



**INFORMATICS
INSTITUTE OF
TECHNOLOGY**

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

**SinEX: Sinhala Question Answering System Using Deep
Learning**

A dissertation by

Mr. Sirimega Kavindu Lanarolle

Supervised by

Ms. Niwarthana Kariyabaduge

Submitted in partial fulfilment of the requirements for the BEng (Hons) Software
Engineering degree at the University of Westminster.

July - 2022

Abstract

Sinhala is the National language used by the major ethnic group in Sri Lanka. As the native language of Sri Lanka approximately 16 million people used it as their mother tongue while it is spoken by 2 million people as their second language. Sinhala language has rich script writing system which is derived from Brahmin script. Sinhala have coexisted from generations, where it strongly influenced Sinhala's grammar, vocabulary and phonology. Due to the complexity and low resources Sinhala language lacks proper tools for natural language processing systems and researches. Comprehension is an important aspect in Sinhala due to its complexity and using an AI based solution in Sinhala comprehension will be useful and effective.

In the perspective of derivation in natural language processing domain there are so many cutting edge systems and researches are found in different languages including Chinese and Japanese for comprehension. Due to the poverty in both linguistic and economic capital, Sinhala language lacks proper tools for natural language processing systems and researches. Furthermore, the existing work in Sinhala comprehension systems are not efficient and effective enough for real world use hence a system that is able perform machine comprehension in Sinhala language is a necessity.

The research uses a novel dataset created from different sources to increase the accuracy and applicability of the deep learning model when used in real world scenarios. Deep learning model is developed using the Bidirectional Encoder Representations from Transformers (BERT) model and the system needs the context and the question from the user's end then the relevant answer will be identified by the trained deep learning model. An evaluation matrix has been implemented in the research to identify the performance of the trained model and it was identified that the model has performed well.