Toxic Comments Classification Web App

Project Owner: Aisha Dubhashi
Developer: Yiwei Sun
QA: Jerry Chen

Motivation

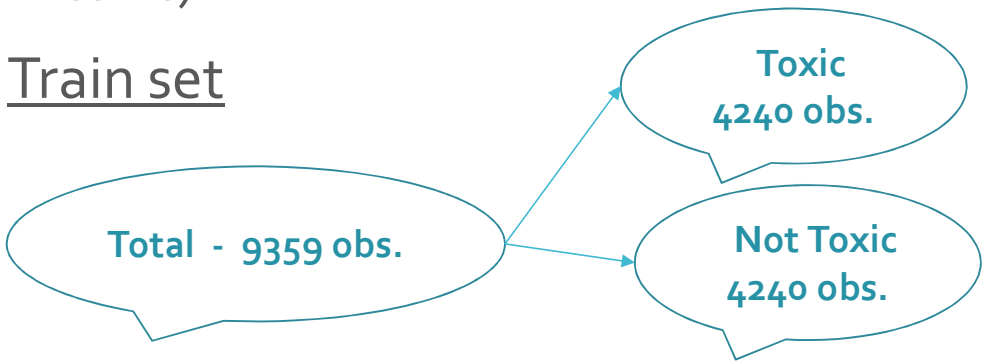
- Background
  Online comments are informative
- Problem
  They can be misleading with negative online behavior
- Solution
  Classify the toxic comment with a Web App

Datasets

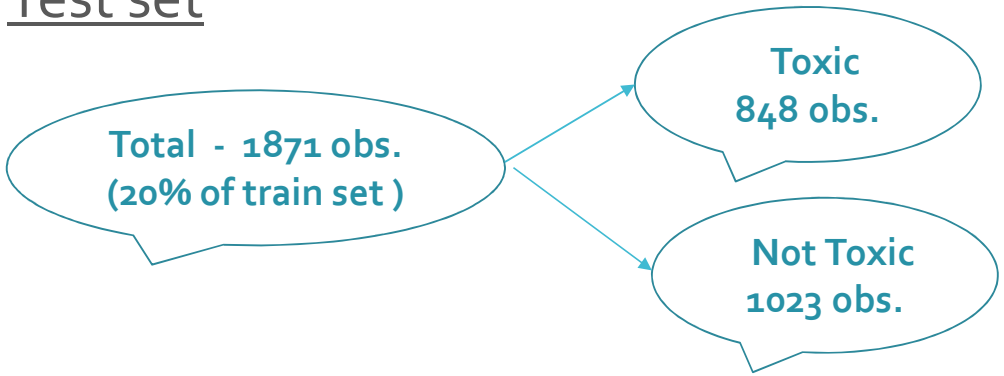
- Datasets

  Wikipedia comments that have been labeled by human raters for toxic behavior (toxic vs. non-toxic)

- Train set



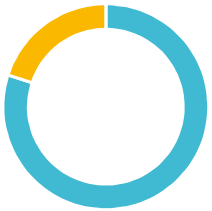
- Test set



Model & Success Criteria

- Model
  - Word & Characters Tokenization and Stemming by TfidfVectorizer() in Python
  - Logistic Regression to predict Toxic (1) or Not Toxic (0)

- Success Criteria
  - Performance of classification
  - Cross Validation Accuracy: 81.74%
  - Test Accuracy: 84.29%



Interesting Insight

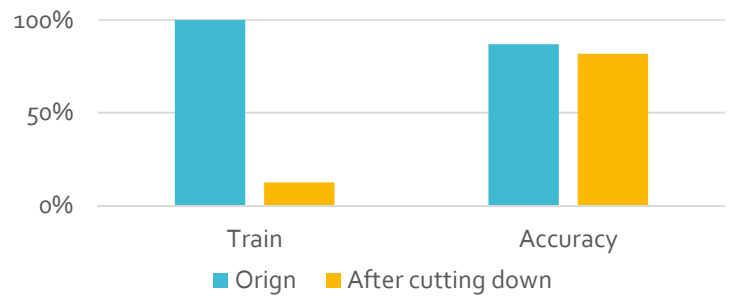
- Data

  The word "Toxic" is not Toxic comment – Not a rude or disrespectful word

- Model

  To increase time efficiency, training set has been cut down to **1/8** and the model 10-fold cv accuracy only decreases **6%**

  - Because of marginal effect for word tokenization



The End

- Thank you very much for listening!

- Contact Info:
  - Name: Yiwei Sun
  - Master of Science in Analytics (MSiA) 2018
  - Email: yiweisun2018@u.northwestern.edu