Toxic Comment Classification Using Machine Learning

L.A.S. Pramodya, R.M.G.U. Rathnayaka, K.K.S. Lahiru and D.R.V.L.B. Thambawita

Computer Science and Technology, Uva Wellassa University, Badulla, Sri Lanka

Comment classification models are available today for “flagging” the comments. However, determining whether or not a comment should be “flagged” is difficult and time-consuming. Another major problem is the lack of sufficient data for training the model, and there are some issues with the available datasets because those are annotated by the human raters and those annotations are dependent on their personal beliefs. Lack of multi-label comment classification model causes for issues of abusive behavior. This paper presents models for multi-label text classification for identifying the different level of toxicity within a comment. In this paper, we use Wikipedia comments which have been labeled by human raters for toxic behavior provided by Kaggle. Comments have been categorized into six categories as toxic, severe-toxic, obscene, threat, insult, and identity-hate. The dataset contains 159572 comments. For data analyzing we use python seaborn library and python matploitlib library. It is understood that the dataset is highly skewed. Most of the comments do not belong to any of the six categories. Researchers used undersampling for majority class to correct the bias in the original dataset. We tested three models: a feed-forward neural network with Keras and word embedding, a Naive Bayes model with Scikit-Learn, and a LightGBM with 4-fold cross-validation. For the neural network, it took 3.5 hours to be trained on Nvidia GeForce 840M which is having 384 CUDA cores, Naive Bayes model with Scikit-Learn took 3 hours where LightGBM with k-fold took 4 hours. Researchersran 100 epochs from each model. At the end of 100 epoch, the neural network gave 0.9930 of validation accuracy and loss was just 0.2714, Naive Bayes model with Scikit-Learn gave 0.9556 validation accuracy and loss was 0.4121 where LightGBM with k-fold accuracy was 0.9000 and validation loss was 0.4263. The neural network gave the best accuracy at the end of the 100th epoch.

Keywords: Comment classifications, Deep neural networks, Machine learning, Naive bayes, LightGBM, Keras, Scikit-Learn