ABSTRACT

Sentiment analysis is a computational study of people's sentiments, attitudes or opinions conveyed in the form of text. Analysis of text information from social media could give interesting insights and results of people's opinions about a product, personality, and service. Learning good semantic vector representations for phrases, sentences, and paragraphs is a challenging and ongoing area of research in natural language processing and understanding.

For this project, I have taken a Large Movie Review Dataset of IMDB, and based on the reviews given by the user, the reviews are classified into a positive review and a negative review.

The first task in this project is to convert movie reviews or words into vector points or data points and then classification is applied to these data points which predicts the given movie review as positive or negative. For classification, I have used four approaches: SVM (Support Vector Machine), Naïve-Bayes, Random Forest, and Bagged Tree approach. These approaches classify the given data points, but they differ to a huge extent in their accuracy. When the data points have been extracted from the words using NLP (Natural Language Processing), then this data is used to train the machine so that the machine can predict the output of the other data which has to be tested which is also known as test data. Thus, the result of the test data will be the output of the machine.

From CV (Cross Validation) scores of different classifiers, I concluded that Random Forest does not suit well in text data, since random forest makes the split based on a randomly selected feature, while text data are sparse vectors and we may select and split based on irrelevant features whereas linear SVM (Support Vector Machine) did the best job of predicting the user's reviews' sentiments or opinions with an accuracy of 88.62%.