Machine Learning Engineer Nanodegree

Capstone Proposal

Ramkumar Narayanan

October 2nd, 2018

Domain Background

Social media serves as a mean to express their thoughts or feelings about different subjects. It can be a political view or about a new product launched in the market or even about a movie.

An average of 2 hours and 15 minutes per day is spend on social networks. Facebook users generate 4 million comments every minute. Every second, on average, around 6,000 tweets are tweeted on Twitter which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year. When we start analyzing the data from social media and perform sentiment analysis we would be able to obtain great deal of insights regarding that particular subject. It is like a crowd funded survey without any formal questionnaire at zero expense.

We are planning to conduct sentiment analysis in our company based on social media #hashtags right after a significant company milestone. I am hoping this project will be a foundation to get started in the domain of NLP which eventually I can apply at my work.

Problem Statement

Discussing things that we care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments. Many companies are working on tools to improve healthy online conversations and curb negative online behaviors like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion).

So the objective is to build a multi-label classification model that is capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate among a list of comments posted online. By doing this we can help online discussion become more productive and respectful.

Datasets and Inputs

The dataset that is going to be used here is list of comments from Wikipedia's talk page edit.

The comments in the dataset falls under one or more categories listed below.
1.Toxic
2.Severe Toxi
3.Obscene
4.Threat
5.Insult
6.Identity hate

The Dataset has 159,571 comments categorized into 6 different categories. If the number under any category is '1' then the comment falls under that category and '0' if the comment doesn't fall under the category. There are no personal information (no pii or spii data) about the user and it is a publicly available dataset in Kaggle. Below is a sample of the data set

Screenshot-1

A pair of jew-hating weiner nazi schmucks.

Record #	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
1	0002bcb3da6cb337	COCKSUCKER BEFORE YOU PISS AROUND ON...	1	1	1	0	1	0
2	001810bf8c45bf5f	You are gay or antisemmitian?	1	0	1	0	1	1
3	00190820581d90ce	FUCK YOUR FILTHY MOTHER IN THE ASS, DRY!	1	0	1	0	1	0
4	0020e7119b96eeeb	Stupid peace of shit stop deleting my stuff ass...	1	1	1	0	1	0
5	0020fd96ed3b8c8b	=Tony Sidaway is obviously a fistfuckee. He lov...	1	0	1	0	1	0
6	0028d62e8a5629aa	All of my edits are good. Cunts like you who r...	1	0	1	0	1	0
7	0036621e4c7e10b5	Would you both shut up, you don't run wikipe...	1	0	0	0	1	0
8	00472b8e2d38d1ea	A pair of jew-hating weiner nazi schmucks.	1	0	1	0	1	1
9	00686325bcc16080	You should be fired, you're a moronic wimp w...	1	0	0	0	1	0
10	006b94add72ed61c	I think that your a Fagget get a oife and burn i...	1	0	1	1	1	1
11	006d11791d76b9f3	REPLY ABOVE:	0	0	0	0	1	0
12	006e87872c8b370c	you are a stupid fuck	1	1	1	0	1	0
13	0086998b34865f93	Fuck you, block me, you faggot pussy!	1	0	1	0	1	0
14	008e0818dde894fb	Kill all niggers.	1	0	1	0	1	1

Id - Unique key to represent a user
Comment_text - Raw comment data

Out of the 159,571 comments

15,294 falls under Toxic
1,595 falls under Severe Toxic
8,449 falls under Obscene
478 falls under Threat
7,877 falls under Insult
1,405 falls under identity_hate
Remaining comments are neutral.

The test set has 153,164 records with just Id and comment_text field.

Solution Statement:

The solution is to come up with the probability of a comment present in the test data falling under any of the above mentioned category. First I need to clean up and pre-process the data by removing punctuations, symbols. emojis(if any) and tokenize the data.
Second use any of the existing pre trained word embedding models like Word2Vec or Glove to obtain vector representations for words
I am planning to experiment by implementing a Convolutional neural net model or a Recurrent Neural Net model and come up with probability of a comment falling under any of the above category.
The training set with be around 145000 comments and 15000 will be the validation set.
The test set has 153,164 comments.

Benchmark Model

It is a kaggle competition and the leader board says the top score is 98% accuracy. Since this is the first time I am trying to build an NLP model I will strive to achieve at least 90% accuracy.

Evaluation Metrics

Since this is a multi-label classification task final evaluation metric is (ROC AUC) score. Area Under the Receiver Operating Characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate.

So according to kaggle the model is evaluated on the mean column-wise ROC AUC. In other words, the score is the average of the individual AUCs of each predicted column.

The output file (test file) must predict a probability for each of the six possible types of comment toxicity (toxic, severe_toxic, obscene, threat, insult, identity_hate). Below is the sample test file

Test file sample:

```
id,toxic,severe_toxic,obscene,threat,insult,identity_hate
00001cee341fdb12,0.5,0.5,0.5,0.5,0.5,0.5
0000247867823ef7,0.5,0.5,0.5,0.5,0.5,0.5
```

Project Design

Step-1: Data Explorations
Before starting the project, I would like to go through the data thoroughly. I need to understand on what basis the comments present in the training data are labelled.
Step-2: Data pre-processing and noise removal:
Apply noise removal technique like Lexicon Normalization or Object standardization to remove texts which are not relevant to the context of the data. For e.g.: URLS, punctuations, acronyms etc.
Step-3: Feature Engineering:
To analyze the preprocessed data, I will convert the texts into features. Depending upon the usage, text features can be constructed using various techniques – Syntactical Parsing, Entities / N-grams / word-based features, Statistical features, and word embedding. I will explore with different technique and see how it affects the output.
Step-4: Neural Net Model
The final step is creating the Recurrent Neural Net model to train using the training data. Based on the learning rate I will tweak the hyper parameters and analyze the result. The final ROC score will be calculated against the test data provided.

Below are the links to the data and some of the materials I referenced for this project.

https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge#evaluation

http://www.letscodepro.com/Twitter-Sentiment-Analysis/

https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/

https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/

http://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics