ABSTRACT

Numerous previous works classified text corpus by topic, sentiment, genre, or author. This investigates a different case of text corpus. The corpus is the tafseer of Holy Quran verses by Al-Jalalayn.

Holy Quran dataset is selected as the corpus for this study because of its content which sometimes is difficult to separate even by human judge. The number of distinctive words is small, but the number of noise words is relatively high. The challenge of classifying the Holy Quran is that there are verses that have implicit meaning. To overcome the lack of ability to recognize implicit meaning in the text, WordNet Thesaurus is used to perform a semantic similarity approach. In this research, several processes to classify a document were performed, which were pre-processing, feature extraction, semantic weighting, classifier training, and evaluation. During feature extraction, produced several features as follows: Term Frequency (TF), Term Frequency–Inverse Document Frequency (TF-IDF), Part-of-Speech Tagging (POSTAG), and Bigram.

The proposed method is performing weight calculation called Document-to-Class semantic similarity. The new measure used in the semantic similarity calculation was a combination of the Wu and Palmer (WUP) method and shortest path semantic similarity method with minor modifications. This was followed with classifier training, where the classification process using a Modified Multinomial Naive-Bayes classifier were performed. The proposed method is to modify the likelihood probability by using a weighted value from a prior process called document-to-class semantic similarity.

During evaluation process, we evaluated the classifier performance using the Holy Quran dataset we created. For comparation, we also used an Amazon review dataset, a Yelp review dataset, and an IMDB review dataset. The measures used in the evaluation process were Accuracy, Precision, Recall, and F1-Measure. The F1-Measures for the Holy Quran dataset using feature combination POSTAG, BIGRAM and TF was 60.5 %. The F1 score for combination POSTAG, BIGRAM and TFIDF was 58.6% and The F1 score for combination POSTAG, BIGRAM and proposed Weighted TF 66.4%.

Keywords: text classification, semantic similarity, feature extraction, Holy Quran *tafseer*