Team: Fei Jia(jf07), Baige Liu(liubaige)

Mentor: Kevin Clark

Problem Description: Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Several platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

Our mission is to build a multi-headed model that's capable of detecting different types of of toxicity like threats, obscenity, insults, and identity-based hate.

Data: A large number of Wikipedia comments which have been labeled by human raters for toxic behavior.

Baseline: GloVe Embedding + LSTM-RNN model.

Evaluation Methodology: We evaluate the baseline model using cross-validation.(Train on 143613 samples, validate on 15958 samples) We achieve a training accuracy of 98.30% and a validation accuracy of 98.23% .

Results: The baseline model works very well. And by manually looking at the test results by ourselves, the model predictions look very reasonable. For our next step, we will try attention model and incorporate character-level modelling.