A DEEP DECISION FORESTS MODEL FOR HATE SPEECH DETECTION

Ndenga K. M.[1]

Kirinyaga University, KENYA

Correspondence: *kmalanga@kyu.ac.ke*

ABSTRACT

Detecting and controlling propagation of hate-speech over social media platforms is a challenge. This problem is exacerbated by extreme fast flow, readily available audience, and relative permanence of information on social media. The objective of this research is to propose a model that could be used to detect political hate speech that is propagated through social media platforms in Kenya. Using Twitter textual data and Keras Tensor Flow Decision Forests (TF-DF), three models were developed that is, Gradient Boosted Trees with Universal Sentence Embeddings (USE), Gradient Boosted Trees, and Random Forest respectively. The Gradient Boosted Trees with USE model exhibited a superior performance with an accuracy of 98.86%, recall of 0.9587, precision of 0.9831, and AUC of 0.9984. Therefore, this model can be utilized for detecting hate speech on social media platforms.

Keywords: - *Hate Speech Detection, Tensor Flow Decision Forests, Gradient Boosted Trees, Universal Sentence Embeddings, National Cohesion and Integration Commission.*