**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 5.258**

***IJCSMC, Vol. 5, Issue. 5, May 2016, pg.639 – 649***

# A Novel Approach for Sentiment Analysis Using Classifiers Naive Bayes, SVM and Modified K-Means

**Rahul Bagga**

**rahulbagga1992@gmail.com**

Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida

**ABSTRACT:**

*Sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, micro blogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in digital forms. Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. In recent years, industrial activities surrounding sentiment analysis have also thrived. We proposed a sentiment analysis system using modified k means and naïve Bayes algorithm.*

## I. INTRODUCTION:

Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. Sentiment analysis, which is also called opinion mining, involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Sentiment analysis can be useful in several ways. For example, in marketing it helps in judging the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features.

There are several challenges in Sentiment analysis. The first is a opinion word that is considered to be positive in one situation may be considered negative in another situation. A second challenge is that people don't always express opinions in a same way. Most traditional text processing relies on the fact

*© 2016, IJCSMC All Rights Reserved* 

that small differences between two pieces of text don't change the meaning very much. In Sentiment analysis, however, "the picture was great" is very different from "the picture was not great". People can be contradictory in their statements. Most reviews will have both positive and negative comments, which is somewhat manageable by analyzing sentences one at a time. However, in the more informal medium like twitter or blogs, the more likely people are to combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context. For example, "That movie was as good as its last movie" is entirely dependent on what the person expressing the opinion thought of the previous model. The user's hunger is on for and dependence upon online advice and recommendations the data reveals is merely one reason behind the emerge of interest in new systems that deal directly with opinions as a first-class object. Sentiment analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. There are few main fields of research predominate in Sentiment analysis: sentiment classification, feature based Sentiment classification and opinion summarization. Sentiment classification deals with classifying entire documents according to the opinions towards certain objects. Feature-based Sentiment classification on the other hand considers the opinions on features of certain objects. Opinion summarization task is different from traditional text summarization because only the features of the product are mined on which the customers have expressed their opinions. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization. Languages that have been studied mostly are English and in Chinese .Presently, there are very few researches conducted on sentiment classification for other languages like Arabic, Italian and Thai. For convenience this paper is organized as follow: section 2 described sentiment classification techniques, section 3 present data sources used for sentiment analysis, section 4 explained tools for sentiment analysis section 5 or final section evaluate comparative result and discussion.

## II. SENTIMENT CLASSIFICATION TECHNIQUES

In these studies, sentiment analysis is often conducted at one of the three levels: the document level, sentence level, or attribute level. In relation to sentiment analysis, the literature survey done indicates two types of techniques including machine learning and unsupervised learning. In addition to that, the nature language processing techniques (NLP) is used in this area, especially in the document sentiment detection. Current-day sentiment detection is thus a discipline at the crossroads of NLP and Information retrieval, and as such it shares a number of characteristics with other tasks such as information extraction and text-mining, computational linguistics, psychology and predicative analysis.

### 1. MACHINE LEARNING

This section provides brief details of the machine learning algorithms used in the experiments.

Multi-Layer Perceptron (MLP): Multi Layer perceptron (MLP) is a feed-forward neural network with one or more layers between input and output layer. Feed-forward means that data flows in one direction from input to output layer (forward). This ANN which multilayer perceptron begin with input layer where every node means a predicator variable. Input nodes or neurons are connected with every neuron in next layer (named as hidden layers). The hidden layer neurons are connected to other hidden layer neuron.[2]

© 2016, IJCSMC All Rights Reserved 

Output layer is made up as follows:

1. When prediction is binary output layer made up of one neuron and

2. When prediction is non-binary then output layer made up of N neuron.

This arrangement makes an efficient flow of information from input layer to output layer. Figure 2 shows the structure of MLP. In figure 4 there is input layer and an output layer like single layer perceptron but there is also a hidden layer work in this algorithm.

Figure 1:MLP

MLP is a back propagation algorithm and has two phases:

Phase I: It is the forward phase where activation are propagated from the input layer to output layer.

Phase II: In this phase to change the weight and bias value errors among practical & real values and the requested nominal value in the output layer is propagate in the backward direction

Main advantages of MLP are It acts as a universal function approximate. MLP can learn each and every relationship among input and output variables. But disadvantages of MLP are MLP needs more time for execution compare to other technique because flexibility lies in the need to have enough training data. It is considered as complex "black box".

## 2. UNSUPERVISED LEARNING

Sentiment analysis is unsupervised learning" because it does not require prior training in order to mine the data. Instead, it measures how far a word is inclined towards positive and negative.

Much of the research in unsupervised sentiment classification makes use of lexical resources available. Kamps et al [5] focused on the use of lexical relations in sentiment classification. Andrea Esuli and Fabrizio Sebastiani [6]proposed semi-supervised learning method started from expanding an initial seed set using WordNet. Their basic assumption is terms with similar orientation tend to have similar glosses. They determined the expanded seed term‟s semantic orientation through gloss classification by statistical technique.

When the review where an opinion lies in, cannot provide enough contextual information to determine the orientation of opinion, Chunxu Wu[7] proposed an approach which resort to other reviews discussing the same topic to mine useful contextual information, then use semantic similarity measures to judge the orientation of opinion. They attempted to tackle this problem by getting the orientation of context independent opinions , then consider the context dependent opinions using linguistic rules to infer orientation of context distinct-dependent opinion ,then extract contextual information from other reviews that comment on the same product feature to judge the context indistinct-dependent opinions. An unsupervised learning algorithm by extracting the sentiment phrases of each review by rules of part-of-speech (POS) patterns was investigated by Ting-Chun Peng and Chia-Chun Shih [8]. For each unknown sentiment phrase, they used it as a query term to get top-N relevant snippets from a search engine respectively. Next, by using a gathered sentiment lexicon, predictive sentiments of unknown sentiment phrases are computed based on the sentiments of nearby known sentiment words inside the snippets. Gang Li & Fei Liu [9] developed an approach based on the k-means clustering algorithm. The technique of TF-IDF (term frequency – inverse document frequency) weighting is applied on the raw data. Then, a voting mechanism is used to extract a more stable clustering result. The result is obtained based on multiple implementations of the clustering process. Finally, the term score is used to further enhance the clustering result. Documents are clustered into positive group and negative group. Chaovalit and Zhou [10] compared the Semantic Orientation approach with the N-gram model machine learning approach by applying to movie

© 2016, IJCSMC All Rights Reserved 

reviews. They confirmed from the results that the machine learning approach is more accurate but requires a significant amount of time to train the model.

## 3. NEGATION

Negation is a very common linguistic construction that affects polarity and therefore, needs to be taken into consideration in sentiment analysis. Negation is not only conveyed by common negation words (not, neither, nor) but also by other lexical units. Research in the field has shown that there are many other words that invert the polarity of an opinion expressed, such as valence shifters, connectives or modals. "I find the functionality of the new mobile less practical", is an example for valence shifter, "Perhaps it is a great phone, but I fail to see why", shows the effect of connectives. An example sentence using modal is, "In theory, the phone should have worked even under water". As can be seen from these examples, negation is a difficult yet important aspect of sentiment analysis. Kennedy and Inkpen [10] evaluate a negation model which is fairly identical to the one proposed by Polanyi and Zaenen [11] in document-level polarity classification. A simple scope for negation is chosen. A polar expression is thought to be negated if the negation word immediately precedes it. Wilson et al. [12] carry out more advanced negation modeling on expression-level polarity classification. The work uses supervised machine learning where negation modeling is mostly encoded as features using polar expressions. Jin-Cheon Na [13], reported a study in automatically classifying documents as expressing positive or negative. He investigated the use of simple linguistic processing to address the problems of negation phrase. In sentiment analysis, the most prominent work examining the impact of different scope models for negation is Jia et al. [14]. They proposed a scope detection method to handle negation using static delimiters, dynamic delimiters, and heuristic rules focused on polar expressions Static delimiters are unambiguous words, such as because or unless marking the beginning of another clause.

## 4. LEVEL BASED CLASSIFIACTION

The sentiment analysis can be performed at one of the three levels: the document level, sentence level, feature level.

Document Level Sentiment Classification: In document level sentiment analysis main challenge is to extract informative text for inferring sentiment of the whole document. The learning methods can be confused because of objective statements are rendered by subjective statements and complicate further for document categorization task with conflicting sentiment. [15]

Sentence Level Sentiment Classification: The sentiment classification is a fine-grained level than document level sentiment classification in which polarity of the sentence can be given by three categories as positive, negative and neutral. The challenge faced by sentence level sentiment classification is the identification features indicating whether sentences are on-topic which is kind of co-reference problem [15]

Feature Level Sentiment Classification: Product features are defined as product attributes or components. Analysis of such features for identifying sentiment of the document is called as feature based sentiment analysis. In this approach positive or negative opinion is identified from the already extracted features. It is a fine grained analysis model among all other models [16].

© 2016, IJCSMC All Rights Reserved 

## III. DATASOURCE

People and companies across disciplines exploit the rich and unique source of data for varied purposes. The major criterion for the improvement of the quality services rendered and enhancement of deliverables are the user opinions. Blogs, review sites and micro blogs provide a good understanding of the reception level of products and services.

Blogs

The name associated to universe of all the blog sites is called blogosphere. People write about the topics they want to share with others on a blog. Blogging is a happening thing because of its ease and simplicity of creating blog posts, its free form and unedited nature. We find a large number of posts on virtually every topic of interest on blogosphere. Sources of opinion in many of the studies related to sentiment analysis, blogs are used. [17]

Review Sites

Opinions are the decision makes for any user in making a purchase. The user generated reviews for products and services are largely available on internet. The sentiment classification uses reviewer's data collected from the websites like www.gsmarena.com (mobile reviews), www.amazon.com (product reviews), www. CNETdownload.com (product reviews), which hosts millions of product reviews by consumers. [18]

Micro-blogging

A very popular communication tool among Internet users is micro-blogging. Millions of messages appear daily in popular web-sites for micro-blogging such as Twitter, Tumbler, Face book. Twitter messages sometimes express opinions which are used as data source for classifying sentiment. [19]

## IV. PROPOSED METHODOLOGY

The proposed architecture of four modules: user interface, log pre-processing, Feature Clustering using Modified K-means, Naïve Bayes Classification, Training and testing using support vector machine for more accurate categorization of opinion. This system can solve irrelevant data and more accuracy by associating Modified K means with naïve Bayes Classification algorithm.

© 2016, IJCSMC All Rights Reserved 

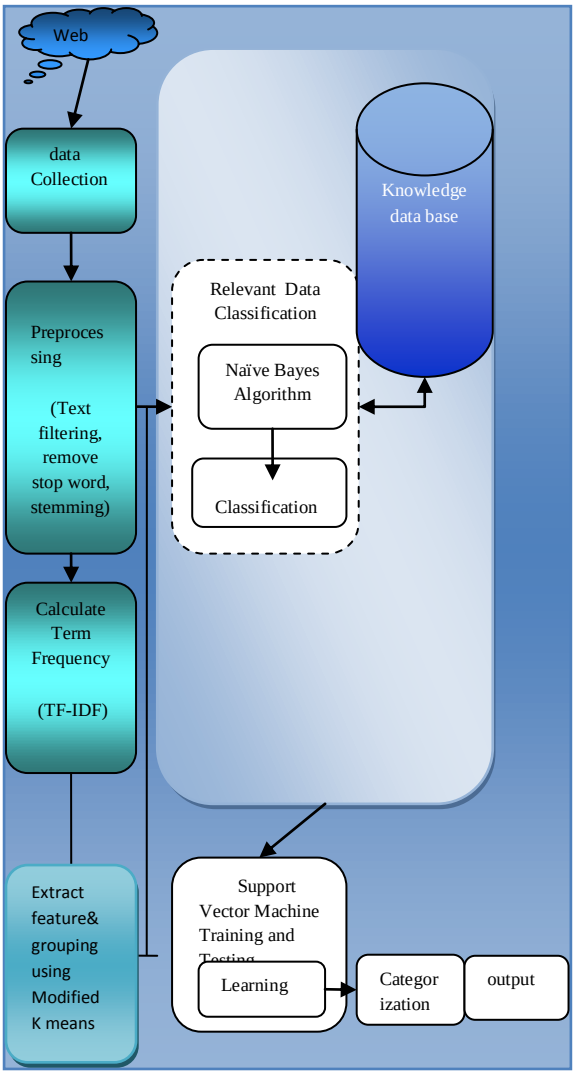


**Figure 2: Proposed System Architecture**

**A. Naive Bayes (NB):** Naive Bayes Classifier uses Bayes Theorem, which finds the probability of an event given the probability of another event that has already occurred. Naive Bayes classifier performs extremely well for problems which are linearly separable and even for problems which are non-linearly separable it performs reasonably well [3]. We used the already implemented Naive Bayes implementation in Weka2 toolkit.

**Algorithm**

**S1:** Initialize P(positive) ¬num - popozitii (positive)/ num_total_propozitii

**S2:** Initialize P(negative) ¬num - popozitii (negative) / num_total_propozitii

**S3:** Convert sentences into words

for each class of {positive, negative}:

for each word in {phrase}

P(word | class) < num_apartii (word | class) 1 | num_cuv (class) + num_total_cuvinte

P (class) ¬P (class) * P (word | class)

Returns max {P(pos), P(neg)}[1]

© 2016, IJCSMC All Rights Reserved 

The above algorithm can be represented using figure 2

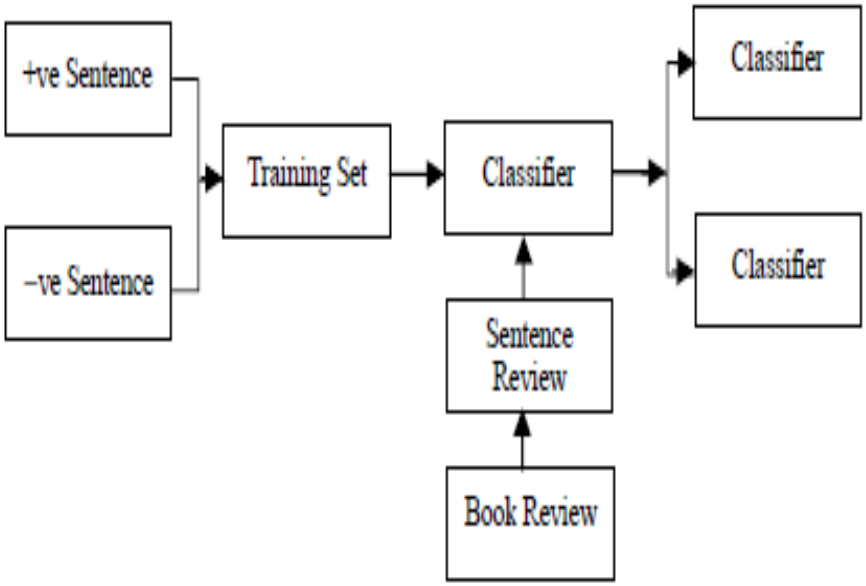


Fig 3:Naïve bayes classification

Major advantages of Naïve Bayes Classification is easy to interpret and efficient computation

**Modified approach K-mean algorithm:**

The K-mean algorithm is a popular clustering algorithm and has its application in data mining, image segmentation, bioinformatics and many other fields. This algorithm works well with small datasets. In this paper we proposed an algorithm that works well with large datasets. Modified k-mean algorithm avoids getting into locally optimal solution in some degree, and reduces the adoption of cluster -error criterion.

Algorithm: Modified approach (S, k), S={x1,x2,…,xn }

Input: The number of clusters k1( k1> k ) and a dataset containing n objects(Xij+).

Output: A set of k clusters (Cij) that minimize the Cluster - error criterion.

***Algorithm***

1. Compute the distance between each data point and all other data- points in the set D

2. Find the closest pair of data points from the set D and form a data-point set Am (1<= p <= k+1) which contains these two data- points, Delete these two data points from the set D

3. Find the data point in D that is closest to the data point set Ap, Add it to Ap and delete it from D

4. Repeat step 4 until the number of data points in Am reaches (n/k)

5. If p<k+1, then p = p+1, find another pair of data points from D between which the distance is the shortest, form another data-point set Ap and delete them from D, Go to step 4.

© 2016, IJCSMC All Rights Reserved 

Table1: Comparison between k-means and modified K means

| Number of Records | Time taken to execute (In millisecond) K-Mean Algorithms | Time taken to execute (In millisecond) Modified K-Mean Algorithm |
|---|---|---|
| 300 | 95240 | 61613 |
| 400 | 116243 | 73322 |
| 500 | 135624 | 103232 |
| 600 | 158333 | 122429 |

## V. RESULT

5. Classification using Modified K means and Naïve Bayes

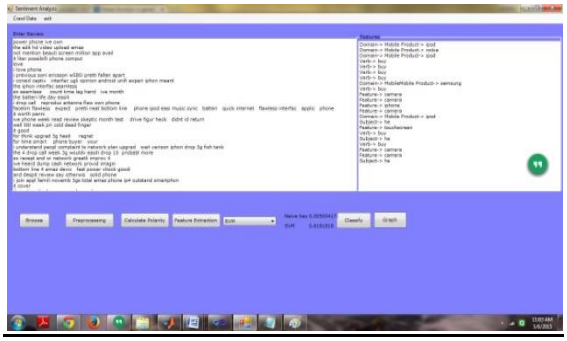

Figure 4: Result

## VI.CONCLUSION

A Sentiment Analyzer tool, although complex is an engaging project. Serving as the perfect Market Analysis tool, both organizations and individuals stand to benefit from this system. As the world is rapidly going mobile, the social media platforms represent a plethora of knowledge and information one can use to his/her benefit. Neglecting such an abundant source of information seems like a daunting task and hence this Sentiment Analysis tool has been developed. We proposed a method using naïve bayes and modified k means clustering and found that it is more accurate than naïve bayes and support vector machine techniques individually. We obtained an overall classification accuracy of 89.01% on the test set of 1000 mobile reviews. The running time of our algorithm is O(n + V log V) for training and O(n) for testing, where n is the number of words in the documents (linear) and V the size of the reduced vocabulary. It is much faster than other machine learning algorithms like Maxent classification or Support Vector Machines which take a long time to converge to the optimal set of weights. It achieved a better or similar accuracy when compared to more complicated models like SVMs, auto encoders, contextual valence shifters, matrix factorization, appraisal groups etc. From our point of view MKM and Naïve Bayes is best suitable for text based classification and SVM for social

© 2016, IJCSMC All Rights Reserved 

interpretation. In future we will be finding out the best result of sentiment analysis by applying other method on social networking reviews.

| **Name of Algorithm** | **Dataset** | **Accuracy(%)** |
|---|---|---|
| Naive Bayes | 500 mobile dataset | 79.66 |
| SVM | 500 mobile dataset | 83.59 |
| NB+Modified K-Means | 500 mobile dataset | 89 |

http://jmcauley.ucsd.edu/data/amazon/links.html

## VII. FUTURE WORK

1. Brand Monitoring: Evaluate and monitor your brand on the basis of customer views and opinions.

2. Campaign Monitoring: Create product campaigns to gain insight about the perception of the people on social media about the product since it spreads virally.

3. Competitive Intelligence: Track and follow your competitors and their customers to check their pain points and areas of improvements to tap their potential.

4. Identifiers: Check the influencers who are talking about your brand in various channels and influencing others.

5. Use bigger data sets in future

Sentiment Analysis has been helpful in giving us an accuracy of 60 to 70% which is actually comparable with a human analyst. It can become a fuel for efforts to shape opinions, attitude and emotions on the web and hence can be used for predictive modelling. It can be used for marketing strategies, Poll Prediction, Government Intelligence, Product Comparison and other numerous business activities to come out with a result which has a bearing on the decision making body. In stark contrast to marketing of the past, today's marketers are measured by how much revenue they bring in per dollar spent. What "sentiment analysis" does is give those marketers an alternative way to measure their effectiveness tracking how customers feel about and how much they are talking about a brand. All that effort on branding, messages, and colour schemes can finally be validated! However, the number of positive mentions you generate is simply a step towards revenue. Thus marketers can track their campaign's performance in real time and get an idea about the general consensus in the market and from hindsight can predict how their future decisions would lead to the conversion of consumer's money into our own bank balance.

**Acknowledgement**

I would sincerely like to thank my HOD **Prof. (Dr.) Abhay Bansal,** Joint Acting Head, Amity School of Engineering & Technology, Director- DICET, Professor & HoD(CSE) and faculty guide, **Mr. Gaurav Raj Assistant Professor,** Department of CSE , Amity University ,Noida for their timely assistance and support in guiding me on presenting this paper. The insights provided by them helped me clear several roadblocks.

*© 2016, IJCSMC All Rights Reserved* 

**REFERENCES**

[1] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Sentiment classification using machine learning techniques." In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86.

[2] "Towards Enhanced Opinion Classification using NLP Techniques", IJCNLP 2011, pages 101–107, Chiang Mai, Thailand, November 13, 2011

[3] Qiang Ye, Ziqiong Zhang, Rob Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", Expert Systems with Applications 36 (2009) 6527–6535.

[4] Rui Xia , Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011) 1138–1152.

[5] Kamps, Maarten Marx, Robert J. Mokken and Maarten De Rijke, "Using wordnet to measure semantic orientation of adjectives", Proceedings of 4th International Conference on Language Resources and Evaluation, pp. 1115-1118, Lisbon, Portugal, 2004.

[6] Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification",Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005.

[7] Chunxu Wu, Lingfeng Shen , "A New Method of Using Contextual Information to Infer the Semantic Orientations of Context Dependent Opinions" , 2009 International Conference on Artificial Intelligence and Computational Intelligence

[8]Ting-Chun Peng and Chia-Chun Shih , "An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and ntelligent Agent Technology JOURNAL

[9] Hu, and Liu, "Mining and summarizing customer reviews", Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 2005,pp. 168–177.

[10] Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, vol. 22, pp. 110–125,2006.

[11] Polanyi and A. Zaenen, "Contextual lexical valence shifters," in Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text,AAAI technical report SS-04-07, 2004.

[12] Wilson et al , Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011) 1138–1152

[13] Jin-Cheon Na , Christopher Khoo, Paul Horng Jyh Wu, "Use of negation phrases in automatic sentiment classification of product reviews", Library Collections, Acquisitions,& Technical Services 29 (2005) 180–191.

[14] Jia, C. Yu, andW. Meng , "The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness", In Proceedings of CIKM ,2009

[15] V. S. Jagtap and Karishma Pawar, Analysis of different approaches to Sentence-Level Sentiment Classification, International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 3, PP : 164-170 1 April 2013

*© 2016, IJCSMC All Rights Reserved* 

[16] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, Clustering Product Features for Opinion Mining, WSDM'11, February 9–12, 2011, Hong Kong, China. Copyright 2011 ACM 978-1-4503-0493-1/11/02...$10.00

[17] Singh and Vivek Kumar, A clustering and opinion mining approach to socio-political analysis of the blogosphere, Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference.

[18] G.Vinodhini and RM.Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey, Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[19] Alexander Pak and Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining

[20] VIDEO LECTURES OF STANFORD https://class.coursera.org/nlp/lecture

[21] SCHOLAR PAPER shahid.shaikh@nu.edu.pk

[22] Opinion Mining and Sentiment Analysis (Foundations and Trends in Information

Retrieval BY BO PANG LILLIAN LEE

[23] Amazon data sets provided by http://jmcauley.ucsd.edu/data/amazon/links.html

*© 2016, IJCSMC All Rights Reserved* 