Yahoo Troll Questions

Brute Force

Chinmay Parekh(IMT2020069)
BTV Sumanth(IMT2020072)

Given Dataset

The training dataset provided to us has three columns:

1.Question id

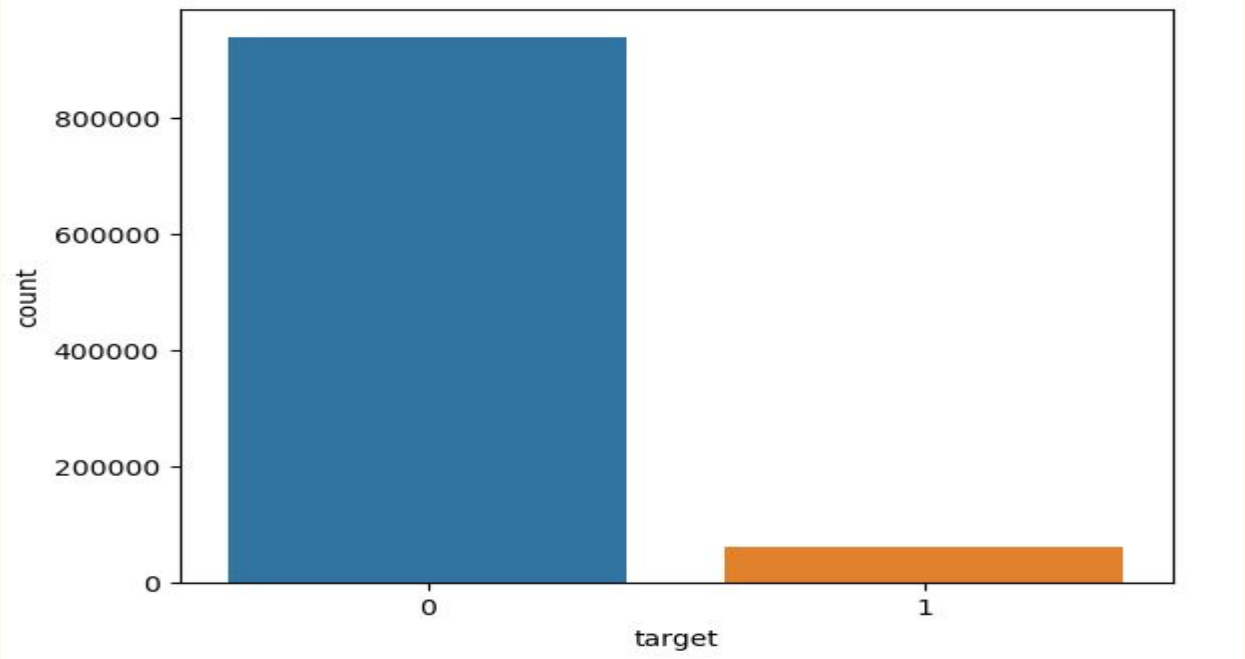
2.Question text

3.Target

It has around 10,00,000 rows

EDA

Count of Trolls



Preprocessing

1. **Tokenisation(Regex tokenizer)**

2. **Remove stop words**

3. **Stemming**/Lemmatisation

4. **Bag-of-words**(using Count-Vectoriser)

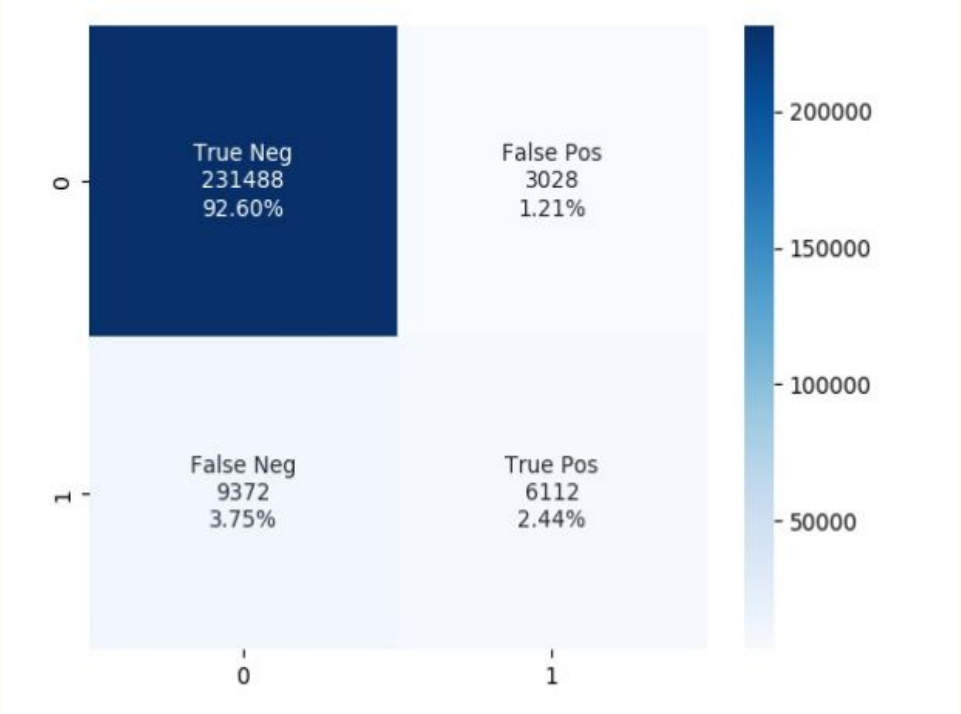
Models

- Gaussian NaiveBayes
- Multinomial NaiveBayes
- Logistic Regression

Logistic Regression

F1 score(for validation data):

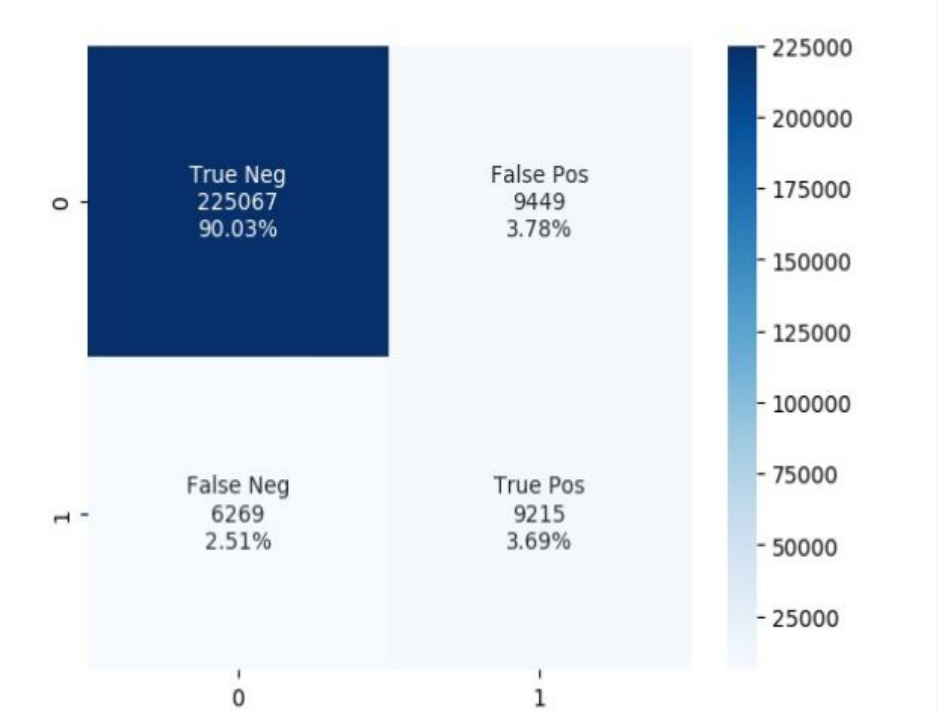
0.7351708178432523



Multinomial NaiveBayes

F1 score(for validation data):

0.7529845851387855



What next?

Tf-Idf

Feature Engineering

- Correct Grammar Usage
- Length of comments

Undersampling to handle biased data using imbalance-learn library

Word2Vec

Instead of removing all stop words,we plan on removing words having a positive sentiment

What next?

- Decision Trees
- XG Boost
- Random Forest
- SVM

Project Link

https://github.com/sumanth2002629/Yahoo-troll-questions-detection/