An Ensemble Approach to Toxic Comment Classification

Matthew Jung & Alex Hua
Simon Fraser University, 8888 University Dr., Burnaby, BC, Canada
{mjj8, aghua}@sfu.ca

Introduction

Motivation

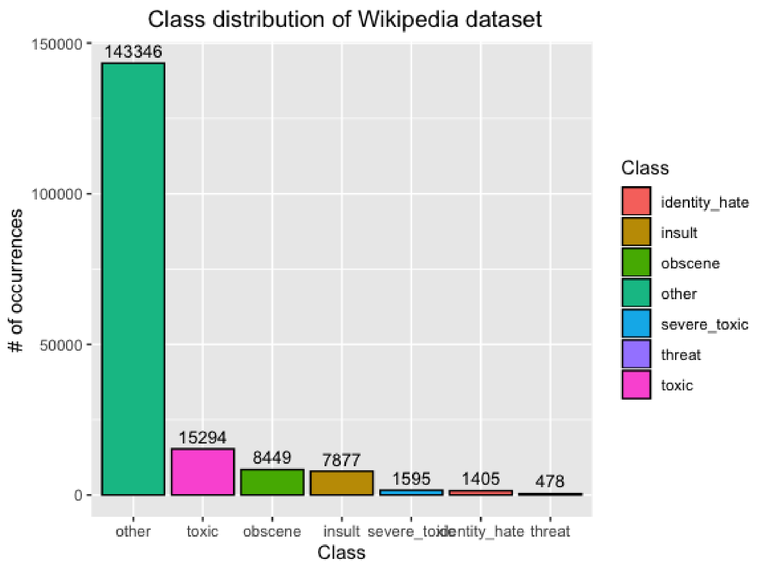
- Toxic comments can be the crux of online platforms and forums
- Automatic classification of toxic comments, such as hate speech, threats, and insults, can maintain peaceful conversations

Contributions

- We propose a novel approach using an ensemble model composed of a bidirectional LSTM (biLSTM) and a Naïve Bayes Classifier
- Our model extends (van Aken, 2018)'s biLSTM methods [1]

Data Set

- We analyze the dataset published by Google Jigsaw in December 2017 for the Kaggle Comment Classification Challenge [2]
- Contains 159,571 Wikipedia comments annotated by human raters as 'toxic', 'severe_toxic', 'obscene', 'threat', 'insult', or 'identity_hate'
- Comments can be associated with multiple classes at once
- Split dataset into train and test: train size = 127,656 (80%). Test size = 31,915 (20%)
- Comments are mostly written in English. Includes some outliers, e.g., in Arabic, Chinese, or German



Comment Example:

You sir are an imbecile, and a pervert.

1 toxic 0 severe_toxic 0 obscene 0 threat
1 insult 0 identity_hate

Methods

Baseline 1: Naïve Bayes Classifier

$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$ Bayes Theorem

$$P(y|x_1, ..., x_w) = \frac{P(y)\prod_{i=1}^{w} P(x_i|y)}{P(x_1)P(x_2)\cdots P(x_w)}$$

Above,

- P(y | x) posterior probability - the probability of y being TRUE given that x is TRUE
- P(y) prior probability - the probability of y being TRUE
- P(x | y) likelihood probability - the probability of x being TRUE given that y is TRUE
- P(x) the probability of x being TRUE

Baseline 2: biLSTM

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$
$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (4)$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (5)$$
$$h_t = o_t \cdot tanh(C_t) \quad (6)$$

Ensemble:

$$P(Embl_{i,l}) = \frac{P(MNB_{i,l}) + P(biLSTM_{i,l})}{2}$$

$$Embl = \{P(Embl_{i,l}) | 1 \leq i \leq length(comments), 1 \leq l \leq length(labels)\}$$

Experiments

- Split input data into training and test sets
- Created embedded language model from GloVe, for biLSTM
- Created biLSTM model
- Created Bag of Words using tf-idf, for Naïve Bayes Classifier
- Created Naïve Bayes Classifier
- Created Ensemble model from the results of biLSTM output and Naïve Bayes Classifier
- Evaluated Ensemble model by calculating the area under the receiver operating characteristic curve (ROC_AUC)

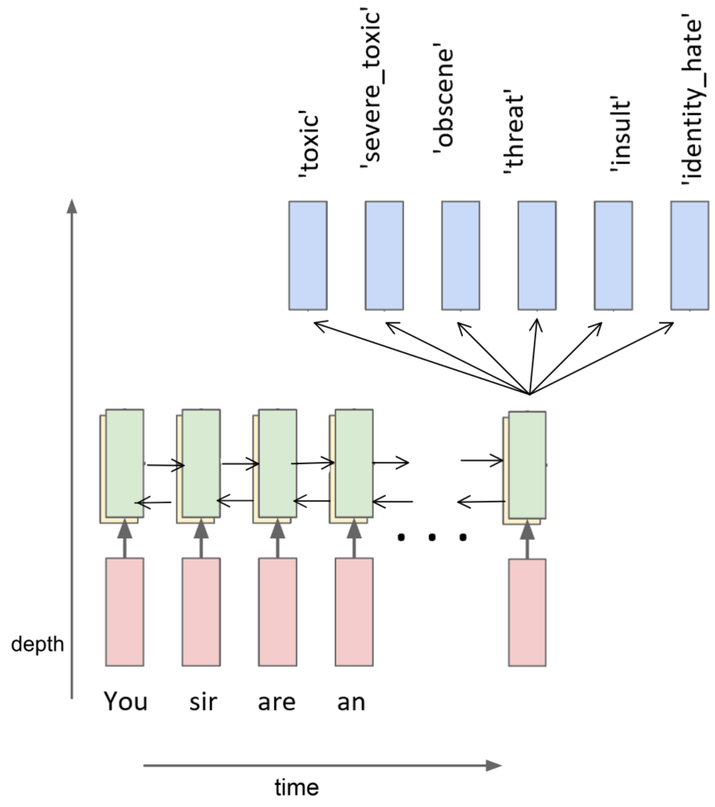
Examples

Baseline 1: Naïve Bayes Classifier

$$P(insult|x_1, ..., x_n)$$

$$= \frac{P(You|insult)P(sir|insult)\cdots P(pervert|insult)P(insult)}{P(You)P(sir)\cdots P(pervert)}$$

Baseline 2: biLSTM



Results

Baseline 1: Naïve Bayes Classifier

Achieved score of 0.9258

toxic	severe_toxic	obscene	threat	insult	identity_hate
0.038310397	0	0	0	0.030952359	0
0	0	0	0	0	0
0	0	0	0	0	0

Baseline 2: biLSTM

Achieved score of 0.9724

toxic	severe_toxic	obscene	threat	insult	identity_hate
0.05996269	0.004004822	0.032855547	0.000965544	0.028431965	0.002759394
0.005259011	8.35E-05	0.001948994	2.15E-05	0.001823005	0.000108382
0.010651516	0.000260439	0.004958904	4.16E-05	0.003635139	0.00023117

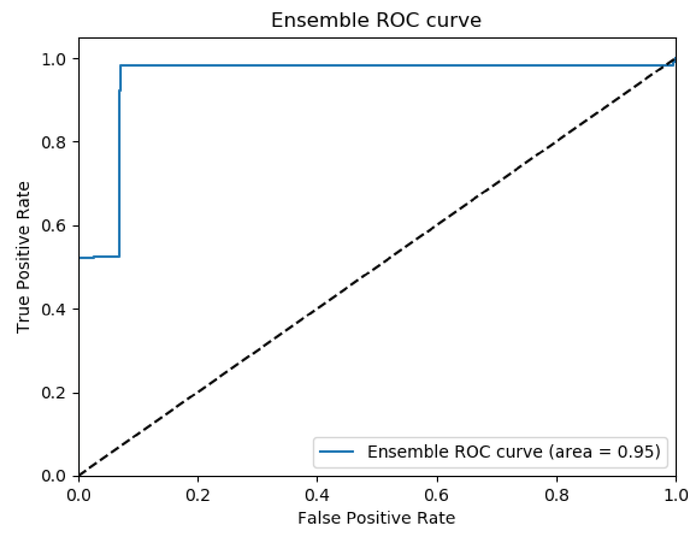
Ensemble:

Achieved score of 0.9519

toxic	severe_toxic	obscene	threat	insult	identity_hate
0.04913654	0.002002411	0.01642777	0.0004827721	0.02969216	0.001379697
0.002629505	0.00004173353	0.0009744970	0.00001076569	0.0009115023	0.00005419117
0.005325758	0.0001302195	0.002479452	0.00002079247	0.001817569	0.0001155850

Model Scores:

Model	ROC_AUC Score
biLSTM van Aken et al. [1]	0.9810
Naive Bayes Classifier (Baseline 1)	0.9258
biLSTM (Baseline 2)	0.9724
Ensemble (biLSTM + MNB)	0.9519



References

[1] van Aken et al. *Challenges for Toxic Comment Classification: An In-Depth Error Analysis.* acl:w18-5105, 2018.

[2] Jigsaw-Alphabet Inc./Kaggle. *Toxic Comment Classification Challenge*, 2018.