ANTI-ASIAN DETECTION PLATFORM (AADP)

HERATH MUDALIGE THINURA LAKSARA KUMARASINGHE

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

Department of Computing

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

2021

Abstract

In the twenty-first century, people have become more digitalized than before. Microblogging platforms are the primary source of the communication network. Microblogging platforms give users the freedom to express their feelings. After coronavirus streaks, there is an increment in microblogging platforms because this is the only way of connecting to people. Since coronavirus started in Wuhan, people thought that china had created the virus and started discriminating against Asians. There is a rise in online discrimination due to specific incidents that happen in 2020. The prevention of hate content is essential before it could get uncontrollable. This research study forces on classifying hate and counter hate speech towards Asians in microblogging platforms.

Anti-Asian Detection Platform helps develops and researchers to classify hate and counter hate speech with the help of transformation learning. Classifier considers not only text content. It also considers emojis, hashtags before classifying. The classification model is built customizing the BERT model and with the help of a new fine-tuning strategy to the BERT model. Anti-Asian Detection Platforms classification model uses multilayer perceptron with BERT model.

The final model is identified after experimenting with available methods. A new hypothesis is introduced for classifying hate and counter hate speech towards Asians in a pandemic situation. A web service is developed to help developers to integrate a model that can identify how user behaviours. The annotated dataset was produced with new hashtags that can be used to fetch hate and counter hate content. Anti-Asian Detection Platform has outperformed similar systems. The classification model was evaluated with recall, precision and F1 score. The following are the results for the categories, hate F1 score is 0.62, the counter hate F1 score is 0.76, and Neutral F1 score is 0.73. Anti-Asian Detection is further evaluated with the help of domain and technical experts.

Keywords: *Natural language processing, Neural networks, Feature selection, Transfer learning*; *Supervised learning*