ID 14

Hatred Comments Detection in Twitter using Deep Learning

KR Hingurage# and DS Vithanage

Department of ICT, Faculty of Technology, University of Ruhuna, Sri Lanka

#hingurage.kr@gmail.com

Social media applications are the most popular web and mobile applications across the globe. In the meantime, however, this has resulted in the emergence of conflict and hatred by making online environments, particularly Twitter, uninviting for users. This issue typically affects individuals, organizations and governments because it can have a reasonable or unreasonable impact on someone's reputation, as well as could trigger discrimination, hostility and violence, which can lead to or include terrorism or atrocious crimes in society. Therefore, an accurate, efficient automatic model to detect and classify hate speech on Twitter is a particularly useful tool for the relevant authorities. This paper describes the detection and classification of hate speech on Twitter by using deep learning. This can fill in the gaps between current models with higher accuracy and reliability. Thus, this research is beneficial in several ways, such as the detection of hate speech in distinct categories, such as in toxic, severe toxic, obscene, threat, insult and identity hate. The developed application used deep-learning algorithms to find the number of occurrences of the words and semantic words. The LSTM model is used to train the data set and to get the probability values. The classes of hate speech were calculated against the training data set and were found to be above 72%. In conclusion, the developed method can help to detect and classify hate speech into six classes on Twitter.

Keywords*: social media, Twitter, hate speech*