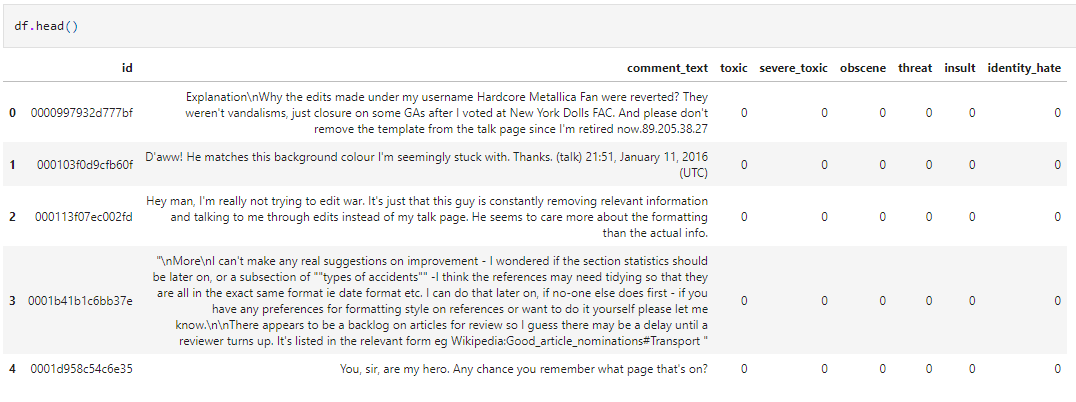
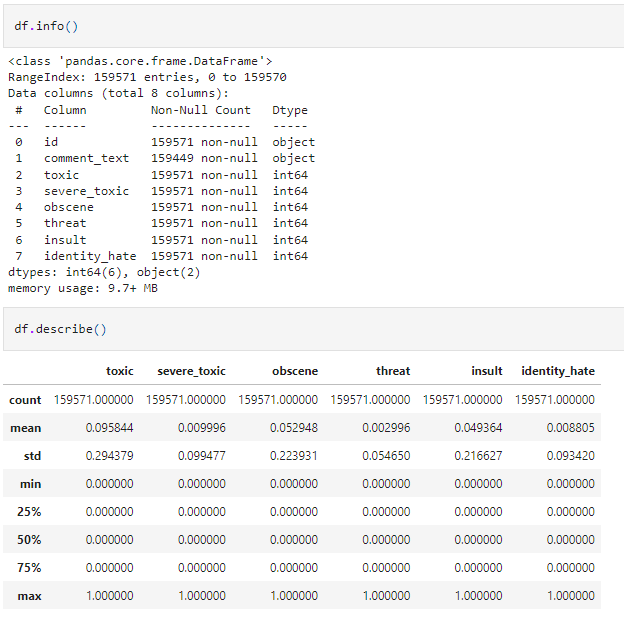
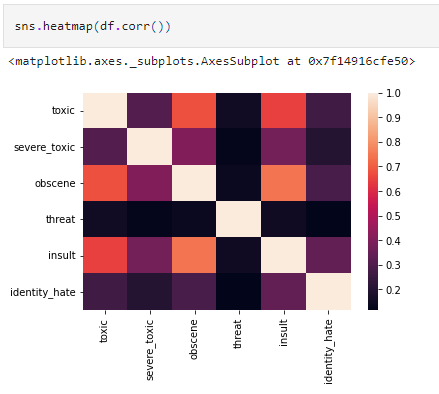
**Business Problem:**

Given a user comment, we need to identify the classes of toxicity it will fall into. This is a Multi-label classification (toxic, severe\_toxic, obscene, threat, insult, identity\_hate)

**Exploratory Data Analysis:**

****

****

****

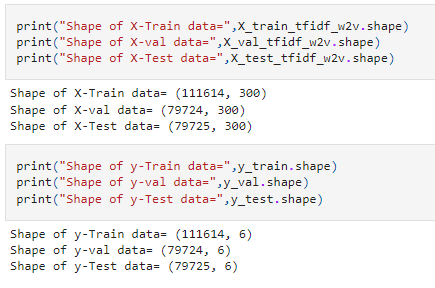
****

**Data preprocessing:**

* Removing special characters.
* Converting everything to lowercase
* Stop word removal
* Lemmatization
* Removing extra space characters
* Removing words with less than 2 characters
* Removing NULL values

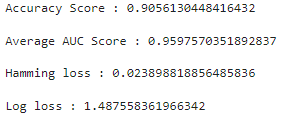
**Vectorization:**

Tf-Idf weighted word2vector

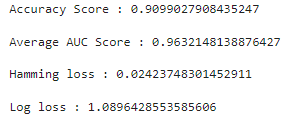


**Baseline Models:**

BinaryRelevance



Label PowerSet



**Neural Networks:**

DNN:

accuracy: 0.9670

avg\_AUC: 0.9728

Hamming\_Loss: 0.0205

Log\_Loss: 1.6014

Bi-directional LSTM:

accuracy: 0.9917

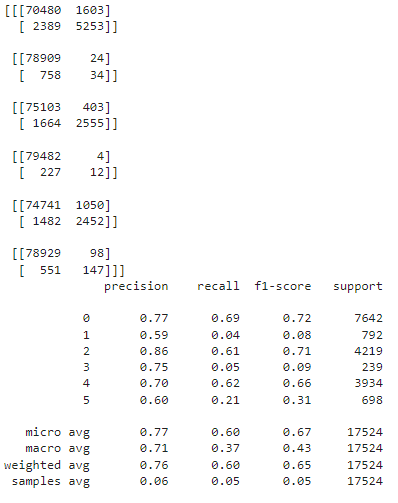
avg\_AUC: 0.5366

Hamming\_Loss: 0.0432

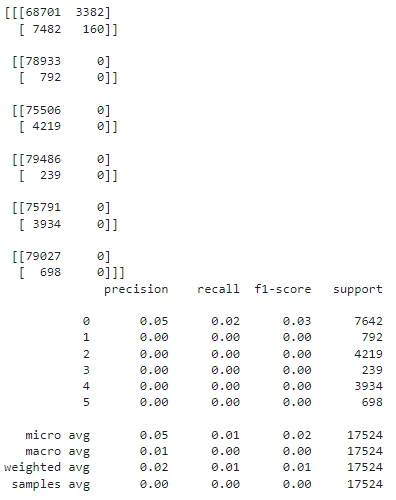
Log\_Loss: 0.4641

**Confusion Matrix:**

DNN:



Bi-directional LSTM:



**Future work:**

* Splitting the train, test and validation datasets using stratification technique.
* Performing Hyper-parameter tuning for the Deep Learning models using Optuna library.
* Using Bert
* …