UNIVERSITY OF WESTMINSTER

INFORMATICS INSTITUTE OF TECHNOLOGY

"ScholarBERT"
Multi-Label Text Classification for Scientific Articles

A dissertation by

P.N.W. Palliyaguru

Supervised by

Mr. Achala Aponsu

Submitted in partial fulfilment of the requirements for the

M.Sc. in Advanced Software Engineering

Department of Computing

May 2022

Abstract

Scientific article publications gained rapid growth in the recent decade due to digitalization. Publications companies, researchers and article readers are concerned about the content discoverability of the articles. This leads to the essential need for efficient extraction of insights from data of the articles. To make the search easier and more relevant, and improves user experience by proper recommendation, it is important to classify article abstract more efficiently.

During the past few years, deep learning pre-training models have led to remarkable breakthroughs for natural language processing. The proposed implementation is a supervised machine learning-based approach. The research was carried out to determine if it was possible to use a multi-labelled article abstract pre-labelled data with existing BERT pre-trained model which were trained on the scientific domain in the form of transfer learning, without compromising the accuracy and performance of the machine learning model. The resulting research outcome was a ScholarBERT deep learning-based pre-training model which is used as a core for an article classification system, in which research domain experts and research have the capability to identify the article categories of given scientific article. An overall accuracy of 82% was achieved during the testing phase of the created ScholarBERT model.

Keywords: Transfer learning, Pre-training models, BERT, supervised learning, Multi label Classification