
INFORMATICS INSTITUTE OF TECHNOLOGY

**INFORMATICS INSTITUTE OF TECHNOLOGY**

In Collaboration with

**UNIVERSITY OF WESTMINSTER**

# SIMILARS: A System to Detect Correlations Between Two Research Papers Using Natural Language Processing

A dissertation by

**Ms. Gayani Silva**

W1715530 – 2018832

Supervised by

**Mr. Prathieshna Vekneswaran**

Submitted in partial fulfillment of the requirements for the BSc (Hons) Computer Science degree at the University of Westminster.

**May 2023**

# ABSTRACT

People who regularly work with academic papers frequently summarize and compare the contents of research papers, but doing so is an exhausting and challenging task. Due to this problem, it may also cause depression or mental stress in students. Currently, there is no system that can compare the meaning of two documents and provide users with a percentage of how similar the two documents are, despite the fact that there are various apps that can be used to compare the words in two papers.

To solve this problem, it is suggested to investigate a novel approach for measuring semantic similarity for scientific paper texts. This approach is developed by tuning a deep learning transformer-based model called SCIBERT for the semantic textual similarity task. The fine-tuning process is done by training the model on the SICKR-STS dataset and optimizing hyperparameters as required. The final model consists of two phases, combining cross encoder and bi encoder techniques to highlight better results than in previous work.

The proposed system, SIMILARS, is evaluated on the test data of the SICKR-STS benchmark dataset to measure its performance. It is an appropriate statistic for evaluating how well the system performed on the STS task, as the dataset includes similarity scores annotated by human experts. On the other hand, the performance of the model is determined by evaluating predicted similarity using the Pearson and Spearman correlation metrics. The final model has improved the Pearson correlation score from 0.65 before fine tuning to 0.91 after fine tuning.

**Keywords** - Natural Language Processing, Semantic Textual Similarity, Information Retrieval, Encoding Architecture, Transformer Based Models