General framework for Opinion Mining**

In purview of above tasks, we can discuss general framework for opinion mining. Overall system is divided into two parts corresponding to tasks as discussed above. The inputs to the opinion mining system are raw reviews available in different formats. Reviews generally contain so many extraneous things which do not participate in decision process, therefore it is important to extract relevant portion from the review which will be used for mining opinions. As a next subtask, we convert the relevant text in a form that is more desirable for tagging. This pre processed text is then tagged using any tagger i.e., Stanford POS tagger from which product features can be identified.

In task 2, identified features and tagged text from previous task are provided input for further processing. With the help of product features, adjectives which are nearby to these features are identified as opinion words. For more detail about this, refer section 3.2.2. Then we calculate semantic orientation of these phrases using different existing methods for example PMI-IR algorithm. The sign of the result decide about its positive or negative-ness. Overall opinion about any document is then can be determined by taking average of all the extracted feature's semantic orientation.

### **2.3 Applications of Opinion Mining:**

Opinion Mining has got a wide range of applications. Nowadays everyone wants to have a look into services/products that he/she is going to encounter. Thus we can point out some key areas where opinion mining is playing very important role as follows:

#### **Shopping:**

OM can play an important role in this field. OM can provide a complete understanding about any item by allowing users on the web to express their views and thus making it easier to decide about any item.

#### **Government:**

Opinion on public policies, about politician etc makes government or politicians to have re-evaluation of their policies and properties. So it provides a good interaction between public and government or politician indirectly.

#### **Education:**

Opinions in terms of feedback make course instructor to revise its syllabus or teaching techniques etc. Academics can know the sentiment on courses based on sentiment analysis of opinion expressed by students. This can help to improve service delivery



and bolster marketing campaigns. In e-learning systems, opinions can be used to evaluate academic institution and academics. Legal researchers can benefit from different opinions that are posted for a legal issue.

### **Marketing:**

Since opinions are available more easily and cheaply on the internet, it has eliminated the need of consultant for surveys. Also according to opinions, companies launch new items that suit more to the customers or take corrective measures if anything is disliked by customers.

### **Entertainment:**

Now a day we have a number of movies, video games, TV serials etc for our entertainment. So it is almost impossible to watch or have look on each one. Here

we can take help from reviews, forums, blogs etc., but again these all are in such a large number that we cannot come to decision very soon. Here opinion mining can be playing very important role to make our time more enjoyable.

### **Research & Development:**

Product reviews can be taken as feedback to improve features and provide a platform for innovation. Web based applications could offer platforms for customers to design products and submit the design to the manufacturers. This approach could significantly assist in making features that are liked by customers [8].

There are other different areas where opinion mining is and supposed to play an important role in near future.

Before going into detail of different methods and approaches employed for finding product's features and semantic orientation at different levels i.e., at document level, sentence level (Feature based) etc., we focus on some important issues related to opinion mining. In order to apply opinion mining, first concern is regarding availability of reviews. Therefore subsection 3.1 deals with different review formats, available on the web, which can be used for mining the opinions. In subsection 3.2, we describe document level approaches which are used for determining overall sentiment of the document. In subsection 3.3, we present more elaborate approaches which analyze the document at sentence level and facilitate us to comment on and analyze the quality of various features of the reviewed product.

### **3.1 Different Review Formats and Feature's types**

Current research in opinion mining is mainly carried out from online product reviews. There are number of ways to express opinion which are available in different formats but we for convenience of applying opinion mining, these formats can be divided into three broad categories, which we can term as Format 1, Format 2 and Format 3. These different review formats may need different techniques to perform feature extraction task. Therefore we present these formats in more detail as follows:

**Format 1** – In this format, the reviewer is asked to describe pros and cons separately and also write a detailed. For example:

**Pros:** Great photos, easy to use, very small

**Cons:** Battery usage; included memory is stingy.

**Detail Review:** I had never used a digital camera prior to purchasing this Canon A70. I have always used a SLR

**Format 2** – This format contains both Pros and cons separately, but separate detailed review is not available. That is, the detailed review is in pros and cons itself. For example:

**Pros:** It's small in size, and the rotatable lens is great. It's very easy to use, and has fast response from the shutter.

**Cons:** It almost has no cons. It could be better if the LCD is bigger and it's going to be best if the model is designed to a smaller size.

**Format 3** – Free format: The reviewer can write freely, i.e., no separation of pros and cons. For example:

I did a lot of research last year before I bought this camera... It kind hurt to leave behind my beloved Nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital. The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out.

For formats 1 and 2, semantic orientations (positive or negative) of the features are known because pros and cons are separated. But in this case, we need to identify product features which have been commented upon. For format 3, we need to identify both product features and semantic orientations [4].

**Types of features:** while going in detail of features, we are encountered with two types of features: Implicit and Explicit. Let us consider an evaluative text (review) be  $r$ . In most general case,  $r$  consists of a sequence of sentences  $r = \langle s_1, s_2, \dots, s_m \rangle$ :

**Implicit Features:** If a feature  $f$  does not appear in evaluative text  $r$  but is implied, it is called an implicit feature in  $r$ . For example, “**size**” is an implicit feature in the following sentence as it does not appear in the sentence but it is implied:

“This camera is too large”.

**Explicit Features:** If a feature  $f$  appears in evaluative text  $r$ , it is called an explicit feature in  $r$ . For example, “**battery life**” in the following sentence is an explicit feature:

“The battery life of this camera is too short”.

Similarly we can say about explicit and implicit opinions [4].

### **3.2 Approaches to find Semantic Orientation of Reviews at Document Level:**

Given a set of evaluative texts  $D$ , a sentiment classifier classifies each document  $d \in D$  into one of the two classes, positive and negative. Positive means that  $d$  expresses a positive opinion. Negative means that  $d$  expresses a negative opinion. For example, given some reviews of a movie, the system classifies them into positive reviews and negative reviews.

The main advantage of sentiment classification is to give a quick determination of the prevailing opinion on an object. The task is somewhat similar but also different from classic topic-based text classification [5], which classifies documents into predefined topic classes, e.g., politics, science, sports, etc. However some shortcomings of the document-level classification that make it less useful are:

- It does not give details on what people liked or disliked. In a typical evaluative text such as a review, the author usually writes specific aspects of an object that he/she likes or dislikes. The ability to extract such details is useful in practice.
- It is not easily applicable to non-reviews, e.g., forum and blog postings, because although their main focus may not be evaluation or reviewing of a product, they may still contain a few opinion sentences. In such cases, we need to identify and extract opinion sentences.

We have studied and analyzed two different approaches for finding overall sentiment of the document, of which first is an unsupervised approach (subsection 3.2.1) and second uses semi-supervised approach (subsection 3.2.2).

#### **3.2.1 Unsupervised Approach for Finding Semantic Orientation**

Unsupervised learning studies how systems can learn to represent particular input pattern in a way that reflects the statistical structure of the overall collection of the input patterns [14].

One form of Unsupervised Learning is clustering. Another is Blind source Separation based on ICA (Independent Component Analysis). Among others the neural networks, Self Organized Map (SOM) and Adaptive Resonance Theory (ART) are commonly used algorithms. PCA (Principal Component), K-means etc are many other models which are used in unsupervised learning. [20]

Turney [15] use an unsupervised learning algorithm for classifying reviews as recommended or not recommended. Algorithm takes reviews as input and gives classification as output. Steps to classify the reviews can be described as follows:

##### **Step 1:**

Firstly part-of-speech tagging is done to make review friendlier for processing. This is required for extracting phrases containing adjectives and opinion words. Two

consecutive words are extracted if their tags conform to any pattern given in the Table 3.1 below.

	First Word	Second Word	Third Word
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

JJ - Adjectives, NN - Noun, RB - Adverbs, VB - Verbs

**Table 3.1**

**Step 2:**

In second step, semantic orientation of the extracted phrases is calculated using PMI-IR algorithm. This algorithm uses mutual information as a measure of the strength of semantic association between two words.

The Point-wise Mutual Information (PMI) between two words, word<sub>1</sub> and word<sub>2</sub>, is defined as [6]:

$$PMI(word_1, word_2) = \log_2 \left( \frac{P(word_1 \wedge word_2)}{P(word_1)P(word_2)} \right)$$

Here, p (word1 & word2) is the probability that word1 and word2 co-occur. If the words are statistically independent, then the probability that they co-occur is given by p (word1)\* p (word2). The ratio is the measure of degree of statistical independence between two words. The log of this gives the amount of Information that we acquire about the presence of one the two words when we observe other.

In order to find out semantic orientation of extracted phrases, we find its semantic orientation with appropriate positive and negative adjectives. The difference between these two values gives semantic orientation of the phrase. For example, if adjective selected are excellent (positive) and poor (negative), then the semantic orientation of the phrase can be calculated as follows:

$$SO (phrase) = PMI (phrase, \text{“excellent”}) - PMI (phrase, \text{“poor”})$$

Using above formula, semantic orientation is calculated for all the phrases identified in step 1. Thus for each extracted phrases, we are able to determine opinion about that phrase (either positive or negative)

### Step 3:

Once semantic orientation for each extracted phrase is available, average semantic orientation of all the phrases is calculated and on the basis of sign the final result, reviews are classified as recommended (positive sentiment) or not recommended (negative sentiment).

Following fig.3.2 shows calculation of semantic orientation and final recommendation for two different reviews (*recommended* and *not recommended*). Both the reviews are of Bank of America taken from Epinions.com.

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
small part	JJ NN	0.053
online service	JJ NN	2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	1.288
well other	RB JJ	0.237
inconveniently located	RB VBN	-1.541
other bank	JJ NN	-0.850
true service	JJ NN	-0.732
<b>Average Semantic Orientation</b>		<b>0.322</b>

An example of the processing of a review that the author has classified as **recommended**

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
little difference	JJ NN	-1.615
clever tricks	JJ NNS	-0.040
programs such	NNS JJ	0.117
possible moment	JJ NN	-0.668
unethical practices	JJ NNS	-8.484
low funds	JJ NNS	-6.843
old man	JJ NN	-2.566
other problems	JJ NNS	-2.748
probably wondering	RB VBG	-1.830
virtual monopoly	JJ NN	-2.050
other bank	JJ NN	-0.850
extra day	JJ NN	-0.286
direct deposits	JJ NNS	5.771
online web	JJ NN	1.936
cool thing	JJ NN	0.395
very handy	RB JJ	1.349
lesser evil	RBR JJ	-2.288
<b>Average Semantic Orientation</b>		<b>-1.218</b>

An example of the processing of a review that the author has classified as **not recommended**

Fig. 3.2

In our opinion, the approach discussed above has following limitations:

- Since this approach takes only some predefined patterns as mentioned in table, it is quite possible that some important phrases which may have strong effect on sentiment determination may be left out. So more thorough set of patterns may lead to more accurate and precise result.
- Also technique based on mutual information between phrases and words “excellent” and “poor”, it is not necessary that these words are perfect to relate phrase’s sentiment with positive or negative.

### 3.2.2 Semi-supervised Approach for Finding Semantic Orientation

The task of separating one review from other can be considered as a genre or style classification problem. It involves identifying subjectivity, which is inherent in unstructured (free text) data. Therefore traditional bag of word approach generally fails in such cases. Most of the time, a classifier based on relative frequency of each part of speech in a document can outperform bag of words approach. Dave et al [7] suggested a semi supervised approach for automatically distinguishing between positive and negative reviews. Their classifier draws on IR techniques for feature extraction and scoring. They start with structured reviews for training and testing and then use classifier to identify and classify review sentences from the web where classification is more difficult.

Overall approach used by them can be understood through following fig 3.1.

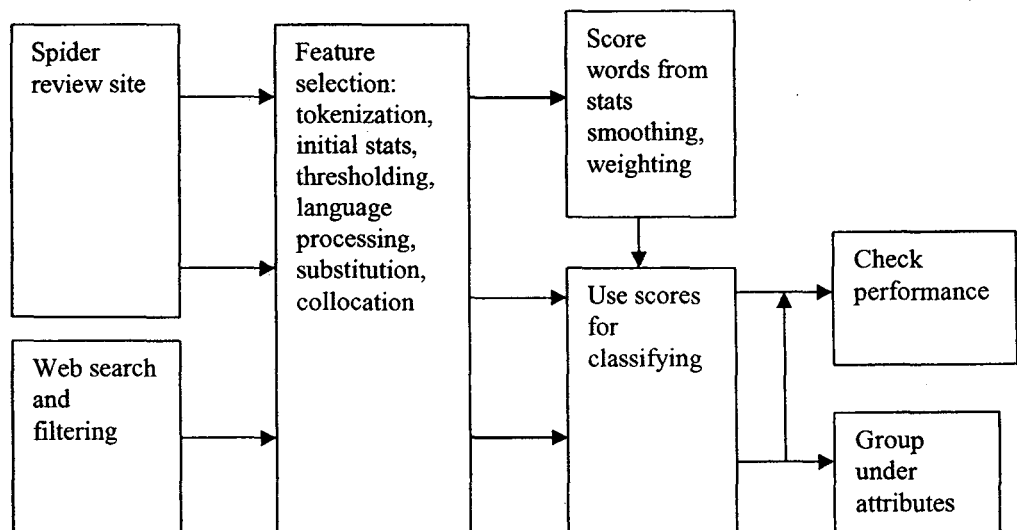


Fig. 3.1

We present here some important issues in automating the process of review classification.

#### Corpus:

First issue in supervised/semi-supervised learning technique is identifying the labeled corpus that can be used for training. Corpora are available from sites like C|net and Amazon which provide a number of reviews for different types of products. As shown in the fig.3 spiders, web search and filtering system are helpful in generating corpora.

#### Feature selection:

Most important module of the system is feature selection module which involves tokenization, transformations, substitutions and other language processing techniques

in order to select appropriate feature. Following are some important tasks in feature selection:

**- Tokenization:**

Starting with a raw document and stripping out HTML tags, documents are divided into sentences, which are optionally run through a parser before being split into single-word tokens.

**- Metadata and statistical substitutions**

In this subtask, similar tokens are grouped and numerical tokens can be replaced with NUMBER.

**- Linguistic substitutions and Language-based modifications**

In this subtask, a linguistic parser parses the document sentence by sentence, yielding part of speech of each word and relationship between parts of sentences. Knowing the part of speech, WordNet can be used for finding similarities of meanings. Further custom thesaurus can also be developed by finding word collocations.

**- N-grams and proximity**

After tokenization and substitution, n-adjacent tokens are combined into n-grams. For example, “this” followed by “is” becomes “this is” in a bi-gram. Authors suggest use of high scoring n-grams [19].

**- Substrings**

Using the idea of n-grams, the authors attempt to identify arbitrary length substrings that provide optimal classification. The classification was achieved with the tradeoff: as substrings become longer and generally more discriminatory, their frequency decreases, so there is less evidence for considering them relevant.

**Thresholding:**

After feature completion, frequency calculation is done to count: number of times each term occurs, number of documents each term occurs in and number of categories a term occurs in. The frequencies are normalized to overcome the skew. Upper and lower limits can be set for each of these measures, constraining number of features required. This improves relevance of the remaining features and reduces the amount of required computation.

**Smoothing:**

Before assigning scores based on term frequencies, smoothing of these numbers can be done by assigning probabilities to unseen events and making the known probabilities less “sharp”. Naïve Bayes [11], Laplace smoothing etc can be used here.



### Scoring:

After selecting a set of features  $f_1, \dots, f_n$  and optionally smoothing their probabilities, scores must assign them. As a baseline approach, following formula can be used to assign scores:

$$\text{score}(f_i) = \{p(f_i|C) - p(f_i|C')\} / \{p(f_i|C) + p(f_i|C')\}$$

Determine  $p(f_i|C)$ , the normalized term frequency, by taking the number of times a feature occurs in and dividing it by the total number of tokens in . A term's score is thus a measure of bias ranging from  $-1$  to  $1$ .

### Reweighting:

One interesting property of the baseline measure is that it does not incorporate the strength of evidence for a given feature. Thus a rare term has the same score as a much frequent one. In order to overcome these limitations, Gaussian weighting scheme, residual inverse document frequency scheme can be used.

### Classification:

Once each term has a score, sum of all scores of the words in document are calculated and sign of the total score can be used to determine a class. In other words, if document  $d_i = f_1 \dots f_n$

$$\begin{aligned} \text{class}(d_i) &= C & \text{eval}(d_i) > 0 \\ &= C' & \text{eval}(d_i) < 0 \end{aligned}$$

Where

$$\text{eval}(d_i) = \sum_j \text{score}(f_j)$$

After having result from all mentioned steps, Dave et al tried to group sentences under attribute for example: to use words matching the `_product`-typeword substitution as potential attributes of a product, for bi-grams starting with "the" and applying some simple thresholds and the same stop-words worked even better.

Though method gives fairly accurate classification of the documents, but still there are some issues which can be eliminated to improve the efficiency of the algorithm:

- A corpus of more finely-tagged documents is needed. Without a set of documents tagged at the sentence or expression level, it is difficult to design for or evaluate extraction performance.

- Find ways to decrease overfitting and generate features more useful for extracting opinions and attributes from web searches.
- Try to improve the efficiency of the algorithms. Discussed substring algorithms take several minutes on even the smaller second test case and require over a gigabyte of memory. There should be ways to make this more reasonable.

### **3.3 Approaches to find Semantic Orientation of Reviews at Sentence Level (Feature Based Classification):**

Approaches to find semantic orientation of reviews at sentence level facilitates us to comment on and analyze the quality of various features of the reviewed product. However these approaches require in-depth analysis of review document. The applicability of these approaches is closely related to format of the reviews. If pros and cons of the review are available then the problem becomes relatively simpler where main objective is to identify the product features which have been commented upon. In second case, where pros and cons are not available, we need to find out the features of the product along with its semantic orientations. Accordingly we discuss two different approaches for determining semantic orientation, approach 1 is for format 1 & 2 (where pros and cons are available) and second approach is for format 3 (free format).

#### **3.3.1 Approach for Finding Semantic Orientation for Format 1 & 2:**

As discussed earlier in format 1 & 2, pros and cons of the review are available; we do not need to calculate the semantic orientation of the product features. Main objective in this approach is to identify the product features for which pros and cons are available. Product features can be identified by using supervised rule learning technique [9] where language pattern can be used to identify features from pros and cons.

In this [3] approach, it is assumed that each sentence segment has at most one product feature and separated by commas, full stops etc. A dataset is prepared by manually labeling (or tagging) a large number of reviews.

Following steps are performed for training:

1. Perform (a) POS tagging and (b) remove digits. For example:

“Battery Usage” → “<N> Battery <N> Usage”

“16MB” → “MB”

Where <N> is for nouns, <V> is for verbs, <Adj> is for adjective and so on.

2. For determining explicit features; replace actual words with [feature]. For example:

“<N> Battery <N> Usage” → “<N> [feature] <N> Usage”

3. For implicit features, replace all those words which indicate the feature with [feature]. For example:

“MB” → [feature]

4. Use n-gram to produce smaller segment from original segment. For example:

“<V> included <N> [feature] <N> is <Adj> stingy” will be broken into smaller one as:

a. ‘<V> included <N> [feature] <N> is’

b. “<N> [feature] <N> is <Adj> stingy”

5. Duplicate tags can be distinguished by sequence numbers. If there are two consecutive words with same tag say <N>, then we assign sequence number to distinguish between them. For example:

“<N1> [feature] <N2> usage”

6. Word stemming is commonly done to reduce a word to its stem.

Once the training is complete, association rule mining can be applied to find the rules; which are used for determining product features. Association rule mining generates rules with minimum support and confidence. Post processing is required to remove unnecessary rules. Final patterns are then used to match with new reviews in order to find product features. Again to refine the result after matching with tagged new review, two strategies, frequent-noun and frequent-term can be employed. It has been found that system developed in this manner can provide good result for finding product’s features.

This approach provides certain flexibility to the users which are helpful in improving the efficiency of the system. Firstly system allows the user to set a value for the maximum length that a pattern could expand. Secondly, it also allows the user to set the maximum length of a review segment that a pattern should be applied to. These two values enable the user to refine the patterns for better extraction. Moreover the user can also add new patterns. But since a lot of manual processing such as numeral replacement, distinguishing duplicate etc is required, it will practically impossible for larger domains. It is assumed that each sentence segment has at most one product feature and separated by commas, full stops etc, this make task more constrained and accuracy is diminished.

### **3.3.2 Approach for Finding Semantic Orientation for Format 3:**

Format 3 is a free text format in which there is a lot of inherent subjectivity. As the reviewers can write freely, it requires use of NLP techniques in order to find out product features and their orientation. Nature of these reviews makes opinion mining a very challenging problem. For such reviews, the task of finding semantic orientation

can be divided into four parts: identifying product features, finding opinion words, finding orientation of opinion words with respect to extracted features and generating feature based review summary.

### **Identification of Product Features:**

A review may contain explicit or implicit features. Determining explicit features is relatively simpler. Simplest approach can be that only noun/ noun phrases are likely to be product features done. Some post processing for example, word stemming, pruning etc is done to refine the final result. The main disadvantage of this approach is that it can only extract explicit features. Following steps can be employed to extract frequent features from given reviews [13]:

1. First of all, POS tagging is done to identify noun/ noun phrases using any POS tagger.
2. After tagging is performed, association rule mining can be applied to give more frequent and converged set of features.
3. Since patterns of noun/ noun phrases resulted from above step may be in any random order and only specific order of phrases are meaningful. So compact pruning can be done which removes all those noun phrases likely to be meaningless.
4. Also single word nouns may be redundant because its superset may exist in the sentences. If they exist individually with minimum *p-support* value, then it is considered as feature otherwise dropped.

After employing above mentioned steps, a fairly good list of frequent explicit features of the given product/service is obtained.

### **Finding Opinion Words:**

In simplest case, Opinion words are adjectives which are nearby product feature. A nearby adjective refers to the adjacent adjective that modifies the noun/noun phrase that is a frequent feature. Following algorithm can be used to identify opinion words:

#### **Algorithm:**

```
for each sentence in the review database
{
  if (it contains a frequent feature, extract all the adjective words as opinion words)
  {
    for each feature in the sentence
      (the nearby adjective is recorded as its effective Opinion)
  }
}
```

#### **Opinion word extraction algorithm**

Once opinion words are identified, they can help in finding infrequent features. Following steps can be used in order to find infrequent features identification:

1. From the given tagged review, opinion words are extracted which are adjectives nearby and modify the noun/ noun phrases that are frequent feature.
2. Then sentences with no frequent features but opinion words are identified and nearest noun/ noun phrases are treated as infrequent features in the given review.

### **Finding Semantic Orientation:**

In this subsection, we discuss the approach to determine semantic orientation of above identified product features and opinion words. Since adjectives share same orientation as their synonyms and opposite orientations as their antonyms, this idea can be used to predict the semantic orientation of an adjective. For this, the synset of the given adjective and the antonym set are searched. If a synonym/antonym has known orientation, then the orientation of the given adjective could be set correspondingly.

To start with, a set of seed adjectives (which have known orientations) can be used. This set further grows by searching appropriate word from WordNet. The procedure of growing the set and for predicting semantic orientations for all the adjectives in the opinion list is shown below:

```
Procedure OrientationPrediction(adjective_list, seed_list)
begin
  do {
    size1 = # of words in seed_list;
    OrientationSearch(adjective_list, seed_list);
    size2 = # of words in seed_list;
  } while (size1 ≠ size2);
end
```

```
Procedure OrientationSearch(adjective_list, seed_list)
begin
  for each adjective wi in adjective_list
  begin
    if (wi has synonym s in seed_list)
      { wi's orientation= s's orientation;
        add wi with orientation to seed_list; }
    else if (wi has antonym a in seed_list)
      { wi's orientation = opposite orientation of a's orientation;
        add wi with orientation to seed_list; }
  endfor;
end
```

### **Predicting the semantic orientations of opinion words**

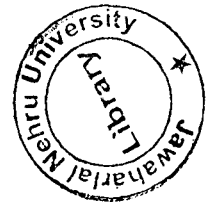
Now after predicting the orientations of opinion words, a step ahead to predict the orientation of whole sentence i.e., positive or negative is described. Bing et al suggest use of dominant orientation of the opinion words in the sentence to determine the orientation of the sentence. That is, if positive/negative opinion prevails, the opinion sentence is regarded as a positive/negative one. But if same numbers of positive and negative opinion words are present in the sentence, predict the orientation using the average orientation of effective opinions or the orientation of the previous opinion sentence. The detailed procedure is described as:

TH-19215

```

Procedure SentenceOrientation()
begin
  for each opinion sentence si
  begin
    orientation = 0;
    for each opinion word op in si
      orientation += wordOrientation(op, si);
      /*Positive = 1, Negative = -1, Neutral = 0*/
    if (orientation > 0) si's orientation = Positive;
    else if (orientation < 0) si's orientation = Negative;
    else {
      for each feature f in si
        orientation += wordOrientation(f's effective opinion, si);
      if (orientation > 0)
        si's orientation = Positive;
      else if (orientation < 0)
        si's orientation = Negative;
      else si's orientation = si-1's orientation;
    }
  endfor;
end

```



```

Procedure wordOrientation(word, sentence)
begin
  orientation = orientation of word in seed_list;
  If (there is NEGATION_WORD appears closely
      around word in sentence)
    orientation = Opposite(orientation);
end

```

### Predicting the orientations of opinion sentences

### **Generating Feature Based Review Summary:**

After all the previous steps, now it is stage to generate the final feature-based review summary, which is straightforward and consists of the following steps:

- For each identified feature, related opinion sentences are put into positive and negative categories as per its opinion sentences' orientations. A count is computed to show how many reviews give positive/negative opinions to the feature.
- All features are ranked according to the frequency of their appearances in the reviews. Feature phrases appear before single word features as phrases normally are more interesting to users. Other types of rankings are also possible. For example, rank features according the number of reviews that express positive or negative opinions.

The precision of above mentioned step to identify frequent features and their orientation is improved by Popescu and Etzioni in [1] by evaluating each noun phrase by computing a PMI score between the phrase and meronymy discriminators associated with the product class e.g., a scanner class. The meronymy discriminators for the scanner class are, "of scanner", "scanner has", "scanner comes with", etc., which are used to find components or parts of scanners by searching on the Web

This chapter deals with proposed work and experiments. In this dissertation, following works were proposed:

- To analyse and compare methods for finding semantic orientation and provide suggestions for improvement
- To propose an algorithm for finding semantic orientation
- Perform some experiments for finding semantic orientation

#### **4.1 Comparative Study of Different Methods for Finding Semantic Orientation:**

In previous chapter, we have discussed and analyzed various approaches and methods for mining the opinions. These methods involve finding product's features, finding semantic orientation at sentence as well as document level with summarization. Since there are a number of approaches for same, it is obvious to have a question that which performs better. However Opinion Mining requires a lot of natural language processing, it is quite difficult to decide about the best one. Each method has its own advantages and disadvantages. Some employ unsupervised learning, some use supervised and some uses hybrid. Some may find semantic orientation at document level and some at sentence level. Based on our study, we present a comparative description of the approaches studied in Table 4.1.



	<b>Bing Liu (Semi-supervised Approach)</b>	<b>Turney &amp; Dave (Unsupervised Approach)</b>	<b>Liu &amp; Minguing(Semi-supervised Approach)</b>
<b>Review Format Used</b>	Format1 & 3	Format 3	Format 2
<b>Product feature Identification</b>	Yes	No	Yes
<b>Opinion words Identification</b>	Yes	Yes	No need
<b>Semantic Orientation at sentence level</b>	Yes	No	No need
<b>Semantic Orientation at document level</b>	Yes	Yes	No need

**Table 4.1**

Possibilities for improving the existing methods can be pointed out in following discussion. Since we are dealing with natural language and which is basically indirect, it is very tough to cope with it. Country, region, religion, age-group, sex, literacy etc are many factors which make our opinion mining task more difficult.

However we can identify a number of important issues, which should be considered to make our result more real and practical:

- **Rating Inconsistency:** In most extreme case, reviewers may not understand the rating system. For example a reviewer gives 1 star rating instead of 5 stars.
- **Ambivalence & Comparison:** Some reviewers may use negative connotation but finally they seem to be satisfied with the product. Some may compare negative experience with one product with positive using other. Some people like to describe about their past experience. For example, about their former cameras etc [10].
- **Sparse Data:** Many reviews are very short, and therefore we must be able to recognize a broad range of very specific features.
- **Skewed Distribution:** It has been found that positive reviews are predominant, thus making classification biased.

Above discussed issues are the major bottlenecks for better results. Rating inconsistency problem can be removed by not having any such system of rating. Ambivalence and comparison are always user dependent so almost difficult to be

eliminated. Sparse data issue can be handled if we take reviews with equal weight in terms of region, religion, ethnic group, age, sex and so on. But it is again a very and tedious task; some tradeoff with accuracy can be tolerated. Last issue also needs similar treatment. Reviews about different products/ services need equal participation for better result. There are also other issues as we mentioned, play important role in such decision-making event.

#### **4.2 Proposed Algorithm for Finding Semantic Orientation:**

We have proposed an algorithm for determining semantic orientation of words of any given review of format 3 (free form text.), which is shown below:

##### **Algorithm:**

**Input:** Raw review from epinions.com, set of predefined rated adjective words

**Output:** Document with semantic orientation (positive/negative)

**Step 1:** For each sentence segment in the given review,

Assign part-of-Speech tag to it.

**Step 2:** For each tagged word in review,

Find words or sentences with idioms, out of domain words etc. and remove

**Step 3:** Based on POS tagging, find the features of the product using Rule\_For\_Extraction() function.

**Step 4:** For each feature obtained in last step,

Find its orientation w.r.t. positive and negative adjectives using PMI\_cal() function

**Step 5:** Find the difference between corresponding positive and negative orientations as:

Semantic Orientation = PMI of phrase with “excellent” – PMI of phrases with “poor”

**Step 6:** Calculate average semantic orientation of the each extracted features of the review.

If average is positive

Then document is recommended

Else

Not recommended

**Step 7:** Stop

**Function Rule\_For\_Extraction**(preprocessed tagged review)

```
{ until end of review is not encountered
  { read three consecutive tag of the review say t1, t2 and t3
    if
      

| <b>word1's tag (t1)</b> | <b>word2's tag (t2)</b> | <b>word3's tag (t3)</b> |
|-------------------------|-------------------------|-------------------------|
| JJ                      | NN or NNS               | anything                |
| RB, RBR, or RBS         | JJ                      | not NN nor NNS          |
| JJ                      | JJ                      | not NN nor NNS          |
| NN or NNS               | JJ                      | not NN nor NNS          |
| RB, RBR, or RBS         | VB, VBD, VBN, or VBG    | anything                |


      Then extract word1 and word2
      t1 = t2, t2 = t3 and t3 = new tagged word;
    }
  }
}
```

**PMI\_cal** (no of hits of phrase with excellent, no of hits of phrase with poor)

```
{ POOR = no of hits for "poor";
  EX = no of hits for "excellent";
  ex = no of hits of phrase with excellent;
  poor = no of hits of phrase with poor;
  Semantic Orientation of Phrase =  $\log_2 [ ( ex*POOR) / ( poor * EX) ]$ 
}
```

### **Brief description of proposed algorithm:**

The method proposed has a number of steps to be performed:

In step 1, we take reviews as input to the algorithm and perform Part-of-Speech tagging in order to identify different parts of the sentences in the given review.

Since among these, some are main indicators of semantic words, this is an important step to proceed further.

In next step, we pre-process the tagged review. This is done to make input free from out of context words. However this step requires more human involvement, but it will always pay in terms of accuracy.

Step 3 basically deals with feature's extraction from the tagged review. This can be explained as follows:

For extracting features, we have number of methods. Some use adjectives as good indicators to the product features. Nearby noun or noun phrases are treated as product features. Some use predefined patterns, which extract the product features if the sequences of words are following the any of the given patterns. We employed last approach to extract the features of the review (product).

In 4<sup>th</sup> step, we employ the function to calculate semantic orientation of the extracted product's features in last step. A number of methods exist for this and its detail is discussed as follows:

To find semantic orientation of the extracted phrases, again a number of methods are being used. Here we used PMI-IR algorithm which finds the association of the extracted words with predefined words "excellent" and "poor". Then differences between these are calculated and sign of average to all the differences decide the class of the review.

Step 5 finds the average of all the semantic orientation of the extracted features of the review. Last step checks for sign of the average semantic orientation and if it is positive, then review is classified as recommended or positive and if its value comes out to be negative, it is termed as negative or not recommended.

### **4.3 Experiments for Finding Semantic Orientation:**

For performing experiment to determine semantic orientation, we downloaded some reviews from Epinions.com of Bank of America. Downloaded reviews are of different times with about equal participation in terms of sex and rating system.

Since on the web, a lot of things are hidden, we could not consider the age group, ethnic and other things in our experiment. However following steps are taken in order to find semantic orientation of reviews:

1. Download number of reviews of Bank of America as fig 4.1.

## Big Bank, Big Money, Big Customer Service Issues

Written: Dec 01 '05

**Product Rating:** ★★☆☆☆ **Pros:** The bank is so big, it has any financial service you could ever hope for

**Cons:** The bank is so big, it is impersonal and makes you feel like a number

**The Bottom Line:** Bank of America has lots of financial services to choose from, but it has grown large to offer decent customer service.

---

### Bryan\_Carey's Full Review: Bank of America

Many large banking operations exist in the USA, competing for the deposits and loans of everyday Americas with smaller, more local banks. One bank that is a giant among America's financial institutions is **Bank of America**, a financial operation headquartered in Charlotte, North Carolina.

#### What Does This Bank Offer?:

Bank of America is a full service bank offering virtually anything you would want in a financial institution. It has checking accounts, savings, home loans, education loans, car loans, business loans, certificates of deposit, credit cards, brokerage services, and much more.

Fig. 4.1

2. We use Stanford Part-of-Speech tagger to tagging purpose. We just taken main body of the review and tagged it for input as in next step as shown in fig 4.2.

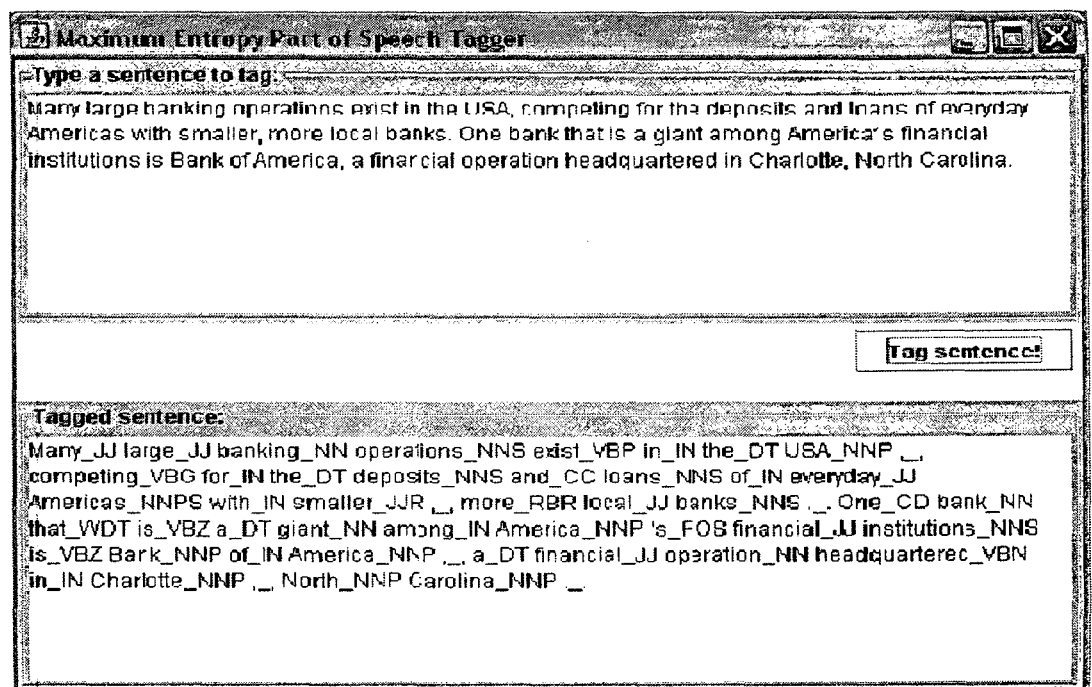


Fig. 4.2

3. Tagged review is then preprocessed in order to remove idioms, out of domain words, example of others etc. from this.
4. From the tagged review, we identified some pre specified patterns of part-of-speech tags. As per patterns, we extracted words representing product features in the review.
5. Once we have a set of extracted words representing opinions in the given review, we find its semantic orientation using PMI-IR method. We search number of results obtained by AltaVista site. For positive inclination, we search with “excellent” and for negative; it is searched with combination with “poor”. Then we calculate semantic orientation of each extracted phrases and take average of all those values and sign of summation decide the orientation of the document. For example, Table 4.2 shows the details of extracted phrases, its POS patterns, Semantic Orientation and Final result of the document.

Extracted Words(phrases)	Part-of-Speech patterns		Semantic Orientation
	First Tag	Second Tag	
large banking	JJ	NN	2.8307
local banks	JJ	NNS	-1.5488
financial operation	JJ	NN	0.5026
different ways	JJ	NNS	-0.2789
personal perception	JJ	NNP	-1.5471
so large	RB	JJ	-1.4898
free number	JJ	NN	3.5780
more personal	RBR	JJ	0.2047
much one	RB	CD	-1.1754
too many	RB	NN	-1.8570
so busy	IN	JJ	-2.6501
.....	.....	.....	.....
Average Semantic Orientation			-ve
<b>Expected Outcome = Not Recommended</b>			
<b>Final Outcome = Not Recommended</b>			
<b>Result = “Correct”</b>			

**Table 4.2**

S. No.	Expected Outcome	Semantic Orientation	Final Outcome	Result
Review 1	Not recommended	-ve	Not recommended	Correct
Review 2	Recommended	-ve	Not recommended	Wrong
Review 3	Not recommended	-ve	Not recommended	Correct
Review 4	Not recommended	+ve	Recommended	Wrong
Review 5	Recommended	-ve	Not recommended	Wrong
Review 6	Not recommended	-ve	Not recommended	Correct
Review 7	Recommended	+ve	Recommended	Correct
Review 8	Not recommended	-ve	Not recommended	Correct
Review 9	Not recommended	-ve	Not recommended	Correct
Review 10	Recommended	-ve	Not recommended	Wrong
Review 11	Recommended	-ve	Not recommended	Wrong
Review 12	Recommended	-ve	Not recommended	Wrong
Review 13	Not recommended	-ve	Not recommended	Correct
Review 14	Not recommended	-ve	Not recommended	Correct
Review 15	Recommended	-ve	Not recommended	Wrong
Review 16	Recommended	+ve	Recommended	Correct
Review 17	Recommended	+ve	Recommended	Correct
Review 18	Recommended	-ve	Not recommended	Wrong
Review 19	Not recommended	-ve	Not recommended	Correct
Review 20	Recommended	+ve	Recommended	Correct
Review 21	Recommended	+ve	Recommended	Correct
Review 22	Not recommended	-ve	Not recommended	Correct
Review 23	Not recommended	-ve	Not recommended	Correct
<b>Total Reviews = 23,</b> <b>Correct Classification = 15, Wrong Classification = 8</b> <b>Final Result = 65.21%</b>				

**Table 4.3**

Table 4.3 shows overall result of our experiment. Result obtained was 65.21% result for Bank of America review spread across time with good mixture of different rate review by both sexes.

The web contains a wealth of product reviews but sifting through them is a daunting task. Ideally an opinion mining tool would process a set of search result for a given item, generating a list of product attributes and aggregating opinions about each of them (poor, mixed, good). However the process of determining appropriate orientation of any document (review) is full of uncertainty about product, subjectivity, imprecision in ratings etc and cannot be handled by traditional text mining methods [18]. Therefore building such a system is a very complicated task. A number of approaches have been applied to determine the polarity of the document or finding overall opinion about any particular service or product.

Categorization in opinion mining is basically done into three parts: at document level, at sentence level and Comparison. Each of this categorization has its own advantages and disadvantages. Opinion mining is much more involved with natural language processing because the reviews we encounter are of free format and have a large amount of inherent subjectivity. Therefore determining semantic aspects of a document based on the position of words in the document and other semantic parameters is very crucial in order to have efficient and more accurate result. Apart from these issues, some other issues like rating inconsistency, ambivalence, comparison, sparse data and skewed distribution make this mining task more complicated [12].

In this work, we tried to explore and analyze the existing methods based on unsupervised and semi supervised learning. We have also tried to provide a general framework in order to deal with opinion mining efficiently. On the basis of different



parameters, we have compared different approaches for finding semantic orientation as well as product features. Based on the studies, we found that semi-supervised techniques are better for format 1 and format 2 where we have smaller sentences and unsupervised learning is more suitable for free format review. Both approaches have their own advantages and disadvantages. After analyzing these, we have proposed an algorithm which can take advantages of both approaches and that can find the polarity of the opinion words in the given review using semantic similarity measures.

Finally we can conclude that this work is an attempt to explore new area of opinion mining. We have been able to identify many research issues with this detailed background. There are a number of directions where we can further work: new approach dealing with skewed and sparse data efficiently, more accurate method to find semantic orientation as well as product feature identification etc.

## References

---

1. Ana-Maria Popescu , Oren Etzioni, “Extracting product features and opinions from reviews”, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing 2005 (pp.339-346)
2. Andera Esuli, Dissertation Abstract “Automatic Generation of lexical Resources for Opinion Mining: Models, Algorithms and Applications”, ACM SIGIR Forum, Volume 42, Issue 2, 2008 (pp. 105-106)
3. Bing Liu , Minqing Hu , Junsheng Cheng, “Opinion observer: analyzing and comparing opinions on the Web”, Proceedings of the 14th international conference on World Wide Web 2005 (pp. 342-350)
4. Bing Liu, “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)”, Springer-Verlag New York, Inc., Secaucus, NJ, 2006(pp. 417-430)
5. Bo Pang , Lillian Lee , Shivakumar Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques”, Proceedings of the ACL-02 conference on Empirical methods in natural language processing 2002 (pp.79-86)
6. Church K. W., Hanks P., “Word association norms, mutual information and lexicography”, Proceedings of the 27<sup>th</sup> Annual Conference of the ACL, 1989 (pp. 76-83)
7. Dave K, D., Lawrence, S., Pennock, D.: ‘Mining the peanut gallery: Opinion extraction and semantic classification of product reviews’, Proceedings of the 12th international conference World Wide Web 2003 (pp. 519-528)
8. H. Binali, V. Potdar and C. Wu, “A State of The Art Opinion Mining and Its Application Domains”, Proceedings of the 2009 IEEE International Conference on Industrial Technology- volume 002009 (pp. 1-6)
9. Hu Minqing, Liu Bing, “Opinion Feature Extraction Using Class Sequential Rules” In Proceedings of the Spring Symposia on Computational Approaches to Analyzing Weblogs, 2006
10. Jin Wei, Ho Hung Hay, Srihari Rohini K., “OpinionMiner: A novel Machine learning System for Web Opinion Mining and Extraction”, Proceedings of the ACM international conference on Knowledge discovery and data mining 2009 (pp. 1195-1203)

11. Kimber M, Han J, "Data Mining: Concepts and Techniques", by Elsevier
12. Kim Soo-Min, Hovy Edauard, "Determining the sentiment of Opinions", In Proceedings of COLING 2004.
13. Minqing Hu, Bing Liu, "Mining and Summarizing Customer Reviews", Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining 2004 (pp. 168-177)
14. Peter Dayan, MIT, "Unsupervised Learning" Appeared in The MIT Encyclopedia of the cognitive sciences, Editors R.A. Wilson & F. Keil
15. P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of the 40th Annual Meeting on Association for Computational Linguistics 2002 (pp. 417-424)
16. Shandilya K. Shaishir, Dr. Jaain Suresh, "Opinion Extraction & Classification of Reviews from Web Documnets", IACC 2009: IEEE International Advance Computing Conference 2009 (pp. 924-927)
17. T. Wilson, J. Wiebe and R. Hwa , "Just how mad are you? Finding strong and weak opinion clauses", AAAI'04: Proceedings of the 19th National Conference on Artificial Intelligence, 2004 (pp. 761-767)
18. V. Hatzivassiloglou and K. Mckeown, "Predicting the semantic Orientation of Adjectives", proceedings of the 35<sup>th</sup> Annual Meeting of the ACL and the 8<sup>th</sup> Conference of the European Chapter of the ACL 1997 (pp. 174-181)
19. Witten IH, Frank E, "Data mining: practical machine learning tools and techniques with Java implementations", ACM SIGMOD Record, v.31 n.1, March 2002
20. Zoubin Ghahramani, "Unsupervised Learning", Gatsby Computational Neuroscience Unit, University College London, UK