Features Extraction and Opinion Mining in Product Reviews

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: E-commerce websites often request the customers to write reviews, which helps the manufacturers to improve the quality of their products and other customers in choosing the right product or service. Each of these reviews may describe the different features of the products. Hence, the customer has to go through a large number of reviews before they can arrive to a fully informed decision on whether to buy the product or not. The manual analysis of such huge number of reviews is practically impossible. To solve this problem and for improving accuracy we need an automated approach which automatically extracts the product features from the reviews and determines if they have been expressed in a positive, negative or neutral way by the reviewers.

Keywords: **opinion mining, polarity classification, Reviews, 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 is the field of study that analyzes people's opinions, sentiments, 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. 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.

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 [14].

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, neutral or negative opinion about the product. This task is commonly known as document-level sentiment classification. This level of analysis assumes 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 [17].

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 assumes that each sentence contains only one opinion [17].

At the Aspect level:

Both the document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level also called feature level. It directly looks into the opinion instead of paragraph, sentences, phrases, and document. The goal is to find polarity on entities and their aspects. The first step is to find the feature and after that classifying whether the review is positive or negative for that feature [17].

III. Related Works

In [2] the proposed algorithm works in two steps, feature extraction and polarity classification. They use association rule mining to identify the most characteristic features of a product. In the second step we develop a supervised machine learning algorithm based polarity classifier that determines the sentiment of the review sentences with respect to the prominent features. They experiments on the benchmark reviews of five popular products show that our classifier is highly efficient.

In [3] paper an aspect based opinion mining system is proposed to classify the reviews as positive, negative and neutral for each feature. Negation is also handled in the proposed system. Experimental results using reviews of products show the effectiveness of the system. They used Amazon website (www.amazon.com) to collect the reviews.In [1] is focus on Techniques, Applications and Challenges of Opinion Mining.

In [4] is based on phrase-level to examine customer reviews. Phrase-level opinion mining is also well-known as aspect based opinion mining. It is used to extract most important aspects of an item and to predict the orientation of each aspect from the item reviews. The projected system implements aspect extraction using frequent itemset mining in customer product reviews and mining opinions whether it is positive or negative opinion. It identifies sentiment orientation of each aspect by supervised learning algorithms in customer reviews.

In [6] paper they are extracting reviews from different ecommerce sites and storing the reviews in MongoDB, one of the NoSQL database. From these review sentences, product features are extracted. The proposed method uses Apriori algorithm for feature extraction. The classification is done on product features based on unsupervised SentiWordNet method. In this method they are taking Adjective, Adverb, Verb, Noun as opinion words and negation rules are used for classification of reviews into positive and negative. They concluded proposed method gives 84% accuracy compared to general SentiWordNet method. The feature summarized reviews helps customers to analyze interesting features on products.

IV. PROPOSED SYSTEM

In this section, provide the outline of approach to solving the problem of product feature extraction and polarity classification for product reviews. In this proposed methodology, initially Mobile (Samsung S5) reviews are collected from e-commerce site (Amazon.com).The collected reviews are stored in the text file in the form of documents and these documents are given as input for pre-processing stages. Pre-processing includes Punctuation Removal, Tokenization, Stop word Removal and POS tagging. Aspect extraction step will give aspects from reviews. Opinion orientation is used to identify whether it is positive, negative or neutral opinion sentence based on aspects.

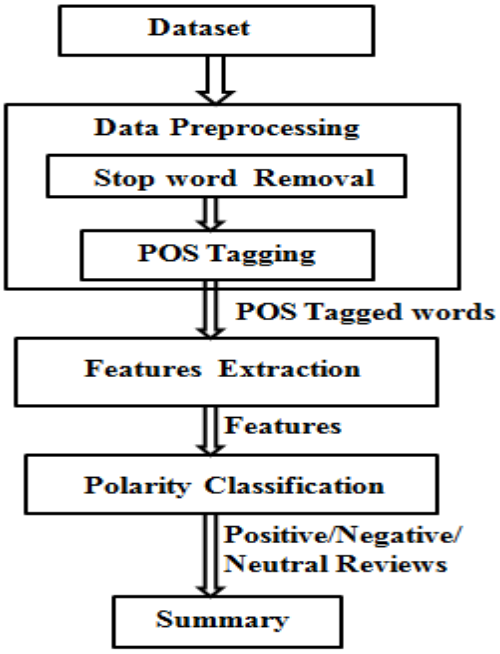

Figure 1: Proposed Architecture of features extraction and polarity classification

The steps of the whole process for Aspect Extraction and Aspect Orientation are described below.

1. Take benchmark dataset of product reviews as input and perform pre-processing.
2. Pre-processing includes Punctuation Removal, Stop word Removal and POS tagging.
3. (POS tagging) Part of speech tagging of all sentences, tag as /NN, /JJ, /VB, /RB for a noun, adjective, verb & adverb.
4. The noun, noun phrases, adjectives, verb and adverb along with their word position are captured in the sentence.
5. The product feature's opinion word extracted from key Adjective phrases is prepared.
6. Find Aspect Orientation as positive or negative using polarity classification algorithm.

1. Data Preprocessing

Data preprocessing performed on a benchmark dataset. A review sentence is given as input to data preprocessing.

- ***Stop Word Removal***

Most frequently used words in English are not useful in text mining. Such words are called stop words. Stop words are language specific functional words which carry no information. It may be of types such as pronouns, prepositions, conjunctions. Stop word removal is used to remove unwanted words in each review sentence. Words like is, are, was etc. Reviews are stored in a text file which is given as input to stop word removal. Stopwords are collected and stored in a text file. Stopwords are removed by checking against Stopwords list.

- ***POS Tagging***

The Part-Of-Speech of a word is a linguistic category that is defined by its syntactic or morphological behavior. Common POS categories in English grammar are: noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. POS tagging is the task of labeling (or tagging) each word in a sentence with its appropriate part of speech. POS tagging is an important phase of opinion mining, it is essential to determine the features and opinion words from the reviews. POS tagger is used to tag all the words of reviews.

2. Aspect Extraction

The sentences that contain feature and opinion words are nothing but opinion sentence. In feature extraction, we are going to extract the features of the product. First extracts noun and noun phrases in each review sentence and store it in a file. Aspects like camera, battery, display, storage, etc. Then Extracts adjectives from reviews and stored it.

3. Polarity Classification

The proposed methodology first determines the number of positive, negative and neutral opinion sentence in reviews using opinion words. Examples of positive opinion words are long, excellent and good and the negative opinion words are like poor, bad etc. and the next step is to identify the number of positive, negative and neutral opinions of each extracted aspect.

3.1 Polarity Classification Algorithms

Machine learning classifiers are implemented to classify polarity. Multiclass classifiers such as Support Vector Machine (SVM) classifier, Decision Tree, k-Nearest Neighbors, Naïve Bayes classifier can be implemented to classify the polarity of reviews.

3.1.1 Support Vector Machine (SVM)

The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes. In the figure2 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 the 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 [21].

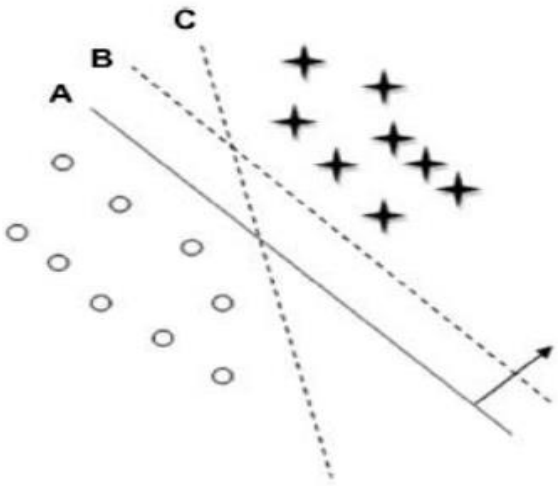

Figure 2: Support Vector Machine

3.1.2 Naïve Bayesian classifier

The Naïve Bayesian classifier works as follows: Suppose that there exist a set of training data, *D*, in which each tuple is represented by an *n*-dimensional feature vector, $X = x_1, x_2, x_3, \ldots, x_n$ indicating *n* measurements made on the tuple from *n* attributes or features. Assume that there are *m* classes $C_1, C_2, \ldots, C_m$. Given a tuple *X*, the classifier will predict that *X belongs* to C_i if and only if: $P(C_i|X) > P(C_j|X)$, where $i, j \in [1, m]$ *and* $i\_ = j$.$P(C_i|X)$ is computed as:[26]

$$P(C_i/X) = \prod_{k=1}^{n} P(x_k/C_i) \qquad \textbf{(3.1)}$$

Naïve Bayes algorithm is easy to implement and requires a small amount of training data to estimate the parameters. It obtained Good results in most of the cases but it has dependencies exist among variables [26].

3.1.3 Logistic classifier

Logistic classifier has a place with the group of classifiers known as the exponential or log-linear classifiers. Like innocent Bayes, it log-linear classifier works by extricating some set of weighted components from the information, taking logs, and joining them linearly (implying that

every element is increased by a weight and afterward included). In fact, logistic classifier alludes to a classifier that characterizes a perception into one of two classes, and multinomial logistic classifier is utilized when arranging for more than two classes. While logistic classifier in this way varies in the way it calculates probabilities, it is still similar to naive Bayes in being a linear classifier. Logistic classifier estimates P(y|x) by separating some set of elements from the input, consolidating them linearly (increasing every element by a weight and

adding them up), and afterward applying a combination function. Logistic classifier has Mean Square error high and doesn't perform well when feature space is too large [5].

V. RESULTS

WEKA, an open source tool is a collection of machine learning algorithms. In the WEKA tool, initial the data set will be loaded. Using Weka tool compares the results of Support vector machine (SVM), Naïve Bayes classification, logistic classifier, K-Nearest Neighbor, and Decision Table.

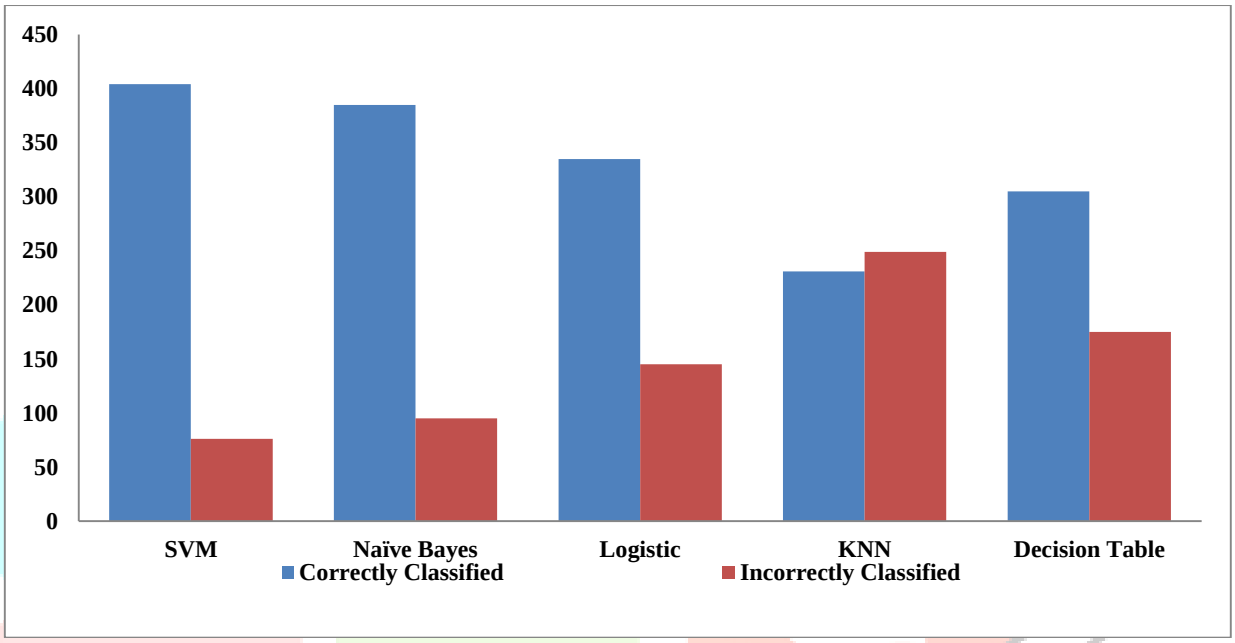

Figure 3: Graph of correctly and incorrectly classified instances by ML classifiers

It can be seen from the figure 3 support vector machine classify a good number of correctly classified instances than other classification algorithms.

Information retrieval measure: This field having different measures like precision, recall, F-measure, accuracy we compare them and analysis their results based on the graph which are shown as below:

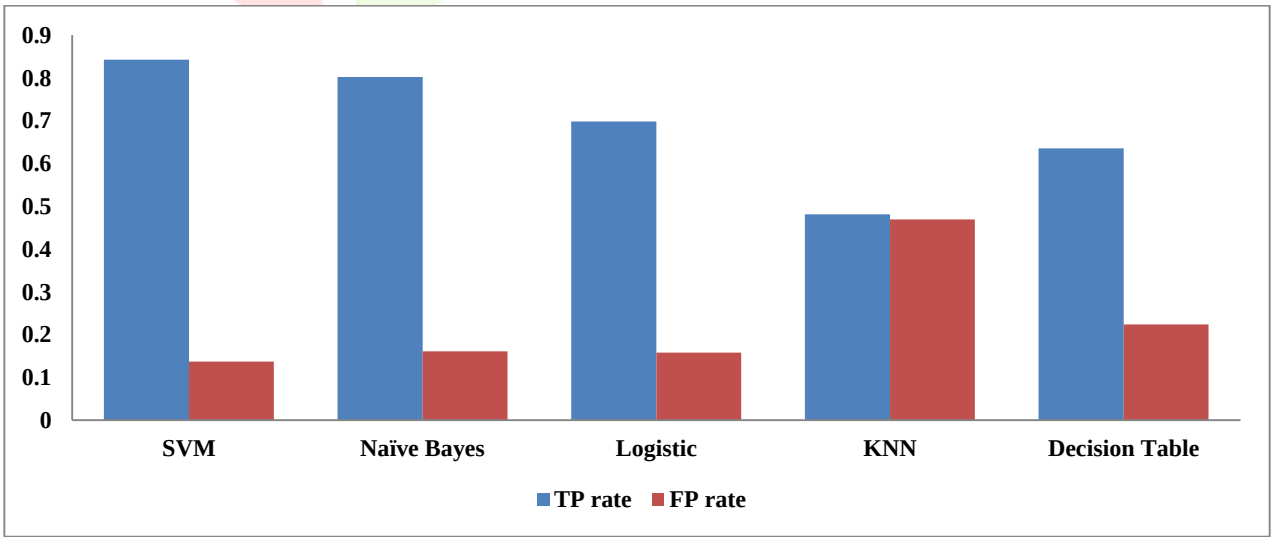

Figure 4: Graph of TP rate & FP rate ML classifiers

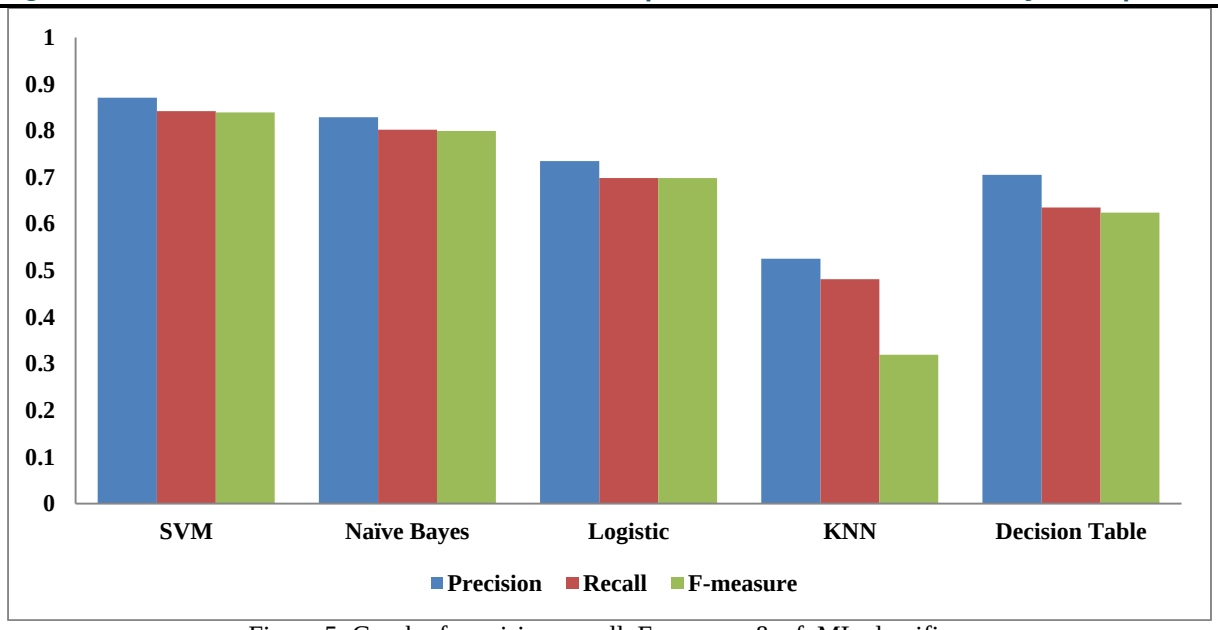

Figure 5: Graph of precision, recall, F measure & of ML classifiers

It can be seen from the figure, the classification precision, recall and F-measure obtained through support vector machine is better than Naïve Bayes and other classifiers.

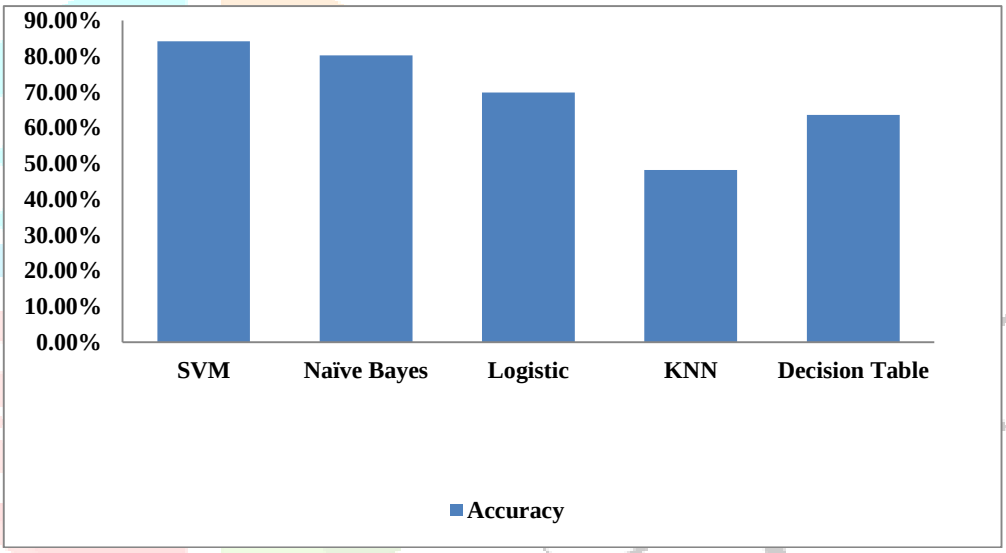

Figure 6: Graph of accuracy & of ML classifiers

It can be seen from the figure, SVM gives the highest number of accuracy.

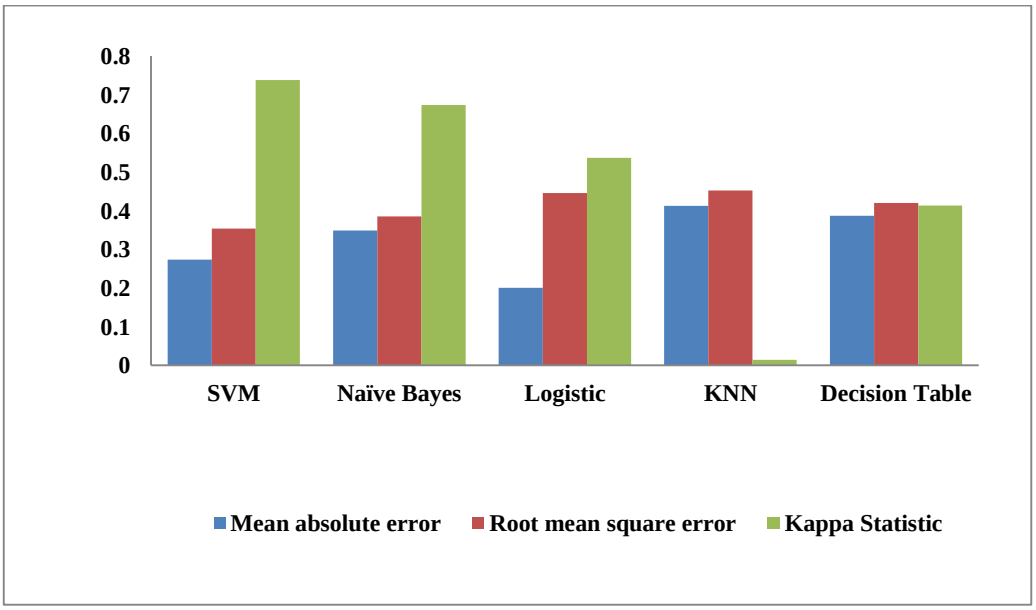

Figure 5: Graph of Kappa Statistic, Mean absolute error & Root mean squared error of ML classifiers

VI. CONCLUSION

Opinion mining is an emerging field of data mining used to extract the knowledge from the huge volume of customer comments, feedback, and reviews on any product or topic etc. Aspect based opinion mining is necessary because nowadays everyone is busy and they don't have time to read all the positive or negative reviews if someone just wants to know about some features of the product. Opinion mining can help users to identify what is good or bad features in a product based on other users reviews and also help manufacture companies to improve their product qualities. Support vector machine gives a good accurate result. In this proposed method analysis is done only on mobile product reviews. In our future work would like to do analysis of different product reviews.

VII. REFERENCES

1. Kazi Mostafizur Rahman and Aditya Khamparia 2016 "Techniques, Applications and Challenges of Opinion Mining" International Journal of Computer Technology and Applications.
2. Siddharth Aravindan, Asif Ekbal 2014 "Feature Extraction and Opinion Mining in Online Product Reviews"13th International Conference on Information Technology IEEE
3. Richa Sharma, Shweta Nigam and Rekha Jain 2014 "MINING OF PRODUCT REVIEWS AT ASPECT LEVEL" International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.3, May
4. A.Jeyapriya , C.S.Kanimozhi Selvi 2015 "Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm" IEEE SPONSORED 2ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEMS(ICECS)
5. Santhosh Kumar K L, Jayanti Desai , Jharna Majumdar "Opinion Mining and Sentiment Analysis on Online Customer Review" International Conference on Computational Intelligence and Computing Research IEEE 2016
6. Anisha P Rodrigues , Dr. Niranjan N Chiplunkar 2016 "Mining Online Product Reviews and Extracting Product features using Unsupervised method" IEEE
7. Sudeshna Sarkar "Opinion Mining" 24th and 26th October 2007
8. TechTarget "opinion mining (sentiment mining)", http://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining
9. INTECH@ A Proposal for Brand Analysis with Opinion Mining
10. Jiawei Han, Micheline Kamber "Data Mining : Concepts and Techniques"
11. https://www.slideshare.net/shitalkr/opinion-mining-62938614
12. http://bitcoincryptocurrency.xyz/bitcoin-for-beginners-ebook-review/
13. Tanvir Ahmad, Mohammad Najmud Doja "Opinion Mining using Frequent Pattern Growth Method from Unstructured Text" International
14. Veena Dubey and Dharmendra Lal Gupta 2016 "Sentiment Analysis Based on Opinion Classification Techniques: A Survey " International Journal of Advanced Research in Computer Science and Software Engineering ,
15. Aman Amani Khalaf Samha 2016 "ASPECT-BASED OPINION MINING FROM CUSTOMER REVIEWS"
16. Kazi Mostafizur Rahman and Aditya Khamparia "Techniques, Applications and Challenges of Opinion Mining" IJCTA 2016
17. Bing Liu "Sentiment Analysis and Opinion Mining" ,April 22, 2012
18. http://www.wideskills.com/data-mining/challenges-in-data-mining
19. http://bitcoincryptocurrency.xyz/bitcoin-for-beginners-ebook-review/
20. Walaa Medhat 2014 "Sentiment analysis algorithms and applications: A survey" Ain Shams Engineering Journal 1093–1113
21. Sonal Meenu "Aspect based Opinion Mining for Mobile Phones" IEEE 2016
22. Vijayshri Ramkrishna Ingale 2016 "Product Feature-based Ratings for Opinion Summarization of E-Commerce Feedback Comments" IJCA 2016
23. Santhosh Kumar K "Opinion Mining and Sentiment Analysis on Online Customer Review"
24. Anju Joshi "Aspect Level Opinion Mining on Customer Reviews using Support Vector Machine" IJARCCE 2017
25. Xing Fan 2015 "Sentiment analysis using product review data" SPRINGER
26. Symposium on Computational and Business Intelligence 2013
27. https://en.wikipedia.org/wiki/Weka