Info about dataset:

You are provided with a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of toxicity are:

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

You must create a model which predicts a probability of each type of toxicity for each comment.

File descriptions

* train.csv - the training set, contains comments with their binary labels
* test.csv - the test set, you must predict the toxicity probabilities for these comments. To deter hand labeling, the test set contains some comments which are not included in scoring.
* sample\_submission.csv - a sample submission file in the correct format

we worked with two attributes – comment\_text and toxic .

Achieved Accuracy:

Accuracy of Naive Bayes Classifier: 93.95680147058823 %

Confusion Matrix For Naive Bayes(Multinominal) Classifier:

TN: 42479 FP: 698

FN: 2195 TP: 2500

Accuracy of Logistic Regression Classifier: 94.60853943850267 %

Confusion Matrix For Logistic Regression Classifier:

TN: 42918 FP: 259

FN: 2322 TP: 2373

Accuracy of Linear Support Vector Classifier: 95.34174465240642 %

Confusion Matrix For Linear Support Vector Classifier:

TN: 42254 FP: 923

FN: 1307 TP: 3388

Accuracy of Decision Tree Classifier: 94.64405080213903 %

Confusion Matrix For Decision Tree Classifier:

TN: 42335 FP: 842

FN: 1722 TP: 2973

Accuracy of Random Forest Classifier: 93.6559993315508 %

Confusion Matrix For Random Forest Classifier:

TN: 43050 FP: 127

FN: 2910 TP: 1785