Profanity word detection for Sinhala language using deep learning

Kumara S.K.H.R.S.[1*] and Jayaneththi J.K.D.B.G.[2]

1Department of ICT, Faculty of Technology, University of Ruhuna, Sri Lanka.
2Department of ICT, Faculty of Technology, University of Ruhuna, Sri Lanka.

In the present world, content censorship is an important concept. When it comes to the Sinhala language, several studies have been conducted on text-based content censorship methods, but not for audio content. The Sri Lankan government prohibits the use of profanity in public media. Therefore, Sri Lankan media companies must check their videos for profanity before telecasting. Till now, this process has been done manually, and it is extremely difficult with long videos and audio clips. This study suggests developing a deep learning model that can automatically find profanity words in Sinhala audio files. The ten profanity words were selected and audio samples from 100 people were gathered. The data was preprocessed, transformed into spectrogram images, and applied to a convolutional neural network (CNN) to develop the profanity filter model. By converting audio files to spectrograms and applying image processing to extract the features from the dataset, the model predicts the profanity words. This paper addresses the procedure of the mentioned process and its capabilities with upcoming updated versions of the final product.

Keywords: profanity detection, Sinhala language, deep learning, audio processing.

*Corresponding author: rhameesha@gmail.com