ADAPTTEXT : A NOVEL TECHNIQUE FOR DOMAIN-INDEPENDENT SINHALA TEXT CLASSIFICATION

Kodithuwakku Arachchige Yathindra Rawya

A dissertation submitted in partial fulfilment of the requirement for the
Bachelor of Engineering (Honours) degree in Software Engineering

Department of Computing
Informatics Institute of Technology, Sri Lanka
in collaboration with
University of Westminster, UK

2021

Abstract

Text classification facilitates the ability to classify text data into multiple categories by assigning labels. It is a core piece of Natural language processing, consisting of a wide range of use cases, including fake-news detection, sentiment classification, hate-speech / cyberbullying detection, user intent classification, news-article classification, and many more.

Sinhala is being the most used language in Sri Lanka, which is morphologically rich and agglutinative. Therefore it is complex compared with languages like English, which has a simple morphology. Furthermore, Sinhala has its own writing system where the solutions developed for English might not be reusable.

Due to these complexities and being a low resource language, there is no proper generic and automated solution to perform Sinhala text classification. Even the current task-specific text classification approaches have not considered the polysemy of words and were not focused on addressing data scarcity issues.

Therefore, AdaptText has been developed as a novel generic architecture and a technique for text classification in Sinhala. In order to measure the efficiency of the proposed novel technique, cross-domain testing and evaluation have been performed with multiple datasets and againt current best performing approaches. AdaptText could address the discussed research gaps and could achieve state-of-the-art results for both binary and multiclass Sinhala text classification.

Keywords - Natural language processing, Knowledge transfer, Classification algorithms, Supervised learning