Kindness Project: Toxic Comment Classifier

Project Owner: Ting-Hsuan (Doris) Lin

Mentor: Dipanjan Sarker

Problems Identification

- Problem:
  - Online profanity such as abusive language and cyberbullying not only cause psychologically negative impacts to individual users but also make platform providers lose users due to the poisoned conversation environment
  - Effectively tacking online hate speech became urgent after governments required social media companies to take down inappropriate content within limited time and imposed penalty to those who failed to follow [link].
- Challenge:
  - Manual monitoring to mitigate online profanity is expensive and inefficient.
  - Online conversation is colloquial with slangs and mis-spelled words, making toxic comment hard to be detected.
  - People are worried about if their freedom of speech is violated when their innocent comments are incorrectly flagged as inappropriate content.

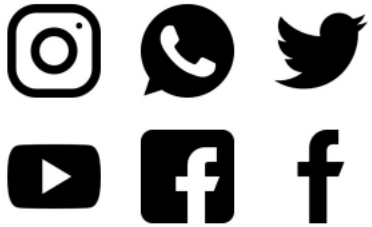
Proposed Solution

- Companies with messaging features and online communication platform can reduce the cost of manual monitoring with machine learning solutions with below benefits:
  - Detect types of toxicity comment data to help online discussion platform improve user experience.
  - Automate manual labeling process by building multi-headed classifiers
  - Provide clear visualization for non-tech audience of content insights of different types of toxic comments
  - However, the current solution to mitigate online profanity is through manual monitoring, which faces the challenges of high costs and low efficiency.

Proposed Solution

- Toxic comment classifier is a natural language processing application that allows online platforms to detect the toxicity of a message based on its textual context, with the long term goal to improve the quality of online conversation.
  - Detect types of toxicity comment data to help online discussion platform improve user experience.
  - Automate manual labeling process by building multi-headed classifiers
  - Provide clear visualization for non-tech audience of content insights of different types of toxic comments

Business Applications



- Social media platform and chat apps
- To spot and remove hate speech, complying with government regulations.

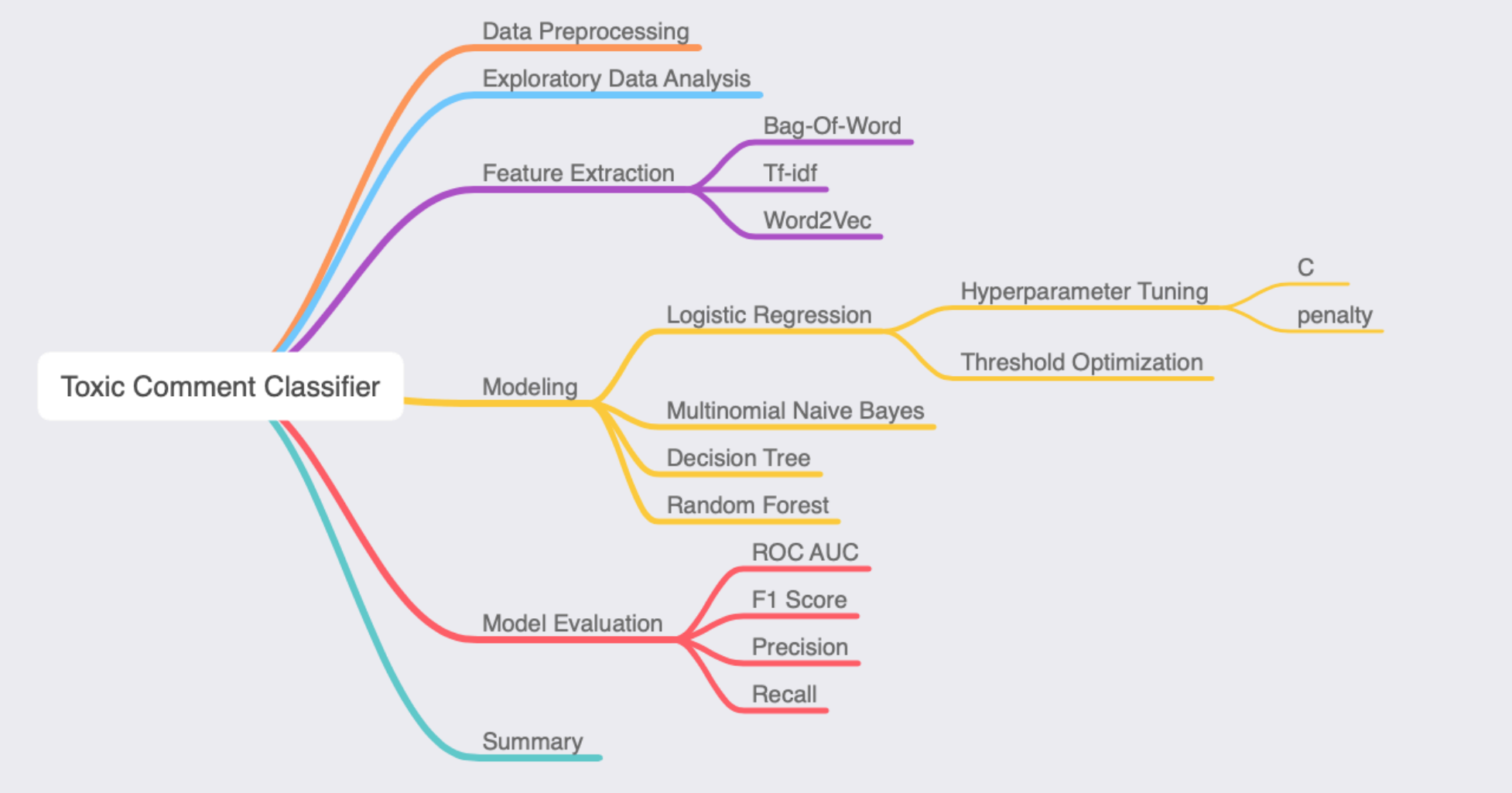
- Video game companies targeting children/ teenagers
- To monitor the conversation in chat windows and block abusive language

- Individuals, for example parents
- To add the classifier to chrome extensions to filter contents.

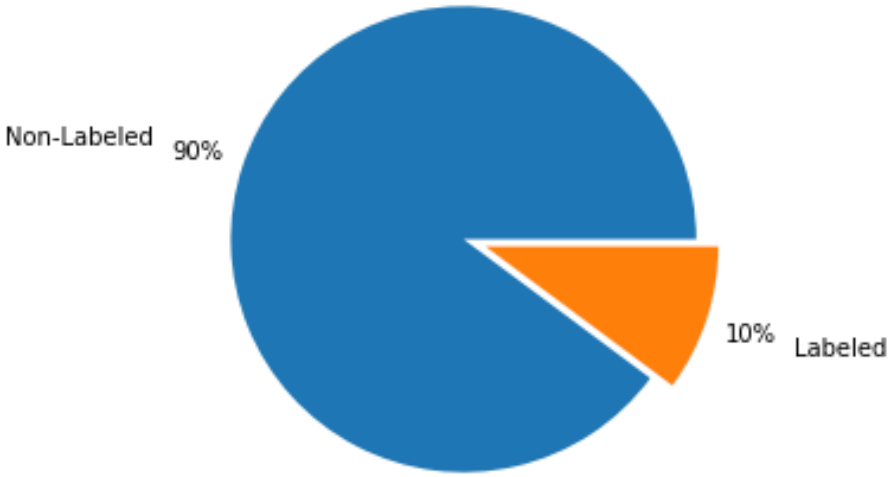
Data Source

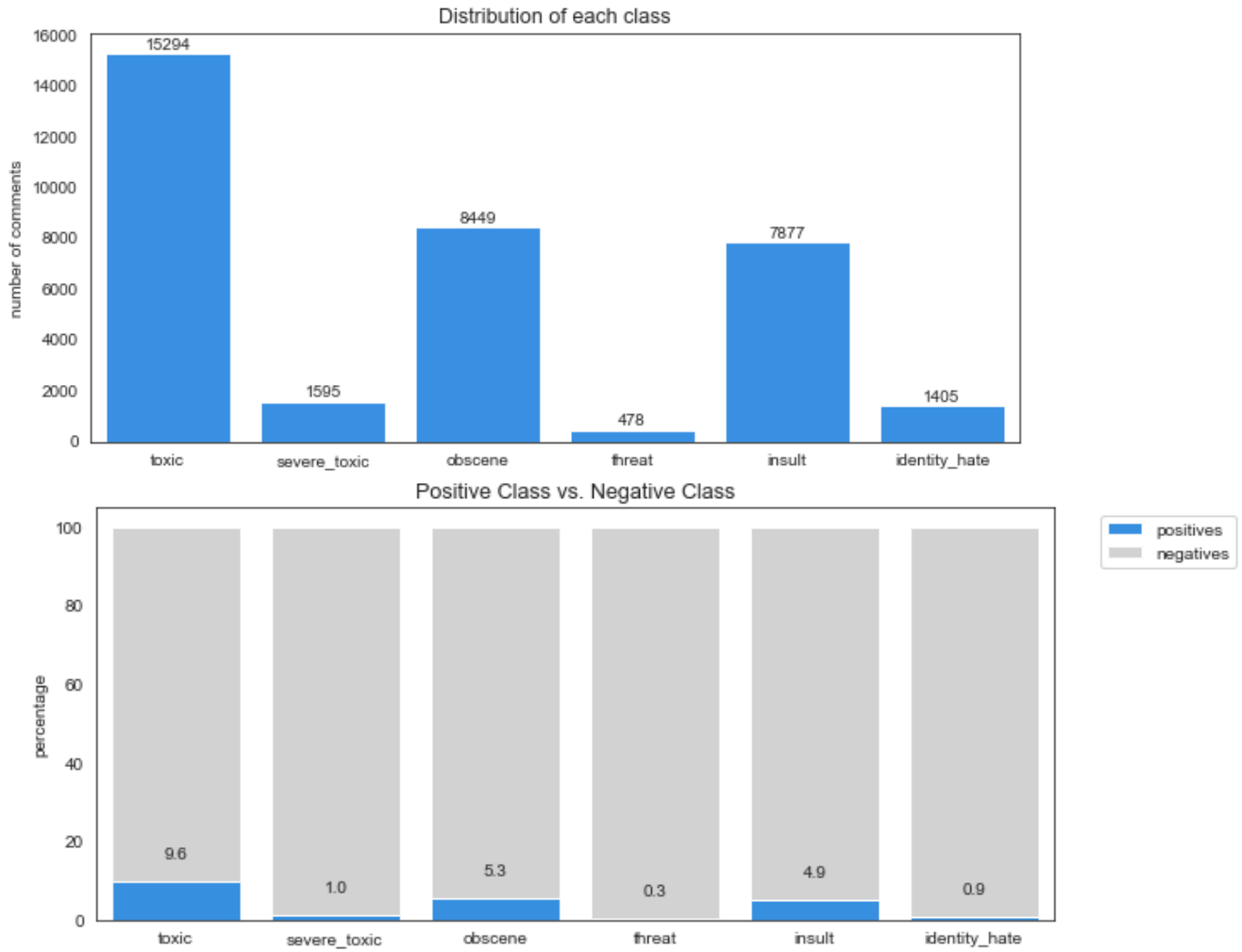
- The data covers 160,000 comments from English Wikipedia talk page edits, dating for 2004 to 2015. It is provided by Google Jigsaw for Kaggle's Toxic Comment Classification Challenge.
- Data are labeled by 5,000 workers with one or more toxic types, including toxic, server toxic, obscene, threat, insult and identity hate.

Project Structure

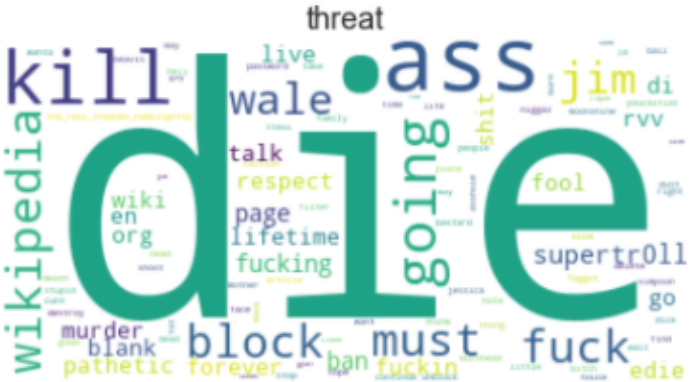


Imbalance in toxic and non-toxic volume





Frequent words reflects common sense

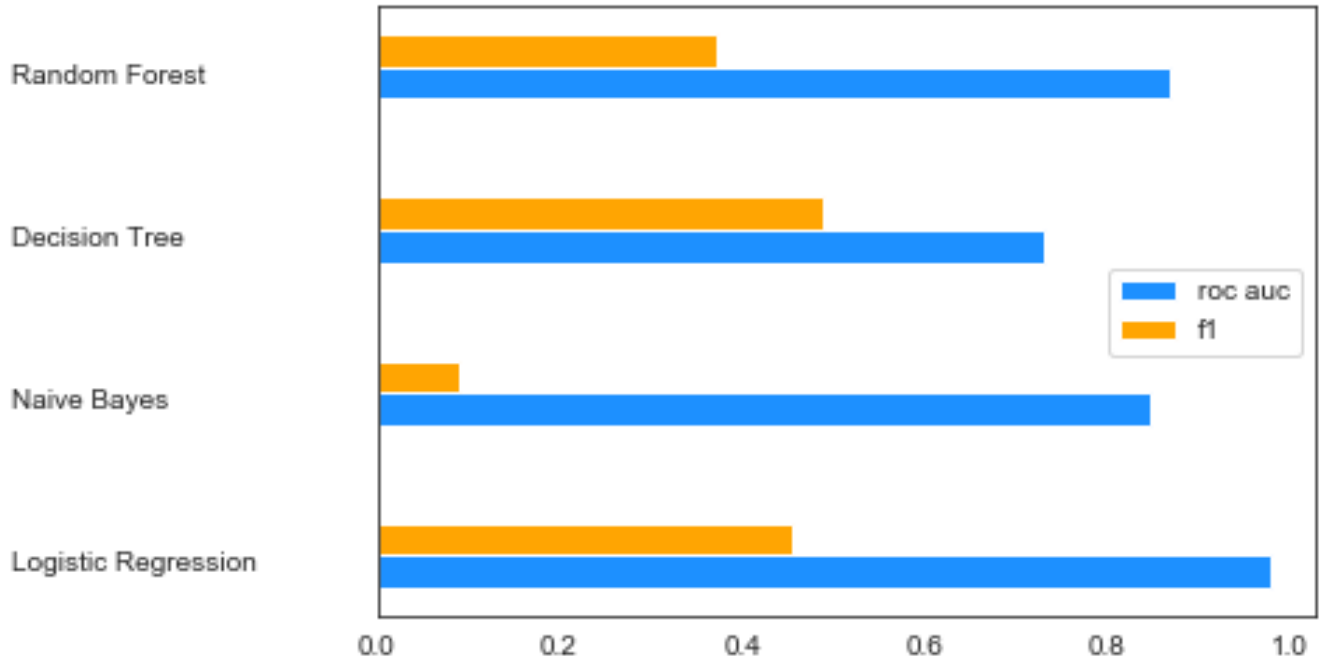


ROC AUC & F1 score serve as key metrics

- The imbalance issue of this dataset and cost of misclassification are considered when choosing suitable metrics for model evaluation.
- ROC AUC and F1 score are the main metrics since both of them are insensitive to the issue of data imbalance and can provide a general summary of the model overall performance.
- Precision and Recall on the other hand provide additional information as secondary metrics when there is a tie. Since the outcome of false alarm is more expensive than miss detection, the false negative error is more tolerable than false positive error. Models with higher precision at the expense of slightly low recall are preferred.

Logistic Regression outperforms others

- Applying Tf-idf to transform content into numerical values to four models.



Model Tuning

- Logistic regression might be prone to be overfitting when there are more independent features than volume of data. In order to avoid that, I focus more on hyper parameters “C” and “penalty” when tuning the model.
- It turns out the model with default hyper parameter setting is the best-performed one. Compared to the benchmark model, which guess the majority class all the time, the model improves by 18% ROC AUC and by 47% F1 score.

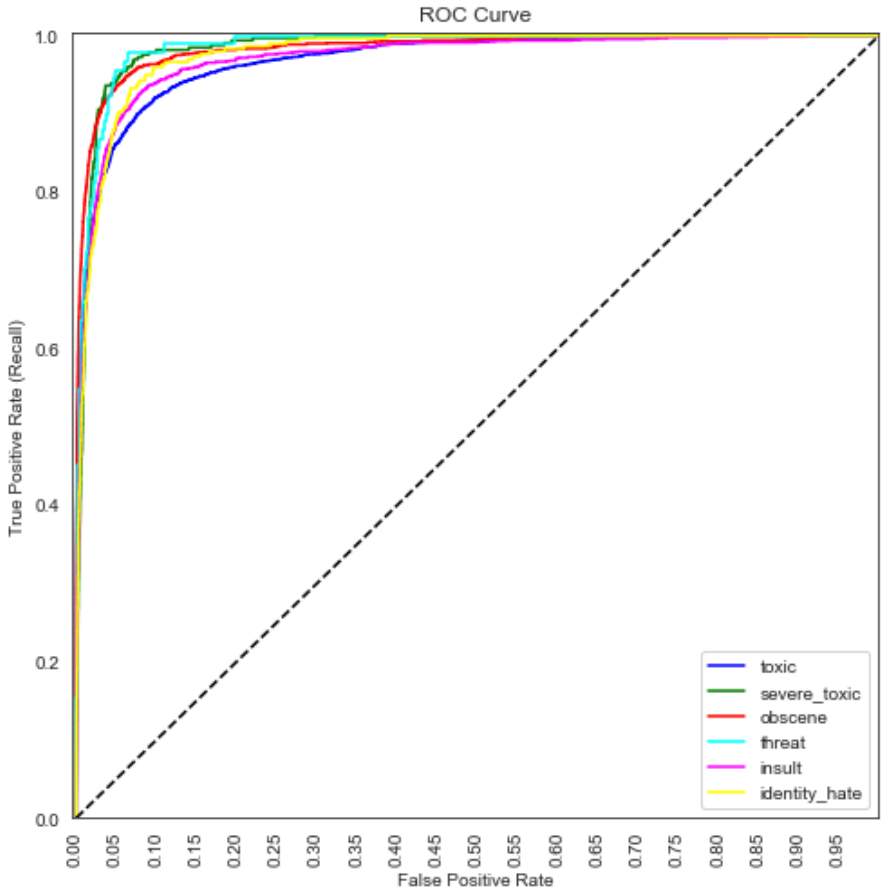
Benchmark model

```
average ROC AUC: 0.5
average f1 score: 0.0
average precison score: 0.0
average recall score: 0.0
```

Optimized model

```
average ROC AUC: 0.6835835414562642
average f1 score: 0.476909265770046 94
average precison score: 0.738031923260121
average recall score: 0.37001984298529633
```

Threshold optimization to reduce false alarms



- The default threshold at 0.5 has already provided good precision score. We can be very confident to trust the result when model predicts a sample Positive. The high precision reduces the risk when we take a suspected comment down from the platform, while it is indeed neutral.
- But 0.4 gives a good balance between precision and recall. While we don't want to piss off the user due to false alarm, we don't want to miss capturing toxic messages and cause negative impact on user behaviors. Thus, 0.4 is the recommended threshold value.

Next Step

- Random forest being the second best-performed model in the project. However, it requires expensive computational power for hyper parameter tuning, I have not gotten a chance to optimize the model. “XGBoost model”, also a tree model but is faster to run, might be an alternative model to use.