“OPINION MINING AND SENTIMENT ANALYSIS”

Major project Submitted in partial fulfillment of the requirements

For the award of degree of

Master of Technology
In
Information Systems

Submitted By

DINESH KUMAR

(Roll No. 02/IS/10)

Under the guidance of

Mr. RAHUL KATARYA

Assistant Professor

Department of Information Technology

Department of Information Technology

Delhi Technological University

Delhi

Session 2010-2012

CERTIFICATE

This is to certify that **Mr. Dinesh Kumar** (02/IS/10) has carried out the major project titled "OPINION MINING AND SENTIMENT ANALYSIS" as a partial requirement for the award of Master of Technology degree in Information Systems by Delhi Technological University.

The major project is a bonafide piece of work carried out and completed under my supervision and guidance during the academic session 2010-2012. The matter contained in this report has not been submitted elsewhere for the award of any other degree.

(Project Guide)

Mr. Rahul Katarya
(Asst. Professor)
Department of Information Technology
Delhi Technological University
Bawana Road, Delhi-110042

Acknowledgements

I take this opportunity to express my sincere gratitude towards **Mr. Rahul Katarya, Assistant Professor** (Information Technology) for his constant support and encouragement. His excellent guidance has been instrumental in making this project work a success.

I would like to thank **Dr. O.P. Verma**, H.O.D of Department of Information Technology for his useful insights and guidance towards the project. His suggestions and advice proved very valuable throughout.

I would like to thank members of the Department of Information Technology at Delhi Technological University for their valuable suggestions and helpful discussions.

I would also like to thank my family and friends, who have been a source of encouragement and inspiration throughout the duration of the project. I would like to thank the entire DTU family for making my stay at DTU a memorable one.

Dinesh Kumar
Roll No. 02/IS/10
M.Tech (Information Systems)
E-mail: dine121sh@gmail.com

ABSTRACT

This Thesis deals with the latest online opinion mining and sentiment analysis technique i.e**.** Opinions are usually subjective expressions that describe people's sentiments, appraisals or feelings toward entities, events and their properties. Opinions are so important that whenever we need to make a decision, we want to hear others' opinions. Individuals try to collect opinions from friends and family members and business executives use the surveys, focus groups, consultants to collect opinions. The opinion not only helps the customer to buy good product, also help the product manufacture to see the pros and cons of their product .and also show the comparison of his product with the author competitor and help the product manufacture to see which of his product future most of the customer like and dislike and he can improve their product future.

Keywords: sentiment, subjective, appraisals, entities

Contents

List of Figures

Chapter 1

INTRODUCTION OF MINING

1.1 Introduction

Mining is basically to extract something from its source or warehouse. As earth is the source of metals and minerals, we can extract those from the earth. And data warehouse is the collection of different types of data; we can extract useful data from the data warehouse. It is called data mining [1].

1.2 Introduction of Web Mining

There are large numbers of information available on web. Web mining is the process of extracting useful, valuable information on World Wide Web or we can say that, it is the application of data mining technique to discover pattern from World Wide Web. It is the integration if information that the user need[2].

1.3 Types of Web Mining

Web mining is of three types:

1. Web usage mining
2. Web content mining
3. Web structure mining

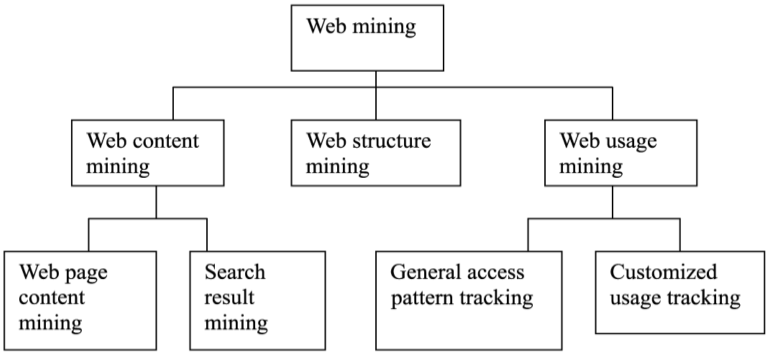


Figure 1.1 Taxonomy of Web Mining

1.3.1 Web Usage Mining

As the name suggests that web usage mining is the process of what we search and use on the internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

Customized usage tracking targeted to specific usage or users. By evaluating a user's sequence of clicks, information about a user (or a group of users) is detected. This could be used to perform pre-fetching and caching of pages [2] [4].

1.3.1.1 Uses of Web usage Mining

- Personalization for a user can be achieved by keeping track of previously accessed pages.
- By determining frequent access behavior for users, needed links can be identified to improve the performance. General access pattern tracking is a type of usage mining that looks at a history of Web pages visited. This type of mining describes what the users generally search on the internet.
- Identifying common access behaviors can help improve actual design of Web pages and make other modifications to the site.
- Patterns can be used to gather business intelligence to improve sales and advertisement.
- Activities of Web usage mining:
- Preprocessing activities help us in processing activities. These activities reformatting the server log data before use.
- Pattern discovery activities find the hidden patterns within the log data. So it is an important portion of mining activities.
- Pattern analysis is the process of analyzing the results of pattern discovery activities.

1.3.1.1.1 Preprocessing

It includes cleaning, user identification, session identification, path completion and formatting of log data. Before processing the log, the data has to be cleansed by removing irrelevant information.

The pages that are visited from one source can be grouped by the server, so that the server can understand the page references from a user. A server also identifies the session. In session identification, the server takes record of the web usage in that session. A session is a set of page references from one source site during one logical period. Login and logoff represents the logical start and end of the session.

1.3.1.1.2 Data Structures

Data Structures are used to keep record of patterns those are identified during the Web usage mining process. One of the basic data structure is tree.

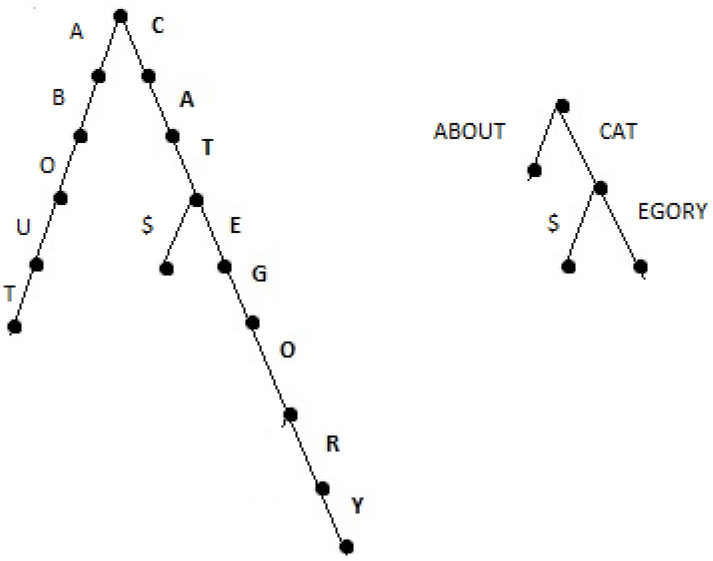


Figure 1.2 Sample Data Structure

With the help of suffix tree, we can find the subsequence in a sequence, and also the common subsequences among multiple sequences.

Pattern Discovery: It is to discover the pattern of a source's usage. A traversal pattern is a set of pages visited by a user in a session.

Pattern Analysis: Pattern analysis is the process of analyzing the results of pattern discovery activities. Once the patterns are identified, they must be analyzed to determine how that information can be used. Web logs can identify patterns that are of interest because of their uniqueness.

1.3.2 Web Content Mining

Web content mining examines the content of Web pages as well as results of Web searching. The content includes textual data as well as multimedia data. Web content mining is further divided into Web page content mining and search results mining.

Web page content mining is traditional searching of Web pages, while Search results mining is a further search of pages found from a previous search. Web content mining can improve the traditional search engines. Web content mining uses data mining techniques for efficiency, effectiveness and scalability.

Web content mining can be divided into:

- Agent based approach
- Database based approach

Agent based approach have software agents like search engines that perform the content mining. Intelligent search agents use techniques like user profiles or knowledge concerning specific domains. Personalize Web agents use information about user preferences to direct their search.

Database-based Approach views the Web data that belongs to a database. There have been approaches that view the Web as a multilevel database, and there have been many query languages that target the Web.

Basically content mining is a type of text mining. Text mining hierarchy shows Simple functions at the top and complex functions at the bottom[3].

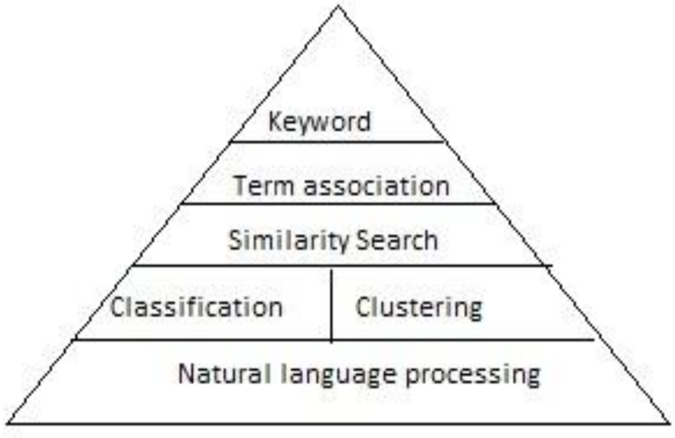


Figure 1.3: Text Mining Hierarchy

1.3.3 Web Structure mining

Web structure mining gives the information from the actual organization of pages on the Web. It is used to classify Web pages and to create similarity measures between documents.

Techniques of Web Structure Mining

- Page Rank
- HITS Algorithm

Page Rank

It is the technique used by Google. It is used to measure the importance of a page and to prioritize the pages returned from a traditional search engine using keyword searching. The Page Rank value for a page is calculated based on the number of pages that point to it.

Given a page p, we use B_p to be the set of pages that point to p and F_p to be the set of links out of p. The Page Rank of p is defined as:

$$PR(p)=cq\in B_p\, PR(q)/N(q)$$

$$0<c<1 \text{ , used for normalization}$$

$$|N_q|=|F_q|$$

Rank sink, is a problem with Page Rank when a cyclic reference occurs.

HITS(Hyperlink-induced topic search) algorithm:

A search engine SE, is used to find a small set, root set (R), of pages P, which satisfy the given query q. This set is then expanded into a larger set, base set (B), by adding pages linked either to or from R. This is used to induce a sub-graph of the Web. This graph is the one that is actually examined to find the hubs and authorities.

R=SE(W, q)
B=R ∪ {pages that link to pages in R}
GB,L= Sub-graph of W induced by B
GB, L'= Delete links in G within same site
Xp= ∑q where q, p L' Yq; Find authority weights
Yp= ∑q where p, q L' Xq; Find hub weights
A=p {p has one of the highest Xp}
H=p p has one of the highest Yp

1.4 Need of Web Mining

In today's life web mining is very important because for every question comes in mind, we take help of internet and search engines. As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. Web mining reduces the cost and time of the users because it gives the information that is really useful for the users. With the recent explosive growth of the amount of content on the Internet, it has become increasingly difficult for users to find and utilize information and for content providers to classify and catalog documents. Traditional web search engines often return hundreds or thousands of results for a search, which is time consuming for users to browse. On-line libraries, search engines, and other large document repositories (e.g. customer support databases, product specification databases, press release archives, news story archives, etc.) are growing so rapidly that it is difficult and costly to categorize every document manually. In order to deal with these problems, researchers look toward automated methods of working with web documents so that they can be more easily browsed, organized, and cataloged with minimal human intervention. And hence, we use web mining.

1.5 Techniques and Algorithms used in web mining

There are several techniques that are used in web mining. These various approaches and techniques are well studied and implemented in different applications and scenarios by research efforts contributed from the expertise of Database, Artificial Intelligence, Information Science, Natural Language Processing, Human Computer Interaction even Social Science. Although these algorithms and techniques are developed from the perspectives of different disciplines, they are widely used and applied in the above mentioned areas simultaneously.[5]

1.5.1 Association Rule Mining

The purpose of finding association rules is to analyze the co-existence relation between items, which is then utilized to make appropriate recommendation. The issue has attracted a great deal of interest during the recent surge in data mining research because it is the basis of many applications, such as customer behavior analysis, stock trend prediction, and DNA sequence analysis. For example, an association rule “apple⇒ strawberry (90%)” indicates that nine out of ten customers who bought apples also bought strawberry. These rules can be useful for store layout, stock prediction, DNA structure analysis, and so forth.[6][7]

Association rule mining problem: The problem of association rule discovery can be stated as follows [6]: Let I = {i1, i2, . . . , ik} be a set of items. A subset of I is called an item set. The item set, tj, is denoted as {x1, x2 . . . xm}, where xk is an item, i.e., xk ∈ I for 1 ≤ k ≤ m. The number of items in an item set is called the length of the item set. An item set with length is called an item set. An item set, ta = {a1,a2, . . . ,an}, is contained in another item set, tb = {b1,b2, . . , bm}, if there exists integers 1≤ i1 < i2 <. . . <in ≤ m, such that a1 ⊆ bi1 , a2 ⊆ bi2 ,. . . , an ⊆ bin . We denote ta as a subset of tb, and tb a superset of ta. The support of an item set X, denoted as support(X), is the number of transactions in which it occurs as a subset. A k length subset of an item set is called a k-subset. An item set is frequent if its support is greater than a user-specified minimum support (min sup) value. The set of frequent k-item sets is denoted Fk. An association rule is an expression A⇒B, where A and B are item sets. The support of the rule is given as support(A⇒B)=support(A∪ B) and the confidence of the rule is given as con f (A⇒B)=support(A∪ B)/support(A) (i.e., the conditional probability that a transaction contains B, given that it contains A). A rule is confident if its confidence is greater than a user-specified minimum confidence (min con f). The associate rule mining task is to generate all the rules, whose supports are greater than min sup, and the confidences of the rules are greater than min con f. The issue can be tackled by a two-stage strategy.

- Find all frequent item sets. This stage is the most time consuming part. Given k items, there can be potentially 2k frequent item sets. Therefore, almost all the works so far have focused on devising efficient algorithms to discover the frequent item sets, while avoiding traversing unnecessary search space somehow.
- Generate confident rules. This stage is relatively straightforward and can be easily completed.

1.5.2 Building Task-Specific User Access Patterns

Input: the discovered session-task preference distribution matrixθ, m user sessions S = {si, i=1, ..., m}, and the predefined threshold µ .

Output: task-specific user access patterns, TAP = {ap_k, k = 1, ··· , t}

Step 1: For each latent task zk, choose all user sessions with θ i, k >µ to construct a user session aggregation Rk corresponding to zk;

Rk = {si | $\theta_{i,\,k}$ > µ, k = 1, ···, t}

Step 2: Within the Rk, compute the aggregated task-specific user access pattern in terms of a weighted page vector by taking the sessions' associations with zk, i.e. $\theta_{i,k}$ into account

$$ap_k = \frac{\sum_{si \in Rk} \theta i,k\ .Si}{|Rk|}$$

where |Rk| is the number of the chosen user sessions in Rk.

Step 3: Output a set of task-specific user access patterns TAP corresponding to t tasks, TAP = {ap_k, k = 1, ···, t}. In this expression, each user access pattern is represented by a weighted page vector, where the weights indicate the relative visit preferences of pages exhibited by all associated user sessions for this task-specific access pattern.[8]

1.5.3 Spectral Co-Clustering Algorithm

Input: The user session collection S and page view set P, and the Web log file Output: A set C = {C1,···, Ck} of k subsets of sessions and page views such that the cut of k-partitioning of the bipartite graph is minimized. [10]

1. Construct the usage matrix A from the Web usage log, whose element is determined by the visit number or duration of one user on a specific page;

2. Calculate the two diagonal matrices Ds and Dp of A;

3. Form a new matrix NA = D−1/2s AD−1/2p ;

4. Perform SVD operation on NA, and obtain the left and right k singular vectors Ls and Rp, and combine the transformed row and column vectors to create a new projection matrix PV;

5. Execute a clustering algorithm on PV and return the co-clusters of subsets of S and P, Cj = (Sj, Pj).[11][12]

1.5.4 Page Gather algorithm

Step 1. Process the access log into visits.

Step 2. Compute the co-occurrence frequencies between pages and create a similarity matrix.

Step 3. Create the graph corresponding to the matrix, and employ clique(or connected components) finding algorithm in the graph.

Step 4. For each cluster found, create new index Web pages by synthesizing the links to the documents of pages contained in the cluster.

In the first step, an access log, containing a sequence of hits, or requests to the Web server, is taken for processing. Each request typically consists of time-stamp made, the URL requested and the IP address from which the request originated. The IP address in this case is treated as a single user. Thus a series of hits made in a day period, ordered by the timestamps, is collected as a single session for that user. The obtained user sessions in the form of requested URLs form the session vector, will be used in the second step. To compute the co-occurrence frequencies between pages, the conditional probability of each pair of pages P1 and P2 is calculated. Pr (P1|P2) denotes the probability of a user visiting P1 if it has already visited P2, while Pr(P2|P1) is the probability of a user visiting P2 after having visiting P1. The co-occurrence frequency between P1 and P2 is the minimum of these values. The reason why using the minimum of two conditional probabilities is to avoid the problem of asymmetrical relationships of two pages playing distinct roles in the Web site. Last, a matrix corresponding to the calculated co-occurrence frequencies is created, and in turn, a graph which is equivalent to the matrix is built up to reflect the connections of pages derived from the log as well.

In the third step, a clique finding algorithm is employed on the graph to reveal the connected components of the graph. In this manner, a clique (or called cluster) is the collection of nodes (i.e. pages) whose members are directly connected with edges. In other words, the sub graph of the clique, in which each pair of nodes has a connected path between them, satisfies the fact that every node in the clique or cluster is related to at least one other node in the sub graph.

Eventually, for each found cluster of pages, a new indexing page containing all the links to the documents in the cluster is generated. From the above descriptions, we can see that the added indexing pages represent the coherent relationships between pages from the user navigational perspective, therefore, providing an additional way for users to visually know the access intents of other users and easily browse directly to the needed pages from the provided instrumental pages.

1.5.5 Backtrack path finding

For a single target page *T*, the user is expected to execute the following search strategy

1. Start from the root.

2. While (current location *C* is not the target page *T*) do

(a) If any of the links from *C* is likely to reach *T*, follow the link that appears most likely to T.

(b) Else, either go back (backtrack) to the parent of *C* with some possibility, or cease with some possibility.

For a set of target pages (T1,T2,···, Tn), the search pattern follows the similar procedure, but after the user identifying Ti, it continues searching Ti+1. In this scenario, the hardest task

is to differentiate the target page from other pages by simply looking at the Web log. For the former case, the content pages are most likely to be the target pages for a user. Given a website of portal site, where there is not a clear separation between the content and index pages, resulting in the difficulty in judging the content pages, counting the time spent on a specific page will provide a useful hint to judge this. Here it is known that the user spent more time than the time thresholds are considered the target page.

To identify the backtrack points, the Web log is analyzed. However, the browser caching technology brings in unexpected difficulty in differentiating the backtrack points, otherwise, the phenomenon of a page where previous and next pages are the same gives the justification to it. In this work, rather than disabling the browser caching, a new algorithm of detecting

Backtrack points was devised. The algorithm is motivated by the fact that if there is no link between P1 and P2, the user must click the “back” button in the browser to return from P1 to P2.

Therefore the detection of the backtrack points is becoming the process of detecting whether there is a link between two successive pages in the Web log.

Chapter 2

OPINING MINING

2.1 What is Opining mining?

In the world, Textual information broadly categories two main parts facts and opinion facts are objective expression about entity. Opining is usually subjective expression about their entity. The concept of opinion is a very broad concept. Opinion is so important that whenever we want to make decision, we hear the opinion of other people. That is not important the individual but also important for the organization .Opining mining is the process used for automatic extracting a knowledge from others opining about a particular problem or topic. It is type of natural language processing to know the mood feeling of the public about a particular product.

In General, Sentiment analysis is to determine the attitude of a speaker/writer with respect to some topic or the overall polarity of a document. This attitude may be his/her judgment or affective state, evaluation.

Opinion mining also called sentiment analysis. It involves building a system to collect and analyze opinions about the product that are made in blog posts, comments, reviews. Automated opinion mining often use machine learning, a part of artificial intelligence.

2.2 Sentiment Mining

Opinion mining or Sentiment analysis is the detailed computational study of opinions, emotions and sentiments expressed in text.

It is an area of research in which efforts are made to make an automatic system to determine human views or opinion from text that is written in natural language. It seeks to determine the view point of the people underlying a text span.

Generally, opinions can be expressed on anything it can be a product, an organization, a service, an event, an individual, or a topic. The term object is to denote the target entity about which comments are made. An object can have a set of attributes (or properties) and a set of components (or parts). Each component can have number of sub-components, further the set of attributes of those also and so on.

Example of a sample Review: “I bought a Samsung Corby-Mobile Phone a few days ago. It is a nice phone and very easy to use. The touch screen is really very soft and very cool. The voice quality is extremely clear too. Although the battery backup is not long, but I can compromise with that for its other features. However, my friend was angry with me as I did not ask him before I bought the phone. He thought that the phone will be very expensive, and he wanted me to return it to the shop. “

Now the question arises: what we should mine or extract from this sample review?

- First thing that we should notice that there are many opinions in this review.
- Also we should notice that the opinions have some targets or objects about which the opinions are expressed.
- Finally, the sources and the holders of the opinions should also be noticed.

Opinion mining tries to mine all these things from a given review text.

2.3 Usefulness of Opinion mining

If you are in marketing field, for example, it can help you to judge the success of a new product launch or ad campaign, and determining which versions of a particular product or service are famous and even identifying which demographics like or dislike particular features of the product or the service. For example, a sample review might be very positive about Nikon digital camera, but it can be specifically negative about its weight. To be able to identify this kind of information from the sample review in a systematic way gives the vendor a very much clearer picture of the public opinion regarding the product or service rather than doing surveys or focus groups, because the sample review is created by the customer.

Opinion mining system built using software that is capable of extracting knowledge from database .it is simple like list of positive and negative words , or as complicated as conducting deep review of the given data to understand the grammar and sentence structure used in the text.

2.4 Challenges in Opinion Mining

There are a number of challenges that exists in opining mining:

- The first is that a word can to be positive in one situation while it may be considered negative in another situation. Let we take the following example: Take a sample word say "long" for instance. If a customer said that his mobile’s battery life is long, it would be a positive opinion. On the other hand, if some other customer said that the mobile's start-up time is very long, it would be a negative opinion. These differences in the meaning of a word mean that an opinion system that is trained to gather opinions/reviews for one type of product or product feature or service may not perform well on another.
- A second challenge in opinion mining is that people have different ways to express opinions about a target object. Most traditional text processing techniques relies on the fact that, with the little difference between two pieces of text, the meaning of that text does not change so much. However this is not true in case of opinion mining. In opinion mining, "the sound quality is good" is very much different from the "the sound quality is not good".
- Many of the reviews have both positive and negative review which is analyzed one by one at a time. However, the more and more informal the medium (for example: twitter, Facebook or blogs), the more likely the people combine different opinions into one single sentence. For example: "Though the lead actor rocked it, the movie bombed” is easy for a human being to understand it, but it is more difficult for a machine to parse it. Sometimes even some other people may have difficulty in understanding the thought of someone else based on a small piece of text because it lacks in context. For example, "That movie was as good as his last movie" is fully dependent on what the person have the opinion about the previous film.

2.5 Sentiment and Subjective classification

Subject classification is treated as sentiment analysis text classification. Two subtopic that have been extensively studied are (1) classification of opinioned document and expressing positive and negative expression (2) classifying a sentence or clause of the sentence as subjective or objective and for subjective sentence and clause classification it show positive ,negative and neutral opinion. The first topic know as sentiment classification and document level classification aim to find out general opinion if the author for example given a product review, it determine whether the review positive or negative. The second topic goes to individual sentence and determine whether sentence express opinion or not and if show, positive or negative opinion.

2.5.1 Sentence Level Sentiment Classification

Sentence-level sentiment analysis has two basic tasks:

Subjectivity classification: Find out sentence is Subjective or objective.

Objective: e.g., I bought a Nokia Phone two days ago.
Subjective: e.g., it is very nice phone.

Sentiment classification: For subjective sentences or clauses, classify positive or negative. Positive: It is very nice phone.

2.5.2 Document level sentiment classification

Given a set of documents D1, determine whether each document shows positive and negative expression.

Supervised learning:

Classification is done on basis of predefined data.

Supervision: Data is labeled with number of pre-defined classes. It is like that a “faculty” gives the guidance (supervision).

Unsupervised learning (clustering)

Class labels of the data are not known.

Given a set of data, the task is to establish the existence of different classes in the data.

2.5.3 Feature Based

Identify the sentiment of the opinion holder based on the features of a particular object. Because a person may like some features and some other person may not like that.

It gives better and in depth analysis of the product than the sentence level and document level sentiment classification.

2.6 Sentiment analysis and Web 2.0

The rise in social media such as Facebook, twitter, blogs and social networks sites has fueled lots of interest in sentiment analysis. With the proliferation of ratings, reviews, recommendations and any other forms of online expressions, online opinions are now turned into a kind of virtual useful currency for businesses which are looking to market their services, products, identify new opportunities in the market and in managing their reputation in the market. As businesses look for automating the process of filtering out the noise, identifying the relevant content, understanding the conversations and auctioning it properly, many of them are now looking to the field of sentiment analysis.

Research is being done towards this aim. Several research teams in universities/colleges around the world are now focusing on understanding the dynamics of sentiment in communities through sentiment analysis. The Cyber Emotions, for instance, recently identified the role of negative emotions in driving social networks discussions Sentiment analysis could therefore help in understanding why some e-communities die or fade away (e.g., MySpace) while others seem to grow without limits (e.g., Face book).

The problem is that in most sentiment analysis algorithms, very simple terms are used to express sentiments for a product or service. However, linguistic nuances, cultural factors and differing contexts make it extremely hard to turn a string of written text into a simple pro’s or con’s sentiment. The fact is that human often disagrees on the sentiment of text that illustrates how big a task it is for computers to get this right. The task becomes harder as shorter is the string.

Chapter 3

LITERATURE SURVEY

3.1 Analyzing and comparing opinion in WEB

The web has become the excellent source of information to gathering a customer review. There are large no of web site containing such opinion, e.g. customer review of product and blogs. Bing liu and all focusing online customer review on product. They make two contributions. First they propose novel frame work for analysis the customer review of competing product. A prototype system is called opinion observer is also implemented.

The system is such that with single glance of its visualization, the user is able to see that the strength and weakness of the product in the mind of the consumer in the term of various product features. This compression useful for both side the customers and the manufacturers. The customers can see all the product feature and consumer opinion about the product he/she wanted to buy which help him/her which product to buy. For product manufactures, the compression enables it to easily get marketing intelligent. Second technique based on language pattern mining purposed to extract product feature pros and cons of particular type of review.[14]

3.2 Two challenging task need to be perform

1. Identifying the product feature that the customer expresses their positive and negative opinion about the product.
2. For each feature identify the opinion whether it is positive or negative, negative opinion mean complain about product feature

There are three review format of web

1. Pros and cons: this review is asking to describe pros and cons separately.
2. Pros, cons and detailed review: this review asks to describe the pros and cons separately and in detail.
3. Free review: this review write freely or we can say that no separate pros and cons,

3.2.1 Opinion Observer works in two stages

Extracting and analyzing the customer review in two step

Step 1:-

1. In this case it connects and automatic downloads all the review on the page.
2. In this step, all the new reviews (which were not analyzed before) of every product are analyzed. Two tasks are performed, identifying product features and opinion orientations from each review. This can be done automatically or semi automatically.

Step 2:-

In this stage, based on the analysis results, different users can visualize and compare opinions of different products using a user interface. The user simply chooses the products that he/she wishes to compare and the systems then retrieves the analyzed results of these products and displays them in the interface.

3.2.2 Mining comparative sentence and relations

A comparative sentence express ordering relation between two set of entity based upon some common future. For example, the comparative sentence “Sony optics better then both Nikon and Samsung” show the comparative relation. The task of comparative mining is to identify the comparative sentence form text and find comparative relation in the comparative sentence.

This problem has many applications for example the product manufacturer want to see the customer opinion compare with its competitor [16].

It has two main tasks:

- Given a set of evaluating text find the comparative sentence from the text and classified it into different classes.
- Extract comparative relation between identify sentence. These include entity and their future that are being compared.

  The relation is express with

(<relation word>, <features>, <entity1>, <entity2>)

For example, we have comparative sentence “Sony optics better then both Nikon and Samsung”

(Better, {optics}, {Sony}, {canon, Nikon})

Both tasks are very challenging.

3.2.3 Types of Comparatives

- Non-Equal Gradable: Relations the type greater or less than that expresses a total ordering of some entities with comparing to certain features. This type also includes user preferences.
- Educative: Relations of the type equal to that state two entities as equal with respect to some features.
- Superlative: Relations of the type greater than or less than all others that rank one entity over the others.
- Non-Gradable: Non gradable Sentences are those which compare features of two or more entities, but do not grade them. The first three types of comparative are called gradable comparatives.[16]

3.3 Structural Opinion Mining for Graph-based Sentiment presentation

Based on analysis of on-line review we observe that most of the sentences have complicated opinion structures and they cannot be well represented by existing methods, such as frame-based and feature-based ones. In this work, the author proposes a new graph-based representation for sentence level sentiment mining. An integer linear programming-based structural learning method is then introduced to produce the graph representations of input sentences.

- One way to investigate the use of graphs for representing sentence level sentiment. The vertices are representing evaluation target, opinion expression, modifiers of opinion. The Edges are representing relations among them. The semantic relations among individual opinions are also included. With the help of the graph all the opinion which ignore by the other also be extracted.

- The author method is supervised structural learning method which takes a sentence as input and the proposed sentiment representation for output. The inference algorithm is based on integer linear programming which helps to concisely and uniformly handle various properties of our sentiment representation. By setting appropriate prior substructure constraints of the graph, the whole algorithm achieves reasonable.[17][18]

3.4 Graph Properties for extracting opinion

Here are some of the properties of graph G either from the definition of relations:

Properties:

1. The first property says that graph is connected and it is without directed cycle. Using individual opinion representation, each sub-graph of G which takes an opinion expression as root is connected and a cyclic. Thus, the connectedness is guaranteed for opinion expressions connected in opinion thread and the acyclic is guaranteed by the fact that if a modifier is shared by different opinion expressions, then edges from them always keep the graph directed acyclic.
2. The second property says that: Each vertex can have one out edge labeled with transition at most. The opinion thread B1 is a directed path in graph.
3. The third property says that graph is sparse. The graph is almost a rooted tree because the average in-degree of a vertex is 1.03 in our corpus. In other words, we can say that the cases that a modifier connects to more than one opinion expression occasionally occur comparing with those vertices that have a single parent. An explanation for this sparseness is that opinions in online reviews always concentrate in local context and have local semantic connections.

3.5 Mining and Summarizing Customer Reviews

Merchant who are selling products on the Web, usually ask his/her customers to review the products that they have purchased from them to verify and improve the quality of their products and the associated services. On basis of these feedbacks by other customers, a customer can decide whether he/she should purchase the product or not. But due to rapid advancement in

e-commerce technology, it is becoming more and more popular, as a result the number of reviews for a product is also growing rapidly.

For a daily use and famous product, the number of reviews or feedbacks can be in hundreds or even in thousands. As a result, it is difficult for the customer to read them and then to make a decision on whether he/she should purchase the product or not. This also creates problems for the manufacturer of the product to keep track and to manage customer opinions/feedbacks and take actions according to customer reviews. For the manufacturer, there are many additional difficulties because many merchant sites also sell the same product manufactured by different companies and the manufacturer himself/herself normally produces many kinds of products.

In this work, we aim to mine and to summarize all the customer reviews/feedbacks for a product. This mining and summarization task is different from traditional text summarization because we need to mine only the features of the product on which the customers have given their feedbacks, their reviews, their opinions and whether the opinions are positive or negative. We cannot summarize the feedbacks/reviews by selecting a part of reviews or rewrite some of the original sentences from the feedbacks to capture the main points as we do in case of the classic text summarization. So it is very much different from classic text summarization.

With the rapid growth in e-commerce technology, many numbers of products are being sold on the Web, and customers are also increasing who are buying products online using e-commerce. In order to increase their customer satisfaction and for better shopping experience, online merchants need to enable their customers to review or to express opinions and give feedbacks on the products that they are purchasing from them. With increase in users becoming comfortable with the Web, and with growing communication technologies, an increasing number of people are writing reviews and giving feedbacks.

As a result, the number of reviews that a product receives grows rapidly. Some popular/daily use products can get thousands of reviews at merchant sites in a short span of time. Furthermore, many reviews/feedbacks are long sentences and contain only a few part in the sentence in which opinions are expressed on the product.

This also makes it difficult and trouble for a customer to read those lengthy reviews in order to make an decision on whether they should purchase the product or not. If he/she only reads a few reviews, he/she may get a biased view. Because of large number of reviews, it is hard for product manufacturers to keep track of their customer opinions about their products and take actions according to those reviews. There exist many other difficulties also which merchant need to handle. So we need to mine and shorten the length of reviews.[19][20]

3.5.1 The task is performing in three steps

(1) To mine product features on which comments are given by the customers.

(2) Identifying only opinion sentences in each review/feedback and then deciding whether the opinion sentence is positive or negative.

(3) Summarizing the results on the product feedbacks.

Our experimental results using reviews/feedbacks of a number of products that are sold online show the effectiveness of the techniques.

3.5.2 Sentiment Analysis and Subjectivity

Textual information in the world can be divided into two main types:

1. Facts
2. Opinions

Facts: Facts are the expressions about an entity, or an event and their properties.

Opinions: In contrast with facts, Opinions are the subjective expressions that describe people’s sentiments, feelings, emotions or appraisals towards other entities, events and about their properties. The concept opinion is a very broad concept. In this paragraph, we are focusing only on opinion expressions that convey effects people’s emotions or sentiments in positive or negative way. In many of the research on textual information processing, it has been focused on mining and retrieval of factual information, e.g. Web search, information retrieval, text clustering, text classification and many other text mining as well as natural language processing tasks. Very little work had been done in the processing of opinions. Still, opinions (subjective expressions) are very important, whenever we need to make a decision regarding anything we want to get other’s opinions.

This is also true for organizations with individuals. The main reason for the lack of focus on opinions is that there was very little opinionated text available before the existence of World Wide Web. Before the WWW, whenever any individual need to make a decision, one typically ask for opinions from one’s friends and families. Whenever an organization needs to find the opinions/emotions or sentiments of the general public or society about the products they are manufacturing and about their services, they always conduct opinion polls from focus groups and they used to conduct surveys. However, after the Web, especially with the rapid growth of the user generated content on the Web in the last few years, the world has been transformed and now everything is changed.

The Web has dramatically changed the way that people express their views and opinions. They can now post reviews of products at merchant sites and express their views on almost anything in Internet forums, discussion groups, and blogs, which are collectively called the user-generated content. This online word of- mouth behavior represents new and measurable sources of information with many practical applications.

Now if one wants to buy a product, he/she is no longer limited to asking his/her friends and families because there are many product reviews on the Web which give opinions of existing users of the product. For a company, it may no longer be necessary to conduct surveys, organize focus groups or employ external consultants in order to find consumer opinions about its products and those of its competitors because the user-generated content on the Web can already give them such information.

However, finding opinion sources and monitoring them on the Web can still be a formidable task because there are a large number of diverse sources, and each source may also have a huge volume of opinionated text (text with opinions or sentiments). In many cases, opinions are hidden in long forum posts and blogs. It is difficult for a human reader to find relevant sources, extract related sentences with opinions, read them, summarize them, and organize them into usable forms.

Thus, automated opinion discovery and summarization systems are needed. Sentiment analysis, also known as opinion mining, grows out of this need. It is a challenging natural language processing or text mining problem. Due to its tremendous value for practical applications, there has been an explosive growth of both research in academia and applications in the industry. There are now at least 20-30 companies that offer sentiment analysis services in USA alone. This chapter introduces this research field. It focuses on the following topics:[21]

3.6 Sentiment and subjectivity classification

This is the area that has been researched the most in academia. It treats sentiment analysis as a text classification problem. Two sub-topics that have been

1. Classifying an opinionated document as expressing a positive or negative Opinion.
2. Classifying a sentence or a clause of the sentence as subjective or objective, and for a subjective sentence or clause classifying it as expressing a positive, negative or neutral opinion. The first topic, commonly known as sentiment classification or document-level sentiment classification, aims to find the general sentiment of the author in an opinionated text. For example, given a product review, it determines whether the reviewer is positive or negative about the product. The second topic goes to individual sentences to determine whether

a sentence expresses an opinion or not (often called subjectivity classification), and if so, whether the opinion is positive or negative (called sentence-level sentiment classification) [22].

Chapter 4

PROPOSED ALGORITHM

There are two types of textual information - Facts and Opinions.

Facts are the objective expressions about entities, events and their properties. Opinions are the subjective expressions that describe people's sentiments, appraisals or feelings or emotions toward entities, events and their properties.

Opinions are very important because whenever we want to make a decision about anything, we would like to hear other's opinion about that. Individuals try to collect opinions from society, friends and family members. And business executives usually do the surveys, focus groups, meet consultants to collect opinions.

Much of the existing research that are currently being done on textual information processing are now focusing on mining and retrieval of factual information, e.g., Web search, information retrieval, text clustering, text classification and many other text mining and natural language processing tasks.

With the rapid growth in Internet and the Web, now it is possible as well as easy to find out the opinions and experiences of those in the vast number of people that are neither well-known professional critics nor our personal acquaintances —that is, they are the people we have never heard of. Conversely, more and more people are making their opinions/suggestions available to people who are stranger to them via the Internet.

By Opinion mining we can study opinions, sentiments and emotions that are expressed in text.

4.1 Concept of Sentiment Mining

Opinion mining or Sentiment analysis is the detailed computational study of opinions, emotions and sentiments expressed in text.

It is an area of research in which efforts are made to make an automatic system to determine human views or opinion from text that is written in natural language. It seeks to determine the view point of the people underlying a text span.

Generally, opinions can be expressed on anything it can be a product, an organization, a service, an event, an individual, or a topic. The term object is to denote the target entity about which comments are made. An object can have a set of attributes (or properties) and a set of components (or parts). Each component can have number of sub-components, further the set of attributes of those also and so on.

Example of a sample Review: “I bought a Samsung Corby-Mobile Phone a few days ago. It is a nice phone and very easy to use. The touch screen is really very soft and very cool. The voice quality is extremely clear too. Although the battery backup is not long, but I can compromise with that for its other features. However, my friend was angry with me as I did not ask him before I bought the phone. He thought that the phone will be very expensive, and he wanted me to return it to the shop. “

Now the question arises: what we should mine or extract from this sample review?

- First thing that we should notice that there are many opinions in this review.
- Also we should notice that the opinions have some targets or objects about which the opinions are expressed.
- Finally, the sources and the holders of the opinions should also be noticed.

Opinion mining tries to mine all these things from a given review text.

4.2 Representation of Opinion

Opinion can be represented as a touple- (O, F, SO, H, T)

where

- O is a target object.
- F is the features of the object O.
- H represents an opinion holder.

- T is the time when the opinions are expressed.
- SO is the sentiment value of the opinion holder H on feature F of Object O at time T.

  SO is positive, negative, more granular rating or neutral.

4.2.1 Types of Opinion

- Direct Opinions: These are the sentiment expressions regarding some objects. Objects can be products, topics, events or persons.
  - E.g., “The picture quality of this mobile camera is great”
  - Subjective sentences.
- Comparisons: Comparisons are the relations used to express similarities or differences about more than one object. Usually used to express an ordering.
  - E.g., “mobile x is cheaper than mobile y.”
  - Subjective or Objective sentences.

4.2.2 Objectives of Sentiment mining

Given an opinionated document,

- To discover all quintuples (O, F, SO, H, T)
- Unstructured Text is converted to Structured Data.

Sentiment Mining Techniques

1. Sentence Level
2. Document level
3. Feature Based

- **Sentence Level Sentiment Classification**

Sentence-level sentiment analysis has two basic tasks:

- Subjectivity classification: Find out sentence is Subjective or objective.
    - Objective: e.g., I bought a Nokia Phone two days ago.
    - Subjective: e.g., it is very nice phone.
- Sentiment classification: For subjective sentences or clauses, classify positive or negative. Positive: It is very nice phone.

- **Document Level Sentiment Classification**

Given a set of documents D1, determine whether each document shows positive and negative expression.

- Supervised learning:
    - Classification is done on basis of predefined data.
    - Supervision: Data is labeled with number of pre-defined classes. It is like that a “faculty” gives the guidance (supervision).
- Unsupervised learning (clustering)
    - Class labels of the data are not known.
    - Given a set of data, the task is to establish the existence of different classes in the data.

- **Feature Based**

Identify the sentiment of the opinion holder based on the features of a particular object. Because a person may like some features and some other person may not like that.
It gives better and in depth analysis of the product than the sentence level and document level sentiment classification

4.3 Unsupervised Review Classification

Generally, Opinion words and phrases are the dominating factors of sentiment classification. Thus, unsupervised learning if done on basis of such words and phrases would be very useful.

The following technique performs classification based on some fixed syntactic phrases that are more probable to be used to express opinions.

The algorithm works in three steps:

Input to the algorithm is written review.

Output is the Classification i.e. positive or negative.

Steps

1. Use part-of-speech tagger to identify opinion phrases.
2. Estimate the semantic orientation of extracted phrase.
3. Assign the given review to a class (either recommended or not recommended)

All these steps are discussed in detail.

Step 1: Extract the Opinion phrases

In this step, phrases that contain adjectives or adverbs are extracted. The reason for doing this is that adverbs and adjective are the good indicators of subjectivity and opinions.

Therefore, the proposed algorithm extracts two consecutive words, where one member of these two is an adjective/adverb and the other one is a context word.

- From the review, two consecutive words are extracted in such a way that they match with patterns given in the following table. Reason behind this, Adjectives & Adverbs are the two good indicators of Opinion.

Table 1 Patterns of POS tags for extracting two-word phrases

	First word	Second word	Third word (not extracted)
1	JJ	NN or NNS	Any thing
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	Any thing

where,

JJ - Adjective

NN - Noun, singular

NNS- Noun, plural

RB - Adverb

RBR- Adverb, comparative

RBS - Adverb, superlative

VB- Verb, base form

VBD- Verb, past tense

VBG - Verb, present participle

VBN - Verb, past participle

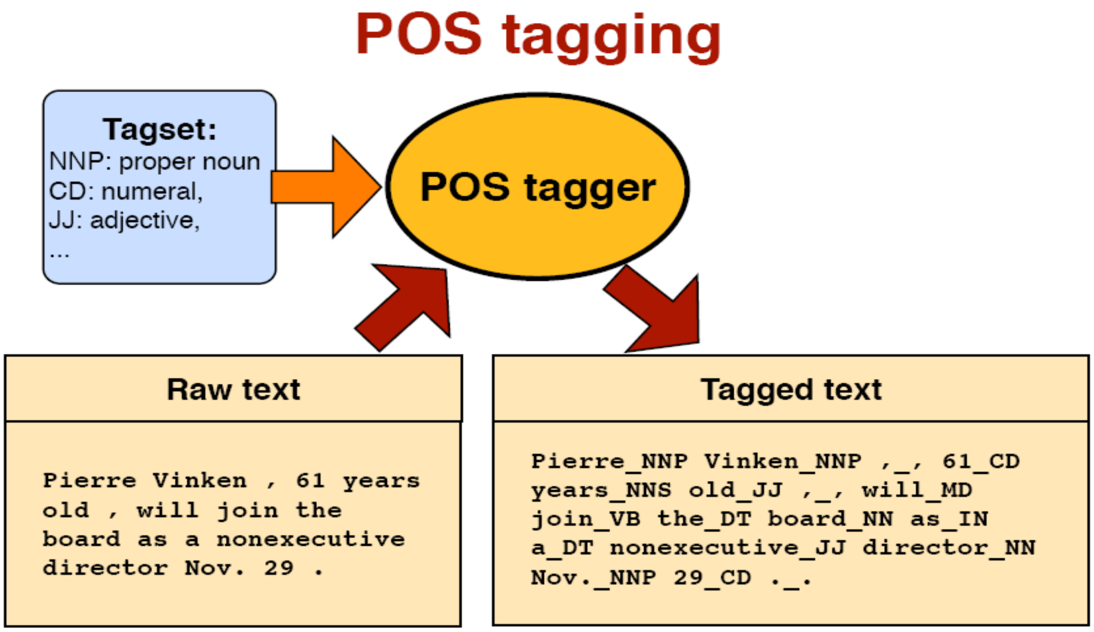
Part-of-Speech Tagging (POS)

Product features are usually nouns or noun phrases in review sentences. Thus the part-of-speech tagging is crucial. We used the NLP processor linguistic parser to parse each review to split text into sentences and to produce the part-of-speech tag for each word (whether the word is a noun, verb, adjective, etc). The process also identifies simple noun and verb groups. The following shows a sentence with POS tags.

```
<S> <NG><W C='PRP' L='SS' T='w' S='Y'> I </W> </NG>
<VG> <W C='VBP'> am </W><W C='RB'> absolutely
</W></VG> <W C='IN'> in </W> <NG> <W C='NN'> awe
</W> </NG> <W C='IN'> of </W> <NG> <W C='DT'> this
</W> <W C='NN'> camera </W></NG><W C='.'> .
</W></S>
```

NLP processor generates XML output. For instance, <W C=‘NN’> indicates a noun and <NG> indicates a noun group/noun phrase. Each sentence is saved in the review database along with the POS tag information of each word in the sentence. A transaction file is then created for the generation of frequent features in the next step. In this file, each line contains words from one sentence, which includes only the identified nouns and noun phrases of the sentence. Other components of the sentence are unlikely to be product features. Some pre-processing of words is also performed, which includes removal of stop words, stemming and fuzzy matching. Fuzzy matching is used to deal with word variants and misspellings.

- Stanford POS Tagger is used for tagging the text
    - http://nlp.stanford.edu/software/tagger.shtml
    - Class Maxent Tagger is used to for tagging the text file.
    - The output is in the form of WORD/TAG e.g. Nice/JJ – means "nice" is Adjective.
    - Once the tagged text is available, find out the two consecutive words which have the tags mentioned in the above table.
- Below diagram describes the POS tagging process.



```
Pierre Vinken , 61 years
old , will join the
board as a nonexecutive
director Nov. 29 .
```

```
Pierre_NNP Vinken_NNP ,_, 61_CD
years_NNS old_JJ ,_, will_MD
join_VB the_DT board_NN as_IN
a_DT nonexecutive_JJ director_NN
Nov._NNP 29_CD ._.
```

Figure 4.1 POS tagging

Step 2: Estimate the semantic orientation of extracted phrase

- Estimate the orientation of the extracted phrases using the Point wise Mutual Information (PMI) measure.
- PMI between 2 words, $word_1$ and $word_2$ can be defined as[24] :

$$\mathrm{PMI}(word_1,\ word_2) = \log_2\left[\frac{\mathrm{p}(word_1\ \&\ word_2)}{\mathrm{p}(word_1)\ \mathrm{p}(word_2)}\right] \longrightarrow \quad 4.1$$

Here,

1. P(word-1 & word-2)   = Probability that both words occurs together.
2. P(word-1)*P(word-2) = Probability of co-occurrence of word1 and word 2, If both words are independent.
3. $\frac{\mathrm{P}(\mathrm{word}\ -1\ \&\ word\ -2)}{\mathrm{P}(\mathrm{word}\ -1) * \mathrm{P}(\mathrm{word}\ -2)}$ = Degree of statistical dependence between words.
4. Log = Gives information of presence of one word when we observe other.

- The semantic orientation of a given phrase is calculated by comparing its similarity to a positive reference word (“excellent”) with its similarity to a negative reference word (“poor”).
- More specifically, a phrase is assigned a numerical rating by taking the mutual information between the given phrase and the word “excellent” and subtracting the mutual information between the given phrase and the word “poor”.
- In addition to determining the direction of the phrase’s semantic orientation (positive or negative, based on the sign of the rating), this numerical rating also indicates the strength of the semantic orientation (based on the magnitude of the number).
- The Semantic Orientation (SO) of a phrase is calculated as :

SO(phrase) = PMI(phrase, “excellent”) – PMI(phrase, “poor”)

when, SO is +ve : phrase is strongly associated with excellent.

SO is –ve : phrase is strongly associated with poor.

- The probabilities are calculated by issuing queries to a search engine and collecting the number of hits.

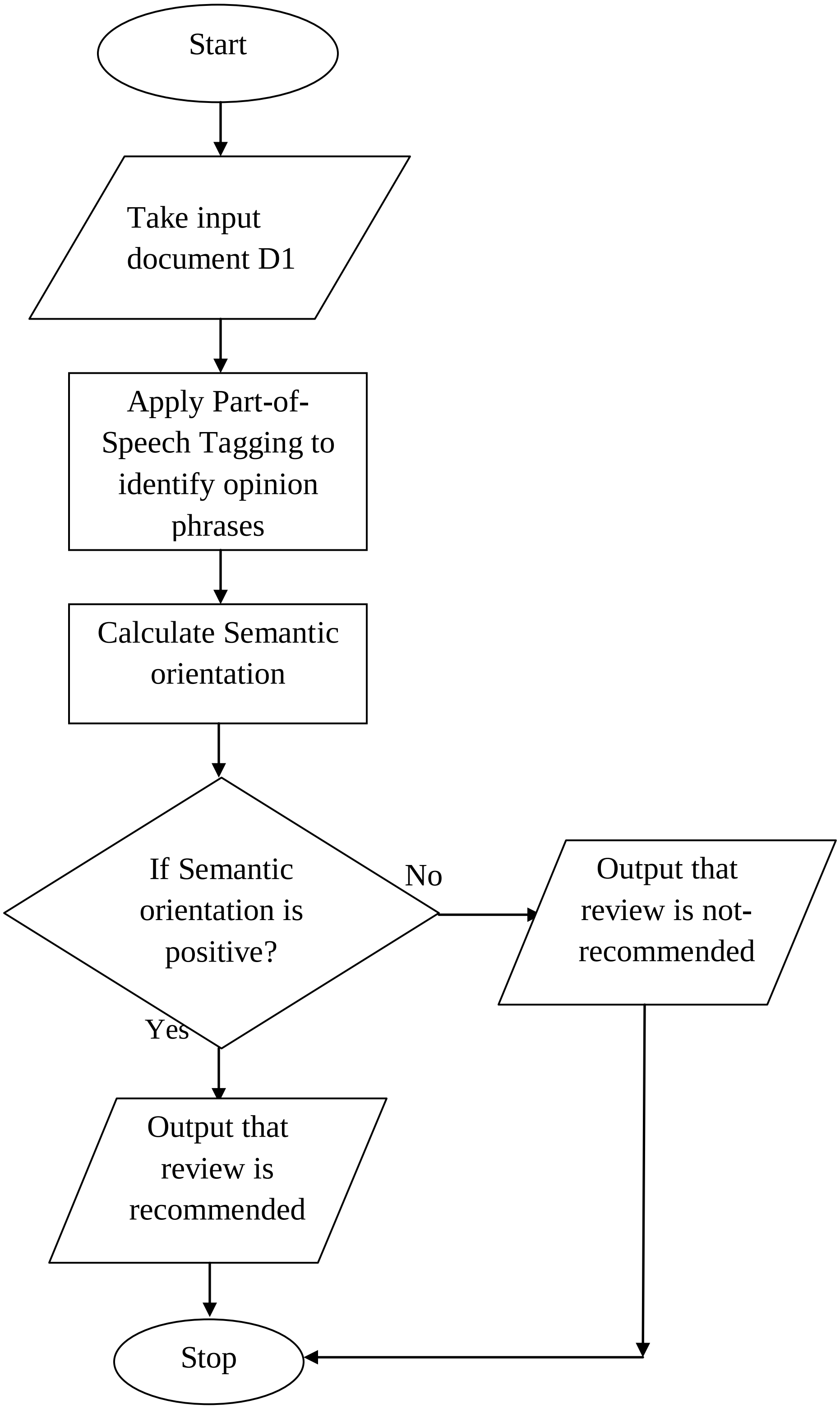
- **Proximity Search** –
    - Search the words such that they are located within ‘n’ words of one another.
    - YINDEX is search engine which allows Proximity search using “**/n**” or “**NEAR**” operator.
    - Query such as “Good Camera **/10** Excellent” will find out the occurrence of “Good Camera” with the word “Excellent”
    - Based on the number of this returned by the search query, the PMI can be calculated as[24] –

$$SO(\mathit{phrase}) = \log_2 \left(\frac{\text{hits}(\mathit{phrase}\ \text{NEAR “excellent”})\ \text{hits}(\text{“poor”})}{\text{hits}(\mathit{phrase}\ \text{NEAR “poor”})\ \text{hits}(\text{“excellent”})} \right) \rightarrow 4.2$$

Step 3: Assign the given review to a class

- Calculate the average Semantic Orientation (SO) of the phrases present in the review text.
- Classify them as recommended or not recommended.
- If the average SO is greater than zero then it is Recommended or Positive review.
- If average SO is less than zero then it is not Recommended or Negative review.

Figure 4.2:Flow-Chart of the whole process:



Start
Take input document D1
Apply Part-of-Speech Tagging to identify opinion phrases
Calculate Semantic orientation
If Semantic orientation is positive?
No
Output that review is not-recommended
Yes
Output that review is recommended
Stop

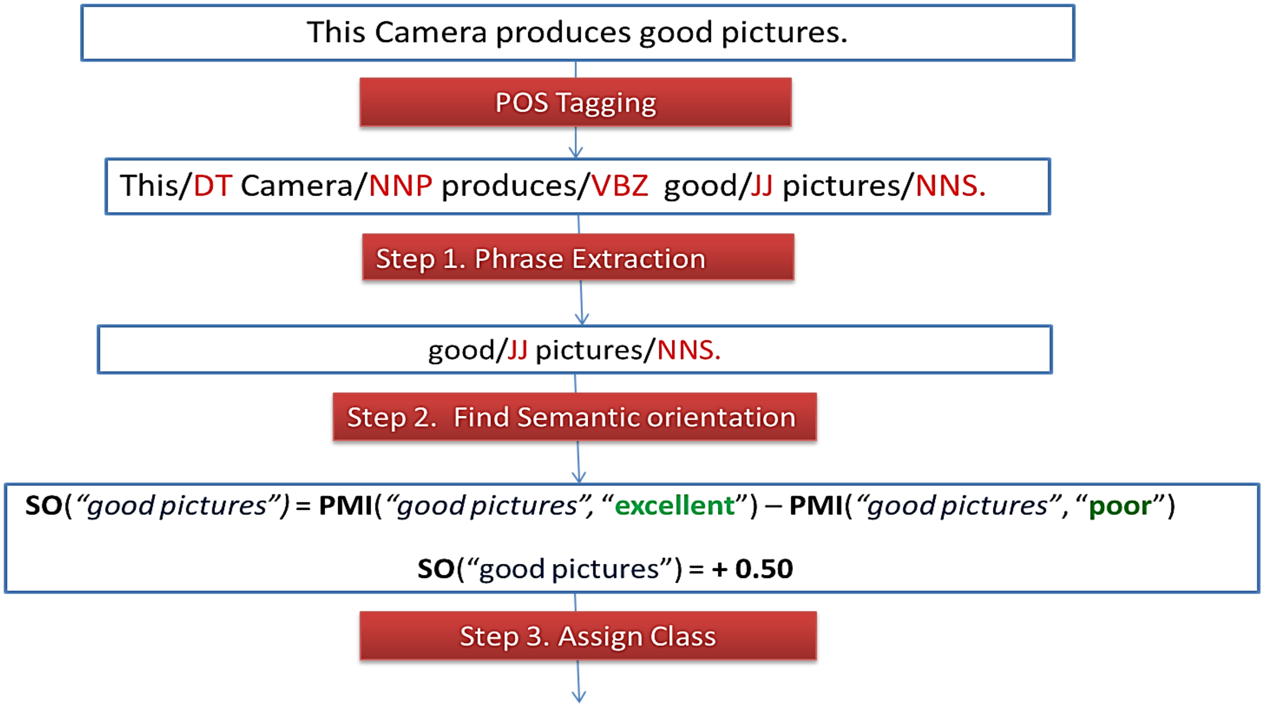
Example:

Let us take an example of digital camera. We will consider the following proposition.

Proposition: " This camera produces good pictures".

Steps:

- First part of speech tagging is applied on it. As a result of POS tagging we get noun that is camera, verb that is good, pictures as adjective.
- Pharse Extraction is done on it which results in some pharses that are likely to show opinions of the people that is good, pictures in our case. It results in some important pharse extraction from the whole sentence and we get good, pictures.
- Semantic orientation is calculated by the formula discussed above .
- Then we will classify it whether it is recommended or not. In our example as semantic orientation is possitive so it is classified into recommended class.


This Camera produces good pictures.
POS Tagging
This/DT Camera/NNP produces/VBZ good/JJ pictures/NNS.
Step 1. Phrase Extraction
good/JJ pictures/NNS.
Step 2. Find Semantic orientation
SO("good pictures") = PMI("good pictures", "excellent") – PMI("good pictures", "poor")
SO("good pictures") = + 0.50
Step 3. Assign Class

Extracted Phrases	Part of speech tag	Semantic orientation	Class
Good Pictures	JJ NN	2.253	Recommended

Figure 4.3. Calculate the average Semantic Orientation (SO)

Chapter 5

RESULT AND CONCLUSION

5.1 Results

Sample-input-1

First we give the document as an input and after that we can apply the Apply Part-of-Speech Tagging to identify opinion phrases and calculate the semantic if the semantic orientation is positive Output that review is recommended else not recommended.

Camera works great, takes nice photos even under dark conditions (using the flash).

It connects seamlessly to the web by wife or 3G, and to my computer using the Bluetooth radio.

I set up my Gmail, hotmail and yahoo accounts with no problem at all.

Nice screen to watch movies and get into Facebook.

Output

	Phrase	Semantic Orientation	Classification
Sentence:1	nice/JJ photos/NNS	0.7828604	(+) Positive phrase
Sentence:2	dark/JJ conditions/NNS	1.3126315	(+) Positive phrase
Sentence:3	Bluetooth/JJ radio/NN	-0.50447255	(-) Negative phrase
Sentence:4	Nice/JJ screen/NN	0.905206	(+) Positive phrase

Average Semantic Orientation = 0.624056339263916

(+) POSITIVE Review!!!

In our input set we are given the positive set of input and see the overall average semantic orientation is positive. output we are clearly see that the camera take nice photo and the semantic orientation is positive and the camera work good even in dark condition .we can clearly see that in result the Bluetooth of camera is not good so that the semantic orientation is positive and the screen is also good. The overall average semantic orientation is positive.

Sample-input-2

Battery life is long.

Camera not good, takes worst photos even using the flash.

It connects slowly to the web by wife or 3G, and to my computer using the Bluetooth radio.

It is hard to setup Gmail and yahoo accounts.

Nice screen to watch movies or get into Facebook.

Output-2

	Phrase	Semantic Orientation	Classification
Sentence:1	nice/JJ battery life/NNS	-0.7828604	(-) Positive phrase
Sentence:2	dark/JJ not good/NNS	-1.3126315	(-) Positive phrase
Sentence:3	Bluetooth/JJ radio/NN	-0.50447255	(-) Negative phrase
Sentence:4	Nice/JJ screen/NN	0.905206	(+) Positive phrase

Average Semantic Orientation = -0.624056339263916

(-) POSITIVE Review!!!

In the case of sample input-2 over all semantic orientation is negative. The battery life is not good and camera not good under dark condition. Bluetooth of the camera is not good but the screen of camera is good. But the overall semantic orientation is negative.

Sample-input-3

> Impressive 15.6 in laptop, offering well build quality, class-leading connectivity, decent video and audio, and plenty of power for the average user. It's fully upgradeable too, including such luxuries as a digital TV tuner and the unique (for a consumer laptop at this price point) option of a Full HD, RGB-LED backlit screen, which makes this an intriguing choice for color-critical work.

Output-3

	Phrase	**Semantic Orientation**	**Classification**
Sentence:1	nice/JJ lap top screen/NNS	0.6828604	(+) Positive phrase
Sentence:2	audio/JJ video/NNS	1.2126315	(+) Positive phrase
Sentence:3	Digital TV tuner/JJ full hd/NN	-0.50447255	(-) Negative phrase
Sentence:4	Rgb/JJ led/NN	0.905206	(+) Positive phrase

Average Semantic Orientation = 2.2962253500
(+) POSITIVE Review!!!

In the case of the same output -3 the overall semantic orientation is positive.

5.2 Conclusion

- Sentiment Mining is the important field of study. It can help the Business to grow if the Opinion of customer is understood properly.
- Review Analysis is based on Natural language processing.
- It is challenging to analyze the human language because it is ambiguous and context sensitive.
- In order to arrive at sensible conclusion, lots of things have to be analyzes such as context, negations and domain of analysis.

5.3 Future work

- **Problems with the previous approach**
  - Sentiment classification at both document and sentence levels are not sufficient,
  - They do not tell what features people like and/or dislike.
  - A positive opinion on an object does not mean that the opinion holder likes everything.
- **Modified Method** – is Feature-based Sentiment Analysis.
- **Feature-based Sentiment Analysis**
  - It provides the users opinion based on the features of the object.
  - We need to Identify object features that have been commented on.
  - Determine whether the opinions on the features are positive, negative or neutral.

Example:
"I bought an i-Phone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me.

Feature Based Summary:

Feature1: Touch screen

Positive

- The touch screen was really cool.
- The touch screen was so easy to use.

Negative

- The screen is easily scratched.
- I have a lot of difficulty from the touch screen.

REFERENCES

[1] Lect. Shital P. Bora, Sbora01@gmail.com., Department of Computer Science and Application, "DATA MINING AND WARE HOUSING", 978-1-4244-8679-3/11/$26.00 ©2011 IEEE

[2] Brijendra Singh1, drbri_singh@hotmail.com, 1Department of computer Science, University of Lucknow, LUCKNOW, INDIA. Hemant Kumar Singh , hemantbib@gmail.com, , Department of computer Applications, AzadIET, Lucknow, INDIA "WEB DATA MINING RESEARCH: A SURVEY", 978-1-4244-5967-4/10/$26.00 ©2010 IEEE

[3] kavita Sharma, kavitasharma_06@yahoo.co.in, Ambedkar Institute of Technology (G.G.S.I.P. University),New Delhi, India, Gulshan Shrivastava, gulshanstv@gmail.com, M.Tech. (Information Security) Ambedkar Institute of Technology(G.G.S.I.P. University), Vikas Kumar, getforvikas@yahoo.in, M.Tech. (Computer Engineering) PDM College of Engineering (M.D. University), Haryana, India," Web Mining: Today and Tomorrow", 978-1-4244-8679-3/11/$26.00 ©2011 IEEE

[4] Sanjay Kumar Malik1, sdmalik@hotmail.com, University School of Information Technology, GGS Indraprastha University, New Delhi, SAM Rizvi, samsam_rizvi@yahoo.com, Deptt. of Computer Science,Jamia Millia Islamia, New Delhi," Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation", 978-0-7695-4587-5/11 $26.00 © 2011 IEEE DOI 10.1109/CICN.2011.97

[5] Show-Jane Yen, Yue-Shi Lee and Min-Chi Hsieh, "An Efficient Incremental Algorithm for Mining Web Traversal Patterns", Proceedings of the 2005 IEEE International Conference on e-Business Engineering (ICEBE'05) 0-7695-2430-3/05 $20.00 © 2005 IEEE.

[6] Sotiris Kotsiantis, Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[7] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, “Mining Association Rules between Sets of Items in Large Databases” , Proceedings of the 1993 ACM SIGMOD Conference
Washington DC, USA, May 2003.

[8] Qingsong Yao, Aijun An and Xiangji Huang, “Mining and Modeling Database User Access Patterns”, FOUNDATIONS OF INTELLIGENT SYSTEMS, Lecture Notes in Computer Science, 2006, Volume 4203/2006, 493-503, DOI: 10.1007/11875604_56.

[9] Tong, R.M. 2001. An operational system for detecting and tracking opinions in on-line discussions. Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification (pp. 1-6). New York, NY: ACM.

[10] Green, N. Rege, M. , Xumin Liu ; Bailey, R., “ Evolutionary spectral co-clustering” , Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA, July 31 – August 5, 2011

[11] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira , “Analysis of representations for domain adaptation”. In Annual Conference on Neural Information Processing Systems 19, pages 137–144, Cambridge, MA, 2007. MIT Press.

[12] J. Blitzer. “Domain Adaptation of Natural Language”, Processing Systems. PhD thesis, The University of Pennsylvania, 2007.

[13] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: “Domain adaptation for sentiment classification?. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 432–439, Prague, Czech Republic, 2007.

[14] Bing Liu, liub@cs.uic.edu, Minqing Hu, mhu1@cs.uic.edu, Junsheng Cheng, Jcheng1@cs.uic.edu, Department of Computer Science , University of Illinois at Chicago,” Opinion Observer: Analyzing and Comparing Opinions on the Web”, WWW 2005, May 10-14, 2005, Chiba, Japan, ACM 1-59593-046-9/05/0005.

[15] Zhai, Y., and Liu, B. Web data extraction based on partial treealignment. WWW’05, 2005.

[16] Nitin Jindal and Bing Liu, Department of Computer Science University of Illinois at Chicago" Mining Comparative Sentences and Relations", Copyright © 2006, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

[17] YuFudan University School of Computer Science anbinWu, Qi Zhang, Xuanjing Huang, LideWu, "Structural Opinion Mining for Graph-based Sentiment Representation" Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1332–1341,

[18] Edinburgh, Scotland, UK, July 27–31, 2011. c2011 Association for Computational Linguistics Ramanathan Narayanan, Bing Liu, and Alok Choudhary. 2009. Sentiment analysis of conditional sentences. In Proceedings of EMNLP.

[19] Minqing Hu and Bing Liu, Department of Computer Science University of Illinois at Chicago, "Mining and Summarizing Customer Reviews", KDD'04, August 22–25, 2004, Seattle, Washington, USA. Copyright 2004 ACM 1-58113-888-1/04/0008.

[20] Bruce, R., and Wiebe, J. 2000. Recognizing Subjectivity: A Case Study of Manual Tagging. Natural Language Engineering.

[21] Bing Liu, Department of Computer Science University of Illinois at Chicago," Sentiment Analysis and Subjectivity", To appear in Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010

[22] G. Qiu, B. Liu, J. Bu and C. Chen. Expanding Domain Sentiment Lexicon through Double Propagation, International Joint Conference on Artificial Intelligence (IJCAI-09), 2009.

[23] S. Sarawagi, "Information extraction," to appear in Foundations and Trends in Information Retrieval, 2009.

[24] Peter D. Turney, "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), Philadelphia, July 2002, pp. 417-424

[25] Raymond Y.K. Lau and C.L. Lai, raylau, chunllai @cityu.edu.hk, "Leveraging the Web Context for Context-Sensitive Opinion Mining", 978-1-4244-4520-2/09/$25.00 ©2009 IEEE

[26] Nikolay Archak, Anindya Ghose, and Panagiotis G. Ipeirotis. Show me the money!: deriving the pricinpower of product features by mining consumer reviews In Pavel Berkhin, Rich Caruana, and Xindong Wu, editors, Proceedings of the 13th ACM SIGKD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007, pages 56–65. ACM, 2007.

[27] Khairullah Khan, Baharum B.Baharudin, Aurangzeb Khan, Fazal-e-Malik, Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia, "Mining Opinion from Text Documents: A Survey", 978-1-4244-2346-0/09/$25.00 ©2009 IEEE 217

[28] Changli Zhang et al., "Sentiment Classification for Chinese Reviews Using Machine Learning Methods based on String Kernel", International on Convergence and Hybrid Information Technology, 2008.

[29] Raymond Y.K. Lau and C.L. Lai Department of Information Systems City University of Hong Kon Tat Chee Avenue, Kowloon Hong Kong "Leveraging the Web Context for Context-Sensitive Opinion Mining", 978-1-4244-4520-2/09/$25.00 ©2009 IEEE

[30] R.Y.K. Lau, D. Song, Y. Li, C.H. Cheung, and J.X. Hao. Towards A Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning. IEEE Transactions on Knowledge and Data Engineering, 21(6):800–813, 2009.

[31] R.Y.K. Lau, D. Song, Y. Li, C.H. Cheung, and J.X. Hao. Towards A Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):800–813, 2009.