International Journal Online of Sports Technology & Human Engineering.
ISSN: 2349-0772 | Volume IV Issue IV August 2017| www.ijosthe.com

Sentiment Analysis of Movie Review using Machine Learning Approach

Rajul Rai[1] Pradeep Mewada[2]
M.Tech Scholar[1]Professor[2]
Department of CSE
TIT, BHOPAL
rajul.rajul.rai@gmail.com

Abstract. With development of Internet and Natural Language processing, use of regional languages is also grown for communication. Sentiment analysis is natural language processing task that extracts useful information from various data forms such as reviews and categorize them on basis of polarity. One of the sub-domain of opinion mining is sentiment analysis which is basically focused on the extraction of emotions and opinions of the people towards a particular topic from textual data. In this paper, sentiment analysis is performed on IMDB movie review database. We examine the sentiment expression to classify the polarity of the movie review on a scale of negative to positive and perform feature extraction and ranking and use these features to train our multilevel classifier to classify the movie review into its correct label. In this paper classification of movie reviews into positive and negative classes with the help of machine learning. Proposed approach using classification techniques has the best accuracy of about 99%.

Keywords: Text Review, Emotions, Feature Extraction, Classification, Emotion recognition.

1 Introduction

Emotions are an important aspect in the interaction and communication between individuals. The exchange of emotions through text messages and posts of personal blogs poses the informal kind of writing challenge for researches. Extraction of emotions from text will applied for deciding the human computer interaction that governs communication and many additional [1]-[3]. Emotions is also expressed by a person's speech, facial and text primarily based emotion respectively. Emotions are also expressed by one word or a bunch of words. Sentence level emotion detection technique plays a vital role to trace emotions or to look out the cues for generating such emotions. Sentences are the essential info units of any document. For that reason, the document level feeling detection technique depends on the feeling expressed by the individual sentences of that document that in turn depends on the emotions expressed by the individual words. The sentiments from various texts and classifies them accordingly into positive, negative or neutral classes.

Sentiment Analysis is an innovation that will be very important in the next few years. High quality data content can be mined with the help of opinion mining. A significant part of the

early research in this field was based on product reviews [1], defining sentiments as positive, negative, or neutral. Most sentiment analysis considers are currently centered around online networking sources, for example IMDB, Twitter [2] and Facebook, requiring the methodologies be customized to serve the rising interest of opinions in the form of text. Besides, performing the phrase-level analysis of movie reviews proves to be a challenging task.

2 Related Work

Godbole, Manjunath and Stevens in their work [5] display a framework that measures positive or negative sentiment to each particular substance in the text corpus. Their framework comprises of two stages, a sentiment acknowledgment stage where opinion expressing elements are resolved and a scoring stage where a relative score for every substance is resolved. In the work by Annett and Kondark [6] it was determined that ML method of sentimental analysis on movies reviews is very fruitful and it was additionally watched that the sort of highlights that are picked dramatically affect on precision of the classifier. As there is an upper bound on the precision level that a reference based approach has as shown in lexical approach.

Pang & Lee work [7] is thought to be a standard in sentimental analysis of movie review. They consider the issue of ordering archives not by topic, but rather by overall sentiment, e.g. deciding if a review is good or bad. They inference, that traditional machine learning methods gives preferable outcomes over human-created baselines. In any case, the three machine learning techniques they utilized (Naive Bayes, Maximum entropy Classification, and Support vector machines) don't give as effective outcomes on sentiment grouping as on traditional classifier based classification. They additionally extricating these portions [8] and executing productive systems for discovering least cuts in graphs; this significantly supports liberality of cross-sentence relevant imperatives, which gives an effective intends to coordinating inter sentence level logical data with customary dictionary of words features.

Singh et al. [9] presents experimental analysis on SentiWordNet approach for performance evaluation for document level sentiment arrangement of Movie audits and Blog posts. Researchers performed variation in semantic features, scoring schemes and thresholds of SentiWordNet approach alongside two most important machine learning approaches i.e. Naive Bayes and SVM. The similar execution of the methodologies for both movie as well as blog reviews is represented through standard execution assessment measurements of Accuracy, F-measure and Entropy.

Tirath Prasad Sahu et al. [10] extracted that features which are strongly effective in deciding the extremity of the movie reviews and used computation linguistic methods preprocessing of the information. Feature impact analysis is also performed by researchers in this paper by computing information gain for each feature to derive a reduced feature set. Six classification techniques are analyzed on this technique and found that Random Forest outperforms an accuracy of 88.95%.

2.1 Machine Learning Approach

For each extracted features of emotion classification algorithm is applied on different set of inputs. Different classifiers are discussed below:

Support Vector Machine (SVM): SVM, a binary classifier is a basic and productive calculation of machine learning calculations, and is generally utilized for design acknowledgment and classification issues. SVM have a very good classification performance compared to other classifiers. SVM uses the kernel function to transform input data into feature space.

K Nearest Neighbor (KNN): A more general version of the nearest neighbor technique bases the classification of an unknown sample on the "votes" of K of its nearest neighbor instead of on solely it's on single nearest neighbor. Among the assorted ways of supervised statistical pattern recognition, the nearest Neighbor is that the most ancient one, it doesn't concerned with a priori assumptions about the distributions from that the training examples are drawn. It involves a training set of all cases. A new sample is classed by calculating the distance to the nearest training case, the sign of that time then determines the classification of the sample. Larger K values facilitate reduce the consequences of noisy points among the training data set, and therefore the alternative of K is usually performed through cross validation.

Neural Network Algorithm: In neural network information and target data are loaded. Input data here may be a matrix of the features extracted from the text inputs. Target data indicates the emotional states of those inputs. Next, the share of input data into three classes particularly training, validation and testing is chosen arbitrarily. The training set fits the parameters of the classifier i.e. finds the best weights for every feature. Validation set tunes the parameters of a classifier that's it determines a stop point for training set. Finally check set tests the ultimate model and estimates the error rate. The default value sets training in 70 % and 15 % each for the remainder. At first the default values are used. Next, the amount of hidden layers is chosen specified, more the number of hidden layers, more complicated the system, higher the result. In conclusion the network is trained several times.

Random Forest: Random forest is found as best model for prediction. It's learning methodology for classification, regression. Multiple decision trees are created at training time and outputting the categories or prediction. Random forest applies bootstrap aggregation technique that decorrelates the trees by showing them totally different training sets. For every tree, a set of all the features are often used. Because the variety of decision tree will increase, the variance of the model is often greatly down and Accuracy will increase. In Random Forest, 2 main parameters are thought of i.e. variety of trees and variety of features they choose at every decision purpose. Accuracy of prediction will increase as a lot of variety of trees creating decisions. RF improves prediction accuracy as compared to single trees. RF handles larger numbers of predictors and it's quicker to predict. RF found to overfit for a few datasets with noisy classification tasks. Large number of trees might create the algorithm slow for real-time prediction.

3 Methodology Used

The proposed algorithm first of all builds a corpus of movie review.

3.1 Preprocessing

Data preprocessing and cleaning step is vital for ensuing analysis. Preprocessing incorporates removal of additional images. Stemming is additionally done as a piece of information preprocessing. The motivation behind the movie reviews pre-preparing stage is to pre-process the review information in such way that it removes unnecessary terms and characters so as to reduce the size of data to efficiently capture the effective results. The initial move towards Pre-processing stage is called tokenization. Thereafter the pre-processing phase removing of unnecessary terms are done such as removal of numeric characters, removing syntactical tokens (e.g. semi-colons, colons), removal of hash tag, etc.

3.2 Post-Processing Phase

Once pre-processing is applied, the pre-processing Module creates the Feature Vector that represents the movie review dataset.
ANFIS based classification phase: This module takes as input matrix V(n*k) and applies a neuro-fuzzy learning algorithm. The learning algorithm works in two sequential steps.
Step 1: Fuzzy C-Means (FCM) clustering
It is applied to generate a collection of clusters where each cluster contains the review file characterized by a similar collection of identifiers (i.e. the terms found in review files after pre-processing). One of the processes to divide data elements into classes is data clustering. In data clustering, items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed. Formally, let V(n*k) me be the n* k reviewfile by dimension matrix containing review file vectors vi such that V = [v1; v2; v3; : : : ; vn], where each vi file vector contains k number of features selected. Let c be the number of clusters and n be the total number of review files such that 2 <= c < n. Matrix V is input into the FCM algorithm which returns a list of cluster centers X = x1; : : : ; xc and a membership matrix U, where each element holds the total membership of a review file vk belonging to cluster ci. FCM updates the cluster centers and the membership grades of each review file using the objective function.

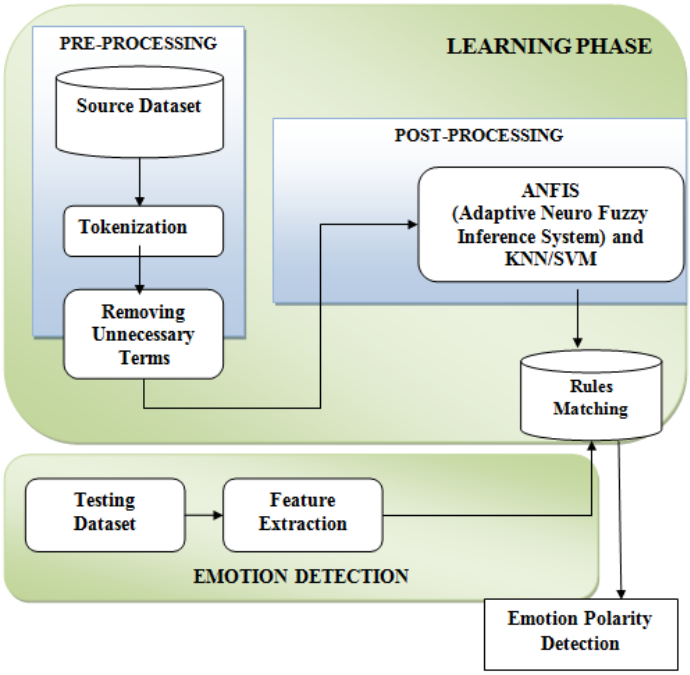

Fig 1. Proposed Polarity Detection Approach

Step 2: Multilevel Classification
Fuzzy clustering is using in inference models based on adaptive neuro fuzzy systems. An ANFIS is a kind of artificial neural network that is based on Takagi Sugeno fuzzy inference system. Since it integrates both neural networks and fuzzy logic principles, it has potential to capture the benefits of the both in a single framework. Then three different classifiers i.e. SVM, KNN and Random Forest (RF) performance is analyzed. The proposed algorithm decides the threshold scoring scheme that will classify the given text into different class of sentiments.

3.3 Detection Phase

In this phase a new review polarity is tested that it is positive or negative. During analysis, a new file is first of all uploaded and further preprocessing phase is carried out. During preprocessing features are extracted. Further multilevel classifiers rules are used to predict the polarity of the reviews.

3.4 Performance Evaluation Parameters

Recognition Accuracy : This measure signifies the recognition accuracy in percentage for each known test text input to the total trained emotional text data and is given by [10]:
Accuracy = correct/predictions
Precision Rate: It is defined as the ratio of correctly recognized emotions for each class to the correctly recognized emotions for all the classes and is given by:

Correctly recognized emotions for a class

$$\text{Precision} = \frac{\text{Correctly recognized emotions for all class}}{\text{}}$$

F-Measure: The F-Measure is the merit of combination of precision rate and recall. The performance of the implementation was evaluated from this factor to obtain the overall performance of the system in term s of correct results i.e. by not considering the wrong recognition observations and is given by:

F-Measure= 2 * [(precision * recall) / (precision + recall)]

4 Result Analysis

On the evaluation of the previous studies on sentimental analysis of movie reviews, the work by Tirath Prasad [10] which is one of the earlier work done in this domain achieve an accuracy of 88.95%. The comparison has been performed on evaluation measures discussed above. It is a clear observation from the Table 1 and fig 2-6 that random forest as well as KNN has the highest value among all other classification techniques where else SVM has the least values.

Table 1. Evaluation Measure

Fuzzy C Mean Clustering				
Classification Technique	Precision	Recall	Accuracy	F_Measure
Existing Work	0.892	0.89	88.95%	0.89
KNN	0.9979	1	99.89%	0.9989
SVM	0.9978	0.9798	98.87%	0.9888
RF	0.9979	1	99.89%	0.9989
K-Mean Clustering				
Classification Technique	Precision	Recall	Accuracy	F_Measure
KNN	0.2278	0.0191	47%	0.0353
SVM	0.4542	0.7739	41.42%	0.5724
RF	0.6629	0.1231	52.39%	0.2077

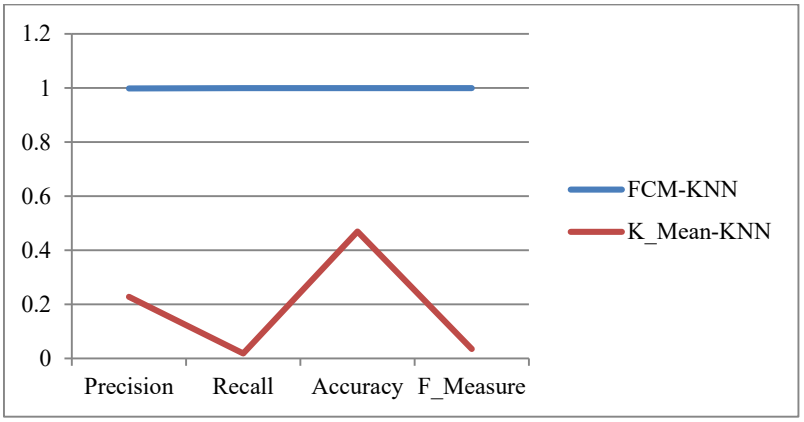

Fig 2. Comparative Evaluation Measure of FCM-KNN and K_Mean-KNN

Graph shows the comparative result analysis of Precision, Recall, Accuracy and F_Measure for FCM-KNN and K_Mean-KNN and concluded that FCM-KNN has better performance as compared to K_Mean-KNN.

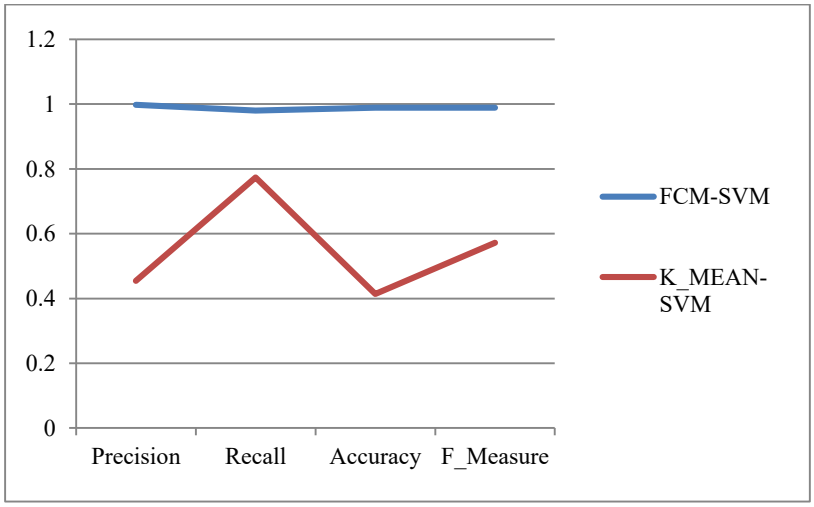

Fig 3. Comparative Evaluation Measure of FCM-KNN and K_Mean-KNN

Graph shows the comparative result analysis of Precision, Recall, Accuracy and F_Measure for FCM-SVM and K_Mean-SVM and concluded that FCM-SVM has better performance as compared to K_Mean- SVM.

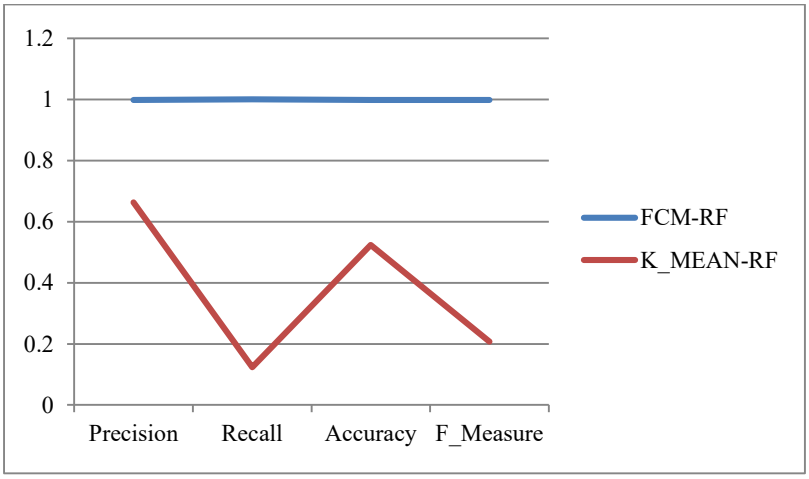

Fig 4. Comparative Evaluation Measure of FCM-KNN and K_Mean-KNN

Graph shows the comparative result analysis of Precision, Recall, Accuracy and F_Measure for FCM-RF and K_Mean-RF and concluded that FCM-RF has better performance as compared to K_Mean-RF.

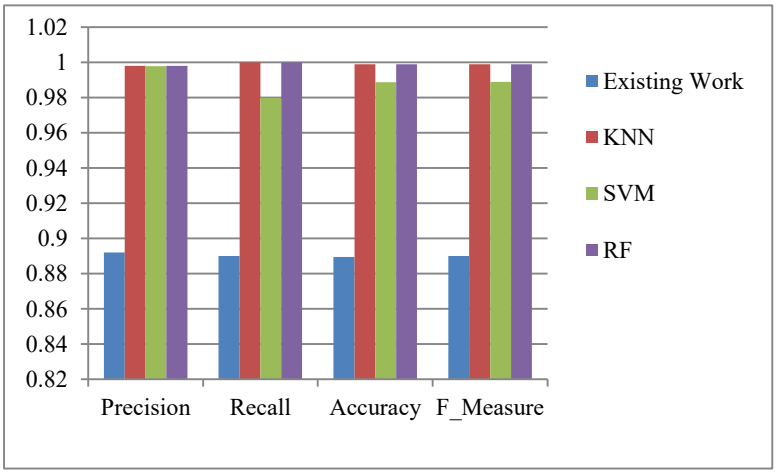

Fig 5. Comparative Evaluation Measure with Existing Work

Graph shows the comparative result analysis of Precision, Recall, Accuracy and F_Measure for FCM-RF has better performance as compared to existing work [10].

4 Conclusion

In business decision making system opinion mining plays a significant role. In this paper, efficient feature extraction is performed that have strong impact while determining the polarity of the movie reviews i.e. positive or negative. Further the feature impact analysis is performed to generate feature set for effective results. Fuzzy clustering based classifier i.e. KNN, SVM and RF approaches are a suitable solution for detecting polarity due to their capability to capture the qualitative and semantic elements of similarity. Among different classification techniques, it is found that the highest accuracy was achieved by Random Forest with an accuracy of 99.89%.

References

1. Amitava Das, Sivaji Bandopadaya, “SentiWordnet for Bangla”, Knowledge Sharing Event -4: Task, Volume 2,2010.
2. Amitava Das, Sivaji Bandopadaya, “SentiWordnet for indian language”, Workshop on Asian Language Resources, pp. 56-63, Beijing, China, 21-22 August 2010.
3. Piyush Arora, Akshat Bakliwal, Vasudev Verma, “Hindi Subjective Lexicon Generation using WordNet Graph Traversal”, IJCLA vol. 3, no. 1, pp. 2539, 2012.
4. Namita Mittal, Basant Aggarwal, Garvit Chouhan, Nitin Bania, Prateek Pareek, “Sentiment Analysis of Hindi Review based on based on Negation and Discourse Relation”, International Joint Conference on Natural Language Processing, pp 45-50, 2013.
5. Godbole, Namrata, ManjaSrinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." ICWSM 7 (2007): 21.
6. Annett, Michelle, and GrzegorzKondrak. "A comparison of sentiment analysis techniques: Polarizing movie blogs." Advances in artificial intelligence. Springer Berlin Heidelberg, 2008. 25-35.
7. Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
8. Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.
9. Singh, V. K., et al. "Sentiment analysis of movie reviews: A new feature based heuristic for aspect-level sentiment classification." Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on. IEEE, 2013.
10. Tirath Prasad Sahu and Sanjeev Ahuja, “Sentiment Analysis of Movie Reviews: A study on Feature Selection & Classification Algorithms”, IEEE, 2016.