SENTIMENT ANALYSIS FOR THE SINHALA LANGUAGE WITH BERT BASED LANGUAGE MODEL

Yalagalage Nemal Shamendu Peiris

A dissertation submitted in partial fulfilment of the requirement for
Bachelor of Science (Honours) degree in Computer Science

Department of Computing

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

2021

Abstract

Sinhala is a low resource language that is spoken by 16 million people in Sri Lanka which is the native language of Sinhalese people. Due to the lack of resources, there are only a minimal amount of researches conducted in the territory of sentiment analysis based on the Sinhala language when compared to other languages like English and Chinese. Most of the existing researches have been conducted by using lexicons and dictionary-based approaches combined with classification algorithms. With the advancements of word embedding and deep learning techniques, recent researches have emerged with utilizing these techniques in the Sinhala language domain for sentiment analysis and text classification tasks. Even more recent developments in the Natural Language Processing (NLP) landscape like Bidirectional Encoder Representations from Transformers (BERT) based language models which have achieved state-of-the-art results for a variety of tasks in the NLP domain haven't been applied to the Sinhala language domain as of now.

Therefore, we introduced a sentiment analysis model for the Sinhala language by using BERT based language model known as Language-agnostic BERT Sentence Embedding (LaBSE). The classification is done using both binary and multiclass dataset consisting of Sinhala news comments. An F1-score of 89.82% for the binary classification and an F1-score of 64.72% for the multiclass classification was achieved by the newly introduced model which surpasses the existing research achievements carried out using deep learning and static word embedding approaches.

Keywords — *Sentiment Analysis, Deep Learning, Language Models*