INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Identifying Hate Speech in Romanized Sinhala in Social Media Comments Using NLP

A Final Project Report by

Mr. Sandaru Kaveesha Munasinghe

(2018161|W1714909)

Supervised by

Mr. Deshan Sumanathilaka

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

July 2022

Abstract

In today's society, the use of harsh language on social media sites is becoming a big concern. The rise of hate speech, in general, is dangerous as it can be threatening, racially motivated, or even ethnically targeted. The use of hate speech is mostly done through social media than it has been said in person. The reason for this is those who have confidence in speaking freely through social media. With confidence, the users that write hate speech as comments are comfortable with doing so, therefore, this causes a flame war in comments with hate speech. The detection of hate speech is very important as this can build a safer environment within social media and adapt to being less influential in using hate speech. Hate speech is multilingual and but most of which is written in Romanized language due to the convenient adaptation of doing so with modern devices than writing in the language. However, considering that a majority writes in Romanized languages, the author chooses to work with Romanized Sinhala to detect hate speech. Currently, much of the research has so far focused on solving it in English. This study aimed to identify hate speech published in Romanized Sinhala and define the gap within the research that is being done.

The main goal of this study is to automatically detect hate content in comments on social media published in Romanized Sinhala using a manually labeled collection of data and NLP technologies. With the use of Deep Learning techniques, a high training accuracy is generated along with a high validation accuracy. BERT text classification is used to ensure success with a training accuracy of 91.99% and a validation accuracy of 95.38%. This proposed solution covers the gap identified which is obtaining high training and accuracy values over the existing hate speech detection system.

Keywords— *Romanized Sinhala hate speech detection, Natural language processing, Deep Learning, NumPy, pandas, BERT, Text Classifications*