Data Mining Event & Application

Topic: Toxic Comment Detection
Data Set: Social Media
Student: Sai Akhilesh Potharaju
Objective: Hate Speech

The data has been collected from
https://conversationai.github.io/research.html

The Hate Speech aims at Removing any textual material containing hate or toxicity, online bullying has become a serious issue in recent years. Removing hate speech makes it easy for everyone to express their opinions freely over the internet. The applications are social networking, online meetings, chatbot training and threatening messages.

The objective is to classify it as belonging to one or more of the following categories: clean, obscene, threatening, insulting, toxic, severely toxic and identity hate.

Dependent Variable : Identity hate
Independent Variables: id, comment text, toxic, severe toxic, obscene, threat, insult, identity hate

The training models that I have used are Binary Relevance, Classifier chains, Support Vector Machines, XGBoost, Extremely Random Trees, Recurrent Neural Networks, and Transfer Learning. The best training models are Decision trees, decision tree ensembling and neural networks.

So the conclusion is classical models like SVM and Logistic Regression fail to achieve a high AUC_ROCscore. Tree-based ensembling methods and Recurrent Neural networks can achieve a very high AUC_ROCscore. Further improvements can be achieved by experimenting with complicated neural networks and the use of state-of-the-art Transformers.