OPINION MINING: CLASSIFICATION, TECHNIQUES, CHALLENGES

Divya Rajgor[1], Prof. Mehul Barot[2]
[1]Research Scholar, Computer Department, LDRP-ITR, Gandhinagar
Kadi Sarva Vishvavidhyalaya
[2]Assistant Prof., Computer Engineering Department, LDRP-ITR, Gandhinagar

Abstract
Opinion mining or sentiment analysis analyses the text written in a natural language about a topic and classify them as positive, negative or neutral based on the human's sentiments, emotion, opinions expressed in it. Nowadays user reviews on blogs, forums, review websites, E-commerce sites are useful for other users to make a decision in planning. The manual analysis of such huge number of reviews is practically impossible. To solve this problem an automated approach of a machine to mine the overall sentiment or opinion polarity form the reviews is needed. Opinion mining applied on data from review sites. We are going to mention the opinion mining techniques, applications, challenges, and classification.
Keyword: opinion mining, sentiment classification, Architecture, machine learning, sentiment analysis, Opinion

I. Introduction

Opinion mining is a type of natural language processing for tracking the mood of the public about a particular product. "What other people think" has always been an important piece of information for most of us during the decision-making process. Opinion mining, also called Sentiment analysis, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g. sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc.[12]

Why opinion mining??

Online opinions have indirect influence on the business of several e-commerce sites. Those sites market their products and the web users go through the reviews of the item before buying that product. Many organizations utilize opinion mining systems to track client audits of products sold online. Opinion mining is an incredible way of maintaining focus on several business trends related to deals administration, status management and also advertising. Pattern prediction is also done using the opinion of the customers.[20]

Types of opinion

1) Direct Opinions: Sentiment expressions on some objects, e.g., products, events, topics, persons.[12] E.g., "The picture quality of this camera is great"

2) Comparisons Opinions:

Relations expressing similarities or differences of more than one object. Usually expressing an ordering.[12] E.g., "Car x is cheaper than car y"

Architecture of Opinion Mining

Opinion mining also called sentiment analysis is a process of finding user's opinion towards a topic or a product. Opinion mining concludes whether user's view is positive, negative, or neutral about product, topic, and event etc. opinion mining and summarization process involve three main steps, first is opinion retrieval, opinion classification and opinion summarization. Review Text is retrieved from review websites. Opinion text in blog, reviews, comments etc. contains subjective information

about topic. Reviews classified as positive or negative review.[5]

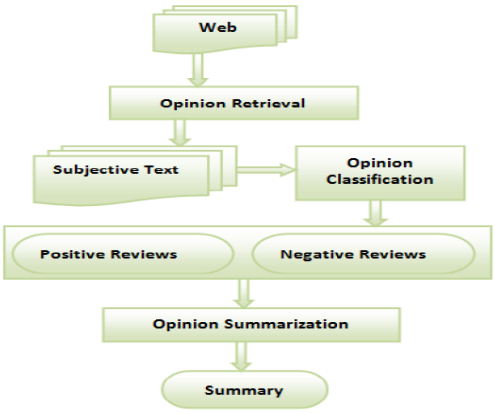

Fig 1: Architecture of opinion mining

Opinion summary is generated based on features opinion sentences by considering frequent features about a topic[5].

II. OPINION MINING CLASSIFICATION

At the document level:

The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment. Overall opinion polarity of the document is calculated and classified as positive or negative. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly known as document-level sentiment classification. This level of analysis Assume that each document (or review) focuses on a single object and contains opinion from a single opinion holder. Thus, it is not applicable to documents which evaluate or compare multiple entities.[11]

At the sentence level:

The task at this level goes to the sentence and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to subjectivity classification, which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions. However, we should note that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions, e.g., "We bought the car last month and the windshield wiper has fallen off." Researchers have also analyzed clauses, but the clause level is still not enough, e.g., "Apple is doing very well in this lousy economy."

This level of analysis assumes that each sentence contains only one opinion [11].

At the Aspect / feature level:

Both the document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level was earlier called feature level. Instead of looking at language constructs such as documents, paragraphs, sentences, clauses or phrases, aspect level directly looks at the opinion itself. It is based on the idea that an opinion

consists of a sentiment (positive or negative) and a target (of opinion). An opinion without its target being identified is of limited use. Realizing the importance of opinion targets also helps us understand the sentiment analysis problem better. For example, although the sentence "although the service is not that great, I still love this restaurant" clearly has a positive tone, we cannot say that this sentence is entirely positive. In fact, the sentence is positive about the restaurant, but negative about its service. In many applications, opinion targets are described by entities and/or their different aspects. Thus, the goal of this level of analysis is to discover sentiments on entities and/or their aspects. For example, the sentence "The iPhone's call quality is good, but its battery life is short" evaluates two aspects, call quality and battery life, of iPhone (entity). The sentiment on iPhone's call quality is positive, but the sentiment on its battery life is negative. The call quality and battery life of iPhone are the opinion targets. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses[11].

Example:

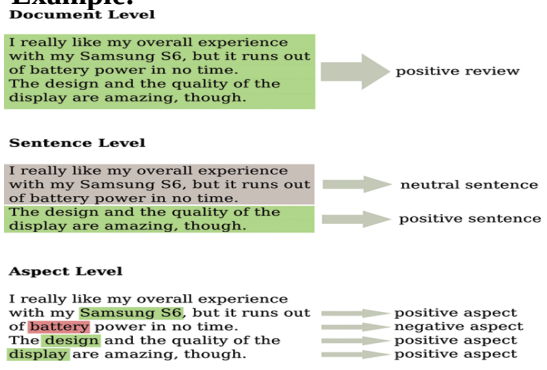

Fig2: Classification of Opinion mining[18]

IV. OPINION MINING TECHNIQUES

Opinion Classification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach. The Machine Learning Approach applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. The text classification methods using ML approach can be roughly divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of labeled training documents.The unsupervised methods are used when it is difficult to find these labeled training documents. The lexicon-based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach. The dictionary-based approach which depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms. The corpus-based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods.[9]

Machine learning approach

Machine learning approach relies on the famous Machine Learning algorithms to solve the opinion mining as a regular text classification problem that makes use of syntactic and/or linguistic features. Text Classification Problem Definition: We have a set of training records D = {X1, X2…. Xn} where each record is labeled to a class. The classification model is related to the features in the underlying record to one of the class labels. Then for a given instance of unknown class, the model is used to predict a class label for it. The hard classification problem is when only one label is assigned to an instance. The soft classification problem is when a probabilistic value of labels is assigned to an instance.[9]

Supervised learning

The supervised learning methods depend on the existence of labeled training documents. In the next subsections, we present in brief details some of the most frequently used classifiers in SA.[9]

Probabilistic classifiers

Probabilistic classifiers use mixture models for classification. The mixture model assumes that each class is a component of the mixture. Each mixture component is a generative model that provides the probability of sampling a particular term for that component. These kinds of classifiers are also called generative classifiers. Three of the most famous probabilistic classifiers are discussed in the next subsections. [9]

Naive Bayes Classifier (NB)

The Naive Bayes classifier is the simplest and most commonly used classifier. Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the bags of words feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)}$$

P(label) is the prior probability of a label or the likelihood that a random feature set the label. P(features|label) is the prior probability that a given feature set is being classified as a label. [9] P(features) is the prior probability that a given feature set is occurred. Given the Naive assumption which states that all features are independent, the equation

Could be rewritten as follows:

$$P(label|features) = \frac{P(label) * P(f1|label) * \ldots.. * P(fn|label)}{P(features)}$$

Bayesian Network (BN)

The main assumption of the NB classifier is the independence of the features. The other extreme assumption is to assume that all the features are fully dependent. This leads to the Bayesian Network model which is a directed acyclic graph whose nodes represent random variables, and edges represent conditional dependencies. BN is considered a complete model for the variables and their relationships. Therefore, a complete joint probability distribution (JPD) over all the variables, is specified for a model. In Text mining, the computation complexity of BN is very expensive, that is why, it is not frequently used. [9]

Maximum Entropy Classifier (ME)

The Maxent Classifier known as a conditional exponential classifier converts Labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely label for a feature set. This classifier is parameterized by a set of X{weights}, which is used to combine the joint features that are generated from a feature-set by an X{encoding}. In particular, the encoding maps each C{(featureset, label)} pair to a vector. The probability of each label is then computed using the following equation:

$$P(fs|label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l))forlinlabels)}$$

Linear classifiers

Given X = {X1, X2, . . ., Xn} is the normalized document word frequency, vector A{a1….an} is a vector of linear coefficients with the same dimensionality as the feature space, and b is a scalar; the output of the linear predictor is defined as p =A.X + b, which is the output of the linear classifier. The predictor p is a separating hyperplane between different classes. There are many kinds of linear classifiers; among them is Support Vector Machines (SVM), which is a form of classifiers that attempt to determine good linear separators between different classes. Two of the most famous linear classifiers are discussed in the following subsections. [9]

Support Vector Machines Classifiers (SVM)

The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes. In Fig. 5 there are 2 classes x, o and there are 3 hyperplanes A, B and C. Hyperplane A provides the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation. Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories . SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly to an inner product space where the classes can be separated linearly with a hyperplane. SVMs are used in many applications, among these applications are classifying reviews according to their quality. [9]

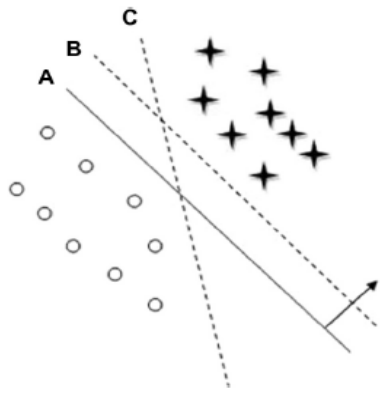

Fig 3: Using Support vector machine on classification problem

Neural Network (NN)

Neural Network consists of many neurons where the neuron is its basic unit. The inputs to the neurons are denoted by the vector overline Xi which is the word frequencies in the *i*th document. There are a set of weights classification problem, it is assumed that the class label of Xi is denoted by yi and the sign of the predicted function pi yields the class label. Multilayer neural networks are used for non-linear boundaries. These multiple layers are used to induce multiple piece wise linear boundaries, which are used to approximate enclosed regions belonging to a particular class. The outputs of the neurons in the earlier layers feed into the neurons in the later layers. The training process is more complex because the errors need to be back-propagated over different layers. [9]

A which are associated with each neuron used in order to compute a function of its inputs f(●). The linear function of the neural network is: pi = A · Xi. In a binary

Decision tree classifiers

Decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data. The condition or predicate is the presence or absence of one or more words. The division of the data space is done recursively until the leaf nodes contain certain minimum numbers of records which are used for the purpose of classification. There are other kinds of predicates which depend on the similarity of documents to correlate sets of terms which may be used to further partitioning of documents. The different kinds of splits are Single Attribute split which use the presence or absence of particular words or phrases at a particular node in the tree in order to perform the split. Similarity-based

multi-attribute split uses documents or frequent words clusters and the similarity of the documents to these words clusters in order to perform the split. Discriminate-based multi-attribute split uses discriminates such as the Fisher discriminate for performing the split. [9]

Rule-based classifiers

In rule based classifiers, the data space is modeled with a set of rules. The left hand side represents a condition on the feature set expressed in disjunctive normal form while the right hand side is the class label. The conditions are on the term presence. Term absence is rarely used because it is not informative in sparse data. There are numbers of criteria in order to generate rules, the training phase construct all the rules depending on these criteria. The most two common criteria are support and confidence. The support is the absolute number of instances in the training data set which are relevant to the rule. The Confidence refers to the conditional probability that the right hand side of the rule is satisfied if the left-hand side is satisfied. Some combined rule algorithms were proposed in. Both decision trees and decision rules tend to encode rules on the feature space, but the decision tree tends to achieve this goal with a hierarchical approach. [9]

Unsupervised learning

The main purpose of text classification is to classify documents into a certain number of predefined categories. In order to accomplish that, large number of labeled training documents are used for supervised learning, as illustrated before. In text classification, it is sometimes difficult to create these labeled training documents, but it is easy to collect the unlabeled documents. The unsupervised learning methods overcome these difficulties. [9]

Lexicon-based approach

Opinion words are employed in many sentiment classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are also opinion phrases and idioms which together are called opinion lexicon. There are three main approaches in order to compile or collect the opinion word list. manual approach is very time consuming and it is not used alone. It is usually combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods. The two automated approaches are presented in the following subsections. [9]

Dictionary-based approach

Presented the main strategy of the dictionary-based approach. A small set of opinion words is collected manually with known orientations. Then, this set is grown by searching in the well known corpora WordNet or thesaurus for their synonyms and antonyms. The newly found words are added to the seed list then the next iteration starts. The iterative process stops when no new words are found. After the process is completed, manual inspection can be carried out to remove or correct errors. The dictionary based approach has a major disadvantage which is the inability to find opinion words with domain and context specific orientations. [9]

Corpus-based approach

The Corpus-based approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus. [9]

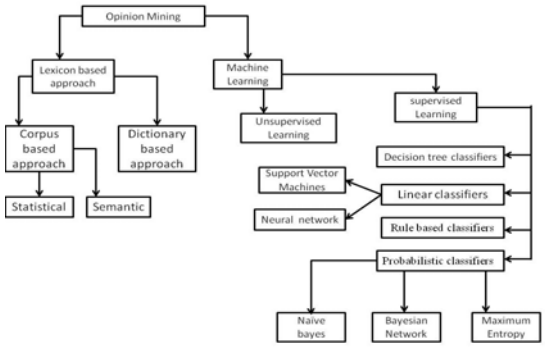

Fig 4: Opinion Mining Techniques[9]

Statistical approach

Finding co-occurrence patterns or seed opinion words can be done using statistical techniques. This could be done by deriving posterior polarities using the co-occurrence of adjectives in a corpus. It is possible to use the entire set of indexed documents on the web as the corpus for the dictionary construction. This overcomes the problem of the unavailability of some words if the used corpus is not large enough. The polarity of a word can be identified by studying the occurrence frequency of the word in a large annotated corpus of texts . If the word occurs more frequently among positive texts, then its polarity is positive. If it occurs more frequently among negative texts, then its polarity is

negative. If it has equal frequencies, then it is a neutral word. The similar opinion words frequently appear together in a corpus. This is the main observation that the state of the art methods are based on. Therefore, if two words appear together frequently within the same context, they are likely to have the same polarity. Therefore, the polarity of an unknown word can be determined by calculating the relative frequency of co-occurrence with another word. This could be done using Point wise Mutual Information. Statistical methods are used in many applications related to Opinion mining. One of them is detecting the reviews manipulation by conducting a statistical test of randomness called Runs test. Latent Semantic Analysis (LSA) is a statistical approach which is used to analyze the relationships between a set of documents and the terms mentioned in these documents in order to produce a set of meaningful patterns related to the documents and terms . [9]

Semantic approach

The Semantic approach gives sentiment values directly and relies on different principles for computing the similarity between words. This principle gives similar sentiment values to semantically close words. WordNet for example provides different kinds of semantic relationships between words used to calculate sentiment polarities. WordNet could be used too for obtaining a list of sentiment words by iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word. The Semantic approach is used in many applications to build a lexicon model for the description of verbs, nouns and adjectives to be used in opinion mining. [9]

V. CHALLENGES

1) **Detection of spam and fake reviews:** The web contains both authentic and spam contents. For effective Sentiment classification, this spam content should be eliminated before processing. This can be done by identifying duplicates, by detecting outliers and by considering reputation of reviewer.[20]

2) **Limitation of classification filtering:** There is a limitation in classification filtering while determining most popular thought or concept. For better sentiment classification result this limitation should be reduced. The risk of filter bubble gives irrelevant opinion sets and it results false summarization of sentiment.[20]

3) **Asymmetry in availability of opinion mining software:** The opinion mining software is very expensive and currently affordable only to big organizations and government. It is beyond the common citizens expectation. This should be available to all people, so that everyone gets benefit from it.[20]

4) **Incorporation of opinion with implicit and behavior data:** For successful analysis of sentiment, the opinion words should integrate with implicit data. The implicit data determine the actual behavior of sentiment words.[20]

5) **Domain-independence:** The biggest challenge faced by opinion mining and sentiment analysis is the domain dependent nature of sentiment words. One features set may give very good performance in one domain, at the same time it perform very poor in some other domain.[20]

6) **Natural language processing overheads:** The natural language overhead like ambiguity, co-reference, Implicitness, inference etc. created hindrance in sentiment analysis too.[20]

VI. APPLICATION DOMAIN

Here we are discussing about few opinion mining application domain.

Shopping (E-Commerce)

Online Shopping is important for people as it deliver package to their door steps because it is easy to get an overall idea about any product by the number of feedbacks mentioned by customers. In this kind of site users go through the feedback before they buy any product. Shopping sites Flipkart, Amazon helps to compare products with all desired description and feedback of customers, information is displayed to the user in a "Graphical User Interface" for easy understanding about the quality/features and services of product. This is what really helps users to get their selection to product want to buy. [18]

Entertainment

For Movie or show or TV programs can easily go through public feedback about any current release or any kind of popular program or movie. Internet movie database (IMDB) which helps the users online to view feedback for movies and other programs. It also helps users to understand the movie or program who are not sure it. [18]

Business

Companies are now asking for feedbacks on their portal which makes savings on marketing budget. Now companies can be able to analysis the data they from the given feedback online in their website. Recommendation to friends and family about products or services to each other which helps to gain an clear picture about the products or services before buy it. [18]

R&D (Research and Development)

In any online shopping portal items feedback are used by industry to improve the quality of the products. Online sites can also ask the customers to provide their own design views to make modification or create new products for customers. For example "Play Store" in which it shows all the applications users has given all sort of feedback/point of view.[18]

Politics

Political people can extract the views on public feedback. During Elections the participated candidates can have an overall knowledge about public opinion which helps to gain an idea about the progress (weakness and strength). [18]

Academics

In any online course, Student's opinion can be used to manage their academic. Student's feedback may help the institute to understand and analysis and make better improvement for student's future benefits. Institution will be aware of their benefits and problems through their opinion in the form on feedback and comments. [18]

Health

User's shares treatment which helps to take precaution while they suffer from headache, fever, high blood pressure, Diabetes and many more which is also called "Home Remedies". Users also share the exercise steps for body building, blood pressure, back pain, migraine etc which helps all to get rid of problems in their life. While doing so users also gives feedback and comment from which is it understandable that particular remedies is working or not to others. [18]

Transportation

Transportation can be any movable items like animals, people, and goods from one place to another as per the requirement. There are many transport modes for example road, air, rail, water, cable, space and pipeline. Transport is necessary to trade between persons, which is essential for the development of civilizations. Transportation like tour and travel, logistic, cab service which is very helpful to people to choose best services mentioned in the feedback and comments from past experience shared by the users. For example from Tour and Travel that is holiday plans it helpful to get an idea about the location and hotels and services provided to them.[18]

CONCLUSION

Opinion miming is an emerging field of data mining used to extract the knowledge from huge volume of customer comments, feedback and reviews on any product or topic etc. A lot of work has been conducted to mine opinions in form of document, sentence and feature level sentiment analysis. There are many drawbacks and difficulties exist like grammar mismatch, writing technique, regional language, fake detection feedbacks etc. This report presents the Various Opinion Mining techniques and applications. It helps users as well as business to improve their Product.

REFERENCES

1. G. Sneka "Algorithms for Opinion Mining and Sentiment Analysis: An Overview" IJARCSSE Volume 6, Issue 2, February 2016
2. Kazi Mostafizur Rahman and Aditya Khamparia"Techniques, Applications and Challenges of Opinion Mining" International Journal of Computer Technology and Applications, 2016
3. Chinsha T C "A Syntactic Approach for Aspect Based Opinion Mining" IEEE ICSC 2015
4. Cristian Bucur "Using Opinion Mining Techniques in Tourism" ScienceDirect 2015
5. P.Kalarani1 "An Overview on Research Challenges in Opinion Mining and Sentiment Analysis" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 10, October 2015
6. Chinsha T C "Aspect based Opinion Mining from Restaurant Reviews " *CACCTHPA-2014*
7. Vijay B. Raut "Opinion Mining and Summarization of Hotel Reviews" ICCICN 2014

8. Heng Ji "SENTIMENT, OPINIONS, EMOTIONS" jih@rpi.edu October 22, 2014
9. Walaa Medhat "Sentiment analysis algorithms and applications: A survey" Ain Shams Engineering Journal (2014) 5, 1093–1113
10. Eivind Bjokelund "A Study of Opinion Mining and Visualization of Hotel Reviews" ACM 2012
11. Bing Liu "Sentiment Analysis and Opinion Mining" ,April 22, 2012
12. Sudeshna Sarkar "Opinion Mining" 24th and 26th October 2007
13. TechTarget "opinion mining (sentiment mining)", http://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining
14. http://www.wideskills.com/data-mining/challenges-in-data-mining
15. Kavita Ganesan, Hyun Duk Kim, "Opinion Mining-A Short Tutorial".
16. Bing Liu "Opinion Mining and Sentiment Analysis: NLP Meets Social Sciences" Department of Computer Science University Of Illinois at Chicago http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
17. INTECH@ A Proposal for Brand Analysis with Opinion Mining
18. Kazi Mostafizur Rahman and Aditya Khamparia "Techniques, Applications and Challenges of Opinion Mining" IJCTA 2016
19. Haseena Rahmath P "Opinion Mining and Sentiment Analysis - Challenges and Applications" International Journal of Application or Innovation in Engineering & Management (IJAIEM) May 2014
20. Veena Dubey and Dharmendra Lal Gupta "Sentiment Analysis Based on Opinion Classification Techniques: A Survey " International Journal of Advanced Research in Computer Science and Software Engineering , June 2016