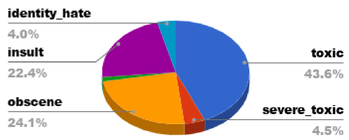
Toxic Comment Classification

Fei Liu Xiaoyan Wu
https://github.com/wxy1224/cs224n_project
Stanford University - Department of Computer Science

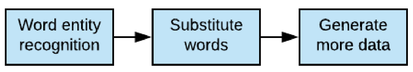
Dataset

- Six labels for each comment
- Training data: 160K comments



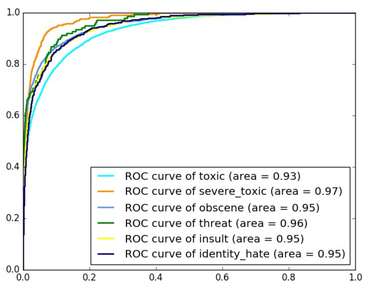
Preprocessing

- Split training and validation data
- Generate embeddings and use pretrained embeddings
- Data augmentation



Benchmark: NB-SVM

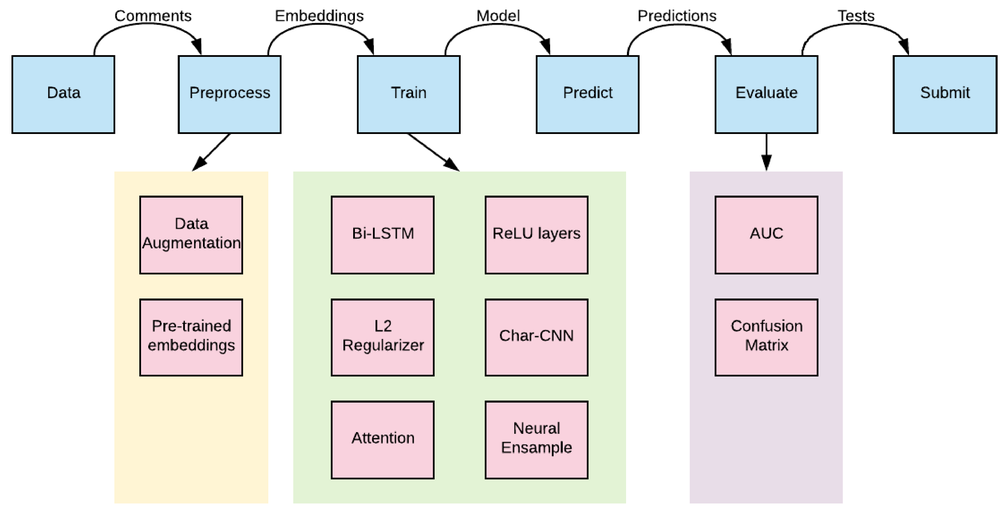
- Naive Bayes features + SVM
- Bottleneck: toxic and threat



Project Description

Many online comments could be toxic, and our project aims to classify these comments into multiple of the following categories: **toxic**, **severe_toxic**, **obscene**, **threat**, **insult**, and **identity_hate**.

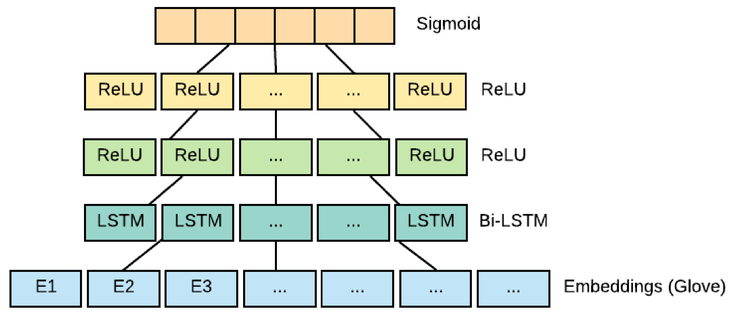
Pipeline



- The system is built based on Bidirectional LSTM.
- ReLU is used in order to capture non-linearity
- Pretrained embeddings are 300-dimension GloVe features
- L2 regularizer to prevent overfitting
- Character-level convolutional networks to handle words that do not appear in dictionaries
- Neural Ensemble to combine multiple models with different configurations

Best Model

- Build on Bi-LSTM
- Implement data augmentation on "PERSON" name entity on "identity_hate" labeled data
- Use small (0.01) L2 regularizer
- Use two ReLU layers
- Use GloVe pretrained embeddings



Accuracy

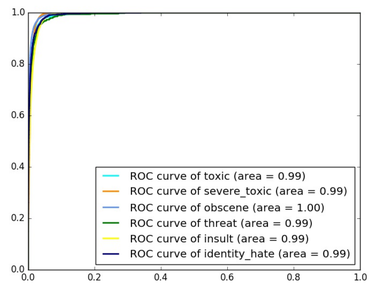
Model Name	AUC Score
NB-SVM	0.7612
Char-CNN	0.4991
Attention	0.7505
L2 Regularizer	0.9648
ReLU Layers	0.9728
Best model	**0.9782**

Table: Best Model AUC Breakdown

Label Name	AUC Score
toxic	0.9918
severe_toxic	0.9907
obscene	0.9946
threat	0.9609
insult	0.9901
identify_hate	0.9712

Error-analysis

- Handling special characters/words not in dictionary



Best Model AUC Curves