ABSTRACT

In this study, researchers intend to conduct a classification of 11th grade high school biology questions using the Naïve Bayes algorithm and Support Vector Machine which have been grouped into 5 topic categories namely Cells, Circulatory System, Defense System, Human Movement System, and Tissues. This research also compares the performance of two classification algorithms, namely Naïve Bayes and Support Vector Machine. This research goes through several stages, the first by going through the data preprocessing stage with the case folding, tokenizing, Stopword Removal, and Stemming processes. Furthermore, the dataset is carried out the TF-IDF process, namely data or term weighting. The SMOTE oversampling method is used by researchers to overcome the imbalance of data from the dataset.Then the use of K-Fold Cross Validation on the dataset with a value of k 10.From the classification results, the performance results obtained using Naïve Bayes classification with SMOTE oversampling have an accuracy of 76%, then Naïve Bayes classification without SMOTE oversampling is 59%, then Support Vector Machine classification with SMOTE oversampling has an accuracy of 70. 59% and Support Vector Machine classification without oversampling SMOTE is the same as using SMOTE which is 70.59%.Based on the accuracy results obtained in this study, the Naïve Bayes algorithm is a better method than Support Vector Machine in classifying questions based on topic categories.

Keywords – ***Problem Classification, Naïve Bayes, Support Vector Machine, SMOTE, oversampling, Cross Validation***