# Sentiment Analysis using Optimal Feature and Ensemble Classifier

**Manpreet Kaur[1] Monika[2]**
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science and Applications
[1,2]K.U., Kurukshetra, India

*Abstract*— Sentiment Analysis means opinion mining. The goal is to identify people's opinions, perceptions, emotions toward entities and their attribute. Opinions are in positive, negative, and neutral in nature. For classifications of these opinions, different classifier can be used. This paper, improve the accuracy of ensemble classifier i.e. random forest and also perform analysis on the large dataset of movie reviews with the help of various classifiers like Support vector machine, k-nearest neighbor, naive bayes on proposed optimal features. The proposed method achieves better accuracy than the previous approaches by using optimal features and TF-IDF for finding the frequency of term occurs in a document and for identifying how the term important is. Here, three metrics are used i.e. F-measure, Precision and recall for evaluating the performance of classifiers.

***Key words:*** Sentiment Analysis, Opinion Mining, Support Vector Machine (SVM), Naive Bayes (NB), Classification, Random Forest, K- Nearest Neighbor (KNN), Sentiments

## I. INTRODUCTION

Sentiment analysis also called opinion mining is used to analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes[1]. It represents a large problem space. Sentiment analysis is most commonly used in industries, but in academia sentiment analysis is frequently used. Sentiment analysis is mainly focuses on the people's opinion which expresses positive or negative sentiments regarding any entity and its attributes. Sentiment analysis is normally applied to reviews and social media for a variety of applications.

Data and information are increases on World Wide Web day by day and people express their feelings, emotions, opinion or attitude towards entity and their attributes over the internet through social media, blogs, ratings and reviews [2]. With the help of sentiments of people towards any entity and product, company analyzes and rates their publicity of things properly with the help of sentiments. People give their sentiments either in positive or negative way. Some sentiments are neutral in nature. Sentiment analysis is beneficial for all small and large organization to enhance their operations, services, product quality etc[3]. Peoples mostly gives their reviews on social media sites like twitter, face book, forums, blogs and online shopping sites etc.

There are three techniques of sentiment analysis i.e. machine learning based approach, lexicon based approach, and Hybrid based approach. In machine learning based is fully automatic and handle large amount of data. It is more beneficial than other approaches. The first approach has three types: supervised learning, unsupervised and semi supervised learning. In lexicon based approach, predefined lists of words associated with the specific sentiment are used. This approach contains two types i.e. dictionary based and corpus based approach and the last approach, is based on combination of both machine learning and lexicon based approach.

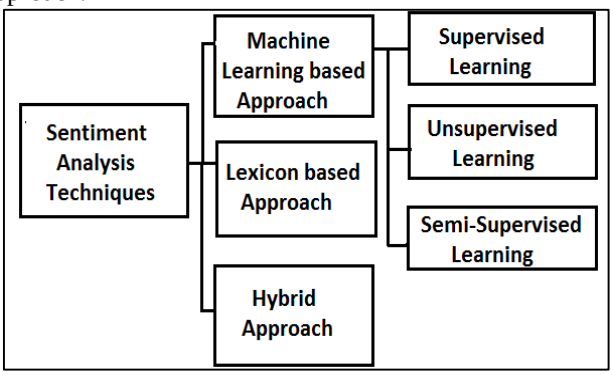


Fig. 1: Sentiment Analysis Techniques

## II. RELATED WORK

In 2014, Luiz F. S. Coletta, Nadia F. F. da Silva, Eduardo R. Hruschka, Estevam R. Hruschka Jr.[4] has identifies the better results and more accuracy by using SVM combined with a cluster ensemble and defines algorithm, named $C^3E$-SL which is used to combine classifier and cluster ensembles. Without using the stand-alone classifier of two. The SVM with cluster ensemble to classifier the tweet messages and finds better results and accuracy.

In 2015, Nur Azizah vidya, Mohammad Ivan Fanany, Indra Budi[5] has presented a paper on "Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers" They overcome the issue of mobile providers to measuring the brand reputations which is based on the customer's sentiment analysis from twitter data about their service quality. They focus on five products like 3G, 4G, short messaging, voice and internet services and also discuss some correlated business insights in a telecommunication service industry.

In 2015, Yun wan, Dr. Qigang Gao[6], has presented a paper on "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis". In this paper, they classifies twitter sentiment about airline services by using an ensemble sentiment classification strategy which is based on majority vote principle of multiple classification methods like SVM, C4.5, naive bayes, Bayesian network, Decision tree and Random Forest algorithms. They applied these six classification methods and the proposed ensemble approach on data set of 12864 tweets, in which 10 fold evaluations is used to validate the classifier.

In 2015, Monisha Kanakaraj and Ram Mohana Reddy Guddeti [7] has presented a paper on "NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifier" They uses NLP to increase the sentiment classification with the addition of Semantic in feature vector and ensemble methods is used for classification. They also raise the exactness of prediction by adding same words and context - sense identities to the feature vectors.

All rights reserved by www.ijsrd.com 

## III. PROPOSED WORK

### *A. Dataset Reading*

Dataset is a collection of different type of data entities which is access either individually or as a whole entity. Dataset learning means collect the data and then build the dataset. In sentiment analysis Datasets are the feedback/ review/comments given by the person regarding any entity which is collected either from manually or pick up online a well formed, organized manner. Here, the large dataset of movie review for performing sentiment analysis are selected. The data is reviews (like positive, negative, and neutral) regarding any entity, persons, etc.

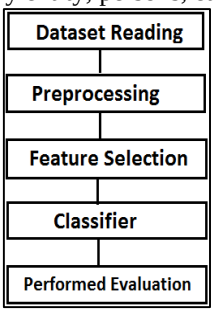


Fig. 2: Workflow of proposed work

### *B. Preprocessing*

Pre-processing is needed to eliminate text noises[8]. Pre-processing means cleaning the datasets like removal of stop words, special symbols, lower case, and stemming. Stop words are the words which have small meaning, such as "and", "the", "a", "an" and similar word. Stemming is the method which contains extra words like "watched" in this word "ed" is extra words. These types of words like stop word, special characters, lower case etc are the irrelevant in the reviews of persons and affect the accuracy. In preprocessing, with the removal of these types of word better accuracy has been achieved.

### *C. Feature Selection*

TF-IDF is stands for term frequency-inverse document frequency used in information retrieval and text mining [9].

*1) Term Frequency*

Which measures how frequently a term occurs in a document. Documents are different in size, and a term would appear in much more times in large document as compare to smaller ones. And term frequency is divided by the document length.

- TF(t)= (number of times term t appear in a document)/ (total number of terms in the documents)

*2) IDF*

**(**Inverse document frequency) which tells how important a term is. While computing TF, all terms are considered equally important. Many terms like "is", "of" and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

- IDF(t)= log_ e (total number of documents/ number of documents with term t in it ).

TFIDF= TF(t)*IDF(t)

*3) VSM*

(Vector space model) is an algebraic model for representing text documents as vectors of identifiers[12]. Vector space model procedure can be divided into three stages. The first phase is the document indexing where content bearing terms are extract from the document text. The second phase is the weighting of the indexed term to improve retrieval of document appropriate to the user and last phase ranks the document with respect to query according to similarity measure.

### *D. Classifier*

Classifiers are used to classify the sentiments of persons which are positive or negative. There are many classifiers like support vector machine, naïve bayes, k-nearest neighbor, ensemble classifier i.e. random forest. SVM algorithm is based on decision plane that defines decision boundaries and decision plane separates group of instances having different class membership [13]. Support vector machine (SVM) are supervised learning model and algorithm which is used to analyze data for classification and regression analysis. SVM used for linear and non linear classification. Then Random Forest classifier which is ensemble learning method for classification, learning and other tasks it operates by constructing a decision tree and giving output in the form of mode of classes or mean prediction of the tree. Next is naïve bayes i.e. probabilistic classifier which is based on bayes theorem [11]. It has strong and naïve independence assumptions and performing well in many complex real world troubles [2]. Naïve bayes algorithm is very efficient and superior in terms of CPU and memory consumptions. It requires small amount of training data. Then k-nearest neighbor is very simple algorithm that stores all available cases and classifies new cases based on the similarity measure [10]. It is non-parametric algorithm means it does not make any assumptions on the underlying data distributions.

### *E. Performed Evaluation*

There are three matrices which is used to determining how well a sentiment analysis system works i.e. precision, recall, f-measure[7].

1) Precision/ Accuracy**:** A measure of how often a sentiment rating was correct.
2) Precision = True Positive / (True Positive + False Positive)
3) Recall**:** A measure of how many documents with sentiment were rated as sentimental. This could be seen as how accurately the system determines neutrality.
4) Recall = True Positive / (True positive + False Negative)
5) F-measure**:** also called F-score, this is combination of precision and recall. The F-measure is very helpful, as it gives us a single metric that rates a system by both precision and recall.
6) F-measure= 2*precision*recall/ (Precision + recall)

## IV. EXPERIMENT/RESULT

The proposed work is implemented on Intel inside® Core(TM) i3 Processor 2.20 GHz, Window 7 Home Basic and used net beans IDE 7.2.1 platform and compare results with existing work[7] and ensemble classifier combining with other different classifiers. In this, the movie review dataset are used for performing the proposed experiment and achieving 100% result as compare to existing work.

All rights reserved by *www.ijsrd.com* 

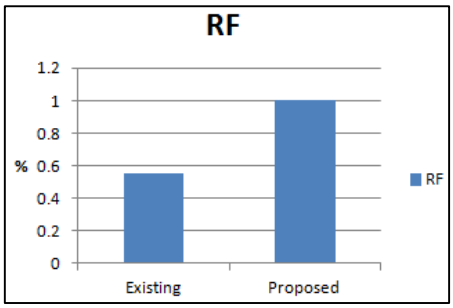


Fig. 3: Comparison of existing and proposed using ensemble classifier.

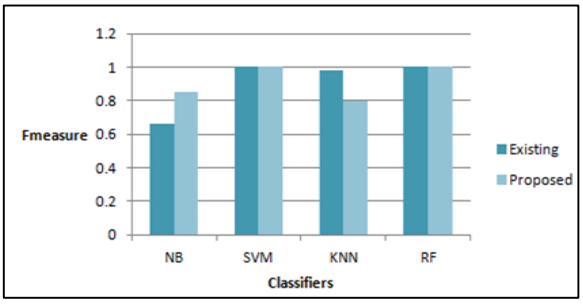


Fig. 4: F-measure of Classifier

In fig 4 we calculate the F-measure of different classifiers and proposed work perform well in naïve bayes, support vector machine, random forest as compare to the exiting work. In proposed work SVM and Random Forest gives the 100% result on selected dataset.

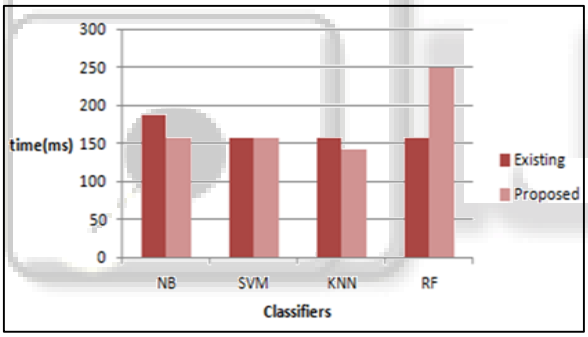


Fig. 5: Classification Time

In fig 5 we calculate the classification time. Classification time is less in case of naïve bayes, SVM, K-nearest neighbour but increase in the case of Random Forest and achieves better accuracy than other classifiers.

## V. Conclusion

Sentiment analysis is used for identifies the opinions/ sentiment of the persons regarding any entity and their attributes. For the classifications of the opinion of persons, different classifier like SVM, naive bayes, KNN and ensemble classifier (random forest) etc are used. In this paper, we perform sentiment analysis on the movie review data, classifies the sentiments (i.e. positive, negative and neutral) of the persons regarding movie with the help of various classifiers. Better results have been achieved with the help of these classifiers. The results show high accuracy in less time. In future the proposed work can be tested on other large datasets and optimization based classification can be used to improve the time complexity of the approach.

## References

[1] Abhinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques", 3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015) ELSEVIER.

[2] Zohreh Madhoushi, Abdul Razak Hamdam, Suhaila Zainudin, "sentiment analysis techniques in recent works", science and Information Conference, July 2015

[3] Pratik Thakor, Dr. Sreela Sasi , "Ontology-based Sentiment Analysis Process for Social Media Content", 2015 INNS Conference on Big Data, Vol 53,2015, ELSEVIER

[4] Luiz F.S. Coletta, Nadia F.F.da Silva, Eduardo R.Hruschka, "Combining Classifications and Clustering for tweet Sentiment Analysis", Brazilian Conference on Intelligent Systems,2014, IEEE.

[5] Nur Azizah vidya, Mohamad Ivan Fanany, Indra Budi, "Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers", The 3rd information System International conference,2015, ELSEVIER

[6] Yu Wan, Dr.Qigang Gao, "An Ensemble Sentiment Classification System of twitter data for Airline Services Analysis", 15th International Conference on Data mining Workshops,2015, IEEE

[7] Monisha Kanakaraj and Ram Mohana Reddy Guddeti , "NLP Based Sentiment Analysis in Twitter Data Using Ensemble Classifier",3rd International Conference on Siganl Processing, Communication and Networking, 2015, IEEE.

[8] Y.Zhang and P.Desouza, "Enhance the power of sentiment analysis," in International Journal of Computer, Information, systems and Control Engineering , Hopkinton, 2014.

[9] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of Sentiment Review using Machine Learning Techniques", 3rd International conference on Recent Trends in computing 2015, ELSEVIER, 2015.

[10] P.Kalaivani, Dr.K.L.Shunmuganathan, "Sentiment classification of Movie Reviews by Supervised Machine Learning Approaches", India Journal of Computer Science and Engneering(IJCSE), 2013

[11] Paulina Aliandu, "Sentiment analysis to determine Accommodation, Shopping and Culinary location on foursquare in kupang city", The Third information System International Conference, ELSEVIER, 2015.

[12] A.B. Manwar, Hemant S. Mahalle, K.D. Chinchkhede, Dr. Vinay Chavan, "A Vector Space Model for Information Reterival: A matlab approach", Indian Journal of Computer Science and Engineering (IJCSE), vol.3 no.2 Apr-may 2012.

[13] Pravesh Kumar Singh, Mohd Shahid Husain, "methodological study of opinion mining and sentiment analysis techniques", International Journal on Soft Computing (IJSC) Vol. 5, No. 1, February 2014.

All rights reserved by *www.ijsrd.com* 