INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Fine-tuning Pre-Trained Deep Bidirectional BERT for Web Page Classification

A Dissertation by

Ambagahage Sumudu Fernando

Supervised by

Mr. Prasan Yapa

Submitted in partial fulfillment of the requirements for the BEng in Software Engineering degree at the University of Westminster.

April 2023

ABSTRACT

With its vast amounts and wide variety of information, the World Wide Web has become one of the richest and most widely available information sources in the current information-driven society. With the advancements in computer networking technologies and with the accessibility and advancements of the internet, the information on the internet is constantly growing at a fast rate which is beneficial for its users in many ways. However, the excessive information on the internet has become the cause of several problems in the areas of web information management, retrieval and integration, web content filtering, parental control systems, and many more. Also, the excessive amount of information can be detrimental to regular users of the internet.

The research project proposes a novel approach for web page classification using Bidirectional Encoder Representations from Transformers (BERT). BERT uses deep bidirectional self-attention to generate contextual representations for text sequences. Unlike unidirectional or pseudo-bidirectional models, BERT learns the context of a word concerning its surroundings rather than the sequence of words and produces accurate results compared to directional models. As one of the few research approaches using deep learning techniques for web page classification, this novel approach could provide a valuable contribution to the research domain.

The research has conducted extensive experiments to identify the optimal conditions for the research project and has achieved satisfactory results in comparison to the other research approaches in the domain of web classification with limited resources and within a limited time frame.

Keywords: Web Page Classification, Web Classification, Web Documents Classification, Web Page Categorization, BERT, BERT Fine-Tuning, Transformers