BIASBLOCKER – A HATE SPEECH DETECTION SYSTEM FOR TRANSLITERATED SINHALA-ENGLISH CODE-MIXED LANGUAGE

Hashini Kodithuwakku

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

School of Computing

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

2023

ABSTRACT

This study proposes a novel system for identifying hate speech in transliterated language that is a mixture of Sinhala and English. Due to the intricacy of the language and the prevalence of code-mixed languages on social media platforms, it is difficult to identify hate speech in these languages.

The proposed novel system uses two pre-trained transformer models to detect hate speech content in Sinhala-English code-mixed, which is first transliterated and then used to train a hate speech detection model. The proposed approach consists of three components: a pre-processing module, a transliteration module, and a hate speech detection module. These components work together to process the input text, transliterate it into Sinhala, and then classify it for hate speech content.

The suggested approach employs a Sinhala-English code-mixed aggregated dataset with hate speech annotations, and then utilizes a pre-trained transformer model to detect hate speech content. The proposed novel solution has outperformed the existing benchmarks for identifying hate speech content in Sinhala-English code-mixed language over 92% in Precision, Recall, and F1-score. The system can be simply modified to accommodate other low-resource code-mixed languages and aid in the identification of hate speech content on social media sites.

Keywords *- Transliteration, Hate Speech Detection, Code-mixed Language, Sinhala-English Code-mixed Language, Deep Learning*

Subject Descriptors –

Natural language processing → Hate speech detection → Code-mixed language processing

Deep learning → Transformers → Cross-lingual text processing

Computing methodologies → Natural language processing → Text classification