Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis

Geetika Gautam
Department of Computer Science & Engg.
Jaypee Institute of Information technology
Noida, India
geetikagautam16@gmail.com

Divakar yadav
Department of Computer Science & Engg.
Jaypee Institute of Information technology
Noida, India
divakar.yadav@jiit.ac.in

***Abstract*— The wide spread of World Wide Web has brought a new way of expressing the sentiments of individuals. It is also a medium with a huge amount of information where users can view the opinion of other users that are classified into different sentiment classes and are increasingly growing as a key factor in decision making. This paper contributes to the sentiment analysis for customers' review classification which is helpful to analyze the information in the form of the number of tweets where opinions are highly unstructured and are either positive or negative, or somewhere in between of these two. For this we first pre-processed the dataset, after that extracted the adjective from the dataset that have some meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning based classification algorithms namely: Naive Bayes, Maximum entropy and SVM along with the Semantic Orientation based WordNet which extracts synonyms and similarity for the content feature. Finally we measured the performance of classifier in terms of recall, precision and accuracy.**

Keywords— Machine Learning, Semantic Orientation, Sentiment Analysis, Twitter

I. INTRODUCTION

The current research paper covers the analysis of the contents on the Web covering lots of areas which are growing exponentially in numbers as well as in volumes as sites are dedicated to specific types of products and they specialize in collecting users' reviews from various sites such as Amazon etc. Even Twitter is an area where the tweets convey opinions, but trying to obtain the overall understanding of these unstructured data (opinions) can be very time consuming. These unstructured data (opinions) on a particular site are seen by the users and thus creating an image about the products or services and hence finally generating a certain judgment. These opinions are then being generalized to gather feedbacks for different purposes to provide useful opinions where we use sentiment analysis.

Sentiment analysis is a process where the dataset consists of emotions, attitudes or assessment which takes into account the way a human thinks [1]. In a sentence, trying to understand the positive and the negative aspect is a very difficult task. The features used to classify the sentences should have a very strong adjective in order to summarize the review. These contents are even written in different approaches which are not easily deduced by the users or the firms making it difficult to classify them.

Sentiment analysis influences users to classify whether the information about the product is satisfactory or not before they acquire it. Marketers and firms use this analysis to understand about their products or services in such a way that it can be offered as per the user's needs.

There are two types of machine learning techniques which are generally used for sentiment analysis, one is unsupervised and the other is supervised [2]. Unsupervised learning does not consist of a category and they do not provide with the correct targets at all and therefore conduct clustering. Supervised learning is based on labeled dataset and thus the labels are provided to the model during the process. These labeled dataset are trained to produce reasonable outputs when encountered during decision- making.

To help us to understand the sentiment analysis in a better way, this research paper is based on the supervised machine learning.

The rest of the paper is organized as follows. Second section discusses in brief about the work carried out for sentiment analysis in different domain by various researchers. Third section is about the approach we followed for sentiment analysis. Section four is about implementation details and results followed by conclusion and future work discussion in the last section.

II. RELATED WORK

In recent years a lot of work has been done in the field of “Sentiment analysis“by number of researchers. In fact work in the field started since the beginning of the century. In its early stage it was intended for binary classification, which assigns opinions or reviews to bipolar classes such as positive or negative. Paper [3] predicts review by the average semantic orientation of a phrase that contains adjective and adverb thus calculating whether the phrase is positive or negative with the use of unsupervised learning algorithm which classifies it as thumbs up or thumbs down review. Some sentiment analyses are based on review of the user summarization system of the product e.g. [4]. In [4] the product feature uses latent semantic analysis (LSA) based filtering mechanism to identify opinion

978-1-4799-5173-4/14/$31.00 ©2014 IEEE

words that are used to select some sentences to become a rich review summarization.

Paper [5] uses a comparison between positive and negative sentences. It extracts information from the Web and manually label the word set which requires a lot of unnecessary effort. Author in [6] has used a rule-based method, based on BaseLine and SVM for sentiment analysis of Chinese document level, which extract the overall document polarity of specific words by a sentiment word dictionary, and adjust it according to the context information. In another work [7], the polarity of the word is being calculated by all the words in the sentence, which can either be positive or negative depending on the related sentence structure. Lakshmi and Edward [8] have proposed to pre process the data to improve the quality structure of the raw sentence. They have applied LSA technique and cosine similarity for sentiment analysis.

Basant Agarwal, et. al. [9] applied phrase pattern method for sentiment classification. It uses part of speech based rules and dependency relation for extracting contextual and syntactic information from the document. In [10] author intended to put forward aspect based opinion polling from unlabeled free form textual customer reviews which do not require customers to answer the questions. M. Karamibekr and A.A. Ghorbani [11] proposed a method based on verbs as an important opinion term for sentiment classification of a document belonging to the social domain. Paper [12] generates a sentiment lexicon called SentiFul which uses and enlarges it through synonyms, antonyms, hyponyms relations, derivation and compounding. They proposed method to distinguish four kinds of affixes on the basis of the role they play for sentiment features namely: propagation, weakening, reversing, and intensifying. These methods assign sentiment polarity which helps in expanding the lexicon to improve the sentiment analysis.

A lot of work has also been done where researchers have explored and applied soft-computing approaches, mainly fuzzy logic and neural works for sentiment analysis. [13] And [14] are such examples of works which are based on the fuzzy logic approach. The main contribution of [13] is that it applied fuzzy domain sentiment ontology tree extraction algorithm. This algorithm constructs fuzzy domain sentiment ontology tree based on the reviews that includes extraction of sentiments words, features of the product and relation among features thus precisely predicting the polarity of the reviews. In [14] authors have designed a fuzzy inference system based on membership functions. By designing membership functions they formulated and standardized the process of quantifying the strength of reviewer's opinions in the presence of adverbial modifier. They applied the method for tri-gram patterns of adverbial modifiers.

III. OUR APPROACH

In our approach we used the twitter dataset and analyzed it. This analyses labeled datasets using the unigram feature extraction technique. We used the framework where the pre-processor is applied to the raw sentences which make it more appropriate to understand. Further, the different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content. The complete description of the approach has been described in next sub sections and the block diagram of the same is graphically represented in Fig. 1

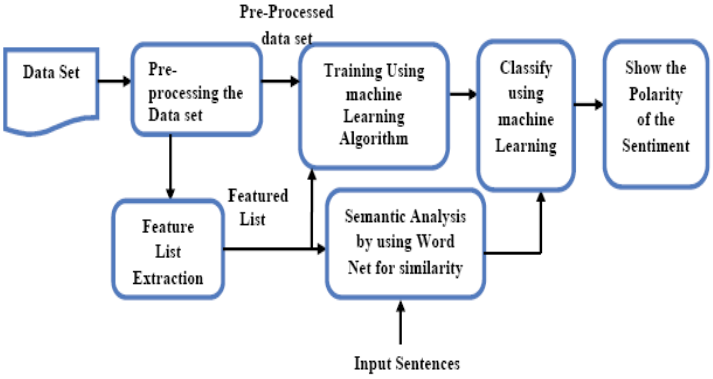
Pre-Processed data set
Data Set
Pre-processing the Data set
Training Using machine Learning Algorithm
Classify using machine Learning
Show the Polarity of the Sentiment
Featured List
Feature List Extraction
Semantic Analysis by using Word Net for similarity
Input Sentences

Fig.1. Diagram of the Approach to Problem

A. Pre-processing of the datasets

The tweets contain a lot of opinions about the data which are expressed in different ways by individuals .The twitters dataset used in this work is already labeled. Labeled dataset has a negative and positive polarity and thus the analysis of the data becomes easy. The raw data having polarity is highly susceptible to inconsistency and redundancy. The quality of the data affects the results and therefore in order to improve the quality, the raw data is pre-processed. It deals with the preparation that removes the repeated words and punctuations and improves the efficiency the data.

For example, "that painting is Beauuuutifull #" after pre-processing converts to "painting Beautiful." Similary, "@Geet is Noww Hardworkingg" converts to "Geet now hardworking".

B. Feature Extraction

The improved dataset after pre- processing has a lot of distinctive properties. The feature extraction method, extracts the aspect (adjective) from the dataset. Later this adjective is used to show the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using unigram model [15]. Unigram model extracts the adjective and segregates it. It discards the preceding and successive word occurring with the adjective in the sentences. For above example, i.e. "painting Beautiful" through unigram model, only Beautiful is extracted from the sentence.

C. Training and classification

Supervised learning is an important technique for solving classification problems. In this work too, we applied various supervised techniques to get the desired result for sentiment analysis. In next few paragraphs we have briefly discussed about the three supervised techniques i.e. naïve bayes, maximum entropy and support vector machine followed by the semantic analysis which was used along with all three techniques to compute the similarity.

- **Naive Bayes**

It has been used because of its simplicity in both during training and classifying stage. It is a probabilistic classifier and can learn the pattern of examining a set of documents that has been categorized. It compare the contents with the list of words to classify the documents to their right category [16].

$$C^* = \text{argmac}_c P_{NB}(c|d)$$

$$P_{NB}(c|d) := \frac{(P(c) \sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)} \qquad -----(1)$$

Class c* is assigned to tweet *d*, where, *f* represents a feature and $n_i(d)$ represents the count of feature f_i found in tweet *d*. There are a total of m features. Parameters P(c) and P (f|c) are obtained through maximum likelihood estimates which are incremented by one for smoothing.

Pre-processed data along with extracted feature is provided as input for training the classifier using naïve bayes. Once the training is complete, during classification it provides the polarity of the sentiments. For example for the review comment "I am happy' it provide Positive polarity as result.

- **Maximum entropy**

Maximum entropy maximizes the entropy defined on the conditional probability distribution. It even handles overlap feature and is same as logistic regression which finds distribution over classes. It also follows certain feature exception constraints [17].

$$P_{ME}(c|d,) = \frac{\exp[\Sigma_i \lambda_i f_i (c,d)]}{\Sigma_{c'} \exp[\Sigma_i \lambda_i f_i (c,d)]} \qquad ------(2)$$

Where, *c* is the class, *d* is the tweet, and □ is a weight vector. The weight vectors decide the significance of a feature in classification.

It follows the similar processes as naïve bayes, discussed above and provides the polarity of the sentiments.

- **Support vector machine**

Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space. The input data are two sets of vectors of size m each. Then every data represented as a vector is classified in a particular class. Now the task is to find a margin between two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SVM also supports classification and regression which are useful for statistical learning theory and it helps recognizing the factors precisely, that needs to be taken into account, to understand it successfully [18].

- **Semantic Analysis**

After the training and classification we used semantic analysis. Semantic analysis is derived from the WordNet database where each term is associated with each other. This database is of English words which are linked together. If two words are close to each other, they are semantically similar. More specifically, we are able to determine synonym like similarity. We map terms and examine their relationship in the ontology. The key task is to use the stored documents that contain terms and then check the similarity with the words that the user uses in their sentences. Thus it is helpful to show the polarity of the sentiment for the users.

For example in the sentence"I am happy" the word ''happy'' being an adjective gets selected and is compared with the stored feature vector for synonyms. Let us assume 2 words; 'glad' and 'satisfied' tend to be very similar to the word 'happy'. Now after the semantic analysis, 'glad' replaces 'happy' which gives a positive polarity.

IV. IMPLEMENTATION AND RESULT

We used Python and Natural Language Tool Kit to train and classify the naive bayes, maximum entropy and support vector machine. In total we used data set of size 19340 out of which 18340 were used for training and 1000 for testing. For training Fig 2. Show the overall flow of processes.

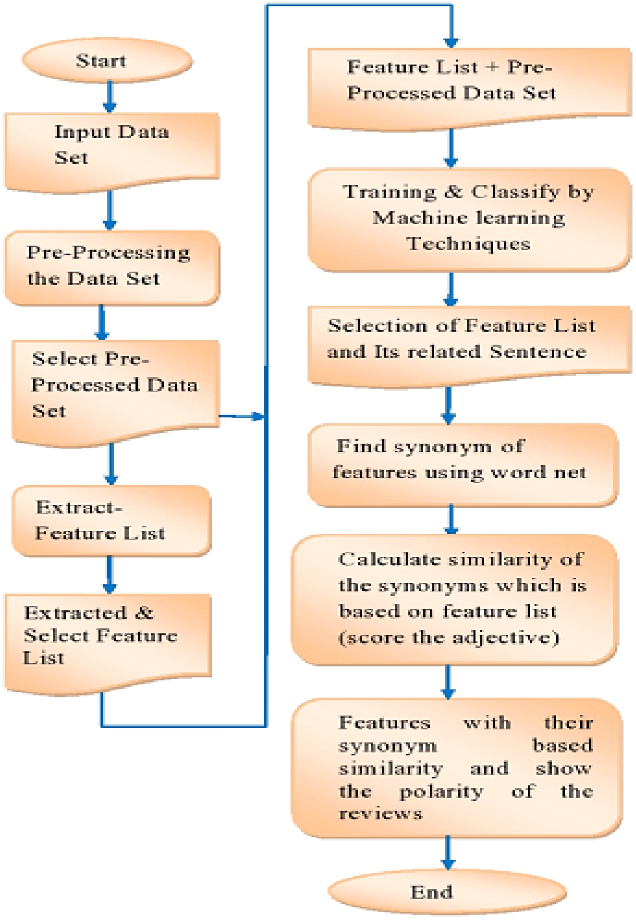
Start
Input Data Set
Pre-Processing the Data Set
Select Pre-Processed Data Set
Extract-Feature List
Extracted & Select Feature List
Feature List + Pre-Processed Data Set
Training & Classify by Machine learning Techniques
Selection of Feature List and Its related Sentence
Find synonym of features using word net
Calculate similarity of the synonyms which is based on feature list (score the adjective)
Features with their synonym based similarity and show the polarity of the reviews
End

Fig. 2. Flow Diagram of the proposed methodology

The description of the process in pseudo code form is shown in fig. 3.

```
Input: Labeled Dataset
Output: positive and negative polarity with synonym of
             words and similarity between words
Step-1 Pre-Processing the tweets:
           Pre-processing ()
           Remove URL:
           Remove special symbols
           Convert to lower:
Step-2 Get the Feature Vector List:
             For w in words:
                    Replace two or more words
                 Strip:
               If (w in stopwords)
                    Continue
                Else:
                   Append the file
              Return feature vector
Step-3 Extract Features from Feature Vector List:
                 For word in feature list
                 Features=word in tweets_words
                 Return features
Step-4 Combine Pre-Processing Dataset and Feature
                           Vector List
              Pre-processed file=path name of the file
              Stopwords=file path name
              Feature Vector List=file path of feature vector
                                 list
Step-5 Training the step 4
             Apply classifiers classes

Step-6 Find Synonym and Similarity of the Feature Vector
                    For every sentences in feature list
                    Extract feature vector in the tweets ()
                    For each Feature Vector: x
                      For each Feature Vector: y
                         Find the similarity(x, y)
                             If (similarity>threshold)
                                        Match found
                         Feature Vector: x= Feature Vector: y
                                   Classify (x, y)
     Print: sentiment polarity with similar feature words
```

Fig. 3. Pseudo code of the process:

A. Results Analysis:

In this section we discuss the results obtained through naïve bayes, maximum entropy, and support vector machine and compared their relative performances on three parameters namely: accuracy, precision and recall where,

- Accuracy is measured in percentage and is computed as:

$$Accurcy = \frac{Tp + Tn}{Tp + Tn + Fn + Fp} \quad ------(3)$$

- Recall positive (p) and Recall negative (n) are the recall ratio and are computed as:

$$Recall\ (p) = \frac{Tp}{Tp + Fn} \quad ------(4)$$

$$Recall\ (n) = \frac{Tn}{Fp + Tn} \quad ------(5)$$

- Precision positive(p) and Precision negative (n) are precision ratio and are computed as:

$$Precision\ (p) = \frac{Tp}{Tp + Fp} \quad ------(6)$$

$$Precision\ (n) = \frac{Tn}{Fn + Tn} \quad -------(7)$$

Table 1, Table 2 and Table 3 show the performance measures of naive bayes, maximum entropy and support vector machine based classifiers respectively in terms of precision and recall. Similarly Table 4 shows the performance of the classifiers terms of accuracy. Fig. 4 shows a comprehensive view of accuracy of the three supervised learning techniques and semantic analysis (WordNet). A comparative measurement on the basis of recall parameter is shown in Fig. 5. Similarly comparative measurement on the basis of precision parameter is shown in Fig. 6.

TABLE I. NAIVE BAYESIAN CLASSIFICATION MEASUREMENTS

Performance Measures (%)	
Positive Recall	91.2
Negative Recall	85.4
Positive Precision	49.3
Negative Precision	39.3

TABLEII. MAXIMUM ENTROPY MEASURMENTS

Performance Measures (%)	
Positive Recall	86.1
Negative Recall	80.0
Positive Precision	40.4
Negative Precision	33.6

TABLE III. SUPPORT VECTOR MACHINE MEASURMENTS

Performance Measures (%)	
Positive Recall	88.3
Negative Recall	83.5
Positive Precision	43.8
Negative Precision	35.7

TABLE IV. ACCURACY COMPARISON

Methods	Accuracy
Naive Bayes	88.2
Maximum Entropy	83.8
Support Vector machine	85.5
Semantic Analysis (WordNet)	89.9

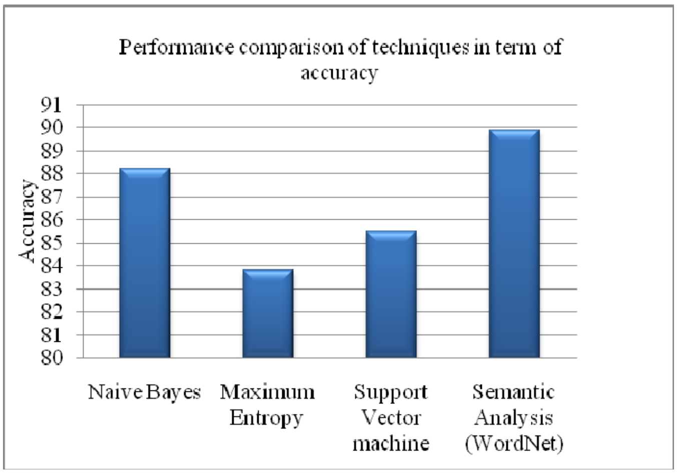

Fig. 4. Performance comparison of techniques in terms of accuracy

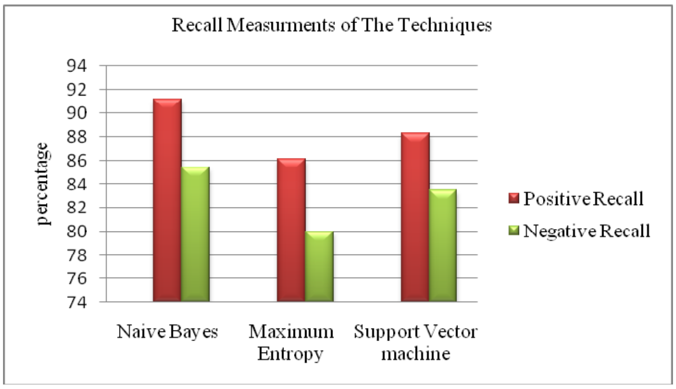

Fig.5. Measurements of positive and negative recall of the techniques

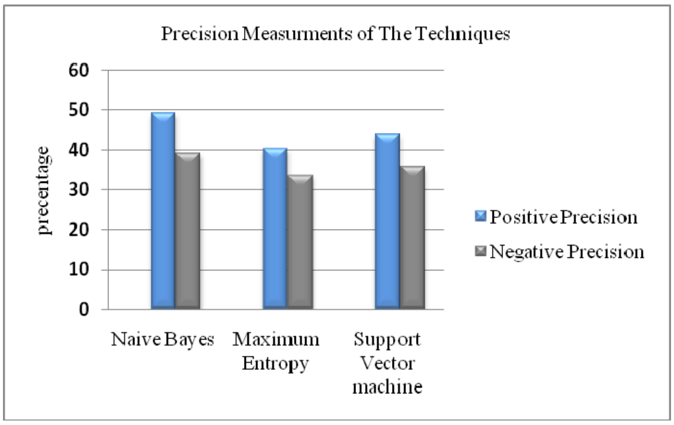

Fig.6. Measurements of positive and negative precision of the techniques

V. CONCLUSION

In this paper, we proposed a set of techniques of machine learning with semantic analysis for classifying the sentence and product reviews based on twitter data. The key aim is to analyze a large amount of reviews by using twitter dataset which are already labeled. The naïve byes technique which gives us a better result than the maximum entropy and SVM is being subjected to unigram model which gives a better result than using it alone. Further the accuracy is again improved when the semantic analysis WordNet is followed up by the above procedure taking it to 89.9% from 88.2%. The training data set can be increased to improve the feature vector related sentence identification process and can also extend WordNet for the summarization of the reviews. It may give better visualization of the content in better manner that will be helpful for the users.

REFERENCES

[1] R. Feldman, ” Techniques and Applications for Sentiment Analysis ,” Communications of the ACM, Vol. 56 No. 4, pp. 82-89, 2013.

[2] Y. Singh, P. K. Bhatia, and O.P. Sangwan, ”A Review of Studies on Machine Learning Techniques,” International Journal of Computer Science and Security, Volume (1) : Issue (1), pp. 70-84, 2007.

[3] P.D. Turney,” Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424, July 2002.

[4] Ch.L.Liu, W.H. Hsaio, C.H. Lee,and G.C.Lu, and E. Jou,” Movie Rating and Review Summarization in Mobile Environment,” IEEE Transactions on Systems, Man, and Cybernetics, Part C 42(3):pp.397-407, 2012.

[5] Y.Luo,W.Huang,” Product Review Information Extraction Based on Adjective Opinion Words,” Fourth International Joint Conference on Computational Sciences and Optimization (CSO), pp.1309 – 1313, 2011.

[6] R.Liu,R.Xiong,and L.Song, ”A Sentiment Classification Method for Chinese Document,” Processed of the 5th International Conference on Computer Science and Education (ICCSE), pp. 918 – 922, 2010.

[7] A.khan,B.Baharudin, ”Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs,” Processed on National Postgraduate Conference (NPC), pp. 1 – 7, 2011.

[8] L.Ramachandran,E.F.Gehringer, ”Automated Assessment of Review Quality Using Latent Semantic Analysis,” ICALT, IEEE Computer Society, pp. 136-138, 2011.

[9] B.Agarwal,V.K.Sharma,andN.Mittal,"Sentiment Classification of Review Documents using Phrase Patterns," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1577-1580, . 2013.

[10] J.Zhu, H.Wang, M.Zhu, B.K.Tsou, and M.Ma,," Aspect-Based Opinion Polling from Customer Reviews," T. Affective Computing2(1):pp. 37-49, 2011.

[11] M.Karamibekr,A.A.Ghorbani,"Verb Oriented Sentiment Classification," Processed of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol (1): pp. 327-331, 2012.

[12] A. Neviarouskaya, H.Prendinger, and M.Ishizuka," SentiFul: A Lexicon for Sentiment Analysis," T. Affective Computing 2(1), pp.22-36, 2011.

[13] L.Liu, X.Nie,and H.Wang," Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis," Processed of the 5th Image International Congress on Signal Processing (CISP), pp. 1620 – 1624, 2012.

[14] R. Srivastava, M. P. S. Bhatia," Quantifying Modified Opinion Strength: A Fuzzy Inference System for Sentiment Analysis," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1512-1519, 2013.

[15] C. Tillmann , and F. Xia, "A phrase-based unigram model for statistical machine translation," Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL, pp.106-108, 2003.

[16] B.Ren,L.Cheng," Research of Classification System based on Naive Bayes and MetaClass," Second International Conference on Information and Computing Science, ICIC '09, Vol(3), pp. 154 – 156, 2009.

[17] C.I.Tsatsoulis,M.Hofmann,"Focusing on Maximum Entropy Classification of Lyrics by Tom Waits," IEEE International on Advance Computing Conference (IACC), pp. 664 – 667, 2014.

[18] M.A. Hearst,"Support vector machines,"IEEE Intelligent Systems, pp. 18-28, 1998.