ABSTRACT

Sentiment Analysis (SA) is the current field of research in text mining field. SA is detecting opinions, sentiments, and subjectivity of text. It is the application of natural language processing techniques and text analytics to identify and extract subjective information from the frequently used sources such as web and microblogs. The main objective of sentiment analysis is to analyse reviews of products and services, and determine the scores of such sentiments. The major problem is that the reviews are mostly unstructured and thus, need classification or clustering to provide meaningful information for future use.

Most of the supervised approaches in sentiment analysis fail to produce effective mining results, when trained in one domain and applied to others. Another major problem is that most of the works concentrate only on detecting the overall sentiment of any document and do not perform in-depth analysis by ignoring the latent topics and the sentiment associated with those topics. This problem is addressed in this research work. Different methods are proposed to overcome the above-mentioned problems. Hence, the objective is to improve the classification accuracy with multi-classifier systems on IMDb and multi-domain datasets. Multiple Classifier or Multi-Classifier Systems (MCSs)(Ensemble Learning) fuse together multiple classification outputs for better accuracy and classification. This can result in slight differences within the results obtained by individual classifiers. When different classifiers are combined in a proper method, the combined result gives the average of best performing classifier within the ensemble of classifiers. It is used to achieve the best possible classification, by increasing the efficiency and accuracy of classification.

This research work presents a survey of several machine learning techniques to enhance the classification accuracy in sentimental analysis. The

accuracy of these methods is examined to assess their performance based on metrics such as precision, recall, and accuracy. Naïve Bayes classifier (NB) with Joint Sentiment Topic (JST) model is proposed to perform topic modeling with sentiment classification. To overcome the problem of domain independence weakly-supervised approach is used. Topic detection is combined with document-level sentiment classification in the joint sentiment-topic model to provide more useful sentiment–topic mining results. Naïve Bayes algorithm is the most commonly used machine learning algorithm. It is a fast method to develop statistical predictive models. Naïve Bayes model is a simple probabilistic model based on the Bayesian theorem, which analyses the relationship between every attribute. The probability of each class is computed based on the number of occurrences of features in the training dataset. The class with the highest probability for the selected feature is chosen as the resultant class. By considering unigrams and bigrams, the Naive Bayes algorithm with the JST model gives better results when compared to the JST with LDA

The second work is the Maximum Entropy Discrimination Latent Dirichlet Allocation (MEDLDA) with the JST model, which provides a unified framework for classification. The main advantage of using MEDLDA is to identify the topics that are not relevant and do not provide useful information for classification. It is a supervised topic model in which the classifier and the latent topic representation are learnt from the text, by integrating the max-margin principle of Support Vector Machine (SVM) with a hierarchical Bayesian topic model (LDA). MEDLDA optimizes a single objective function on a set of marginal constraints. Classification by MEDLDA model directly optimizes the margin without normalization, which makes learning easier compared to other supervised LDA models. MEDLDA model can be applied to both classification and regression tasks.

As the third approach, the ensemble method is proposed to improve machine learning results by integrating different machine learning models into one optimal predictive model. Ensemble model is a model in which two or more related but different analytical models are executed and the results are averaged into a single score to improve the accuracy of data mining application. Bagging, boosting, stacking and voting are some of the frequently used ensemble methods. Among different ensemble methods voting and averaging are the popularly used ensemble methods. Voting is best suited for classification tasks and averaging is generally applied to regression.

Therefore, the ensemble model with the majority voting technique is employed for improving classification accuracy in sentiment analysis. The proposed ensemble classifier uses the results of the previous work along with Adaboosting in SVM for classification. The classification results of these algorithms are given as inputs to the majority voting rule. The prediction for test sentences by each classifier is taken and the final output prediction is declared the one that has received more than half of the votes. The proposed approach gives better accuracy in classification than the individual machine learning algorithms and also the previously proposed hybrid methods.

The performance of the proposed methods is analysed in terms of accuracy, precision, and recall. Naïve Bayes method with the JST model combines the topic and topic sentiment and gives better results compared to the existing models. It is the first step in topic modeling proposed for the classification of combined sentiment and topic detection. In MEDLDA with JST, accuracy is higher when compared to the basic method with reduced complexity in learning. It is a unified framework that combines topic modeling with the max-margin principle of the SVM classifier. The final ensemble classifier is a combination of three classifiers. Classification accuracy alone is not enough to decide on sentiment documents. So, metrics such as precision and recall are also considered as performance measures. It produces 5 to 10% increase in classification accuracy with the other proposed and existing semi-supervised classifiers.