# 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=

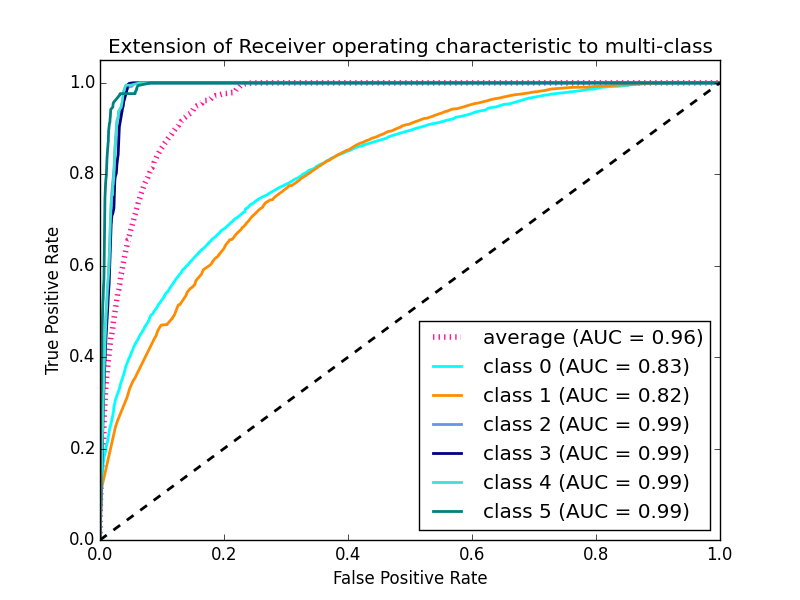
Hypothesized

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:



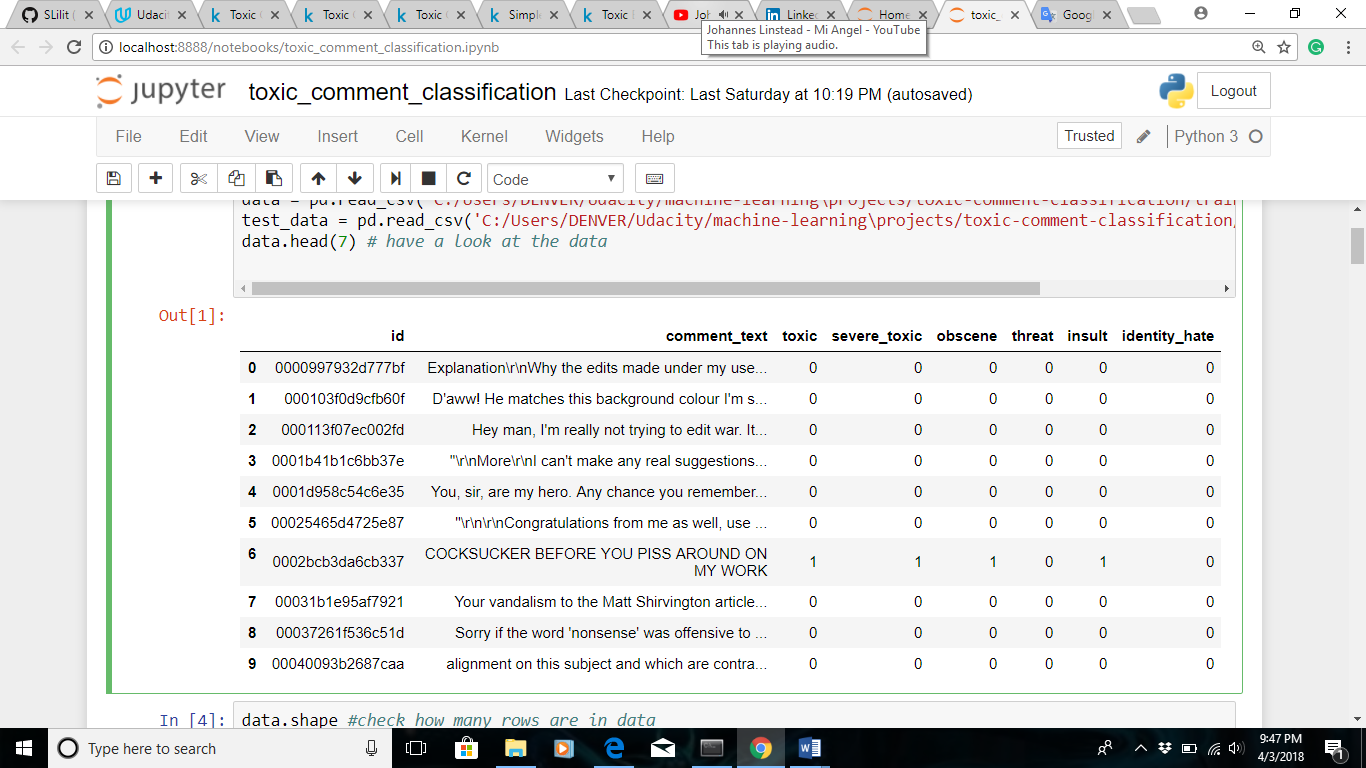
We want to have higher rate recognize toxic comments and lower rate assigning nontoxic comments as toxic. That’s why we will use this score because it shows how well our algorithm is doing determining toxic comments (tp) and not toxic comments (fp) and because it is not changing when the data distribution is changed. Higher 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

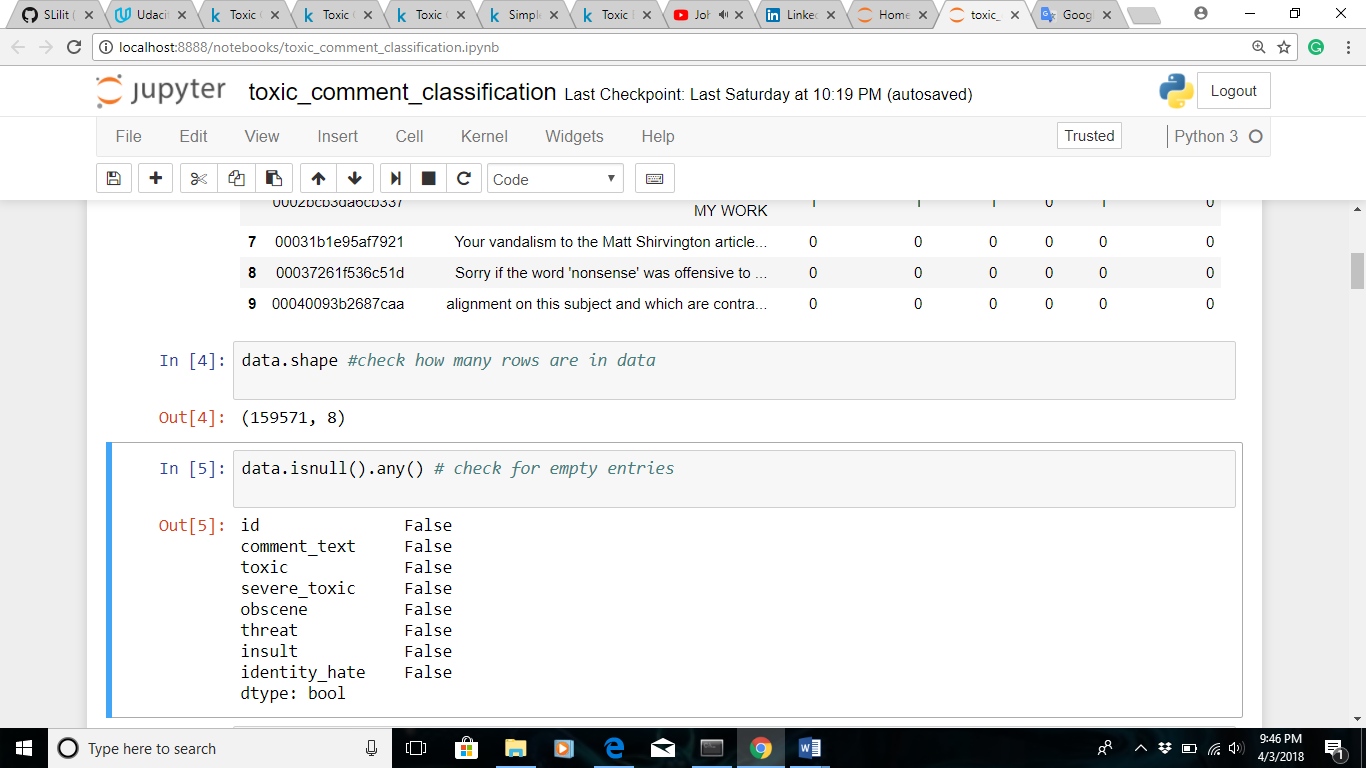
First, we will have a look at our data:



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

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

- \_Have you visualized a relevant characteristic or feature about the dataset or input data?\_

- \_Is the visualization thoroughly analyzed and discussed?\_

- \_If a plot is provided, are the axes, title, and datum clearly defined?\_

### Algorithms and Techniques

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

- \_Are the algorithms you will use, including any default variables/parameters in the project clearly defined?\_

- \_Are the techniques to be used thoroughly discussed and justified?\_

- \_Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?\_

### Benchmark

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

- \_Has some result or value been provided that acts as a benchmark for measuring performance?\_

- \_Is it clear how this result or value was obtained (whether by data or by hypothesis)?\_

## III. Methodology

\_(approx. 3-5 pages)\_

### Data Preprocessing

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

- \_If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?\_

- \_Based on the \*\*Data Exploration\*\* section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?\_

- \_If no preprocessing is needed, has it been made clear why?\_

### Implementation

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

- \_Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?\_

- \_Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?\_

- \_Was there any part of the coding process (e.g., writing complicated functions) that should be documented?\_

### Refinement

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

- \_Has an initial solution been found and clearly reported?\_

- \_Is the process of improvement clearly documented, such as what techniques were used?\_

- \_Are intermediate and final solutions clearly reported as the process is improved?\_

## IV. Results

\_(approx. 2-3 pages)\_

### Model Evaluation and Validation

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

- \_Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?\_

- \_Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?\_

- \_Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?\_

- \_Can results found from the model be trusted?\_

### Justification

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

- \_Are the final results found stronger than the benchmark result reported earlier?\_

- \_Have you thoroughly analyzed and discussed the final solution?\_

- \_Is the final solution significant enough to have solved the problem?\_

## V. Conclusion

\_(approx. 1-2 pages)\_

### Free-Form Visualization

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

- \_Have you visualized a relevant or important quality about the problem, dataset, input data, or results?\_

- \_Is the visualization thoroughly analyzed and discussed?\_

- \_If a plot is provided, are the axes, title, and datum clearly defined?\_

### Reflection

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- \_Have you thoroughly summarized the entire process you used for this project?\_

- \_Were there any interesting aspects of the project?\_

- \_Were there any difficult aspects of the project?\_

- \_Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?\_

### Improvement

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

- \_Are there further improvements that could be made on the algorithms or techniques you used in this project?\_

- \_Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?\_

- \_If you used your final solution as the new benchmark, do you think an even better solution exists?\_

-----------

\*\*Before submitting, ask yourself. . .\*\*

- Does the project report you’ve written follow a well-organized structure similar to that of the project template?

- Is each section (particularly \*\*Analysis\*\* and \*\*Methodology\*\*) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?

- Would the intended audience of your project be able to understand your analysis, methods, and results?

- Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?

- Are all the resources used for this project correctly cited and referenced?

- Is the code that implements your solution easily readable and properly commented?

- Does the code execute without error and produce results similar to those reported?