**Capstone Project**

**Machine Learning Engineer Nanodegree**

**Toxic Comments Classification**

**Mohammed Almohammedsaleh**

# Definition

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Criteria** | **Meets Specifications** | | --- | --- | | Project Overview | Student provides a high-level overview of the project in layman’s terms. Background information such as the problem domain, the project origin, and related data sets or input data is given. | | Problem Statement | The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made. | | Metrics | Metrics used to measure the performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem. | |

## Project Overview

#### Project domain

Nowadays, the majority of individuals are active on at least one social networking platform. Not just adults, but even teenagers use them. Nonetheless, some people abuse these platforms to vent their wrath and rage via bullying and the use of filthy language. That is why it is critical to avoid this sort of behavior on the platform by banning anyone who use this language. It will be a tiresome task to force humans to read each remark and determine whether or not it is offensive. That is where Machine Learning comes in help; we can train a model to determine which comment is harmful.

#### Dataset and input

The dataset was obtained for free from a Kaggle competition page. It contains a large number of comments taken from Wikipedia talk page.

The dataset is divided into three csv files:

* train.csv: containing 150000 comments for training.
* test.csv: containing another 150000 comments for testing (without labels).
* test\_labels.csv: containing labels for test.csv entries.

## Problem Statement

This project will design a machine learning model that will evaluate a huge set of comments and then utilize the trained model to predict the level of toxicity of a given comment. The level of toxicity will fall into one or more of the following types:

* toxic
* severe\_toxic
* obscene
* threat
* insult
* identity\_hate

## Evaluation Metrics

To evaluate our model, we will use the following metrics:

* Prediction accuracy score: which describe how will the model is predicting the testing data.
* Hamming Loss: which describe the percentage incorrect predictions.

In my proposal I chose accuracy score to be my evaluation metrics to the model. However, after further research I found that accuracy score does not describe the whole story. Instead, I will be using precision, recall, and f1-score to see how the model is performing for each class. To achive that, metric.classification\_report will be used.

# Analysis

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Criteria** | **Meets Specifications** | | --- | --- | | Data Exploration | If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics of the data or input that need to be addressed have been identified. | | Exploratory Visualization | A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined. | | Algorithms and Techniques | Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem. | | Benchmark | Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained. | |

## Data Exploration

Data Input Fields:

* id: comment id, will not be used
* comment\_text: contains the raw text for the comment
* toxic: binary field indicating whether the comment is toxic
* sever\_toxic: binary field indicating whether the comment is extremely toxic
* obscene: binary field indicating whether the comment is obscene
* threat: binary field indicating whether the comment is a threat
* insult: binary field indicating whether the comment is an insult
* identity\_hate: binary field indicating whether the comment consist of identity has

The dataset contains 159571 data entry. However only 16225 comments are classified to at least one class. That indicates that there is an imbalance in the dataset

## Exploratory Visualization

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 2:

Figure 1: Data distribution between classes

Figure 1 shows that ‘toxic’ class takes the majority of the dataset entries, but only ten percent of data is classified as shown in figure 2.

## Algorithms and Techniques

## Benchmark Model

BinaryRelevance algorithm provided by scikit-multilearn library will be my benchmark model for this project.

# Methodology

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Criteria** | **Meets Specifications** | | --- | --- | | Data Preprocessing | All preprocessing steps have been clearly documented. Abnormalities or characteristics of the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified. | | Refinement | The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary. | | Implementation | The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed. | |

## Data Preprocessing

## Refinement

## Implementation

### Solution Statement

This is a multi-label classification problem, where we want to classify each comment into all of the six categories. Hence, a comment can be toxic, severe toxic, obscene, threat , insult, and identity hate at the same time. Since these categories are not mutually exclusives.

My approach to solve this problem is to train a machine learning model with a multi-label classification algorithm, after processing and splitting the data into training and testing sets. Then use this model to predict a comment provided by the user and return an output as a JSON object that contains all the labels that describe the comment.

# Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| | **Criteria** | **Meets Specifications** | | --- | --- | | Justification | The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem. | | Model Evaluation and Validation | The final model’s qualities—such as parameters—are evaluated in detail. Some type of analysis is used to validate the robustness of the model’s solution. | |

## Justification

## Model Evaluation and Validation

# Conclusion