# Feature Extraction Techniques and their Role in Sentiment Analysis: A Survey

Binita Verma
B.U Bhopal

Ramjeevan Singh Thakur
MANIT, Bhopal

Shailesh Jaloree
S.AT.I Vidisha

**Abstract.** Web is a huge repository of facts and opinions available for people in the world about a product. Social networking sites on internet have become an essential part for everyone, which grows at rapidly. Consumers give their reviews about product which are useful for others. Sentiment analysis is used to study the people's orientation for making better decision which leads to advancement of business. Analysis of sentiment of an entity is defined in term of positive, negative or neutral. In this paper we compare the existing feature extraction techniques which are used in sentiment analysis and also discussed challenges, applications of sentiment analysis.

***Keywords***: *Sentiment analysis, Feature extraction techniques, Social media.*

## I. INTRODUCTION

Human beings are social. They need to connect with others and expand their connections. Social network means an individual can connected with other persons, share their ideas, values, trade, anything. With the rapid increase of social media sites [1] like Facebook, Orkut, Myspace, Google plus, Twitter, Instagram, etc. There is a huge amount of data present online day by day. peoples are now relying on online product reviews sites for exchange their personal experience and knowledge [2][3]. This kind of information gives a clear picture about the opinion of end users. The opinion may be positive, negative or neutral in nature. Positive thoughts have a positive effect on society and negative thoughts create negative effects. Social media have a large volume of blogs, posts, and reviews etc. [4][5]. Fig.1 shows the process of opinion mining and sentiment analysis.

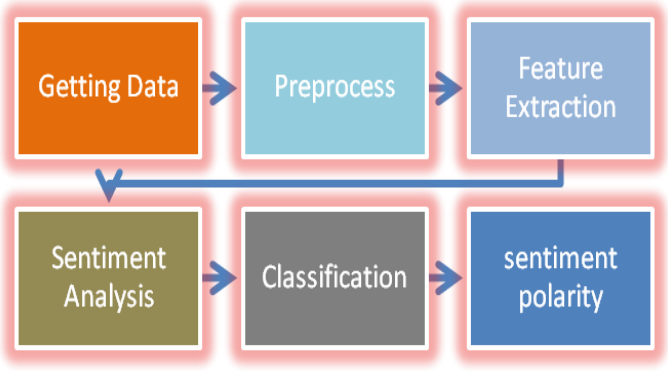


Fig. 1 Process of Opining Mining and Sentiment Analysis

## II. RELATED WORK

There is a lot of work has been done in this field.

Zhang et al. [6] worked Chinese sentiment analysis, proposed a rule based approach including two phases. L. Qu et al. [7] proposed a regression method which is based on bag of opinion model, for review rating prediction from sparse text pattern. Pang et al. [8] proposed machine learning techniques which gives results better than human generated baselines. Barbosa et al. [9] designed two phase automatic sentiment analysis method for classifying tweets, like tweet are positive negative or neutral. Agarwal et al. [10] also proposed a model for classifying sentiments into +ve –ve or neutral. Tree kernel model perform better than unigram model and feature based model. Pak and paroubek [11] developed a model to classify the tweets as objective, negative and positive. They created a twitter corpus by using twitter API and automatically annoting those tweets using emotions. They like multinomial naïve bayes method and pos tags. The training set they used was less efficient as it contain tweets having emoticons only. Parikh and Movassate [12] implemented two models maximum entropy model and naïve bayes bigram model to classify tweets and they found that naive bayes was worked better than maximum entropy model. Hatzivassiloglou and Wiebe [13] studying the effect of dynamic semantically oriented and gradable adjectives to predict subjectivity. Zhuang et al. [14] worked on sentiment summarization for movie reviews. Aspects are extracted and form a cluster manually with labeled dataset. Supervised techniques are expensive for labeled data. Therefore unsupervised or semi-supervised techniques are used to address this issue.

## III. SENTIMENT ANALYSIS

Sentiment analysis is to determine the attitude of a writer or speaker for a given topic [15], can also be applied to audio video and images [16].

*Preprocessing of Web Data*

Raw data is highly susceptible to inconsistency and redundancy. Preprocessing of dataset is used to clean the noisy text, by following steps.

i. Tokenization
ii. Sentence Parsing
iii. Stopword Removal
iv. Stemming and Lemmatization

v. Spell Correction

***Challenges in Sentiment Analysis***

Sentiment analysis is dealing with various issues such as [17],

i. Binary classification: review's polarity is classified as positive negative by ignoring neutral. This type of problem arises when sentiment classification is based on machine learning algorithm. Opining mining that consider only +ve , -ve will not have good accuracy. Nowadays the classification is considered in 5 possibilities: strong positive, positive, strong negative, negative and neutral. It improves the accuracy of opinion mining.

ii. Polarity shift: polarity shift means sentiment of the sentence is calculated in different way from the polarity actually expressed in the sentence for eg. "I don't like this bike". On the other hand "good but it's not my style".

iii. Data Sparsity : the problem is caused due to the character limit in twitter or blogs. Due to this people will not express their opinion clearly.

***Applications of Sentiment Analysis***

Sentiment analysis has many applications in various areas:

i. Support in decision making: customer can use sentiment analysis in order to purchase a product [18].

ii. For business: business people spend a lot of money to gather public opinion through survey, consultants.

**iii.** Predictions and trend analysis: sentiment analysis enables one to predict market trends by tracking public views. It is also helpful in elections to know the expectations of the people.

## IVFEATURE EXTRACTION TECHNIQUES USED IN SENTIMENT ANALYSIS

### *A. Weighing And Aggregation Scheme*

In this technique , V.K Singh et al. used [19] SentiWordNet based scheme for both aspect level sentiment classification and document level sentiment classification.

In Document-level sentiment classification, it works on linguistics and scores. There are two linguistics feature selection schemes. First, we only extract adjectives and any adverbs preceding the selected adjective. Next, we extract both adjectives and verbs along with any adverbs preceding them. An algorithm of adverb+adjective combination is represented in Fig 2.

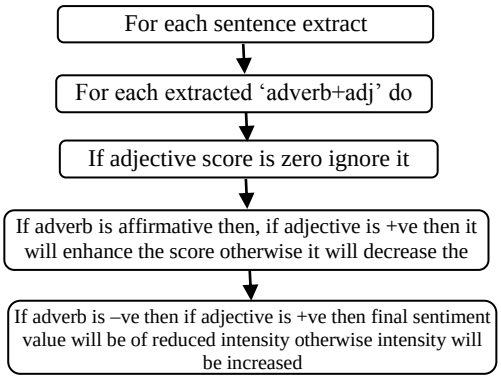


Fig. 2 Algorithm of 'Adverb+Adjective' combination

In aspect-level sentiment classification checks aspects in the review and finds opinion for it. After identification of an aspect, adjectives or adverb+adjective combines are checked for 5-gram backward or forward based on their occurrence. Then sentiment polarity for these terms is computed using the SentiWordNet.

### *B. Neutral /Polar /Irrelevant Classification Model*

This model [20] worked on Twitter micro-blogging tweets. Firstly tweets are preprocessed then classify as +ve , -ve and irrelevant based on their emotional content. Fig 3 present the classification process; Twitter API is used to manually collect the training dataset. In preprocessing it extracts the training dataset from the tweets, then preprocessor transform collected data into feature vectors. The performance of the classifier based on preprocessing techniques used. In early stages, we remove the neutral and irrelevant data results it increase the accuracy of positive and negative classification.

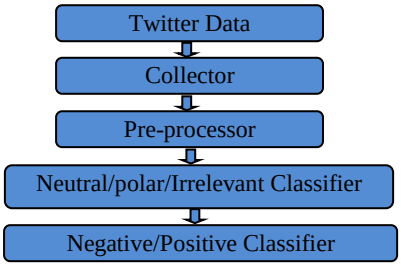


Fig. 3 The Classification Process [17]

The sentiment analysis of the preprocessed tweet data has been carried by different classifiers like naive bayes, SVM, Random forest, J48 and SMO.

### *C. Total Weighted Score Computing Method*

This method presents to predict the semantic orientation of reviews. It integrates grammatical knowledge and takes topic correlations into account. In this method, features are extracted and similarity between these features and the topic is computed [21], then the final score is

computed by assign different weights to the polarity scores of different adjectives. Fig. 4 shows the structure of semantic orientation of predicting system. It has eight main steps.

1. Firstly, Stanford parser is used to parse movie reviews.
2. In this step, use grammatical knowledge by extracting adjectives and features from reviews with their grammatical relationship. The modified features and adjectives are found.
3. All adjectives are divided into five groups based on WordNet.
4. Compute similarity between the features and the topic is done on the basis of algorithm.

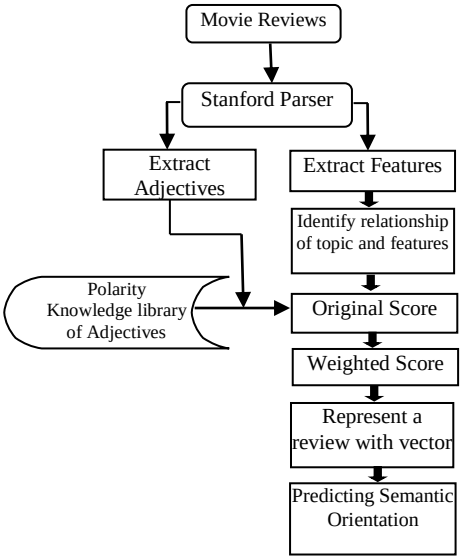


Fig. 4 The structure of semantic orientation predicting system [21]

5. Based on polarity library the value of original score of every noun is compute and similarity between features and topic.
6. Calculate weighted score
7. In this step predict the semantic orientation of movie review.
8. In the last step generate polarity labels for review.

### D. *Aspect Extraction method*

S. Poria et al. [22] proposed a method for extracting both explicit and implicit aspects from opinionated text. This framework leverages on common sense knowledge and on the dependency structure of sentence. The method is fully unsupervised and depends on the accuracy of the dependency parser and opinion lexicon. Aspect can be expressed indirectly through implicit aspect clue (IAC). Fig. 5 shows the process of aspect extraction algorithm. In explicit aspect extraction algorithm, the corpus used by Hu and liu, 2004 [23] and semeval 2014 dataset. For implicit aspect extraction algorithm they use the corpus developed by Cruz-Garcia et al. [24], 2014 , manually labeled each IAC and their corresponding aspects for opinion mining. Preprocessing is a key step for aspect parsing. It has two steps firstly sentence is obtained through Stanford dependency parser and secondly elements are processed by means of Stanford lemmatizer for each sentence. Feature extraction can be done by aspect parser. Implicit aspect lexicon, opinion lexicon and some other rules are also applied.

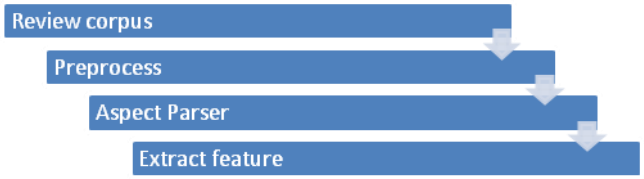


Fig. 5 The process of aspect extraction algorithm

### E. *Intrinsic And Extrinsic Domain Relevance Approach*

By using syntactic rules, Zhen hai et al [25] extracted candidate features from review corpus. Opinion features are domain specific and at the same time not overly generic means domain independent via the intercorpus statistics IEDR criteria are identifies as shown in fig 6.

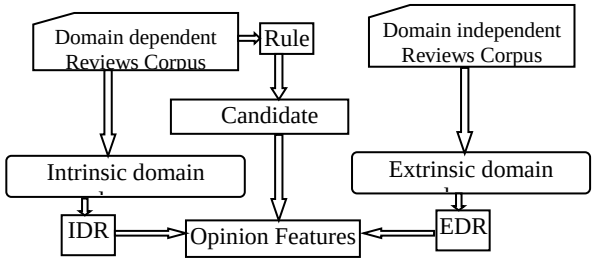


Fig 6. IEDR Workflow [25]

In candidate feature extraction process, firstly sentences in review corpus are analyzed by dependency parsing, then nouns are obtained and checked if it had subject verb dependency relationship or verb object dependency relationship or predict object dependency relationship.

Then intrinsic domain relevance and extrinsic domain relevance are calculated. IDR is domain specific candidate feature and EDR is a domain independent candidate feature. Candidate feature with intrinsic domain relevance score greater than intrinsic threshold and extrinsic domain relevance score less than another threshold are extracted.

### F. *Multi class Bootstrapping algorithm*

Chunliang Zhang and Jingbo Zhu [26] learn Aspect related terms to address the issue of aspect identification in the reviews with several predefined aspects. First apply a single class bootstrapping to learn the ARTs for each aspect. There are many ARTs learned by SCB method may co-occur in multiple ART sets, they proposed to modify each ART's

importance value by considering its ambiguity degree. So this algorithm determines a sentences most relevant aspect by comparing the overall scores of the total aspect-related terms for each aspect. The multi-class bootstrapping learning process is shown in fig 7.

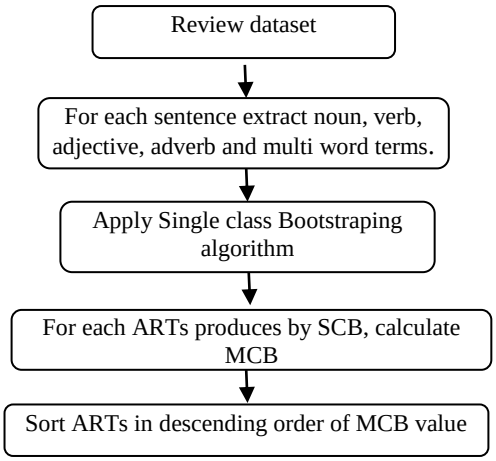


Fig 7 Multi-Class Bootstraping Learning

G. *Feature based opinion summarization:*

Minqing Hu and Bing Liu [23] proposed a set of technique for mining and summarizing product reviews based on data mining and natural language processing method, it provide a feature based summary of large number of customer reviews sold by online product. The task is performed in three steps first; identify features of the product that customers have expressed their opinion. Second identify opinion sentences in each review and deciding whether each opinion sentence is positive or negative third, summarizing the results.

In fig. 8, Given the input, system download all the reviews and put in database. Then find the features that many people have expressed their opinions on. After that, the opinion words are extracted using the resulting frequent features and semantic orientation of the opinion words are identified with the help of WordNet. By using the extracted opinion words,

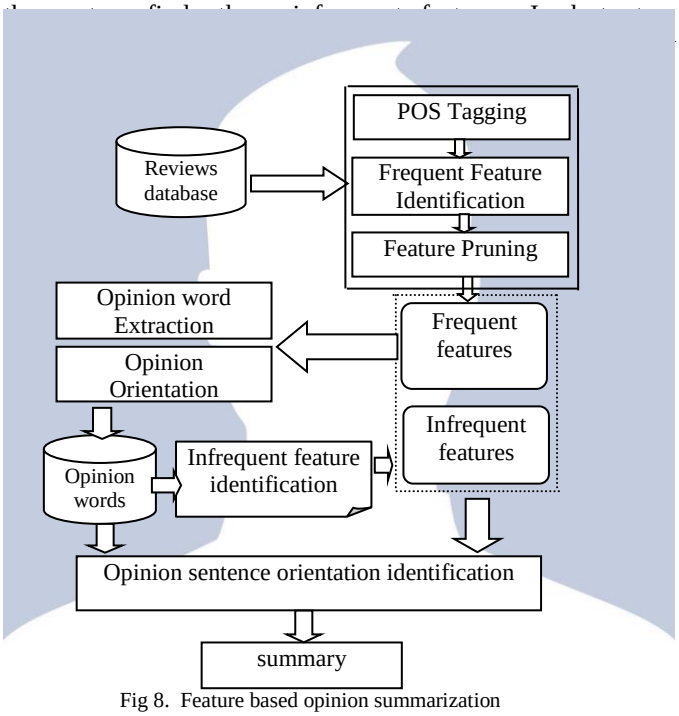


Fig 8. Feature based opinion summarization

Table 1. Comparison of feature extraction techniques used in Sentiment Analysis [27]

| **Method** → | **Feature based opinion summarization [23]** | **Total weighted score computing method [21]** | **Neutral/polar/irrelevant classifier method [20]** | **Aspect Extraction method [22]** |
|---|---|---|---|---|
| **Paper** | Mining and summarizing customer review, 2004 | Predicting The Semantic Orientation Of Movie Reviews, 2010 | Opinion Mining and Sentiment Analysis On A Twitter Data Stream, 2012 | A rule based approach to aspect extraction from product review, 2014 |
| **Class** | +ve , -ve | +ve , -ve | Neutral , polar , irrelevant | +ve , -ve |
| **Dataset** | Amazon.com, c\|net.com | Movie | Movie | Semeval 2014 |
| **Used lexicon** | WordNet | WordNet | None | Senticnet 3 |
| **Type** | Sentence level | Sentence level | Sentence level | Aspect level |
| **Feature extraction** | Adjective | Noun, adjectives | Classifier used | Parser with aspect rules |
| **Parser used** | POS tagger | Stanford parser | NA | Stanford dependency parser |
| **Use** | NA | Identify semantic polarity of review | Analyze performance of various classifying algorithms | Aspect extraction from product review |
| **Prons** | Domain independent | Automatically identifies and extracts opinions | Accuracy is improved | Obtain higher detection accuracy |
| **Cons** | Feature can be explicit or implicit in a sentence | Adjective are not used | High skewness was present in training set | NA |

**Table 1. (Continued…)**

| Method | Intrinsic/extrinsic domain relevance approach [25] | Weighing and aggregation scheme [19] | Multi-class Bootstrapping algorithm [26] |
|---|---|---|---|
| **Paper** | Identifying features in Opinion mining via Intrinsic and Extrinsic domain relevance, 2014 | Sentiment Analysis of Movie Reviews, 2013 | Multi class bootstrapping learning aspect related terms for aspect identification, 2009 |
| **Class** | +ve , -ve | +ve , -ve | +ve , -ve |
| **Dataset** | Cell phone, hotel | Twitter | Restaurant reviews |
| **Used lexicon** | Likelihood ratio test based semantic association method | SentiWordNet | None |
| **Type** | Document level | Aspect level | Aspect level |
| **Feature extraction** | Dependency parser along with rules | Adverb+ adjective combination ,Adverb+ adjective+verb combination | Noun, verb adjective, adverb |
| **Parser used** | Language Technology Platform | POS Tagger | POS tagger |
| **Use** | Feature extraction based on the IEDR feature filtering criterion | Sentiment profile creation with summary | Aspect identification |
| **Pros** | Feature extraction is done using domain independent corpus, F measure and accuracy improved | Document level sentiment classification produce accurate result | Needs no labeled data |
| **Cons** | Less successful dealing with extraction of infrequent features, not extract non-noun opinion features | Aspect level sentiment classification is restricted to the domain | Not identify multiple aspects contained in one sentence |

## IV. CONCLUSION

Sentiment analysis extracts people's opinion in an automatic manner. It counts four tasks: identify opinion, feature extraction sentiment classification and finally result. Feature Based opinion summarization work with customer reviews of 5 product sold online and the technique are highly effective. Total weighted score computing method is a simple technique to use. Rule Based approach exploits common sense knowledge and sentence dependency trees to detect both explicit and implicit aspects and obtain higher detection accuracy. Weighing and aggregation scheme works on aspects and produces sentiment profile. Multi-Class Bootstrapping algorithm works on aspect related terms need no labeled data, achieve good performance in comparison of state-of-the-art machine learning techniques. Neutral/polar/irrelevant analyses tweets to determine their polarity. Intrinsic extrinsic domain relevance approach is an inter-corpus statistics approach to opinion feature extraction based on the IEDR feature filtering criterion. It produces better result to compare with others as it is not domain specific. This method shows feature extraction performance improvement as compared to other methods in sentiment analysis.

## V. REFERENCES


[1] Pushpendra Kumar and Ramjeevan Singh Thakur, "Recommendation system techniques and related issues: a survey", International Journal of Information Technology, Vol.10 (4), pp. 495–501, 2018.

[2] Pushpendra Kumar and R. S. Thakur, "A Framework for Weblog Data Analysis Using HIVE in Hadoop Framework", In: Proceedings of International Conference on Recent Advancement on Computer and Communication, Lecture Notes in Networks andSystems 34,(2018), https://doi.org/10.1007/978-981-10-8198-9_45

[3] Th. Belt, L. Engelen, S. Berben, L. Schoonhoven, "Definition Of Health 20 and Medicine 20: A Systematic Review", J Med Int Res, Vol.12, issue 2, Pp 6-8, 2010.

[4] Vinod Kumar, Pushpendra Kumar and R.S. Thakur, "A brief Investigation on Data Security Tools. and Techniques for Big Data", International Journal of Engineering Science Invention, Vol. 6(9), PP. 20-27, 2017.

[5] Vishal A. Kharde, S. S. Sonawane, "Sentiment Analysis Of Twitter Data: A Survey of Techniques", International Journal of Computer Applications, Vol.139, No.11, Pp.975-8887, April 2016

[6] C. Zhang, D. Zeng, J. Li, F. Y. Wang, and W.Zuo, "Sentiment Analysis Of Chinese Documents: From Sentence To

Document Level", J. Am. Soc. Information Science and Technology, Vol.60, No. 12, Pp. 2474-2487, 2009.

[7] L. Qu, G, Ifrim and G. Weikum, "The Bag of Opinion method for Review rating Prediction from sparse Text Patterns", Proceeding 23rd International Conference Computational Linguistics, Pp. 913-921, 2010.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? : Sentiment Classification using Machine Learning Techniques", Proceeding Conference Empirical Methods in Natural Language Processing, Pp. 79-86, 2002.

[9] L. Barbosa and J. Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data", Coling 2010: Poster Volume, Pp. 36-44, Beijing, 2010.

[10] Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 workshop on Languages in Social Media, Pp. 30-38, 2011.

[11] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", In Proceedings of 7th Conference on International Language Resources and Evaluation, Pp.1320-1326, 2010.

[12] R. Parikh and M. Movassate, "Sentiment Analysis of User-Generated twitter Updates using various Classification Techniques", CS224N Final Report, 2009.

[13] V. Hatzivassiloglou and J.M. Wiebe, "Effects of Adjective orientation and Gradability on sentence Subjectivity", Proceeding 18th Conf. Computational Linguistics, Pp.299-305, 2000.

[14] L. Zhuang, F. Jing and X. Zhu, "Movie review mining and summarization", In proceedings of the 15th ACM international conference on Information and Knowledge management, 2006.

[15] B. Pang and L. Lee, "Opining mining and Sentiment Analysis", Found Trends Inform Retrieve, Pp. 1-35, 2008

[16] Walaa Medhat, Ahmed Hassan, and Hoda Korashy, "Sentiment Analysis algorithm and applications: A survey", Ain Shams Engineering Journal, Vol. 5, Issue 4, Pp.1093-1113, 2008.

[17] Saurabh Dorle and Nitin N. Pise, "Sentiment Analysis Methods and Approach: A Survey", International Journal of Innovative Computer Science and Engineering, Vol. 4 issue 6 Pp. 7-11, 2017.

[18] H. Kour, V. Mangat, and Nidhi, "A Survey of Sentiment Analysis techniques", Internaional conference on I-SMAC, IEEE, 2017.

[19] V.K. Singh, R. Piryani, A. Uddin and P. Waila, "Sentiment Analysis of Movie Reviews: A new feature based heuristic for aspect level sentiment classification", International multi-conference on automation computing communication control and compressed sensing, Pp.712-717, 2013.

[20] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath and A. Perera, "Opinion Mining and Sentiment Analysis on a Twitter Data Stream" , The International Conference on Advance In ICT for Emerging Regions, Pp.182-188, 2012

[21] L. Gongshen, L. Huoyao, L. Junn And L. Jiuchuan, "Predicting The Semantic Orientation Of Movie Reviews", Seventh International Conference On Fuzzy Systems And Knowledge Discovery, Pp. 2483-2488, 2010.

[22] S. Poria, E. Cambria, Lun-Wei Ku, Chen Gui, A. Gelbukh "A Rule-Based Approach to Aspect Extraction from Product Reviews", Preceedings of the second workshop on NLP for Social Media, Pp-28-37, Ireland, 2014.

[23] Minqing Hu and Bing Liu "Mining and Summarizing customer Reviews", In Proceedings of the 10th ACM SIGKDD International conference on knowledge discovery and data mining, Pp. 168-177, USA ,2004.

[24] I. C. Garcia, A. Gelbukh and G. Sidorov, "Implicit aspect indicator extraction for aspectbased opinion mining", International Journal of Computational Linguistics and Applications, vol. 5, issue. 2, Pp. 135-152, 2014.

[25] Z. Hai, K. Chang, Jung-Jae Kim, And Christopher C. Yang, "Identifying features in Opinion mining via Intrinsic and Extrinsic domain relevance", IEEE Transactions on Knowledge and Data engineering, Vol.26, No.3, Pp.623-634, 2014.

[26] C. Zhang and J. Zhu "Multi-Class Bootstraping Learning Aspect Related Terms for Aspect Identification", International Conference on Natural Language Processing and Knowledge Engineering, Dalian IEEE, Pp. 1-6, 2009

[27] S. Pasarate, R. Shedge. "Comparative Study Of Feature Extraction Techniques Used In Sentiment Analysis", International Conference on Innovation and Challenges in Cyber Security 978-1-5090-2084-3 IEEE, 2016.