Abstract

More than 80% of the information on the virtual world is stored in text form, so text mining has a higher potential than data mining. Sentiment Analysis is the new branch on the text itself, aimed at detecting the polarity in a text, sentence, paragraph, or document. Sentiment analysis is good for analyzing the related movie review that clearly expresses an opinion on a film, such as positive or negative. The study aims to analyze performance sentiments on the Internet movie database of 50,000 film reviews using feature selection information Gain and Decision Tree Classifier. Information Gain can reduce irrelevant features and decision tree has good accuracy in a lot of related research. In order to improve performance and accuracy, it was used by both feature extraction and n-gram models to get features with good grammatical structures at movie review. Testing results are divided by n-gram parameters and the highest performance will be used in comparisons between DT methods with DT(Pruning), SVM, MNB, and RF using IG *Threshold*. The best accuracy obtained by DT classification was 72.69% with the parameters of unigram, bigram and trigram combinations and IG Thresholding 0.1.

Keywords: sentiment analysis, movie review, information gain, decision tree, N-gram, threshold