Informatics Institute of Technology

In Collaboration With

University of Westminster, UK

*University of Westminster, Coat of Arms*

# Neural Machine Translation of Legal Documents for Low Resource Languages

A dissertation by

Ms. Kankanamalage Hiruni Malsha Perera

W1742214 / 2018500

Supervised by

Mr. Lakshan Costa

July 2022

Submitted in partial fulfilment of the requirements for the

BSc(Hons) Computer Science Degree at the University of Westminster

# ABSTRACT

With the rise of globalization and digitalization, the exchange of foreign languages and documentation composed using such languages are commonly exchanged across the globe. Due to such reasons, translation and interpretation of such documents have become crucial. The legal industry is one such domain in which this is of utmost importance. Specifically, in the Sri Lankan context, on most occasions legal documents must be made available in both Sinhala and English. But due to its prolix nature, time and time again it has proven to be a difficult task.

In recent years, neural machine translation systems have taken on the stronghold in the field of automatic translation. With the introduction of self-attention, the utilization of pretrained languages models to withstand the limitations faced by domain specific tasks and lack of data for low resource languages has become prevalent. Such a usecase will be explored in the this research in the field of legal translation.

In conclusion this research will provide a stepping stone for the exploration of Encoder-Decoder architectures and pretrained language models to further improve neural machine translation for domain specific corpora in the case of low resource languages. It will also determine the aspects of capturing semantics present in lengthy prolix text present in legal language.

**Keywords**: Machine Translation, Self-Attention, Language Translation, Low Resource Languages, Neural Networks