Offensive Language Detection
IS 6713

Eduardo Jones

Objective

- Create a Machine Learning model that can detect offensive language in tweets

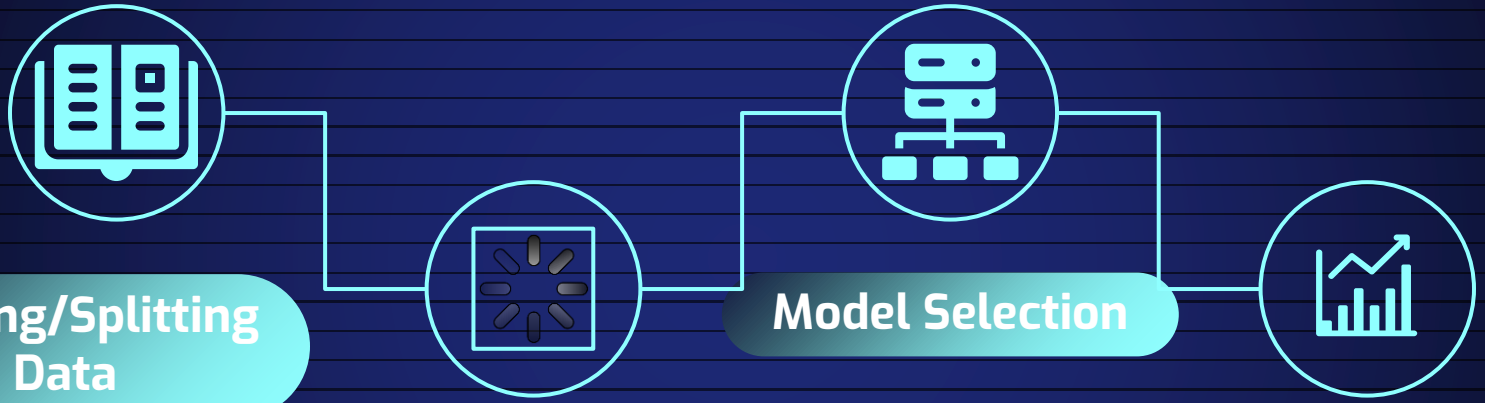
- Model seeks to label whether the tweet is :

  Not Offensive (NOT) : Posts that do not contain offensive or profanity

  Targeted Insult (TIN) : Posts containing insult/threat to an individual, A group, or others

  Untargeted (UNT) : Posts containing non targeted profanity and swearing Posts with general profanity are not targeted, but they contain non-acceptable language

OUR PROCESS



Loading/Splitting Data

Feature Engineering

Model Selection

Model Assessment

Loading/ Splitting Data

- Python for-loop to read each row of the datasets.
- Incorporated the CSV library and used the .reader method for our train/test datasets.
- Appended each row of the two datasets to separate empty lists to create the dataset matrices.
- For training and validation, we used a train_test_split method to train each model on a training dataset, allowing us to score our model on predictions for both the training and validation set.

Feature Engineering

- Number of offensive words: useful in identifying offensive tweets.
- Number of target words: determine an offensive tweet and a targeted tweet.
- Number of fully capitalized words: incorporate a measure of intensity to predict offensive language.

Score with Rule Based System

- Used arbitrary rules to label Tweets
- If there was at least one offensive word and one target word, then it would be labeled as “TIN”.
- One offensive word but no target words, then it would be labeled “OFF”.
- Micro f1 score → .6588

Score with CountVectorizer Features

- Utilized GridSearchCV grid with CountVectorizer and LinearSVC models.
- Best parameters: C: .1 , n_gram range(1,1),
- Micro f1 score → .7222

Score with Both Stacked Rule Based Features and CountVectorizer

- Utilized GridSearchCV grid with CountVectorizer and LinearSVC models.
- Parameters: {'C':[0.01, 0.1, 1.]}, cv =5
- Appended our features to the test data.
- Micro f1 score → .731

Error Analysis (False Positive) Example

- 95913 @USER “Trump will blame it on the immigrants for coming and 'liberals' for noticing. He likely thinks Head Start/HIV are wasteful spending. Congress can no longer pass 'broad' budgets. They will need to itemize every $ and forbid diversion without Congressional permission”
- Actual label: NOT
- Predicted Label: TIN

Error Analysis (False Negative) Example

- 21047 @USER He is still eating and talking about p**sy on carter V. Trust me
- Actual label: TIN
- Predicted label: NOT