ABSTRACT

This thesis discusses handling imbalanced data in classifying multi-label text in the Bukhari hadith. This research uses the ensemble stacking method because it provides good evaluation results in cases of imbalanced data compared to the single classifier. The proposed method generally combines several predictions from the base learner/single classifier method to get better classification prediction results. In the evaluation, several experimental scenarios are applied to observe the impact of base learner combinations as well as the number of features used in the classification. In addition, the experiment using the proposed method is also analyzed compared to the baseline, which is not applying the ensemble stacking method and makes comparisons with the previous studies. The experiment results show that the combination of three base learners (Random Forest, Gaussian NB, and SVM), which is applied to two thousand max features, gives the best evaluation results compared to other experiments. This combination of base learners resulted in 83,14 % accuracy and 42,92 % of f1-score.

Keywords: classification, multi-label, imbalanced data, ensemble stacking, accuracy, f1-score