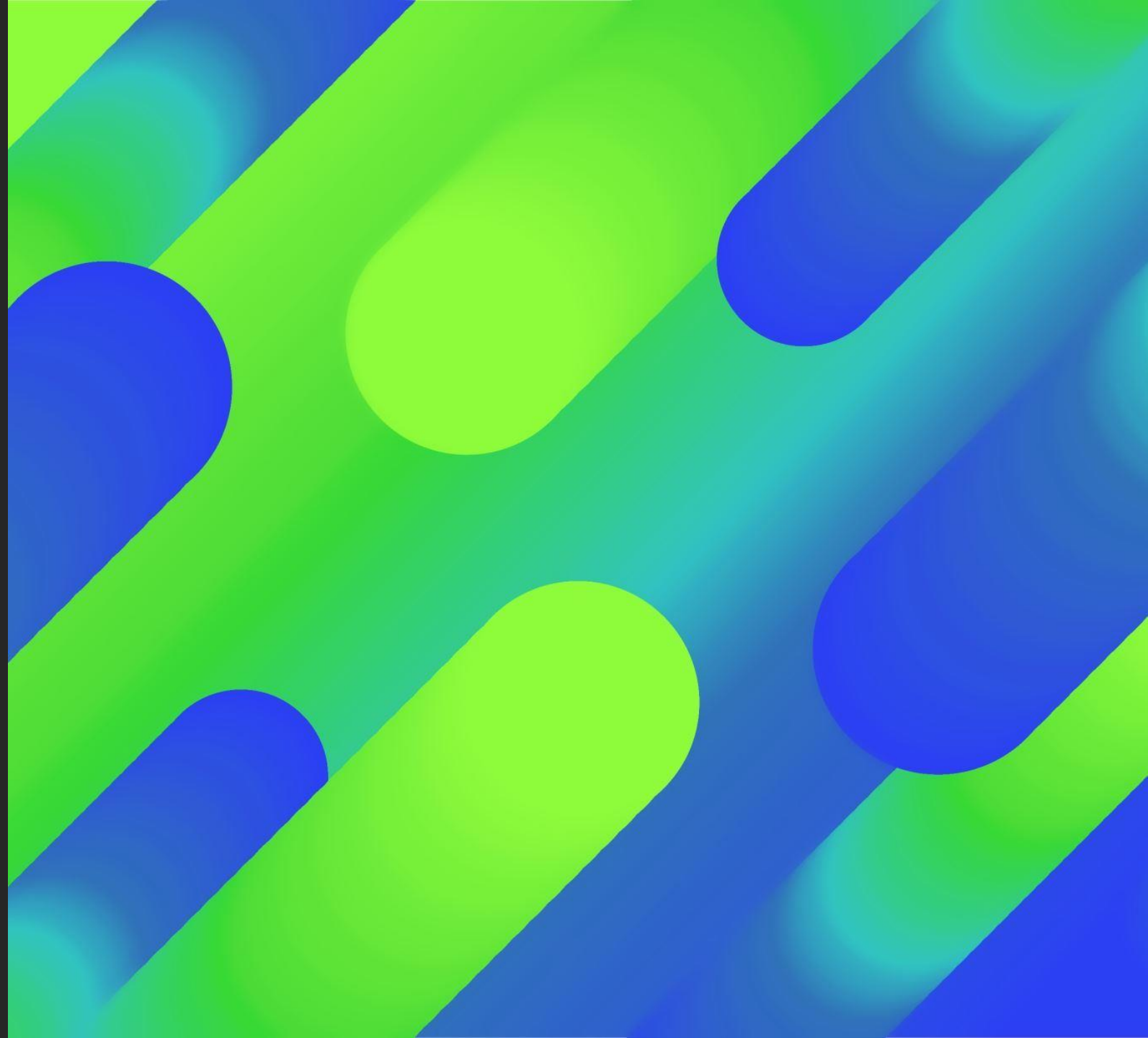
Yahoo Troll Questions

Brute Force

Chinmay Parekh(IMT2020069)

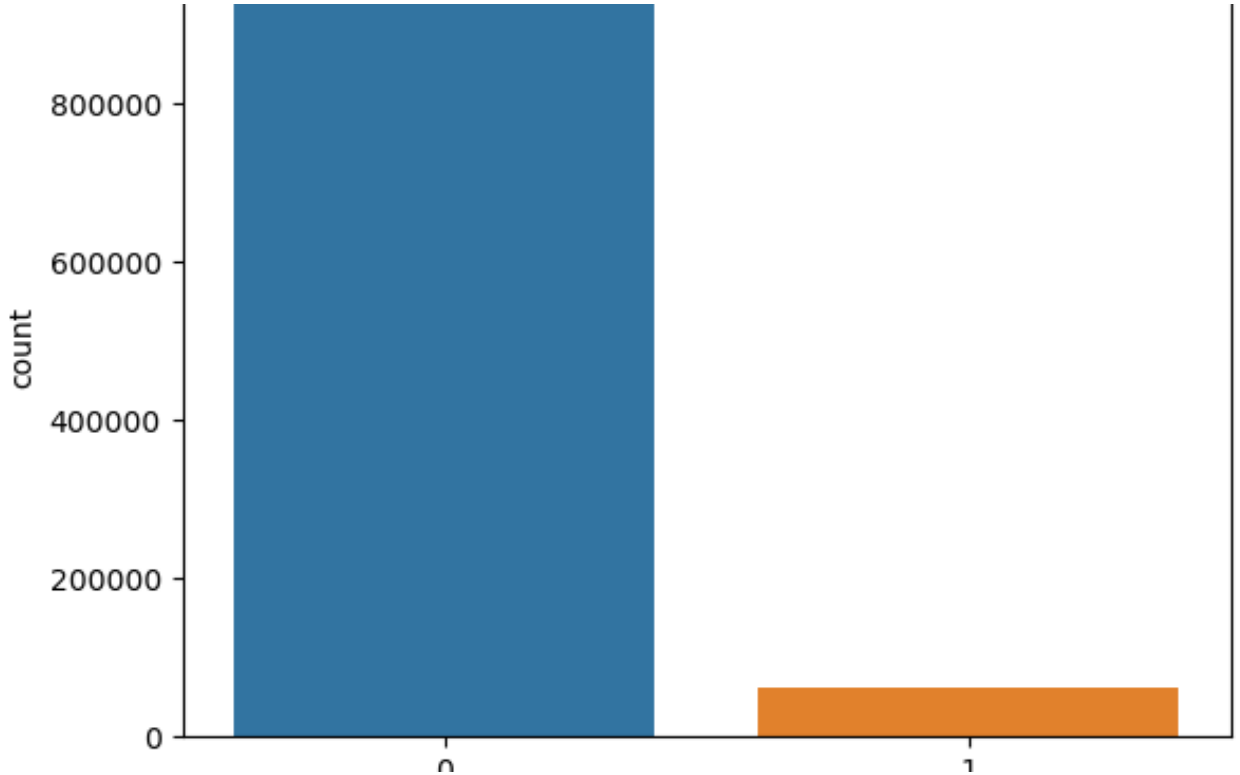
BTV Sumanth(IMT2020072)



Given Dataset

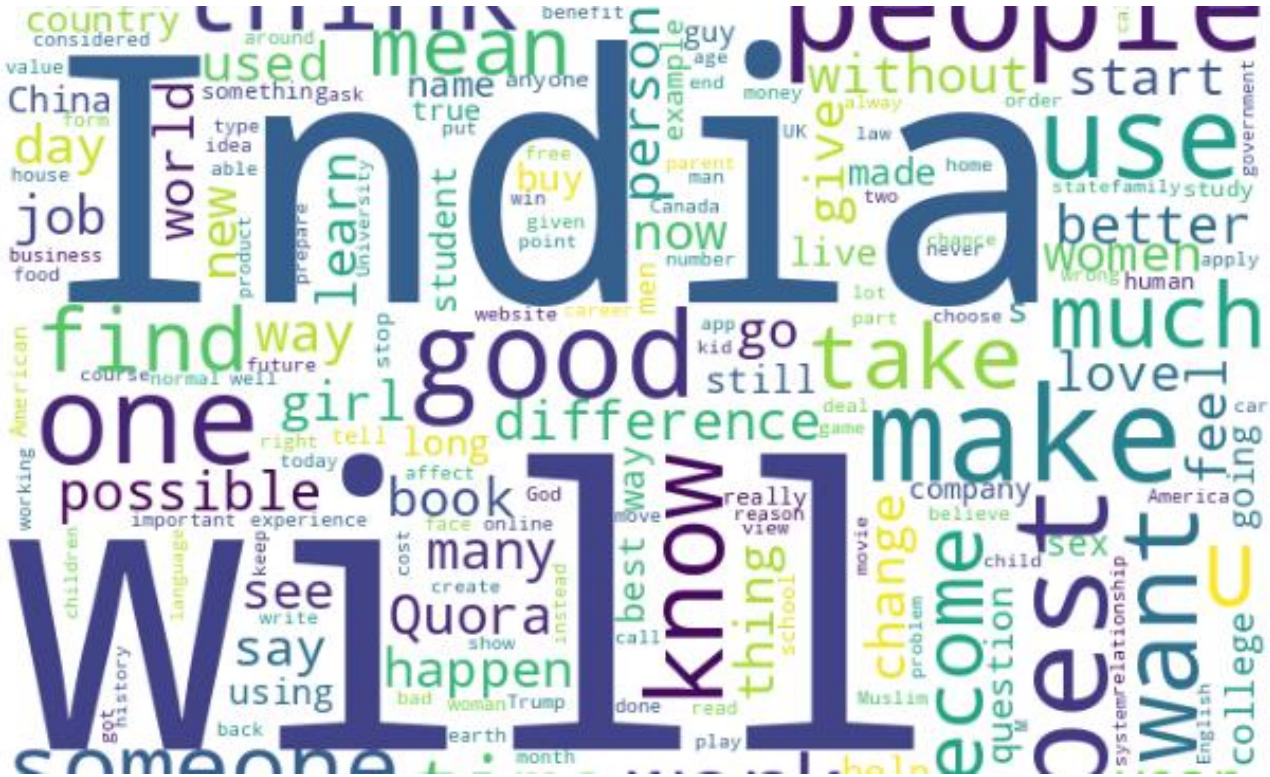
- The training dataset provided to us has three columns:
- 1.Question id
- 2.Question text
- 3.Target
- It has around 10,00,000 rows

Exploratory Data Analysis



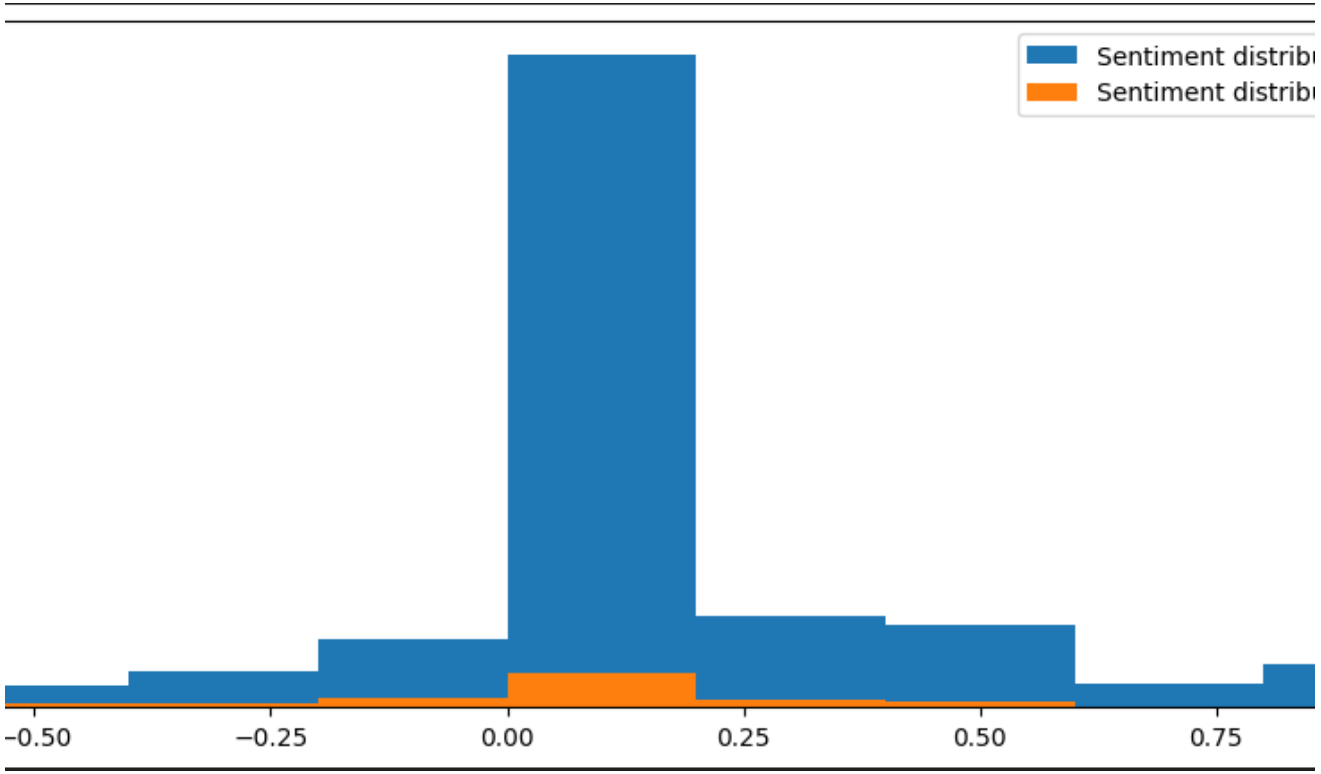
- Count of Trolls

Word Cloud



- Most frequently used words in the dataset

Sentiment Analysis



- We performed sentiment analysis to get an idea of the troll sentiments.

Preprocessing

- For the sake of experimenting, we performed a lot of pre processing.
- We tried the following:
- Tokenization(Regex tokenizer)
- Removal of stop words
- Stemming/Lemmatization
- Identifying non english words.
- Auto-correct sentences
- TFIDF Vectorizer

Observation

Greater the preprocessing, lesser is the F1 score!

Training

- Models:
- Logisitic Regression
- Multionomial Naïve Bayes
- GridsearchCV
- Naïve Bayes
- SVC with different kernels
- XGBoost
- Decision Tree

Final Result

- For the final submission, we used the following:

model = LogisticRegression(max_iter=3000,C=0.25, solver='lbfgs', penalty = 'l2',class_weight={0: 0.4, 1: 0.8},dual=False,intercept_scaling=1000)

We also made use of count vectorizer along with n-gram_range = (1,3), this increased our F1 score significantly.