

Toxic Comment Classification Challenge

Solutions by Divyank Singh

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Toxic Text Classification

A method of classifying a comment into toxic or not-toxic depending upon its inherent features

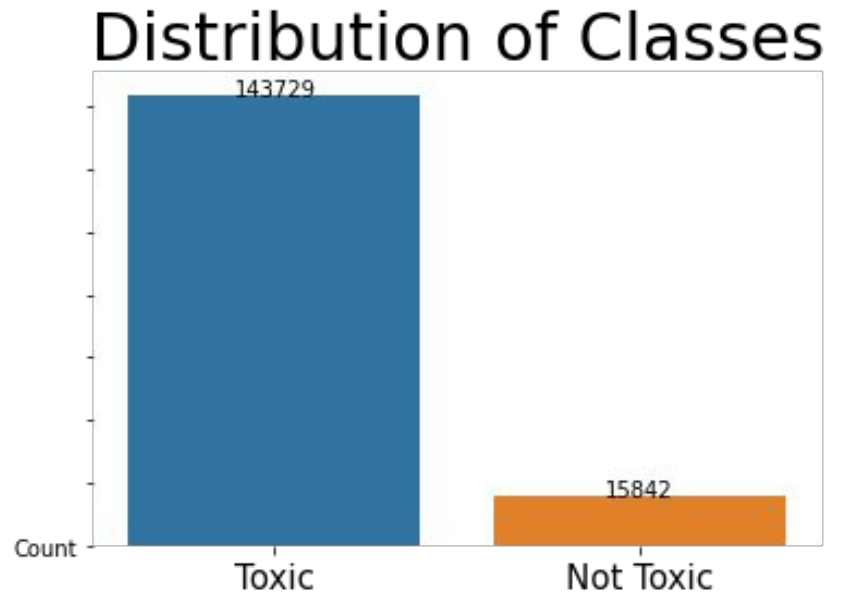
About the challenge

- A kaggle competition where the dataset consists of a large number of wikipedia comments labeled by human raters for toxic behaviour.
- The six labels:
 - Toxic
 - Severe-toxic
 - Obscene
 - Threat
 - Insult
 - Identity hate

Data Analysis

Significant class imbalancing

{0:143729, 1:15842}



Steps Followed

1. Data preprocessing
 - a. Remove numbers
 - b. Strip punctuation
 - c. Remove single characters
 - d. Cleaned multiple spaces
 - e. Vectorization
2. Resampling
 - a. Oversampling minority classes

Resampled data

1 143729

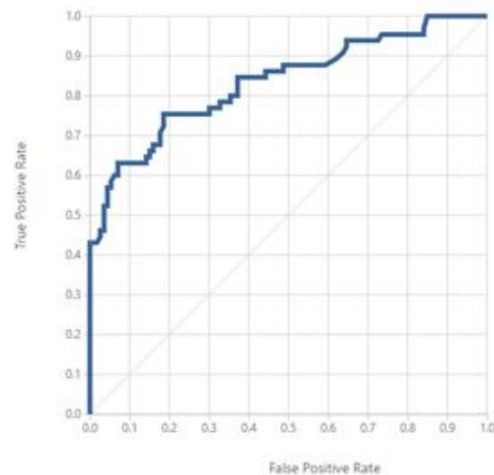
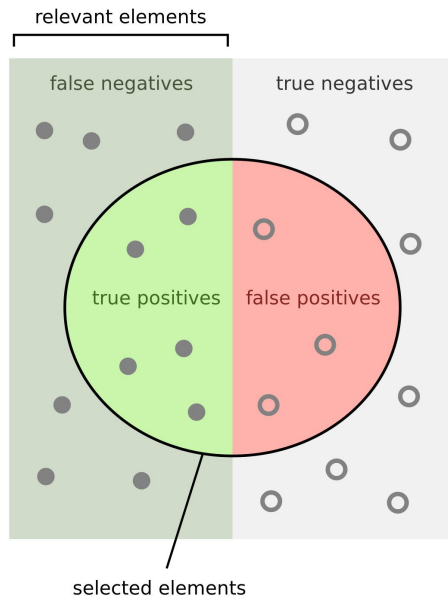
0 143729

Name: label, dtype: int64

Algorithms



Evaluation Metrics



How many selected items are relevant?

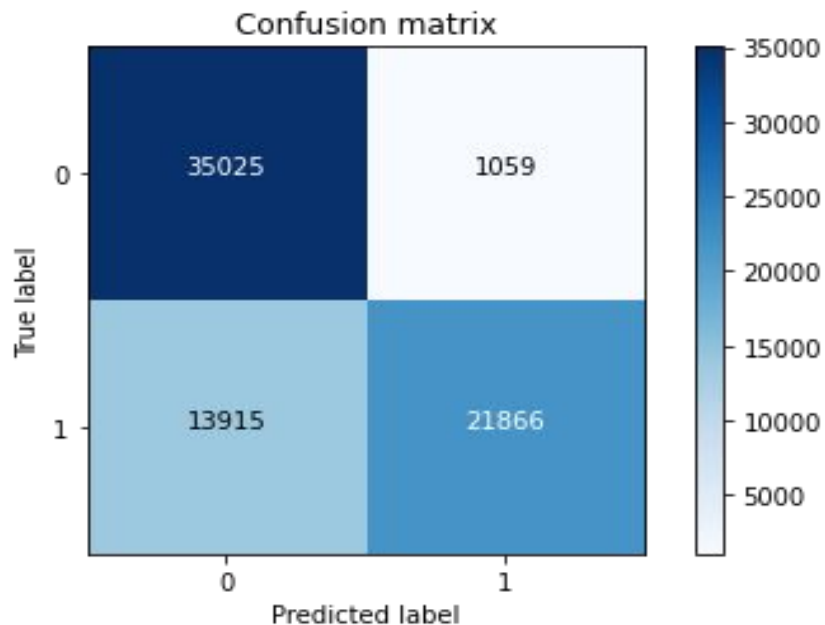
$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Decision Trees

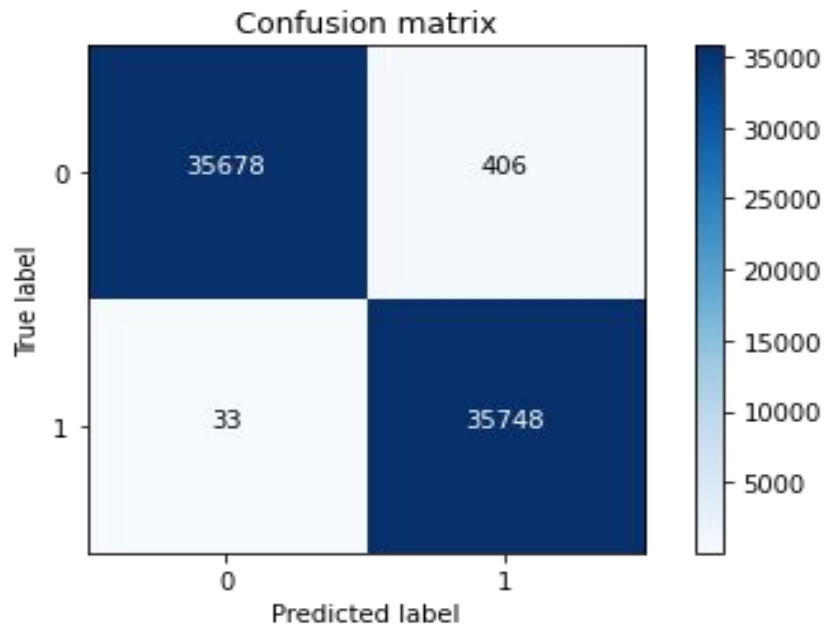
Performance



- Accuracy - 0.79
- Precision - 0.95
- Recall - 0.61
- F1 score - 0.74
- Area under ROC curve - 0.79
- Cross Validation scores - [0.79372782
0.79224936 0.79431921 0.79363726
0.79233271]

Random Forest

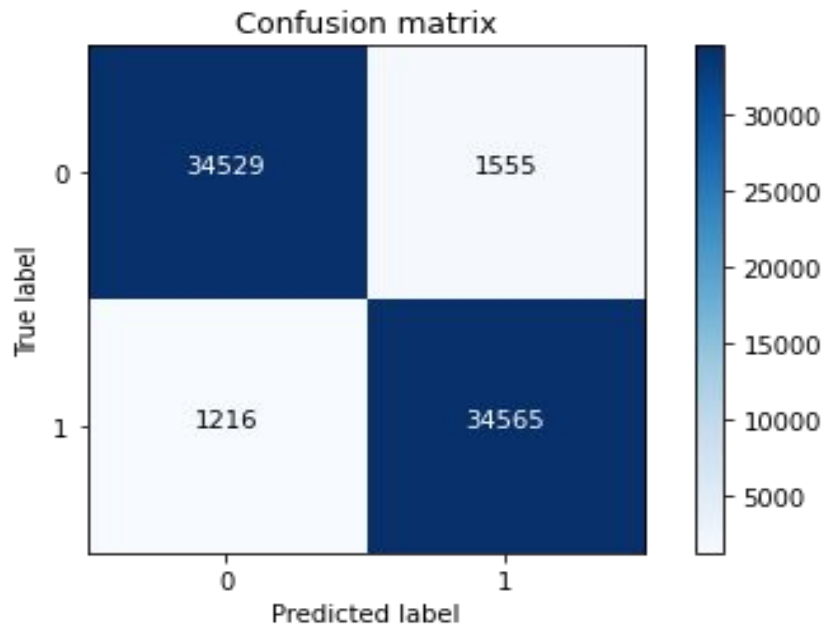
Performance



- Accuracy - 0.99
- Precision - 0.98
- Recall - 0.99
- F1 score - 0.99
- Area under ROC curve - 0.99
- Cross Validation Score - [0.99443401
0.99476449 0.99486885 0.99415561
0.99441652]

Logistic Regression

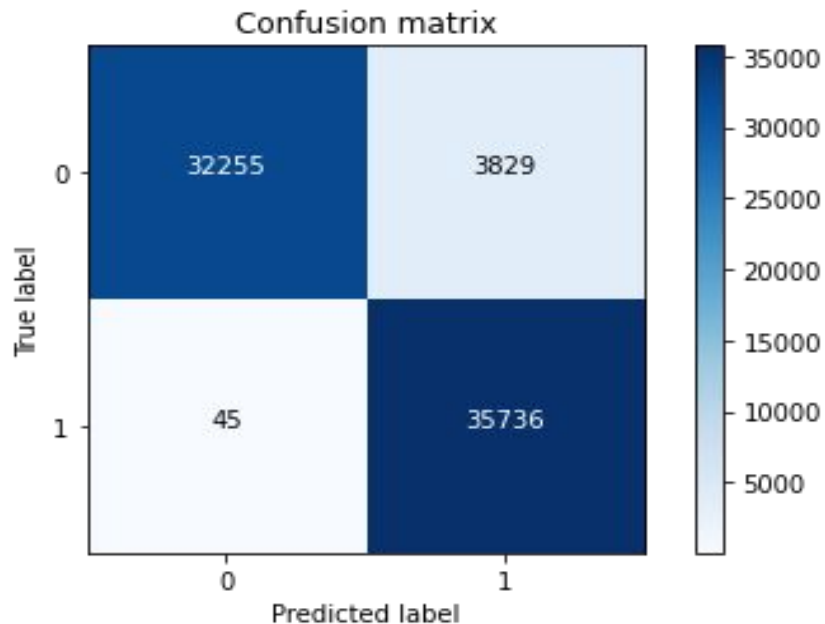
Performance



- Accuracy - 0.96
- Precision - 0.95
- Recall - 0.96
- F1 Score - 0.96
- Area under ROC curve - 0.96
- Cross Validation score - [0.96098588
0.96143811 0.96265567 0.96168096
0.96260284]

K Nearest Neighbor Classification

Performance



- Accuracy - 0.94
- Precision - 0.90
- Recall - 0.99
- F1 Score - 0.94
- Area under ROC curve - 0.94
- Cross Validation Score - [0.89854241 0.92727684 0.93002505 0.92231828 0.89916683]

Hyper Parameter Tuning

Grid Search on Logistic Regression

- Tuned Logistic Regression Parameters: {'C': 31.622776601683793}
- Best score is 0.9817190729406505

Randomized Search on Decision Tree

- Tuned Decision Tree Parameters: {'criterion': 'entropy', 'max_depth': None, 'max_features': 25}
- Best Score is 0.9686667225337686

Scope for future

A hand holding a smartphone is shown in the background. A bright blue rectangular overlay is positioned in the center of the phone's screen. On this overlay, the text 'WHICH SOCIAL MEDIA PLATFORM HAS THE MOST TOXIC COMMENTS?' is written in white, bold, uppercase letters. Below the text, there is a thin white horizontal line. The background of the phone screen shows some blurred social media icons, including a blue 'in' logo for LinkedIn.

**WHICH SOCIAL MEDIA
PLATFORM HAS THE MOST
TOXIC COMMENTS?**
