ID 22

Classification of Research Papers Based on their Subject Field Using an Ensemble Machine Learning Approach

GKE Dilushika[1#] and RAHM Rupasingha[1]

[1]Department of Economics and Statistics, Faculty of Social Sciences and Languages, Sabaragamuwa University of Sri Lanka

#<hmrupasingha@gmail.com>

Research papers are the most important documents for researchers and scholars. Scholars face the challenge of sifting through relevant research data to find papers that align with their educational interests. This abundance of information can complicate the process of identifying valuable insights across various research types. Therefore, finding related research papers becomes a time-consuming process. Here an automatic classification of research papers helps researchers to do their research easily and effectively. This paper introduces a novel method to classify research papers based on subject fields using machine learning. Unlike previous approaches, which rely on abstracts, limited subjects, or individual algorithms, this study analyzed full paper content and expand classification to five disciplines using an ensemble learning approach by combining four individual algorithms. With a dataset of 2000 papers, preprocessed and extracted feature vectors using Term Frequency-Inverse Document Frequency (TF-IDF). The study used five machine learning algorithms namely NavieBayes, Random Forest, Decision Tree (J48), SVM, and ensemble learning. Employing Naive Bayes and ensemble learning, our results demonstrate high accuracy, with ensemble learning surpassing individual algorithms in 5-fold cross-validation. The performance of the classification system was evaluated using metrics such as accuracy, precision, recall, and F-measure, as well as error rates. Results indicate that Naive Bayes exhibited the highest accuracy among individual algorithms, while ensemble learning, particularly through the Majority Voting combination rule outperformed individual algorithms with an accuracy of 94.20%. This study underscores ensemble learning-based machine learning's efficacy in enhancing research paper classification processes and accessing relevant research.

Keywords: *machine learning, ensemble learning, classification, subject field, academic paper analysis*