The 2nd Faculty Annual Research Session
Faculty of Applied Science
University of Vavuniya, Sri Lanka
September 15, 2021

Benchmarking performance of text classifiers for imbalanced data

Upeksha, L.
Department of Physical Science
University of Vavuniya
nishadiniliyanage@gmail.com

Yasotha, R.
Department of Physical Science
University of Vavuniya
yaso9yaso@gmail.com

ABSTRACT

A goal of a text classifier is to classify text documents into defined categories automatically. Classification belongs to supervised learning, where the targets are also provided with the input data. Traditional classification methods perform poorly on imbalanced data, especially among the classes and small samples in each class. When developing a new machine-learning algorithm, it is not easy to demonstrate its performance concerning the sample size. The number of samples influences the model training; although machine learning is beneficial for better performance, creating a large-scale human-coded data set is also costly. This research aims at a minimal human-labelled dataset to be used to train classifiers. A collection of 18828 newsgroup posts on twenty different topics were used for performance benchmarking. The dataset contained imbalanced data with most of the classes of 900 samples; however, classes of 'alt.atheism', 'talk.politics.misc', and 'talk.religion.misc' had 559, 543, and 440, respectively. Fourteen different classifiers benchmarked for the performance of text classifiers measured in terms of F1 score; Passive Aggressive (0.979), Random Forest (0.978), Perceptron (0.976), Elastic Net penalty (0.974), Linear SVC with L2 penalty (0.974), Multinomial Naïve Bayes (0.969), Complement Naïve Bayes (0.966), Ridge Classifier (0.965), SGD classifier (0.954), Linear SVC with L1 penalty (0.935), and Linear SVM with L1 (0.945) observed to be higher performance. Relatively, the performances of Bernoulli Naïve Bayes (0.804), and KNN (0.613) were low. The Nearest Centroid (0.319) was found to be the lowest level of classifier in this experiment.

Keywords: Newsgroups dataset, Passive Aggressive, Supervised learning, Text classifier performance.