Toxic Comment Classifier

Hien Duong
Eric Furukawa
Keuntae “Kurtis” Kim

Introduction

Classification of comments

“Don't mean to bother you”

Toxic: 0, Severe_Toxic: 0, Obscene: 0, Threat: 0, Insult: 0, Identity_Hate: 0

“IT WASNT VANDALISM, DICKHEAD”

Toxic: 1, Severe_Toxic: 0, Obscene: 1, Threat: 0, Insult: 1, Identity_Hate: 0

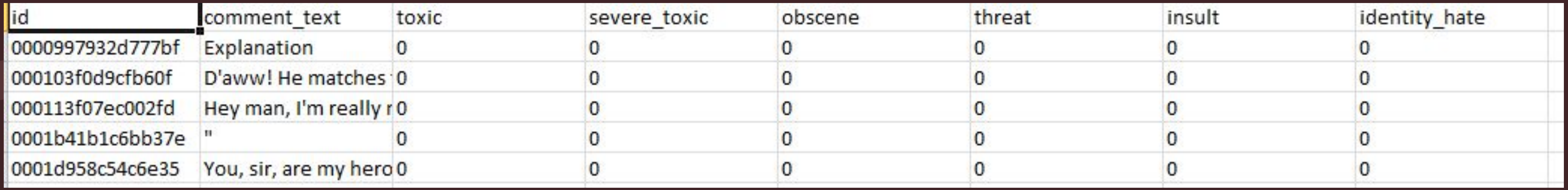
Dataset

Kaggle Toxic Comment Challenge Data set

159572 Training Set (Multi labeled)

153168 Test Set (Multi labeled)

Limited to one third for RAM performance in demo



id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0000997932d777bf	Explanation	0	0	0	0	0	0
000103f0d9cfb60f	D'aww! He matches	0	0	0	0	0	0
000113f07ec002fd	Hey man, I'm really r	0	0	0	0	0	0
0001b41b1c6bb37e	"	0	0	0	0	0	0
0001d958c54c6e35	You, sir, are my hero	0	0	0	0	0	0

Methods and Procedures

Read Test and Train data and vectorize Train data for word counts

Sklearn Naive Bayes Classifier

Our implementation Naive Bayes Classifier

Do binary classification for each class

Results

Report accuracy and f-macro

Accuracy was very high! Reflects high correlation in words and toxicity labels

F-macro was also high due to undersampling implemented to combat class imbalance

Future Works