INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

SinhalaTextGenie

A Sinhala Language Model Trained To Predict and Generate Words

A dissertation by

Mr. Mohamed Rashad

Supervised by

Mrs. Jayamini Liyanage

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

May 2023

Abstract

As the use of Sinhala in the digital space continues to increase, there is a growing need for advanced Sinhala NLP tools that can enhance user experience and contribute to the language's development. However, Sinhala lags behind privileged languages in terms of NLP tools, lacking reliable features such as automated content generation, next word predictions, chatbots, and auto-replies. To address this gap, it is crucial to develop and contribute to the advancement of reliable Sinhala language models, which can promote the growth of the language in the digital realm.

The transformer models in the space of language modeling have proven to give promising results. Despite this no research has been conducted in developing a Sinhala transformer model for text generation. Therefore this work trains a decoder only transformer model on 300000 of diverse sinhala sentences which contains 9 words on average and shows the impressive results it can achieve. 5 decoder layers and 4 attention heads were used to compile the model.

This study shows the accurate predictions of the Sinhala transformer language model that has been developed.

Keywords: NLP, transformer model, text generation, decoder-only transformer model,, deep learning, word prediction