ABSTRACT

Al-Quran is a holy book that covers various areas that can guide Muslims in their life. The Al-Quran could be categorized as a multi-label text to ease the Muslims learning of the Al-Quran. This thesis research discusses the text classification in the case of multi-label datasets on the English translation of the Al-Quran and focuses on handling the imbalanced dataset problem. The classification of this research is built by proposed methods which are Ensemble methods and Resampling. Because of the prior study's successful handling of the imbalanced dataset, those models were selected. Improving poor performance is the objective of this study. Data preprocessing was the first step in this study, followed by resampling and ensemble methods in classification as the main contribution, and the last is measurement of the classification prediction. The way ensemble methods work in classification is to combine the n-base learners to be a strong learner. The F1-score and Hamming Loss are used in this study to evaluate performance. The proposed methods have enhanced the performance of the f1-score and hamming loss based on several experiments. Based the several experiments, the proposed methods have improved the f1-score performance. The random forest can improve the f1-score by 16%, while XGBoost can improve by 14% from the previous study. On the other hand, Resampling gives differences in the result but is not significant. The best performance from several scenario tests can be generated from a combination of the ensemble, Resampling, use of base learners, and k-fold numbers. The combination of Bagging (Random Forest) and oversampling can provide a performance of 50.12% the f1-score and 0.1002 the hamming loss.

Keywords: Imbalanced, Ensemble, Bagging, Boosting, Multi-label, Classification, Resampling