# Opinion Mining and Sentiment Analysis: A Review on Approaches and Techniques Used

Neeraj Kumar[1], Prof. (Dr.) Saroj Hiranwal[2], Vijay Kumar Sharma[3]
[1]*Research Scholar, Rajasthan Institute of Engineering and Technology, Jaipur*
[2]*Professor, Rajasthan Institute of Engineering and Technology, Jaipur*
[3]*Assistant Professor, Rajasthan Institute of Engineering and Technology, Jaipur*

*Abstract*- **Over the last two decades, the emergence of web 2.0 has made it possible for the internet users to express their opinion for a particular product and services by admitting ratings and reviews on that particular blogs or social media websites. Almost all commercial organizations use the analysis of their customers view over their goods and services over their official blogs and mine the opinion of their customers for the goodwill of their customers choice and for the profit of their own organization as they completely believe that there future is totally dependent on customer's satisfaction. Comment section of the websites is flooded with tons of comments every second on the popular platforms that it became uneasy for a person who is totally dependent on those online reviews from various posts to make decisions over that product and services because of the mixed opinion shares by internet users to get over this situation it became an important task to mine the opinion of the reviews and categorize them as spam if the reviews seems to be fake. This paper contains all the aspect related to opinion mining in terms of its classification, various mining techniques that is prevailing in the study of this field. A brief survey has been done on the techniques used for sentiment analysis.**

*Index Terms*- **NBA, Opinion Mining, Sentiment Analysis, SVM.**

## I. INTRODUCTION

The world wide web is flourished with thousands of e-commerce websites and many other online platform that facilitate users to share their views on the selected product their might be a chance that the user is not genuine and sharing or texting a fake review of the product[1].There are some companies who even hire people to write positive reviews for their product. This self-promotion is termed as opinion spam [2] in which not a genuine reviews is in boxed in the comment section but a good reviews is written for the high selling of not that very good product. This misleading is commonly practiced now a day because of cut throat competitions in this commercialized world.

Opinion mining is a discipline in Natural Language Processing to pursue the study about people's feelings, attitude and emotion in textual as well as audio/video form toward events, products, individuals and services [3]. Substantially, opinion mining or sentiment analysis eases to get the attitude of the writer based on the sentence or as a whole document. The Web is full of amorphous information spread across in form of blogs, tweets, reviews etc. The task of analyzing various sentiments is not an easy task for the researchers and thus mining of these opinion or sentiment became a crucial task. Sentiment Analysis or opinion mining take the text as a document level or sentence level and finds the polarity of the text and distinguish them in the form of positive(represents happiness, joy, satisfaction), negative(represents anger, sadness, anxiety, sorrow ) and neutral(includes both positive and negative text in whole documents). Scores is made on the basis of the polarity of the sentiments [4]. Sentiments have to go through two processes. First is to check the polarity as positive, negative or neutral. Second is pinpointing the subjective and objective (facts) of the text to be analyzed [5]. Due to popularity among linguistic researchers sentiment analysis is often coined to many different terms as opinion mining, sentiment analysis, sentiment extraction, information gathering [6]. These days it has become a common practice for every internet users to share their views over multiple platforms like blogs, e-commerce, feedback forums, tweets, social networking sites. These shared views and thoughts make use of decision making strategy for the organization [7].

## II. COMPONENTS OF OPINION MINING

Opinion mining is categorized into three main components mainly:

### *A. Opinion Holder*

It holds the opinion or attitude of an individual or an organization. On account of online journals and reviews, assessment holders are those people who compose these surveys or blogs.

### *B. Opinion Object*

Opinion object is providing a platform on which opinion holder is expressing their opinion.

### *C.* Opinion Orientation

Opinion orientation points if the object is positive, negative or neutral which is shared by opinion holder.

## III. DIFFERENT APPROCHES OF SENTIMENT ANALYSIS

Based on the mindset of an individual performing sentiment analysis, various approaches ranging from keyword based, concept based and lexical affinity based exists. These are as follows.

### *A. Keyword based Approach*

This approach deals with building word lexicon. The analysis is done on the basis of each individual affect words of the sentence like “happy”, “sad”, “tired”, “sorrow”, “joy” [8]. There are some steps so as to create lexicon for this approach. First is to take some root word and some additional words can be added by taking some linguistic heuristic in to account. In another way, lexicon can be created by taking the root words and some extra words can be added based on the number of times it appears in the whole text. [9]. But the main concern of this approach is its drawbacks. First drawback is that it will show its inability in recognizing of the affect if some negation in the sentence appears [7]. For example in sentence 1.1 can be classified using keyword based approach while sentence 1.2 cannot be classified using this approach.

*It is raining good today.* ... *(1.1)*
*It is not raining good today* ...*(1.2)*

Secondly, it completely relies on affect word. If there are no affect words in the sentence then this approach will not guarantee the analysis of the sentence even the sentence gives strong emotions. [8]

### *B. Concept based Approach*

The concept based methodologies utilize web ontologies and semantic systems to accomplish semantic content investigation. Consequently, these methodologies help the framework in extricating the applied and full of feeling data from common Natural Language opinion. These methodologies principally depend on verifiable significance or highlight related with natural language ideas. Thus, these methodologies are superior to the methodologies which utilize keywords and word co-occurrence tallies. Concept based methodologies can identify the sentiments superior to syntactical strategies. These methodologies can likewise discover multi-word articulations even the articulations don't pass on any feeling unequivocally. The concept based methodologies principally depend on the information bases. It is challenging for the linguistics of natural language words if there is no nearly words that is acknowledged by the humans but not by the opinion systems. As the learning bases contain just average data related with ideas, along these lines, it limits their capacity to deal with semantic varieties. In this way, their settled portrayal, at last, places limits on surmising of semantic and full of feeling highlights related with ideas [10].

### *C. Lexical Affinity*

Lexical Affinity approach is marginally further developed than keyword based approach. This approach relegates a probabilistic 'partiality' to self-assertive words for a specific feeling as opposed to just identifying influence words in the content. For instance, a likelihood of 75% can be appointed to "terrific" to demonstrate a negative effect, comparable in 'terrific shot' or 'terrific ideas'. The probabilities doled out to words are typically prepared from phonetic corpora. However, this approach is superior to anything keyword based approach, be that as it may, this approach has two issues. To start with issue is that lexical Affinity approach primarily works at the word-level and can undoubtedly be deceived by sentences like given in (1.3) and (1.4) [8].

The Batsmen hit a terrific shot. … (1.3)

It was a terrific situation amidst jungle. … (1.4)

In sentence (1.3), word "terrific" is in normal form and show positive way while in sentence (1.4) the word "terrific " speaks to other word detects *fear* and therefore negative. Second issue is that lexical affinity probabilities are impacted by a specific area as recommended by the source of the linguistic corpora. Along these lines, a reusable and area free model can't be created [8].

In addition to the above approaches researchers work on three level for mining the opinion of various reviews which are as follows-

*Document Level*- The whole documents containing text is taken into account as a one and single polarity is defined and expressed in terms of positive, negative and neutral.[5][11] has performed at document level.

*Sentence Level*- The sentence level investigation concentrate on dissecting the records at sentence level. The sentences are examined independently and named objective, negative or positive. The general archive along these lines has an arrangement of sentences with each sentence being set apart with its comparing extremity. The work given by [12][13][14][15] was at sentence level.

*Aspect Level*- This gives the fine-grained and much deeper analysis than above two levels. It searches for the phrase present in the given document and classify them accordingly as positive, negative or neutral. Earlier it was called as Feature level extraction. Work done by [16][17] are remarkable.

## IV. DIFFERENT TECHNIQUES USED IN SENTIMENT ANALYSIS

Over the recent decade analysts have endeavored to concentrate on numerous particular assignments of sentiment analysis. Prior numerous analysts have concentrated on doling out sentiments to records by utilizing diverse techniques like machine learning technique, rule based technique and feature extraction. These techniques are discussed as under.

*A. Machine Learning Techniques*: This technique is further classified into two categories that are supervised and unsupervised techniques.

*Supervised Technique*: Supervised techniques can be executed by building a classifier. This classifier is prepared by example which can be physically named based on usual terms in the archives or can be gotten from client created client named online source [9]. Naïve Bayes Classifier (NBC), Support Vector Machines (SVM) and Maximum Entropy are for the most part utilized supervised techniques. Supervised techniques perform superior to unsupervised techniques [5]. From supervised techniques, SVMs perform better if both positive and negative words are available in the blog reviews. In this manner, SVMs are more suitable for sentiment characterization [18]. In any case, a Naïve Bayes classifier might be more reasonable when preparing informational collection is little in light of the fact that SVMs requires a vast informational collection in request to construct a classifier having high caliber. A concise depiction of Naïve Bayes Classifier and Support vector machines is given as takes after.

*(i) Naïve Bayes Classifier*

Naïve Bayes Classifier depends on Bayesian hypothesis and helpful when the scope of the data sources is high. Notwithstanding its effortlessness, Naïve Bayes Classifier performs superior to other order techniques [19].

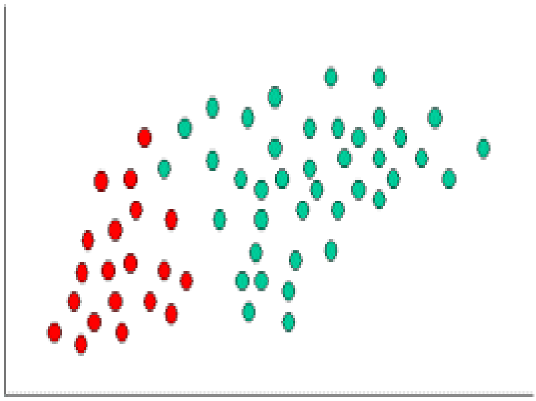

Fig.1.1 Demo of Naïve Bayes Classifier [20]

As appeared in the Figure 1.1, articles can be named either RED or GREEN. Primary undertaking is the distinguishing proof of class of new protests. This choice can be gone up against the premise of existing items. Since, there are twofold the quantities of

GREEN items than RED as indicated Figure 1.1. Along these lines, it can be suspected that new questions will more probable have a place with GREEN class. This confidence is known as the earlier likelihood in the Bayesian investigation. Earlier probabilities take a shot at the premise of past understanding. For this situation, earlier probabilities are the level of GREEN and RED articles. Assume, there are 60 objects, 40 of which are GREEN and 20 are RED. In this way, earlier probabilities for GREEN and RED articles will be as given in (1.5) and (1.6) [19].

*Primary probability for GREEN ∝ (No. of GREEN objects/Total no. of items)*

∝ (40/60) …..(1.5)

*Primary probability for RED ∝ (No. of RED objects/Total no. of objects)*

∝ (20/60) …..(1.6)

*(ii) Support Vector Machines*

Support Vector Machines deal with the possibility of choice planes that determine choice limits. An arrangement of items having a place with various class enrollments are isolated by choice planes [19]. A case to represent the idea of straight SVMs is appeared in Figure 1.3(a). In this illustration, the items either have a place with GREEN class (or RED class).The isolating line indicates the choice limit. On the correct hand side of the limit, all articles are GREEN and to one side hand side of limit, all items are RED. Another protest (white circle) will be named GREEN on the off chance that it tumbles to the correct side of the limit or named RED in the event that it tumbles to one side of the limit.

A classifier that segments an arrangement of items into their individual spaces with a line is called linear classifier and parceling with a bend is known as hyperplane classifier [19]. A case of hyperplane classifier is appeared in Figure 1.2(b).

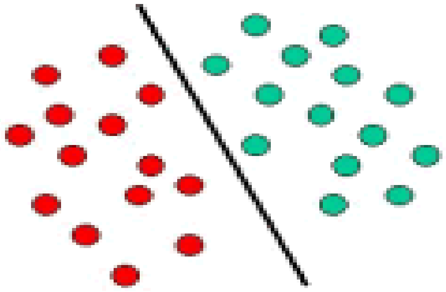

Fig. 1.2(a) Example of linear SVM [20]

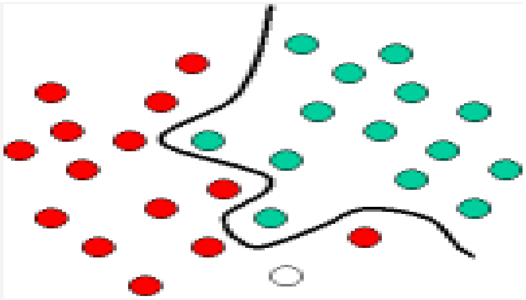

Fig. 1.2(b) Example of hyperplane SVM [20]

Figure 1.4 demonstrates the fundamental idea driving Support Vector Machines. In this figure, unique articles are mapped applying an arrangement of scientific capacities known as pieces.

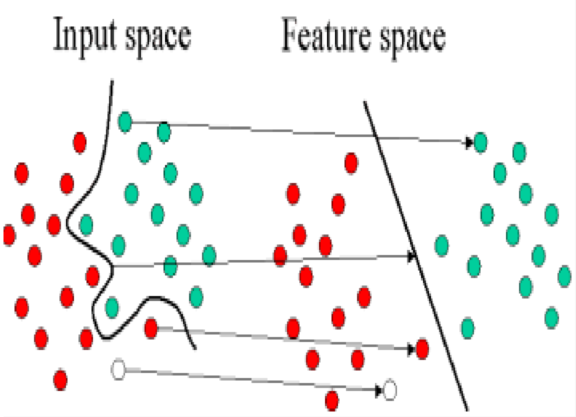

Fig. 1.3: Mapping of objects in SVMs [20]

This procedure of redesigning the articles is known as mapping or change. The figure demonstrates that the mapped objects are directly divisible [19]. In this manner, locate an ideal line as opposed to building the mind boggling bend that can isolate the GREEN and the RED objects.

*Unsupervised Technique:* In unsupervised system, classification of the sentiment analysis is performed. In this system, the features of a given content are analyzed against word vocabularies whose assumption esteems are chosen preceding their utilization [9]. Various leveled bunching and incomplete grouping are for the most part used calculations of unsupervised strategy. The two calculations are talked about as takes after.

*Hierarchical Clustering*

This algorithm divides the item into trees and each node demonstrates a cluster. There may be no or more sub-nodes of the tree and the solution arises as its tree’s nature. [20]

*Partial Clustering*

In partial clustering calculation, objects are parceled. Items can change the groups on the premise of disparity. K-means clustering algorithm is generally utilized algorithm of partial clustering algorithm [20].

*Feature Extraction*

This technique uses the feature of the product and based on its overall feature analysis and polarity classification can be done. The steps include feature extraction, prediction and classification. Several techniques such as POS Tagging, Stemming and Stop word removal are applied in extracting the features of a review on that very domain [21]. In [5] it has been explained that term frequency is preferred more than term presence for better feature extraction.

## IV. RELATED WORK

### *A. Supervised Technique*

For supervised technique, Pang et al.[5] used Support Vector Machine(SVM), Naïve Bayes Classifier and Maximum Entropy for sentiment classification among three machine learning techniques. An accuracy of 82.9% was detected by SVM technique as compared to Naïve Bayes Classifier and proved to be the best technique among all. A conclusion was made in which it was discussed that topic classification is much easier than sentiment classification. As well as it was also challenged that feature based extraction give better sentiment classification. Feature extraction is best achieved when word frequency is encountered repetitively. Presence of uni-gram or N-gram is present in the frequency. It is claimed that uni-gram perform better solution than bi-gram.

*Dave et al.*[22] work over product review and claimed that bi-gram and tri-gram performance gives a good result with an accuracy of 87%.

*Pang et al.*[23] proposed a graph based method using machine learning approach and showed a pre-processing for sentiment classification. As a result an improved solution with great accuracy was measured

*Cui et al.*[18] in this paper strongly claimed SVM technique give good result than Naïve Bayes Classifier. In the same paper he also acclaimed that SVM perform better when the training data is huge with multiple positive and negative reviews while Naïve Bayes Classifier perform better for the small training data set.

*Chen et al.*[24] has worked on the online book reviews of The Da Vinci Code taking the source from Amazon.com and used three techniques that is SVM,NBC and decision trees. The accuracy achieved by using these three techniques was 84.59%.

*Boiy et al.*[25] used SVM, Naïve Bayes Classifier and maximum entropy on movies reviews and car product reviews and claimed an accuracy of 90.25% for sentiment analysis and classification.

*Annet et al.*[26] used the same training data set that was used by [15s] for the movies reviews using SVM, NBC and decision tree with tools like WordNet which gave a much better accuracy of 75% than earlier work.

*Ye et al.*[27] used data set of travelers and get the reviews of tourists from yahoo.com and used N-gram, NBC and SVM and got an accuracy of 85.14%.

*Paltoglou et al.*[28] has used SVM for the movies reviews and achieved an accuracy of 96.90%.

### *B. Unsupervised Technique*

*Hu et al.*[29] proposed a basic technique by utilizing the adjective equivalent word set and antonym set in WordNet to locate the semantic introductions of adjective words.

*Pang et al*[30] used Parts of Speech(POS) techniques for getting the sentiment analysis and classification by extracting some unknown phrases from various blogs by drawing the features of each phrase.

*Li et al.*[31] used k-means clustering algorithm and developed an approach to cluster the data set into positive and negative groups by achieving an accuracy of 70%.

## V. CONCLUSION

Opinion Mining and Sentiment Analysis is the process through which the feeling, thoughts, attitude and reaction of a personal can be determined. Today overall decision and business plan of an organization is completely dependent on the reviews and opinion of the public through which one can decide over its product and services. Websites, blogs on various genres is full of reviews, comments, reaction and opinions. There might be no fields spared where there is no opinion on any topic. From social network sites to tweets and general comments, people begin to flood the comment section by giving there opinion and suggestion over that platform. But the opinion is not true always and there may be some fake opinion and reviews over a particular product or services. To deal with this situation a number of tools and techniques are used.

This paper deals with the survey of various approaches and techniques including Supervised and unsupervised techniques used in sentiment analysis and classification in terms of positive, negative or neutral. As well as many previous works using these techniques with their findings are discussed.

## ACKNOWLEDGMENT

I would like to thank Professor (Dr.) Saroj Hiranwal, Principal, Rajasthan Institute of Engineering and Technology, Jaipur for giving me opportunity to work on this research area. I am grateful to Vijay Kumar Sharma, Assistant Professor, Department of Computer Science and Technology for guiding me on minute of this work.

## REFERENCES

[1] A.Mukherjee,B.Liu,N.Glance Spotting fake reviewer groups in consumer reviews In Proceeding of 21st International Conference on World Wide Web, Leyone, France , pp.191-200, April 2016

[2] Myle Ott, Yejin Choi,Claire cardie, Jeffery T. Hancock Finding Deceptive Opinion Spam by Any Stretch of the Imagination In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – vol.- 1 pp. 309-319, Portland, Oregon, June 19 - 24, 2011

[3] Bing Liu Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers May 2012.

[4] Mukherjee S, Bhattacharyya Sentiment Analysis: A Literature Survey, IIT Bombay, Mumbai,2013.

[5] Pang B, Lee L, Vaithyanathan S Thumbs up? Sentiment classification using machine learning techniques. Proc. ACL-02 Conf. on Empirical methods in natural language processing, vol.-10, pp.-79-86.

[6] Karamibekr M, Ghorbani A Sentiment analysis of social issues. Int. Conf. on Social Informatics, pp.-215-221.

[7] Pang B, Lee L 2008 Opinion mining and sentiment analysis. Foundations and trends in information retrieval, vol. 2 pp.-1-135.

[8] Cambria E 2013 An Introduction to Concept-Level Sentiment Analysis. Advances in Soft Computing and Its Applications, Springer Berlin Heidelberg, pp.- 478-483.

[9] Gebremeskel G 2011 Sentiment Analysis of Twitter posts about news. Master's Thesis, University of Malta.

[10] Cambria E, Schuller B, Xia Y, Havasi C 2013 New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, pp.- 15-21.

[11] P. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews In Proc. Of 40th Annual Meet. Assoc. Comput. Linguist 2002 pp.-417-423.

[12] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 246–253, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[13] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing, EMNLP '03, pp. 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[14] T. W. Intelligent and T. Wilson. Annotating opinions in the world press. In In SIGdial-03, pp. 13–22, 2003.

[15] M. Hu and B. Liu. Mining and summarizing customer reviews. In KDD, pp. 168–177, 2004.

[16] A. Agarwal, F. Biadsy, and K. R.Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams, 2009.

[17] T. Wilson. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of HLT-EMNLP, pp. 347–354, 2005.

[18] Cui H, Mittal V, Datar M 2006 Comparative experiments on sentiment classification for online product reviews. American Association for Artificial Intelligence, vol: 6 pp. 265-1270.

[19] Lewicki P, Hill T 2006 Statistics: methods and applications. Tulsa, OK. Statsoft.

[20] Bair E 2013 Semi-supervised clustering methods. Wiley Interdisciplinary Reviews: Computational Statistics, vol. 5 pp. 349-361.

[21] Minging Hu, Bing Liu Proceedings of the 19th national conference on Artificial Intelligence pp. 755-760

[22] Dave K, Lawrence S, Pennock D M 2003 Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Proc. 12th Int. Conf. on World Wide Web, pp. 519-528.

[23] Pang B, Lee L 2004 A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proc. 42nd Annual Meeting on Association for Computational Linguistics.

[24] Chen C, Ibekwe-SanJuan F, SanJuan E, Weaver C 2006 Visual analysis of conflicting opinions. IEEE Symposium on Visual Analytics Science and Technology, Baltimore, Maryland: United States, pp. 59-66.

[25] Boiy E, Hens P, Deschacht K, Moens M F 2007 Automatic Sentiment Analysis in On-line Text. Proc. ELPUB2007 Conference on Electronic Publishing, Vienna, Austria, pp. 349-360.

[26] Annett M, Kondrak G 2008 A comparison of sentiment analysis techniques: Polarizing movie blogs. Advances in Artificial Intelligence, Springer Berlin Heidelberg, pp. 25-35.

[27] Ye Q, Zhang Z, Law R 2009 Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications, pp. 6527-6535.

[28] Paltoglou G, Thelwall M 2010 A study of information retrieval weighting schemes for sentiment analysis. Proc.48th Annual Meeting of the Association for Computational Linguistics, pp. 1386-1395.

[29] Yang Y, Pedersen J O 1997 A comparative study on feature selection in text categorization." Int. Conf. on Machine Learning, pp. 412-420.

[30] Peng T C, Shih C C 2010 An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs. IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology, vol. 3 pp. 243-248.

[31] Li G, Liu F 2010 A clustering-based approach on sentiment analysis. Int. Conf. on Intelligent Systems and Knowledge Engineering (ISKE), pp. 331-337.