# Machine Learning Engineer Nanodegree

## Capstone Project

Lilit Sargsyan

April 5th, 2018

##TOSIC COMMENT CLASSIFICATION

## I. Definition

### Project Overview

This project comes from the problem that abusive language, threats, and harassment can poison online conversations and as a result make it difficult to share your thoughts online about the things you care about. The threat of online abuse and harassment results on many people stop to express themselves and give up seeking different opinions. According to Pew Research Center1, 27% of American internet users chose not to post something online after seeing someone being harassed. Toxic language makes it hard to discuss important issues.

The Conversation AI2 team, a research initiative founded by Jigsaw3 and Google are working on tools to help improve online conversation. Their research aims to help increase participation, quality and empathy in online discussions. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful, or otherwise are likely to make participant to live the conversation). One of approaches to solve this problem involves people facilitating discussions, but this is time consuming and requires a large workforce, that’s why we need machine learning methods to do it.

1 <http://www.pewinternet.org/2017/07/11/online-harassment-2017/>

2 <https://conversationai.github.io/>

3 <https://jigsaw.google.com/>

### Problem Statement

The goal of this project is connected with finding toxic comments in online discussions. The final model will be able to detect toxic comments and assign to it a probability for each type of toxicity which are the following:

* Toxic
* Severe\_toxic
* Obscene
* Insult
* Identity\_hate

For this we will use data from Kaggle.com Toxic Comment Classification Challenge1. It is a public dataset of comments from Wikipedia’s talk page edits. Dataset contains large number of Wikipedia comments whit the id which have been labeled by 6 toxicity sub-types (reasons why something might be considered toxic). The labeled annotations are based on asking 5000 crowd-workers to rate Wikipedia comments according to their toxicity (likely to make others leave the conversation).

1 <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>

### Metrics

The evaluation metrics for my algorithm is ROC AUC1 (area under receiver operating characteristic curve). ROC is a set of {tp rate, fp rate} where tp is the true positive rate (positives correctly classified divided by total positives) and fp is true negative rate (negatives incorrectly classified divided by total negatives).

True class

**p** **n**

|  |  |
| --- | --- |
| True  Positive | False  Positive |
| False  Negative | True  Negative |

**Y** tp rate=

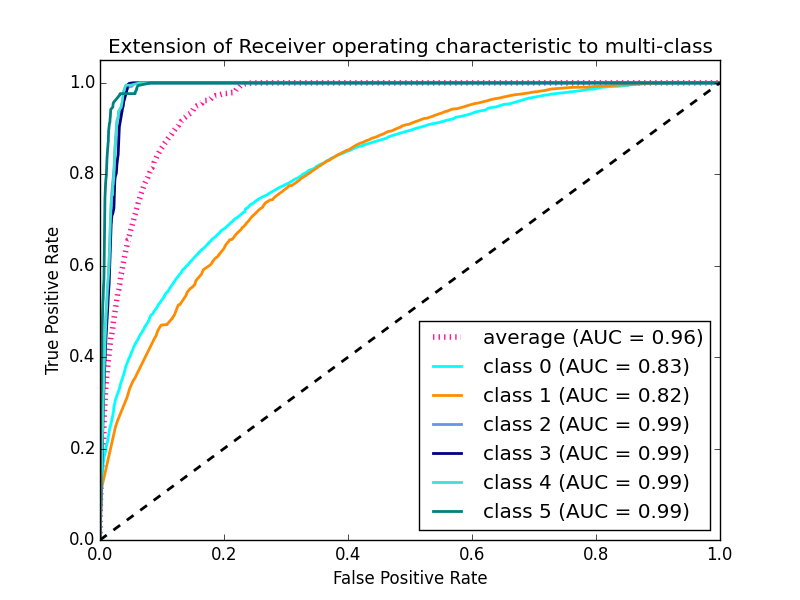
Hypothesized

class

**N** fp rate =

Column totals: **P N**

ROC graphs are two-dimensional graphs in which tp rate is plotted on the Y axis and fp rate is plotted on the X axis. AUC is the area under ROC curve. Because we have multiclass classification with 6 classes the ROC graph will be something like this:



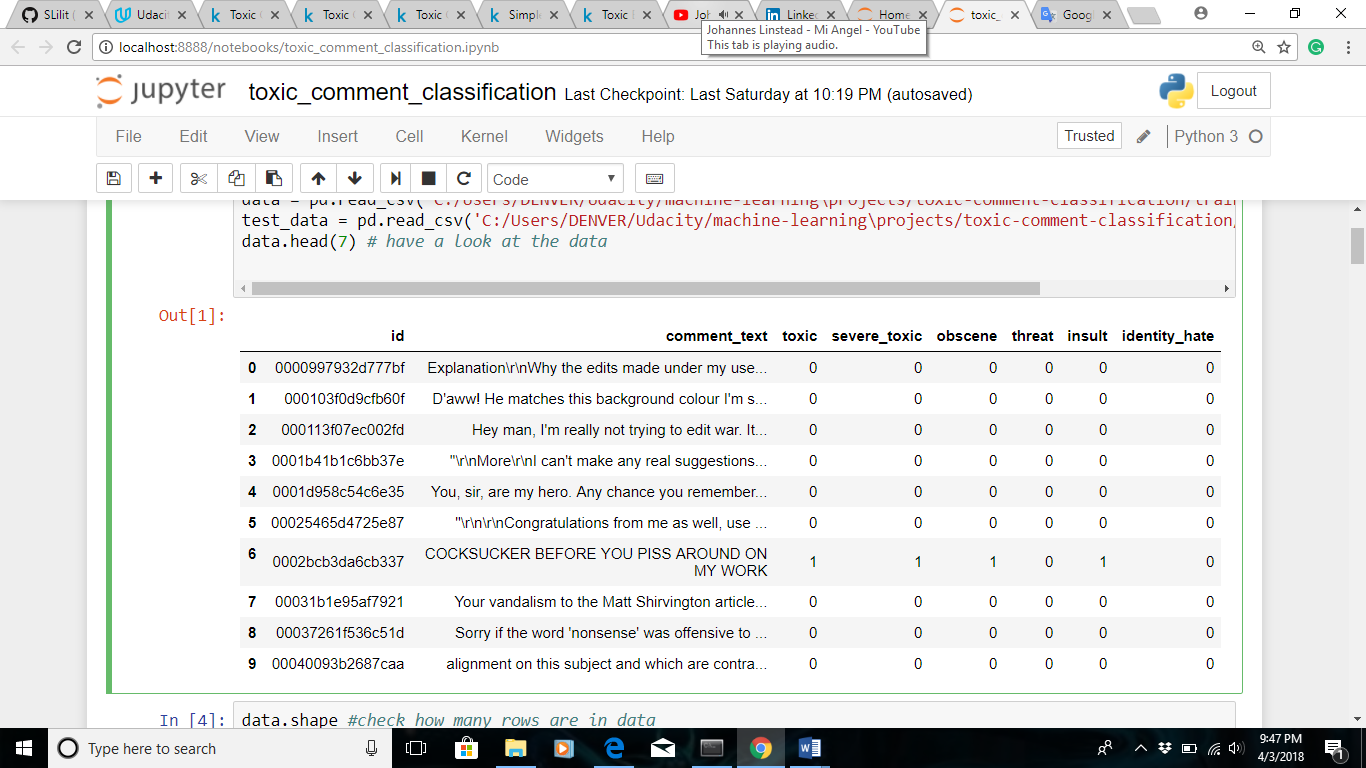
We want to have higher rate for recognizing toxic comments and lower rate assigning nontoxic comments as toxic. That’s why we will use this score b it’ll shows how well our algorithm is doing determining toxic comments (tp) and not toxic comments (fp), also because ROC curves are insensitive to changes in class distribution. If the proportion of positive to negative instances changes in a test set, the ROC curves will not change. Higher ROC AUC score means more toxic comments recognize as toxic and less nontoxic comments as toxic.

1 <http://people.inf.elte.hu/kiss/13dwhdm/roc.pdf>

## II. Analysis

### Data Exploration

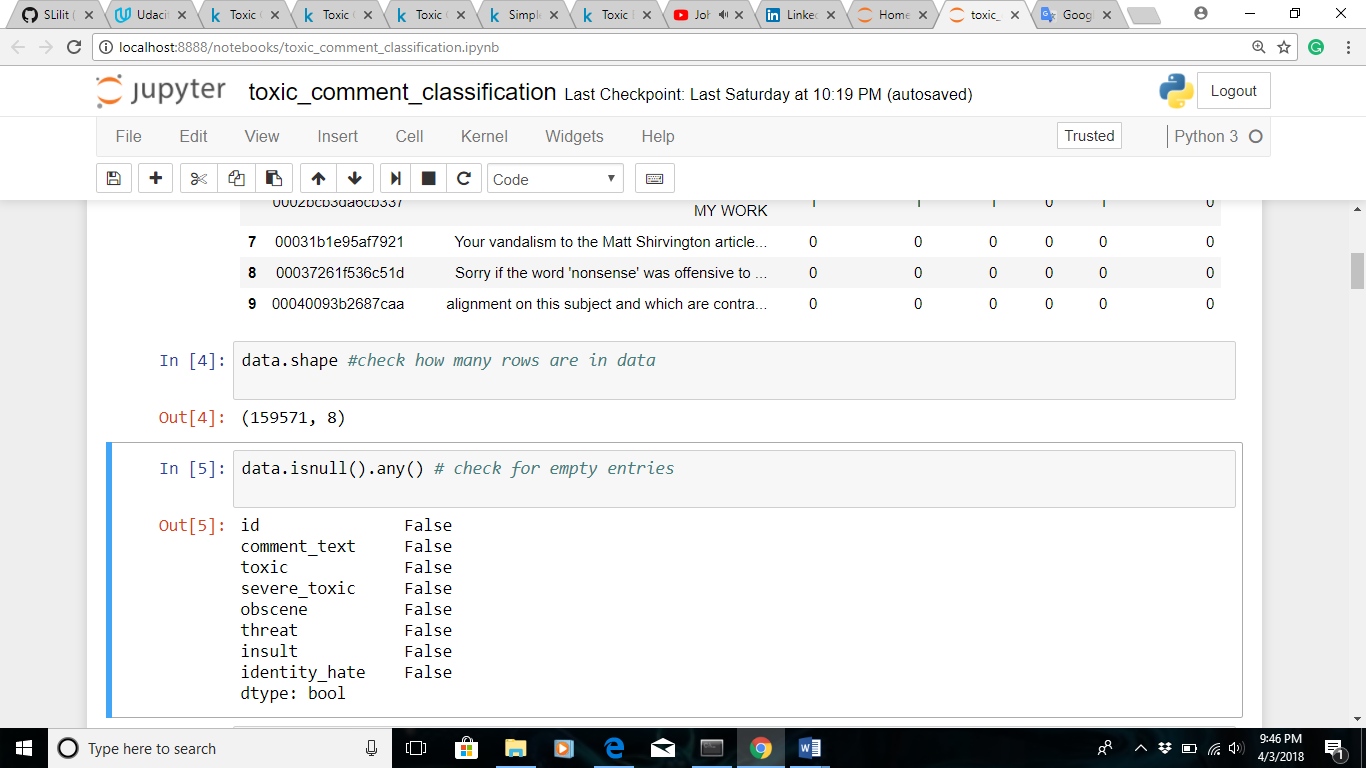
First, we will have a look at our data:



So our data consist of id column, which we don’t need and will get rid of later, comment\_text column, which is our training data, and 6 columns for toxicity types and these will be our labels. From the labels columns we can see that one comment can be assigned to the several toxicity types or any type at all.

In the next step we will have a look at shape of our data and check if there are empty entries (rows with no information).

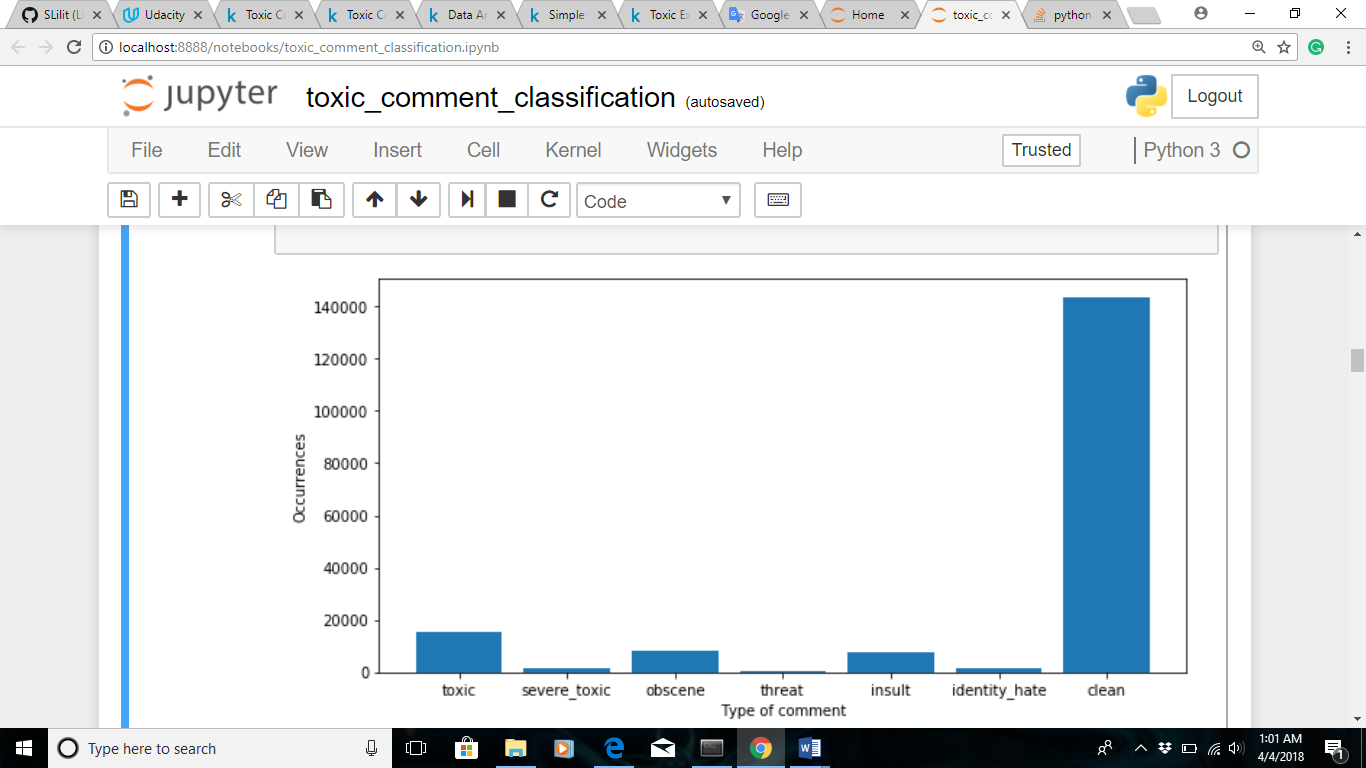
So as we can see from the image below there are 159571 rows with no missing values.



### Exploratory Visualization

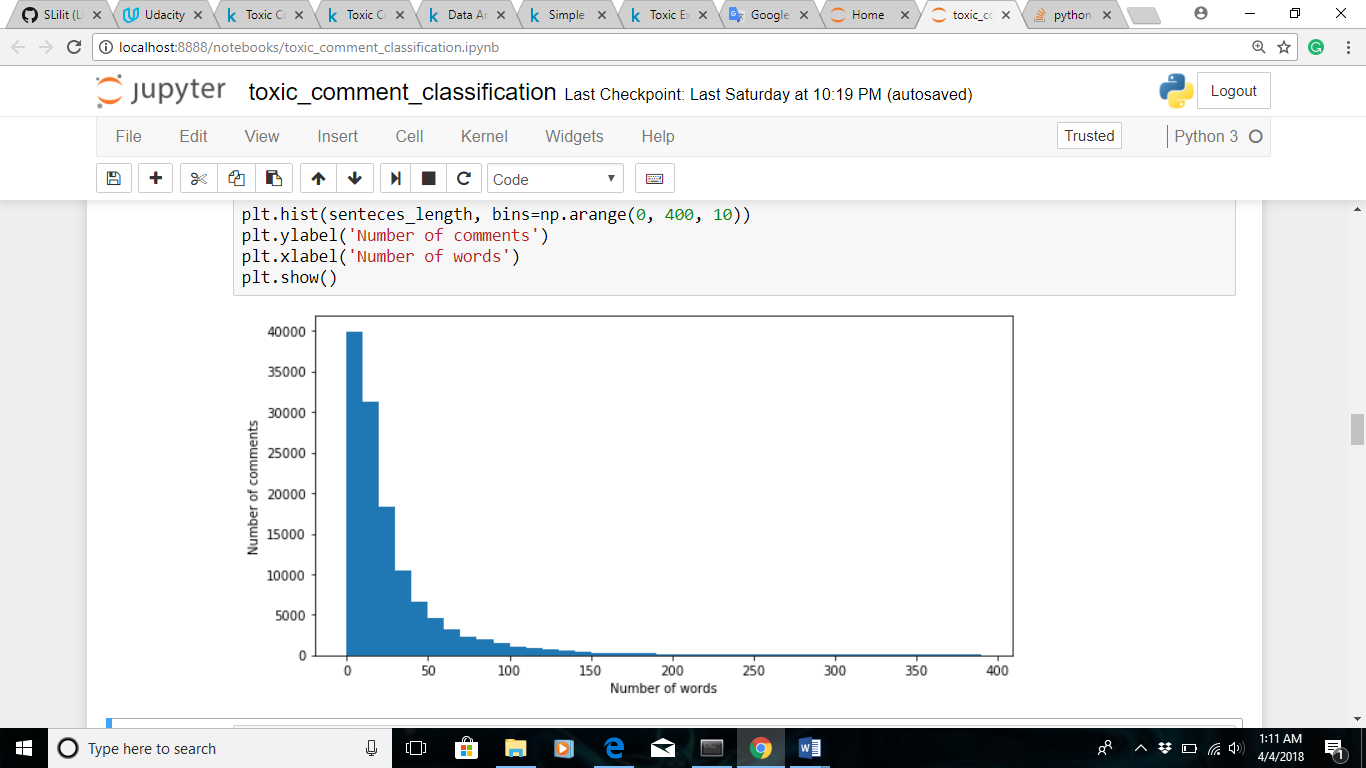
To visualize data distribution first we will add new column for not toxic comments. This column will be one when all the other label columns are zero and zero otherwise. Now let’s have a look to a comments distribution

bars by their types:



As we can see from here our data is skewed and primarily consist of nontoxic comments. Most of toxic comments are of type toxic that’s because other type of toxic comments (like severe\_toxic) can also be type of toxic. So we have very unbalanced data even among toxic comments. This mean if we predict clean for each comment our result still won’t be so bad in term of accuracy that’s why we using ROC AUC(how it’s calculated is describe in Metrics section).

Next visualization that we will need is for data preprocessing (we will describe it deeper in the III. Methodology part). We will plot a histogram to see how many word are there in each comment:



As we can see most of the comments contain less than 50 words, but because most of this comments are normal we will keep length of 150 words for each comment to insure that there is enough toxic word to train on our model.

### Algorithms and Techniques

Deep Learning methods are proving very good in text classifications, achieving state of the art results for most of NLP(Natural language processing) tasks. The classifier is Word Embedding + RNN (Recurrent Neural Networks).

* First layer will be hidden Embedding1 layer.

In an embedding words are represented by dense vectors where a

vector represents the projection of the word into a continuous vector space. The position of a word within the vector space is learned from text and is based on the words that surround the word when it is used.

For Embedding layer 3 arguments required to be specified, these arguments are inpute\_dim, output\_dim, input\_length (will go deeper here in Implementation section).

* Second layer will be LSTM (Long Short-Term Memory) layer.

LSTM network trained using Back Propagation Through Time and overcomes the vanishing gradient problem. As such, it can be used to create large recurrent networks that can be used to address sequence problems (we not only need the words in comments but also their sequences to get better prediction results). Parameters to be tuned are units (smart neurons), dropout, recurrent\_dropout, return\_sequences.

* The third layer is GRU3 layer.

GRU is similar to LSTM but there are some differences one of which is that GRU exposes its full memory content that is seen or used by other units whereas LSTM uses controlled exposure. The parameters to be tuned is the same as with LSTM.(more about GRU layers can be found [here](https://arxiv.org/pdf/1412.3555v1.pdf))

* The forth layer is GlobalMaxPool1D layer.
* The last layer will be Dense layer to with 6 output layers for each toxicity type and with the ‘sigmoid’ activation
* And depending on how well is our model doing we will add Dropout layers.
* After this next step will model compiling and fitting. hyperparameters to be tuned here are: optimizer, loss, metrics, epochs, batch\_size, validation\_split.
* We’ll use callbacks to save weights with the best loss and print ROC AUC after each epoch.
* Then we’ll do prediction on test data with the loaded best weights from previous step.
* Get mean column wise ROC AUC for our model and keep tuning till we get our benchmark.

1 <https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/>

2 <https://machinelearningmastery.com/sequence-classification-lstm-recurrent-neural-networks-python-keras/>

2 <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

3 <https://arxiv.org/pdf/1412.3555v1.pdf>

### Benchmark

So far there are publicly available models served through Perspective API for this problem to determine toxic comments but they don’t show types of toxicity. So benchmarks for this project are:

* Be able to distinguish toxic comments from normal ones and show probability for each toxicity type for that comment.
* The leader board on Kaggle shows achieving ROC AUC score of 0.97 and higher which will be the target of our model.

## III. Methodology

### Data Preprocessing

We keep to kind of data one for our model and one for evaluating our data in kaggle.

For our model first we divide 'comment\_text' column of our data into training and testing subsets and 'toxic', 'severe\_toxic', 'obscene', 'threat', 'insult', 'identity\_hate' columns as training and testing labels using train\_test\_split from sklearn.model\_selection library keeping 20% of data for testing. Now we have following subsets:

* comments\_train (80% of data to train)
* labels\_train (80% of labels to train)
* comments\_text (20% of data to test)
* labels\_test(20% labels to test)

For kagle we keep all data for training and then use testing data from kaggle to do predictions and check our score at kaggle. We can’t use testing data from kaggle to evaluate our model because it doesn’t have labels.

Farther we preprocessing comments\_train, comments\_test, kaggle\_train and kaggle\_test data sets.

We start data preprocessing by data cleaning. Split data using text\_to\_words\_sequence() function from keras.preprocessing\_text library

which by default does three things:

* Splits words by space.
* Filters out punctuation.
* Convert text to lower case.

After this remove numbers and common words using stopwords from nltk.corpus library.

Next step after data cleaning is tokenizing the comments i.e. turning sentences into numbers sequences where each number is word’s index representation. WE using Tokenizer from keras.preprocessing\_text library for this. Here we have to define number of unique words (num\_words) in our dictionary, we set this to 20000.For this we first fit comments\_train

to Tokenizer (to get index for each word) than use text\_to\_sequences to turn sentences to numbers sequences.

The last step in data preprocessing is to make all comments of same length and we already have discussed what length will be good for our data in exploratory visualization section. For this we use pad\_sequences from keras\_preprocessing\_sequence with maxlen = 150. This will fill short sentences with zeroes in the end and cut long sentences from the end to make both short and long sentences of length 150.

<https://machinelearningmastery.com/prepare-text-data-deep-learning-keras/>

### Implementation

Load training data into the data described in previous sections and testing data for kaggle submission into the test\_data (we can use this data for our model we can just make prediction on it for submission).

After data loading, data exploration and exploratory visualizations (this steps are thoroughly described in Analyses section) we do data preprocessing as described in above section. After this data is ready to be trained.

Next we implement our architecture as described in Algorithms and Techniques section but we will go deeper about it here. These are the steps we take:

* First layer of our neural network is Embedding layer. For this layer we need to give three parameters which are:
* inpute\_dim: This is the size of the vocabulary in the text data, in our case it’s 20000(look data preprocessing section when we chose maximum number of words for tokenizing)
* output\_dim: This is the size of the vector space in which words will be embedding (we set this parameter equal to 130)
* input\_length: This is the length of input, in our case this is 150 (look data preprocessing section when we pad comments to the same length)
* Next comes LSTM layer with 100 units (dimensionality of output layer), return\_sequences set to True, dropout set to 0.3, and reccurent\_dropout set to 0.2.
* Next GRU layer with similar arguments as LSTM layer exempt with units of 80.
* Next MaxPool1d with pool\_size=2
* Next is GlobalMaxPool1D layer.
* After this Dropout layer.
* And the last layer is Dense layer with output dimension of 6 because of our toxicity types are 6. Activation is ‘sigmoid’. We use ‘sigmoid’ instead of ‘softmax’ because softmax forces the total predicted probabilities to some up to 1, meaning it won’t be able to predict not toxic comments or multiple toxicity comments.
* We create roc\_auc\_callback function to use in callbacks to show ROC AUC score after each epoch.
* Declare checkpointer to save weights with the best val\_loss. For this we use ModelCheckpoint from keras.callbacks
* Compile the model using ‘adam’ for optimizer, ‘binary\_crossentropy’ for loss function because it’s independently optimizes each class and we have multi-label scenario, and for the metric we use ‘accuracy’.
* Fit the model using 10% of training data as validation data, epochs equal to 4 and batch\_size of 30.

This is the architecture of the model:

Input Embedding LSTM

(None, 150) (None, 150 (None, 150

130) 100)

Dropout Dense

GRU (None, 80) (None, 60)

(None, 80)

Dense Dropout

(None, 6) (None, 60)

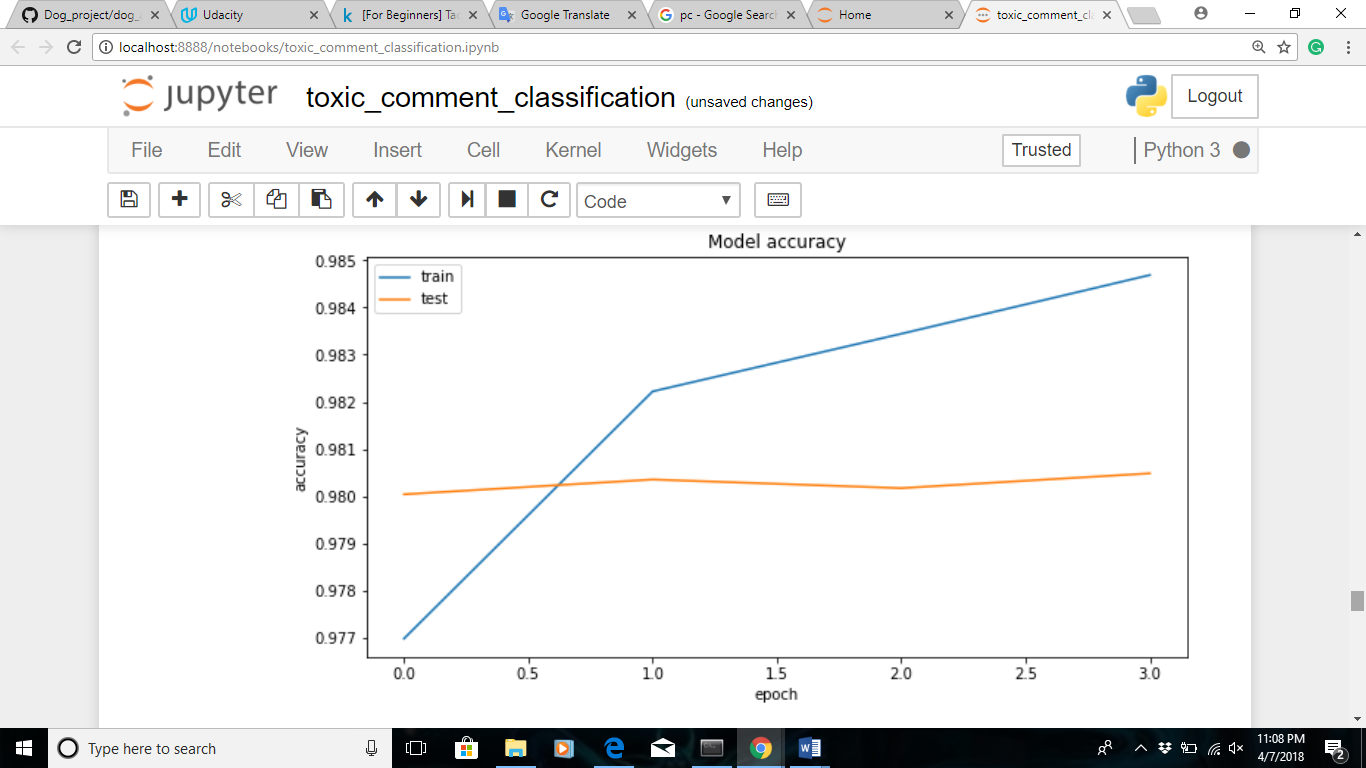
After this comes model tuning and evaluation steps which we will discuss in the next sections.

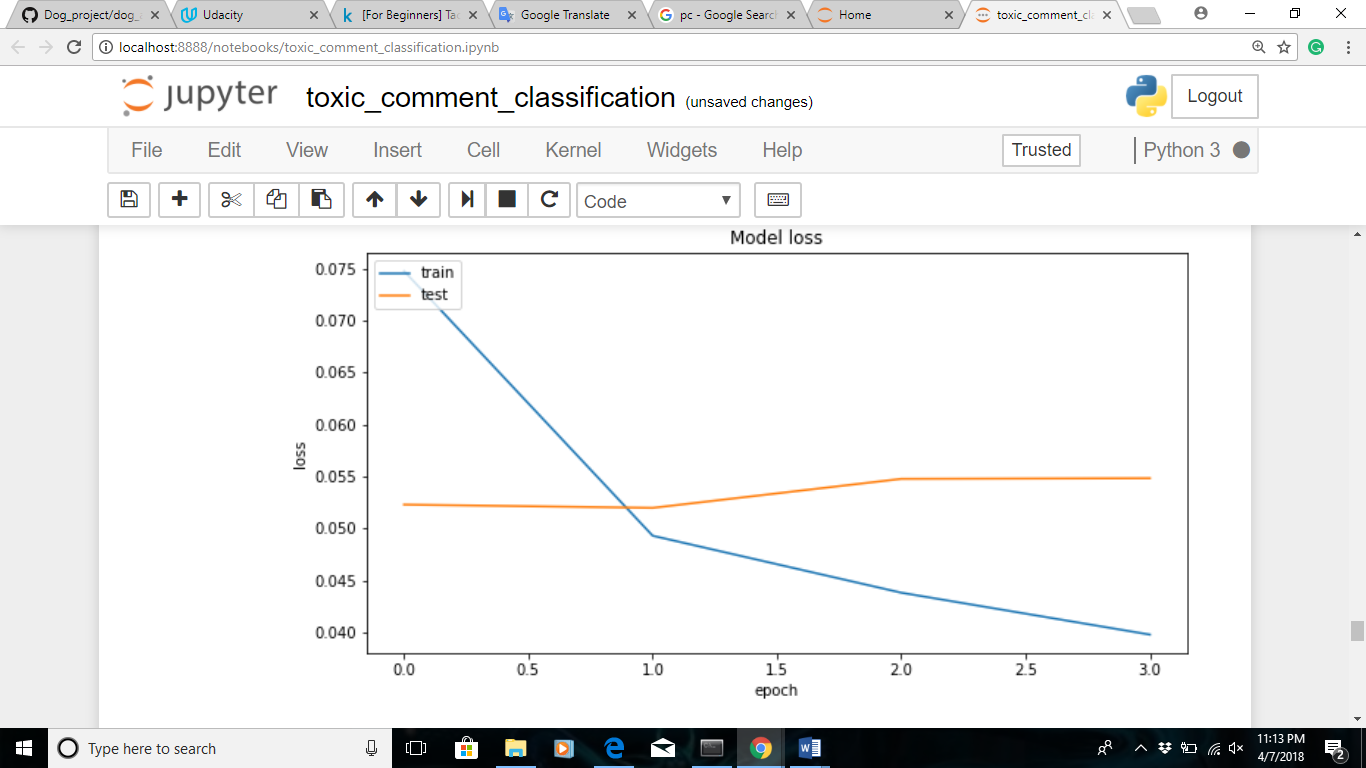
### Refinement

As mentioned in the benchmark our first goal is to predict probability for each 6 type of toxicity. We achieve this by using Dense layer with output dimension of 6 (for each type) and ‘sigmoid’ activation (for probabilities).

Second achieving ROC AUC score 0.97 and higher. One of improvements was connected with the architecture. Before we use CNN layer after Embedding but that was giving us ROC AUC not higher than 0.96. Instead we use LSTM as second layer and that gave us better result.

These are the plots of training the training/validation accuracy and losses.





Divergence indicates overfitting. Which we address by adding Dropout layers and reducing model complexity.

<https://machinelearningmastery.com/display-deep-learning-model-training-history-in-keras/>

## IV. Results

### Model Evaluation and Validation

We used validation set to evaluate our model during the training and testing set for final evaluation.

We achieve mean column wise ROC AUC score of 0.975 using the architecture described in the Implementation section and preprocessed data described in the Data Preprocessing section. This architecture and hyperparameters where chosen because they performed best among the tried combination. Figure below displays ROC curves for each toxicity type for our model, along with corresponding AUC values. In contrast to this the value of useless model is 0.5 and displayed as a diagonal line. Model with this value (AUC=0.5) can be thought as it having the equivalent classification ability of flipping a coin.

