Capstone project 2: Toxic Comment Classification Challenge

1. Introduction

The goal of the project is to build a multi-headed model that is capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate better than perspectives' current models. The data of the project is from the Kaggle competition, its original source is the Wikipedia comments which have been labeled by human rater for toxic behaviors. The types of toxicity are: Toxic, Severe toxic, Obscene, Threat, Insult, and Identity hate.

The research question is to predict the probability of each type of toxicity for each comment. This project is an example of multi-label text classification, since the predicted properties of each comment are not mutually exclusive. The potential audience of this project could be Policy makers, regulators, parents and school officers, who try to prohibit and monitor the abuse or harassment of the online environment.

2. Data

1) Data source:
Competition on Kaggle

2) Data description:
With one explanatory variable which is the text comments, and 6 independent variables describing the types of the toxic comments.

3) Data wrangling steps:

Step1. Understand the data: column types, data frame shape, correlation between the variables.

Step2. Remove outliers

Step3. Imputing missing values

3. Data exploration and storytelling

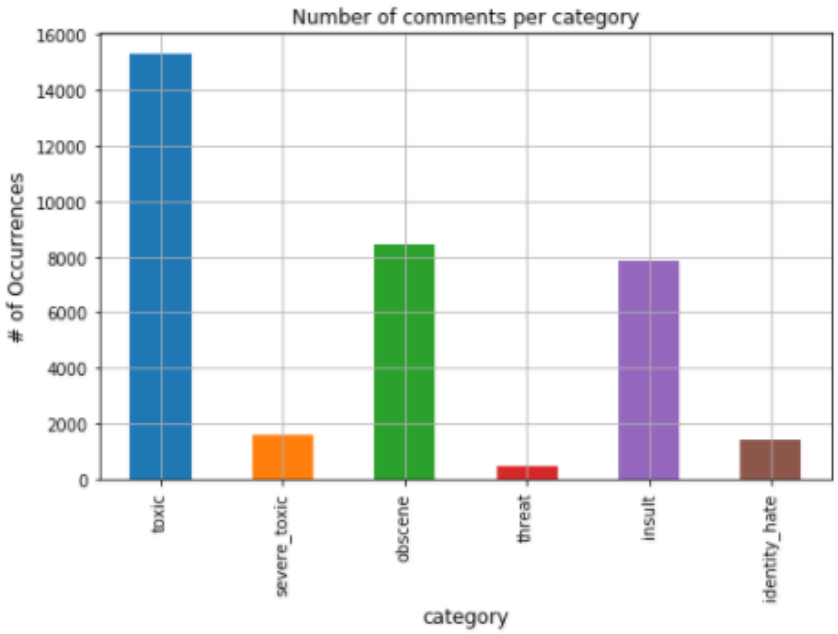
1) Correlation between the types of comments

	toxic	severe_toxic	obscene	threat	insult	identity_hate
toxic	1.000000	0.308619	0.676515	0.157058	0.647518	0.266009
severe_toxic	0.308619	1.000000	0.403014	0.123601	0.375807	0.201600
obscene	0.676515	0.403014	1.000000	0.141179	0.741272	0.286867
threat	0.157058	0.123601	0.141179	1.000000	0.150022	0.115128
insult	0.647518	0.375807	0.741272	0.150022	1.000000	0.337736
identity_hate	0.266009	0.201600	0.286867	0.115128	0.337736	1.000000

The above figure shows that comments being toxic is highly correlated with comments being obscene. (0.67) Also, comments being insult is highly correlated with comments being obscene (0.74).

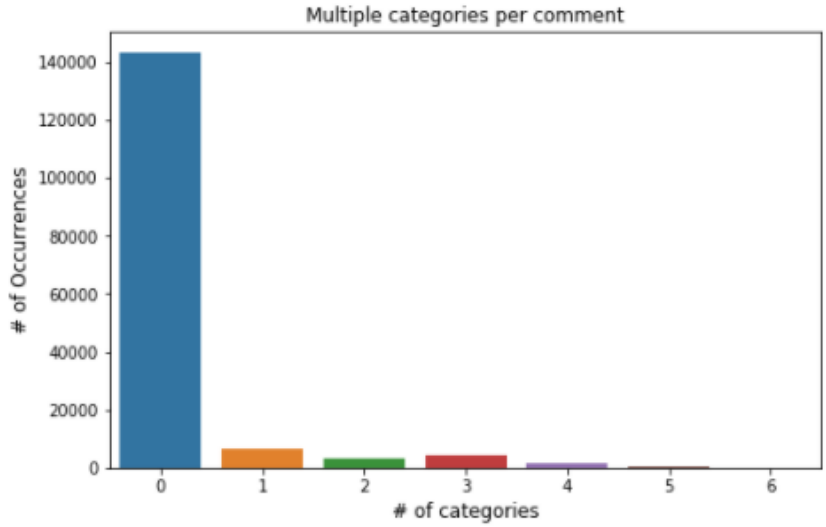
2) Out of all the 159571 comments, below are the number of comments in each type of toxicity.

	category	number_of_comments
0	toxic	15294
1	severe_toxic	1595
2	obscene	8449
3	threat	478
4	insult	7877
5	identity_hate	1405



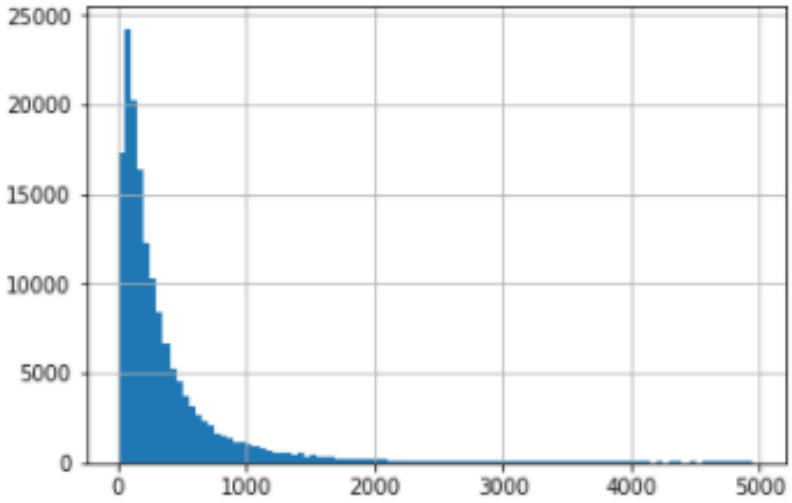
Comments that are labeled as bad are about one tenth of the total comments. Out of all the types of comments, 15294 comments are labeled as toxic, 8449 are labeled as obscene, and 7877 are labeled as insult.

3) How many comments have multiple labels?



89.8% of the comments have no labels. For the rest 10.2% of the comments, most of them have only one label.

4) The distribution of the number of words in comment texts



5. Machine learning and Hyper-parameter tuning

1) Method used: Naïve Bayes, LinearSVC, Logistic Regression

Test accuracy for Naïve Bayes is as follows:

```
... Processing toxic
Test accuracy is 0.9193300290548624
... Processing severe_toxic
Test accuracy is 0.9899922140564765
... Processing obscene
Test accuracy is 0.9516132095178412
... Processing threat
Test accuracy is 0.9971324939706413
... Processing insult
Test accuracy is 0.9517461402609241
... Processing identity_hate
Test accuracy is 0.9910366698949847
```

Test accuracy for LinearSVC is as follows: （The best parameter is C=0.1）

```
... Processing toxic
Test accuracy is 0.955278300005697
... Processing severe_toxic
Test accuracy is 0.9908087886211284
... Processing obscene
Test accuracy is 0.9769080309158928
... Processing threat
Test accuracy is 0.9972844148198788
... Processing insult
Test accuracy is 0.9692929983478608
... Processing identity_hate
Test accuracy is 0.9916443532919349
```

Test accuracy for Logistic Regression is as follows: （The best parameter is C=10）

```
... Processing toxic
Test accuracy is 0.9597789551643594
... Processing severe_toxic
Test accuracy is 0.9905619172411174
... Processing obscene
Test accuracy is 0.9782373383467213
... Processing threat
Test accuracy is 0.9974173455629617
... Processing insult
Test accuracy is 0.9706792760971534
... Processing identity_hate
Test accuracy is 0.9921380960519569
```

2) Stacked model（Using the Majority Voting Method）
   - The stacked model used the majority voting of the three models.
   - The accuracy score for the stacked model is better than the accuracy score for the three different models respectively.
   - Prediction as made using the stacked model, and the result was submitted to Kaggle.