Toxic Comment Classification

Interim Report

March 20, 2018

Anindya Paul	Irene Betts-O'Rourke	Muzi Li
akpaul@bu.edu	ireneb@bu.edu	marlonli@bu.edu

Project Topic

The goal of our project is to identify unwanted and harmful comments on Reddit. To achieve this goal we'll use data from a Kaggle Challenge (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data) which identifies toxic, severe toxic, obscene, threat, insult, and identity hate comments from other normal comments. We used about 10% of the labelled data to verify our accuracy and plan to apply the model on the comments made to the following Reddit topics:

Gun Control:
https://www.reddit.com/r/politics/comments/85dyz5/guns_dont_kill_people_men_and_boys_kill_people/

Abortion:
https://www.reddit.com/r/AskReddit/comments/75z6oy/whats_your_opinion_on_abortion/

The Alt Right Subreddit being banned:
https://www.reddit.com/r/DebateAltRight/comments/5ri7ka/altright_subreddit_banned/

Donald Trump Winning the election:
https://www.reddit.com/r/The_Donald/comments/5bzjv5/donald_j_trump_declared_the_winner/

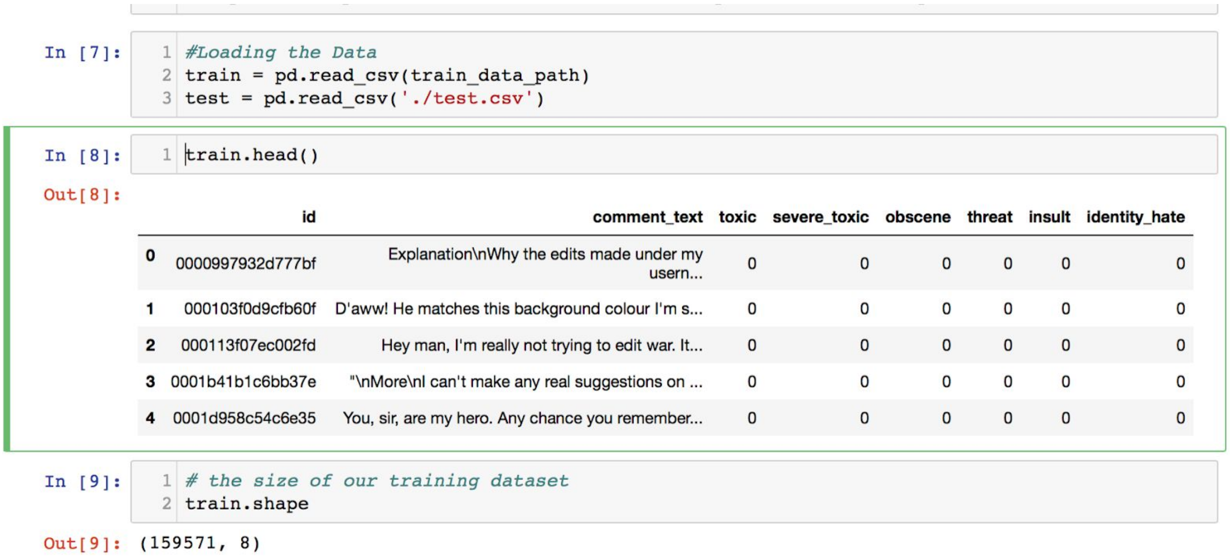
We will evaluate the results and determine accuracy based on specificity and sensitivity and then use the model to analyze commenting patterns across topics to determine which provide the most volatile discussion.

Our final report will include our code, results of our classification, our manual labelling of comments, and the accuracy of our classification.

Training

At present we are dealing with the dataset from Wikipedia comments to test our preprocessing method and how to train our data.

There are two csv datasets, one is labeled by human whether it's toxic or not, and another one is not. First we need to have an initial knowledge about our data source, so we transform them into panda dataframe.



```
In [7]:  #Loading the Data
         train = pd.read_csv(train_data_path)
         test = pd.read_csv('./test.csv')
```

```
In [8]:  train.head()
```

Out[8]:

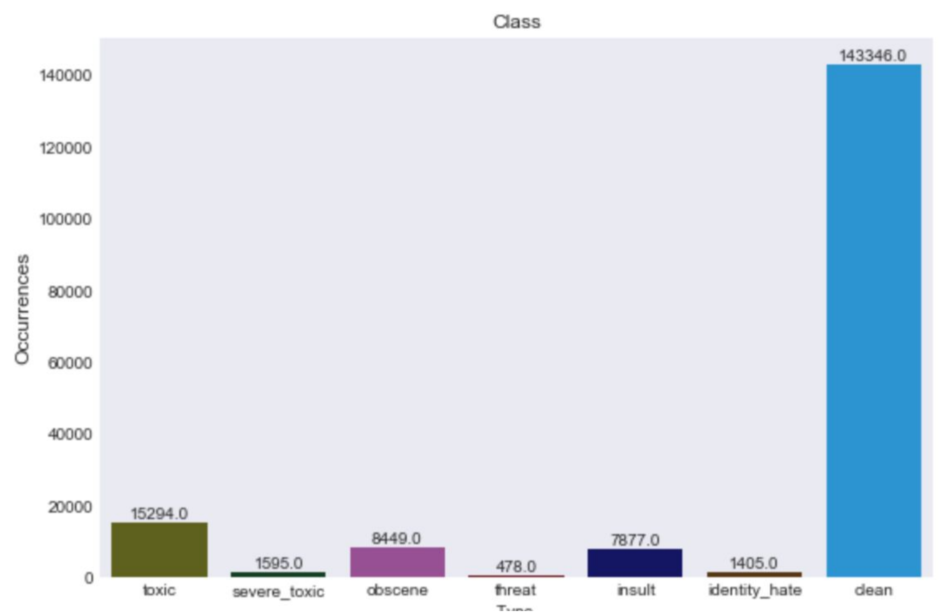
	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

```
In [9]:  # the size of our training dataset
         train.shape

Out[9]:  (159571, 8)
```

We can see that each comment own an unique id and is sorted by several labels, just like what we did in homework of restaurants in Las Vegas, but here we allow overlapping about these labels.

```
plt.show()
```

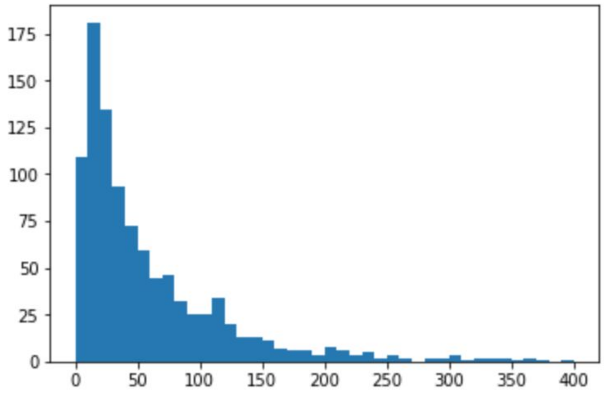


We have more than 150000 comments but most of them (> 80%) are not toxic, so it's very unbalanced, which will happen in most of the dataset of comments, so we need to really care about this point when we want to train our algorithm.

We have so far tried two algorithm, logic regression and Bayes. The mean column wise log loss of logic regression is 0.132922781554

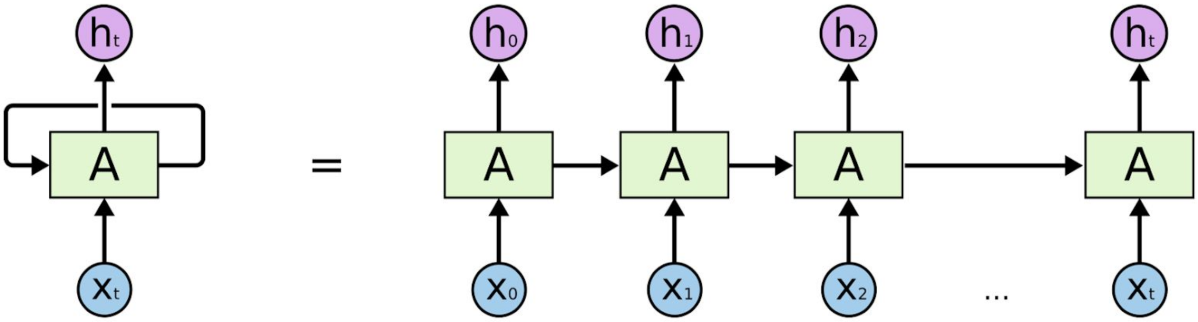
However, for the naive Bayes, the loss is 0.768501747768, which is much worse than the former one, because the naive Bayes is not so valid(since labels here can not be independent to each other), so we are not going to choose the bayes one.

Next, to test the LSTM(long short time memory) method, we tokenized our data. At present, we use the Keras API to help us do the process, but later on we may create our own dictionary via tensorflow. Since most of our comments drop into the range of 0 - 200, we choose 200 as our max length of words.

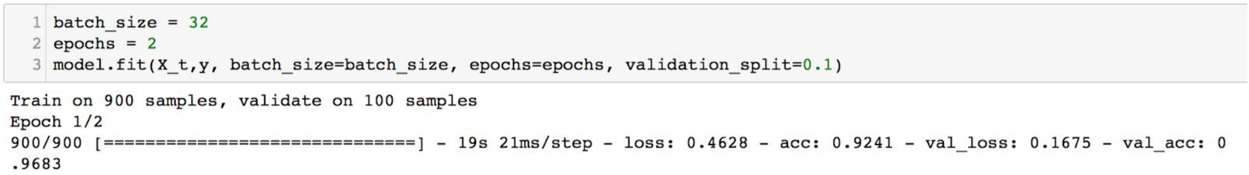


Then we pass it to our Embedding layer, where we project the words to a defined vector space depending on the distance of the surrounding words in a sentence. Embedding allows us to reduce model size and most importantly the huge dimensions we have to deal with, in the case of using one-hot encoding to represent the words in sentence.

We feed this Tensor into the LSTM layer, setting the LSTM to produce an output that has a dimension of 60 and want it to return the whole unrolled sequence of results.Since LSTM or RNN works by recursively feeding the output of a previous network into the input of the current network, and take the final output after X number of recursion.



We plan to set aside 10% of the labeled data for testing the accuracy of our model. As of now, we have received an accuracy of 92% based on training of 900 samples and testing on 100 samples, which we think is a good start.



```
batch_size = 32
epochs = 2
model.fit(X_t,y, batch_size=batch_size, epochs=epochs, validation_split=0.1)
```

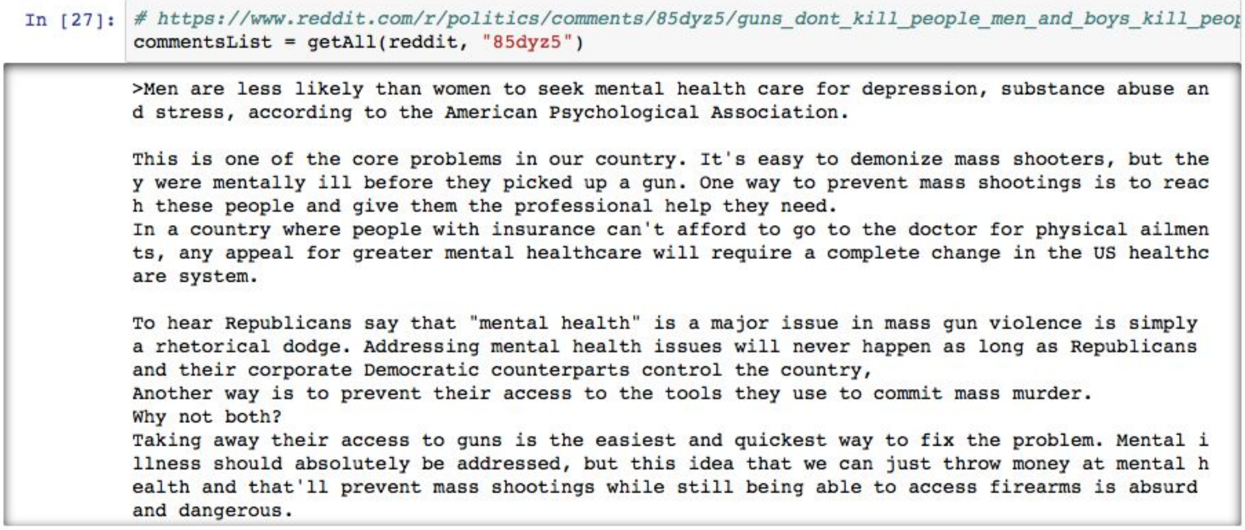
```
Train on 900 samples, validate on 100 samples
Epoch 1/2
900/900 [==============================] - 19s 21ms/step - loss: 0.4628 - acc: 0.9241 - val_loss: 0.1675 - val_acc: 0
.9683
```

Our code related to the training is available in GitHub:
https://github.com/Marlon666/CS506-Toxic-Detection

Extract data (comments) from Reddit

To extract comments from Reddit we'll use "PRAW: The Python Reddit API Wrapper" (https://praw.readthedocs.io/en/latest/index.html).

Below is a screenshot of comments corresponding to the discussion https://www.reddit.com/r/politics/comments/85dyz5/guns_dont_kill_people_men_and_boys_kill_people/



```
In [27]: # https://www.reddit.com/r/politics/comments/85dyz5/guns_dont_kill_people_men_and_boys_kill_peo
         commentsList = getAll(reddit, "85dyz5")
```

```
>Men are less likely than women to seek mental health care for depression, substance abuse an
d stress, according to the American Psychological Association.

This is one of the core problems in our country. It's easy to demonize mass shooters, but the
y were mentally ill before they picked up a gun. One way to prevent mass shootings is to reac
h these people and give them the professional help they need.
In a country where people with insurance can't afford to go to the doctor for physical ailmen
ts, any appeal for greater mental healthcare will require a complete change in the US healthc
are system.

To hear Republicans say that "mental health" is a major issue in mass gun violence is simply
a rhetorical dodge. Addressing mental health issues will never happen as long as Republicans
and their corporate Democratic counterparts control the country,
Another way is to prevent their access to the tools they use to commit mass murder.
Why not both?
Taking away their access to guns is the easiest and quickest way to fix the problem. Mental i
llness should absolutely be addressed, but this idea that we can just throw money at mental h
ealth and that'll prevent mass shootings while still being able to access firearms is absurd
and dangerous.
```

Our code to extract comments is available in GitHub:
https://github.com/anindyapaul/CS506_Project

Next Steps

We’ll continue to work on improving our classification on the Kaggle data. Once we have reached about 95% accuracy, we will use the model to classify comments on Reddit.

As mentioned above, one of the challenges we have is that the number of normal (non-toxic) comments greatly outnumber the number of toxic comments. So we expect to face challenges in improving our accuracy.

We may try the following to improve the accuracy and efficiency of our method:

1. Using pre-trained models to boost accuracy and take advantage of existing efforts
2. Try to find wiser way of setting Hyper-Parameters
3. Introduce early stopping to avoid overfitting.

Beyond the Reddit topics mentioned above, if we find time, we’ll also classify comments on other discussions at https://www.reddit.com/r/politics/ as we think political discussions are likely to have toxic comments over discussions on other topics. If we are successful in classifying Reddit comments, we may use our model on other online discussion forums and social media websites too in future.