Text Preprocessing and Classification Using Machine Learning Technique

Amit Kumar Dewangan[1], S. M. Ghosh[2], A. K.Shrivas[3]
[1,2,3]Dr. C. V. Raman University, Bilaspur (C. G.) India.

ABSTRACT
Sentimental analysis is the method of finding sentiment such as positive or negative from a text data. In this paper we have used some feature selection techniques such as Mutual information, Information gain and TF-IDF to select features from high dimensionality data set. These methods are evaluated over dataset consists of 2000 user-created movie reviews archived on the IMDb (Internet Movie Database) web portal and is known as "Sentiment Polarity Dataset version 2.0" . The reviews are equally partitioned into a positive set and a negative set (1000+1000). The machine learning play very important role for preprocessing and classification of data. The classification is performed using support vector machine(SVM), Random Forest, Random Tree, Naïve Bayes, Bayes Net and J48. We have used ensemble model to achieve high accuracy provided by WEKA tool.

Keywords: Classification, Feature Selection, Cornell Movie Dataset.

I INTRODUCTION

In this modern scenario data is part of human life. It is a necessary element of human daily life. Our environment is bounded from the data and information. Data is any type like structured, semi-structure or unstructured. Structured data like numerical values of attributes and unstructured data is like audio, video, image, text, text with numbers etc. Due to these reasons internet is one of the essential part of human life. The information in it covers a wide range of areas such as academic information, feedback or opinion about products, comments about social issues etc. It helps people to think and make decision in many things. Majority of people always listen to others opinion before taking a final decision. Sentimental analysis is one of the research areas [11].

The analysis of sentiments may be document based where the sentiment in the entire document is summarized as positive or negative. It can be sentence based where each and every sentence, having sentiments, in the text is classified. Sentiment analysis can be phrase based where the phrases in a sentence are classified according to the polarity based on some patterns of their occurrence. Sentiments are classified as positive (denotes a state of happiness, bliss or satisfaction on part of the writer) or negative (denotes a state of sorrow, dejection or disappointment on part of the writer)[7].

II RELATED WORK

The papers are basically focused on encapsulating the movie reviews at characteristic level so that user can find easily that which character of the movie they liked or disliked. In this paper, the author has two different methods are implemented for finding subjectivity of sentences and then rule based system is used to find feature-opinion pair and finally the orientation of extracted opinion is revealed using two different methods. Initially the proposed system uses SentiWordNet approach to find out orientation of extracted opinion and then it uses the method which is based on lexicon consisting list of positive and negative words [1].

It illustrates that comparison of the efficiency of the different classifiers focusing on numeric and text data. Datasets from IMDb and 20newsgroups have been used for the purpose. Current work mainly focuses on comparing different algorithms such as Decision Stump, Decision Table, K-Star, REPTree and ZeroR in the area of numeric classification, and evaluation of the efficiency of Naive Bayes classifier for text classification. In this paper, we have used WEKA tool to evaluate and analysis of datasets [2].

In this paper, authors have analyzed the Movie reviews using various techniques like Naïve Bayes, K-Nearest Neighbor and Random Forest [3].

Three classification models are used for text classification using Waikato Environment for Knowledge Analysis (WEKA). Opinions written in Roman-Urdu and English are extracted from a blog. These extracted opinions are documented in text files to prepare a training dataset containing 150 positive and 150 negative opinions, as labeled examples. Testing data set is supplied to three different models and the results in each case are analyzed. The results show that Naive Baycsian outperformed Decision Tree and KNN in terms of more accuracy, precision, recall and F-measure [4].

The other authors have focused on the classification of opinion mining techniques that conveys user's opinion i.e. positive or negative at various levels. The precise method for predicting opinions enable us, to extract sentiments from the web and foretell online customer's preferences, which could prove valuable for marketing research. Much of the research work had been done on the processing of opinions or sentiments recently because opinions are so important that whenever we need to make a decision we want to know others' opinions [5].

This research concerns on binary classification which is classified into two classes. The classes are positive and negative. The positive class shows good message opinion while the negative class shows the bad message opinion of certain movies. This justification is based on the accuracy level of SVM with the 10-Fold cross validation and confusion matrix. The hybrid Partical Swarm Optimization (PSO) is used to

improve the election of best parameter in order to solve the dual optimization problem [6].

This paper extends our ideas pertaining to Sentiment Analysis to the regional language Kannada, spoken mainly in Karnataka, a state in southern part of India. They have explored the usefulness of semantic approaches and machine learning approaches, used predominately on English language data set, from Kannada web documents. They found the average accuracy of machine learning approaches to be better than the average accuracy of semantic learning approaches for Kannada data set [7].

In this paper, they proposed an approach to understand situations in the real world with the sentiment analysis of Twitter data base on deep learning techniques. With the proposed method, it is possible to predict user satisfaction of a product, happiness with some particular environment or destroy situation after disasters. Recently, deep learning is able to solve problems in computer vision or voice recognition, and convolutional neural network (CNN) works good for image analysis and image classification. The biggest reason to adopt CNN in image analysis and classification is due to CNN can extract an area of features from global information, and it is able to consider the relationship among these features. The above solution can achieve a higher accuracy in case of analysis and classification [9].

III PROPOSED ARCHITECTURE

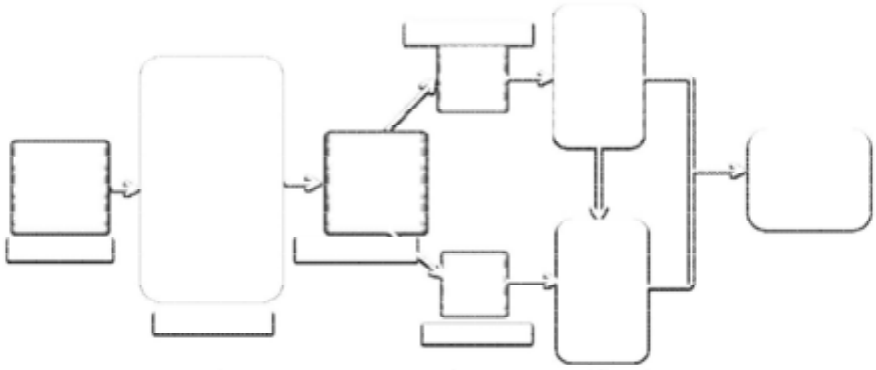

Fig.1 Proposed model for text classification

Figure 1 shows that proposed architecture of research work.In this work, we are using movie review Sentiment Polarity Dataset version 2.0" (http://www.cs.cornell.edu/People/pabo/movie-review-data)[9] dataset. In first step, applied different preprocessing techniques like stemmer, tokenizer, stopwords remover and pruning on the movie review data se to remove the noise and inconsistent data and prepared the smooth dataset. Now. We have divided the dataset into training and testing where training dataset is used for trained the classifier and testing data is used for test the trained classifier. Finally calculated the various performances of classifiers like accuracy, precision, recall and F-measures.

IV METHOD AND MATERIAL

(a) **Dataset-** The dataset consists of 2000 user-created movie reviews archived on the IMDb (Internet Movie Database) web portal at http://reviews.imdb.com/Reviews and is known as "Sentiment Polarity Dataset version 2.0"(http://www.cs.cornell.edu/People/pabo/movie-review-data)[9]. The reviews are equally partitioned into a positive set and a negative set (1000+1000).

(b) **WEKA Tool-** WEKA stands for Waikato Environment Knowledge Analysis(http://www.cs.waikato, ac.nz/~ml/weka/) [9] , it is a collection of various data mining algorithms and tools for in depth analysis.The programming language of WEKA is Java and its distribution is based on GNU (General Public License). There are mainly three uses of WEKA. First the analysis of data mining algorithm; second for generation of model; and last for comparison of various data mining algorithm in order to choose best as predictor.

First thing is to import the dataset from database to weka. Text data is import using TextDirectoryLoader component. To perform the preprocessing in WEKA, we used the StringToWordVector filter from the package weka.filters.unsupervised.attribute. This filter allows to configure the different stages of the term extraction. Configure the tokenizer (term separators), specify a stop-words list and choose a stemmer.

V RESULT AND DISCUSSION

In this experiment first we have applied the StringToWordVector to calculate attribute and we got 47163 in the dataset. After that we apply stremmer, StopWords and Tokenizer for preprocessing. We have used WEKA data mining tool [10] for analysis of Sentiment Polarity Dataset version 2.0" (http://www.cs.cornell.edu/People/pabo/movie-review-data)[9].WEKA is an open source software tool which contains various classification techniques used in this research work. This research work have used decision tree techniques like Naïve Bayes, SMO, Bayes Net, Random Forest, Random Tree and C4.5/J48 for analysis and classification of movie review with 70-30%, 75-25% and 80%-20% training-testing data partition. The accuracy of Naïve Bayes, SMO, Bayes Net, Random Forest, Random Tree and C4.5/J48 as shown in table 1 where Naïve Bayes gives better classificationaccuracy as 79.725%. To achieve the better classification accuracy, we have ensemble the individuals trained classifiers and achieved the best accuracy 81.50% of accuracy with proposed ensemble of Naïve Bayes, BayesNet and Random Forest as shown in table 1. Fig. 5 and Fig. 6 show that accuracy graph with 70-30% and 80-20% training and testing partition of ensemble models. We have also calculated Precision, Recall and F-Measure for the best ensemble model.In case of 70-30% training-testing partition the average Precision , recall and F-measure are 0.815, 0.185 and 0.185 respectively while in case of 80-20% training testing partition , the average Precision, recall and F-measure are 0.816, 0.185 and 0.185. Finally we concluded the our proposed ensemble model gives better perforce of movie review for users.

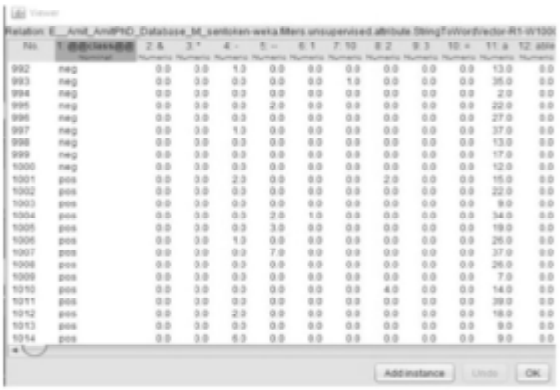

Fig. 2 Total number of word count present in dataset.

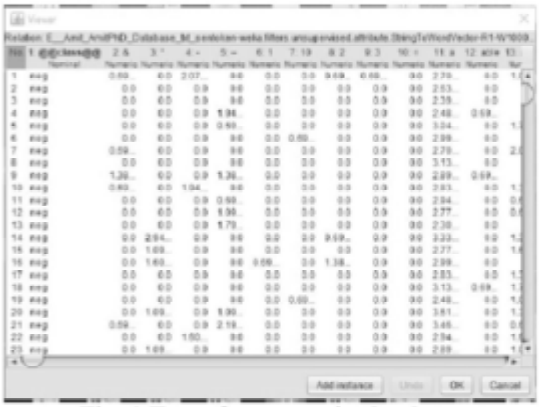

Fig. 3 Term frequency in the dataset.

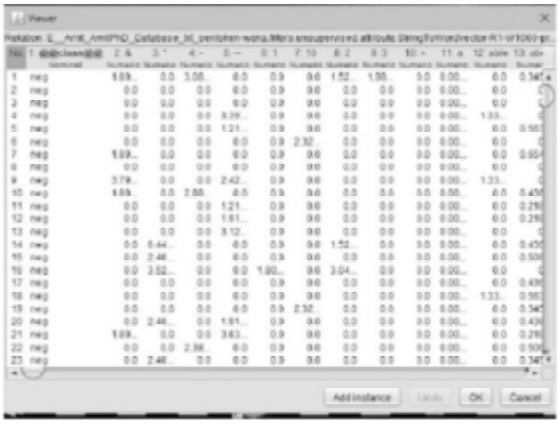

Fig. 4 TF-Idf of the dataset.

Table 1
Accuracy of individuals and ensemble models (in percentage).

Sno.	Model name	split		
		70-30 %	75-25 %	80-20 %
1	J48	65.67	65.00	64.25
2	RF	78.83	76.40	75.50
3	RT	56.50	59.00	57.75
4	SMO	76.00	76.20	75.50
5	NAIVABAYSE	78.67	78.00	79.25
6	BAYSENET	77.17	78.60	77.75
7	NB+BN	81.00	79.80	80.50
8	NB+RF	79.17	77.80	79.50
9	BN+RF	78.83	78.80	77.75
10	NB+BN+RF	81.50	80.60	81.50

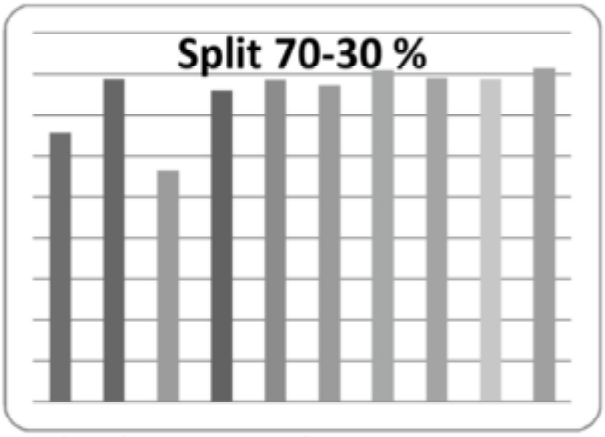

Fig 5. Accuracy table of Ensemble model at 70-30%.

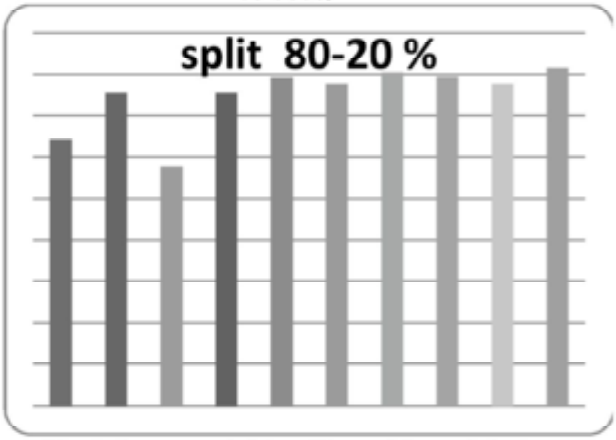

Fig 6. Accuracy table of Ensemble model at 80-20%.

VI CONCLUSION

Text recognition is the very important tasks to get conclude opinion of document, social media post, any script etc. Classification techniques play major role to identify and categorize the sentiments about document, social media post, any script etc. In this research work, proposed ensemble models likeensemble of Naive Bayes, Bayes Net and Random Forest which gives better results compares to individuals and other ensemble model. An optimization of features play important role to develop computationally efficient model. The proposed ensemble model gives satisfactory results with few numbers of features and recommended as classifier for classification of sentiments analysis (positive or negative).

REFERENCES

[1] Sharma S. and Kaur G (2016) , Review Paper On Sentiment Classification Of Movies Review , International Journal of Engineering Applied Sciences and Technology , 2 (1) , PP 67 – 72.

[2] Kumar N., Mitra S, Bhattacharjee M and Mandal L. (2018), Comparison of Different Classification Techniques Using Different Datasets , Proceedings of International Ethical Hacking Conference, PP 261-272.

[3] PalakBaid P. Gupta A. and Chaplot N. (2017), Sentiment Analysis of Movie Reviews using Machine Learning Techniques , International Journal of Computer Applications, 179 (7), PP 45-49.

[4] Bilal M., Israr H., Shahid M. and Khan A. (2015) , Sentiment classification of Roman-Urdu opinions using Naı¨ve Bayesian, Decision Tree and KNN classification techniques , Journal of King Saud University Computer and Information Sciences , 28 , PP 330-344.

[5] Mishra M. and Jha C. K. (2012), Classification of Opinion Mining Techniques , International Journal of Computer Applications , 56 (13), PP 1-6.

[6] Basari A.S.H., Hussin B., Ananta I.G.P and Zeniarja J. (2013) , Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization , Elsevier , Procedia Engineering 53 , PP 453 – 462.

[7] Kumar K. M. A., Rajasimha N., Reddy M. , Rajanarayana A. and Nadgir K. (2015), Analysis of Users' Sentiments from KannadaWeb Documents , Elsevier, ScienceDirect , Procedia Computer Science 54 , PP 247 – 256.

[8] Liaoa S. ,Wangb J. , Yua R. , Satob K. and Chengb Z. (2017) , CNN for situations understanding based on sentiment analysis of twitter data , Elsevier , Procedia Engineering , 111 ,PP 376–381.

[9] www.cs.cornell.edu/people/pabo/movie-review-data/

[10] http://www.cs.waikato.ac.nz/ml/weka/

[11] Shahana P.H, Bini Omman B. (2015), Evaluation of Features on Sentimental Analysis, Elsevier, Science Direct, Procedia Computer Science 46, PP 1585 – 1592.