INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

SINHALA NEWS SUMMARIZATION SYSTEM

A dissertation by

Mr. Pamudu Prabathiya Bandara

W1761216

2019198

Supervised By

Ms. Sulari Fernando

This Project Proposal is submitted in partial fulfilment of

the requirements for the BEng (Hons) Software Engineering degree

at the University of Westminster

May 2023

ABSTRACT

In this modern era, an enormous amount of data is publicly available on the internet. However, nowadays people don't have much time to read all those information because of their busy work schedules. To overcome that issues, automatic text summarization can be utilized to automatically summarize lengthy articles. Even though NLP text summarization has been evolving in past decades, few research attempts were conducted on the Sinhala text summarization domain since it is a low resource language.

This research is dedicated to the Sinhala text summarization in order to create condensed summaries by excerpting the crucial segments from the source news article. In order to achieve this task TextRank algorithm which is a graph based unsupervised learning technique was used. In this research rather than using a statistical feature to identify the important and relevant information, deep learning techniques was adopted to identify the input text more accurately. To identify the semantic relationships SinBERT model and sentence transformer were adhered with the TextRank algorithm.

The quantitative testing was performed by using ROUGE metrics and F1 score to evaluate the performance of the created solution. For the testing task, machine generated summaries were compared with the human created manual summaries based on the assumption that manual summaries are perfect. The created solution obtained better results for the F1 score when compared to the other Sinhala text summarization projects.

Keywords: Automatic Text Summarization, Sinhala news Summarization, Extractive summarization