Toxic Comment Classification

Choenden Kyirong, Jasleen Kaur, Jing Wen
ckyirong@sfu.ca, jka185@sfu.ca, jwa268@sfu.ca

SFU SCHOOL OF COMPUTING SCIENCE

TASK AND MOTIVATION

BACKGROUND

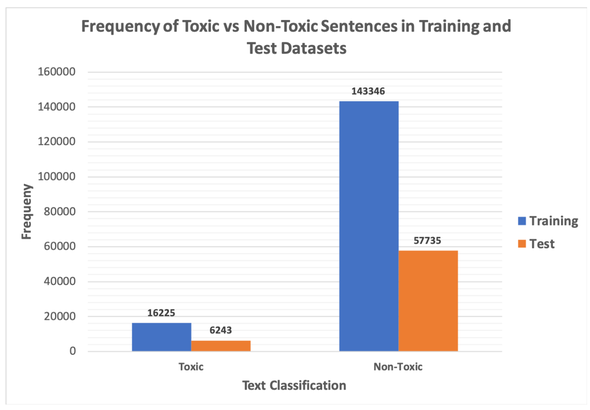
In this day and age, the usage of online human interactions has tremendously increased. Unfortunately, with this increase, online discussion forums have seen a rise in issues of online harassment. Moderators of online forums are not able to control and identify every such negative situation that arises.

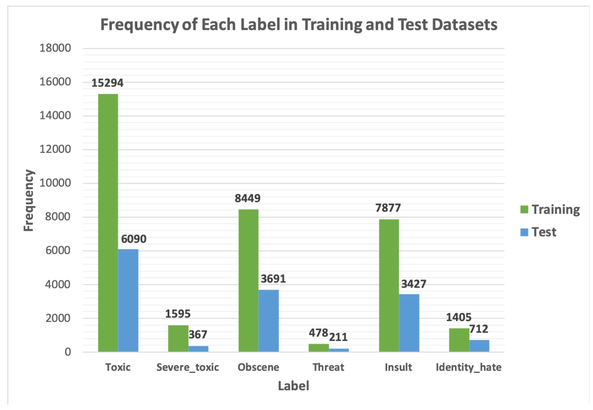
GOAL: Increase accuracy to identify toxic comments in order to improve online conversations by correctly labeling different texts according to their toxicity levels.

- Labels used: toxic, severe_toxic, obscene, threat, insult and identity_hate.

DATASET

- Based on and used datasets provided by Kaggle's "Toxic Comment Classification Challenge" to help identify and classify toxic online comments.
- Training dataset with 159,572 sentences
- Testing dataset with 63,978 sentences
- Pre-trained GloVe vectors from Twitter with 27 billion tokens with 100 and 200 dimensions





PROPOSED METHODS

- **Baseline**
  - SVM model (Khieu and Narwal, 1993)
- **Data Pre-processing** was done for all methods attempted by tokenizing and padding representation
- **LSTM** with 100D embeddings

$$Embeddings \rightarrow LSTM \rightarrow h1 \rightarrow sigmoid$$

- **GRU** with 200D embeddings

$$Embeddings \rightarrow GRU \rightarrow dropout \rightarrow h1 \rightarrow sigmoid$$

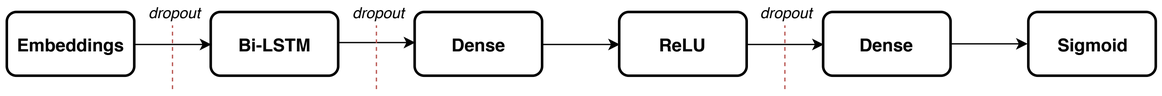
- **Enhanced LSTM** with 200D embeddings

$$Embeddings \rightarrow dropout \rightarrow LSTM \rightarrow dropout \rightarrow ReLU(h1) \rightarrow dropout \rightarrow h2 \rightarrow sigmoid$$

DATA PRE-PROCESSING

1. Delete samples of test data that include -1 in labels
2. Clean up strings
   - "I am 47" → "I am ##"
   - "Aren't" → "Are not"
   - "Hello World!" → "hello world!"
3. Term Frequency Map
   - "Hello hello hello world world the"
     - {"hello" : 1 , "world": 2, "the": 3}
4. Tokenization and Padding of Max Sentence Length (100)
   - "Hello Hello Hello world world the" → [0 0 ... 0 1 1 1 2 2 3]

MODEL ARCHITECTURE



Epochs = 2, Batch Size = 75, Optimizer = Adam, lr = .001, LSTM Hidden = 512, h1= 512, h2 = 256, Dropout = 0.5

EXAMPLE OUTPUT

Comparing two different sentences for each of the six possible types of comment toxicity:

[toxic, severe_toxic, obscene, threat, insult, identity_hate]

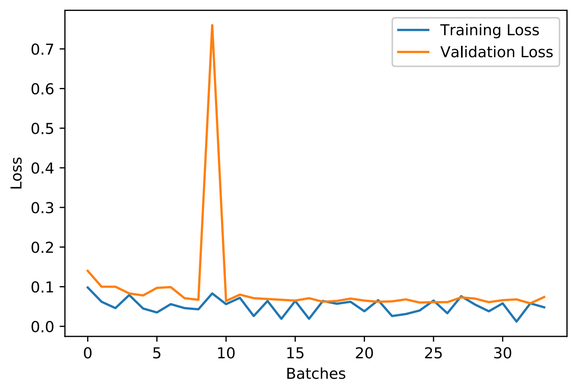
Sentence	Probability for Each Label	Binary Labels
"I hate you. You are stupid"	[.9 .35 .5 .1 .24 .33]	[1 0 1 0 0 0]
"I love you. You are amazing"	[.01 .001 .01 .01 .01 .01]	[0 0 0 0 0 0]

EVALUATION AND ANALYSIS

MODELS STATISTICS:

Method	Accuracy	Correctly Identified Toxic Sentences
LSTM	0.90118	6
GRU	0.84335	1883
Enhanced LSTM	0.8869	2129

LOSS GRAPH FOR ENHANCED LSTM MODEL:



- Baseline SVM model had accuracy of 0.833 (Khieu and Narwal, 1993)
- LSTM model had highest accuracy, but had lowest number of correctly identified toxic sentences
- GRU model had lowest accuracy, but drastically outperformed LSTM in correctly identifying toxic sentences
- Enhanced LSTM had higher accuracy than GRU and had highest number of correctly identified toxic sentences
- Accuracy = $\frac{\#\text{ of correct}}{\#\text{ of sentences}}$

FUTURE WORK

- Implementation of Naive Bayes-SVM(NBSVM)
- NBSVM performs well on snippets and longer documents, for sentiment, topic and subjectivity classification[5]

REFERENCES

[1] Chu, T., Jue, K., Wang, M.L. (2017). Comment Abuse Classification with Deep Learning.

[2] Khieu, K., Narwal, N. (2017). "Detecting and Classifying Toxic Comments".

[3] Pennington, Jeffrey et al. "Glove: Global Vectors for Word Representation." EMNLP (2014).

[4] "Toxic Comment Classification Challenge." Kaggle, Kaggle, 2017, www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[5] Wang, Sida Manning, Christoper. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. 90-94.

[6] Zheng, Heng. "PyTorch Starter." Kaggle, Kaggle, 14 Dec. 2018, www.kaggle.com/hengzheng/pytorch-starter.