Interpreting Code

With Natural Language Processing

Eileen Cai

Data Scientist

The problem

Problem

Spread of child sexual abuse material online and other harmful content, such as, cyber bullying, drug sales and hate speech.

Solution

Guardian of virtue: upholding digital ethics.

- Use machine learning to automatically detect harmful texts.

Impact

Social impact on teen's mental health.
Companies benefit from:

- Sentiment Analysis
- Hate Speech Detection
- Content Moderation

Preprocessing Procedures

Train / Test Split

Split data into train & test for modeling.

Re-sampling

To counter for class imbalance.

Count Vectorization

Transform text into numbers.

	count	hate_speech_count	offensive_language_count	neither_count	class	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...

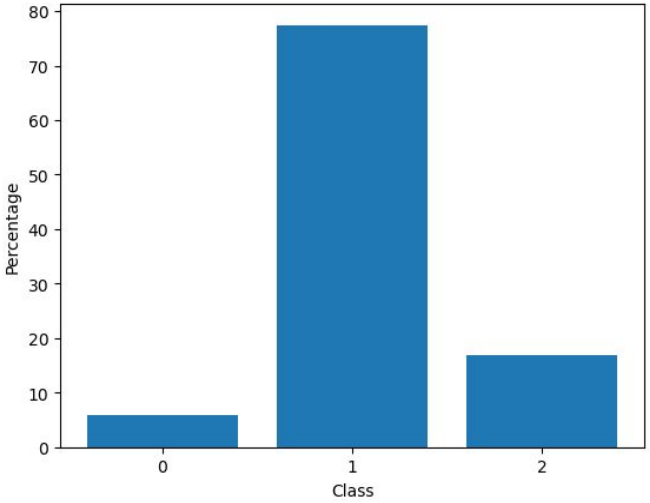
Target: class
Feature: tweet

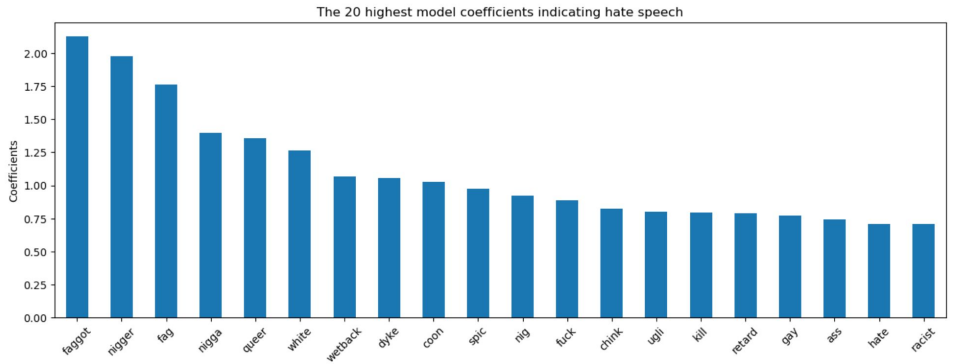
Data source: Huggingface hate_speech_offensive

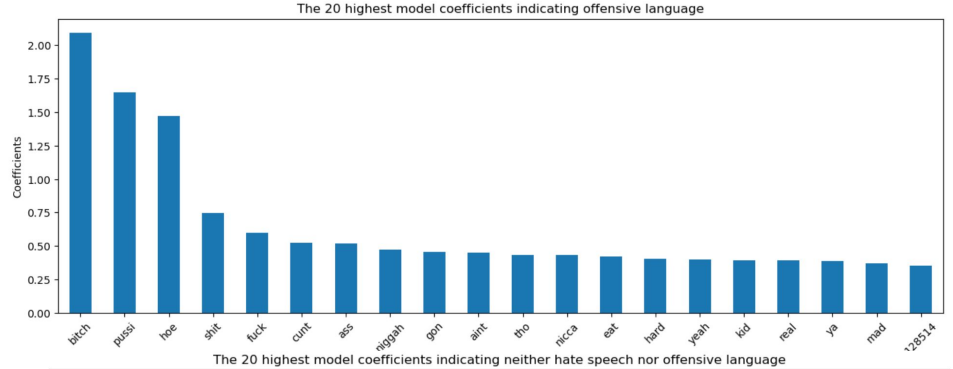
EDA

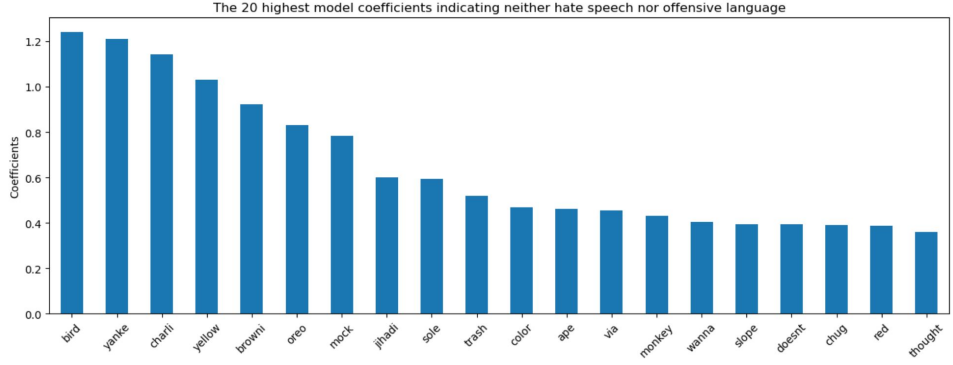
Class Imbalance

- class 0 (6%) - hate speech
- class 1 (77%) - offensive language
- class 2 (17%) - neither







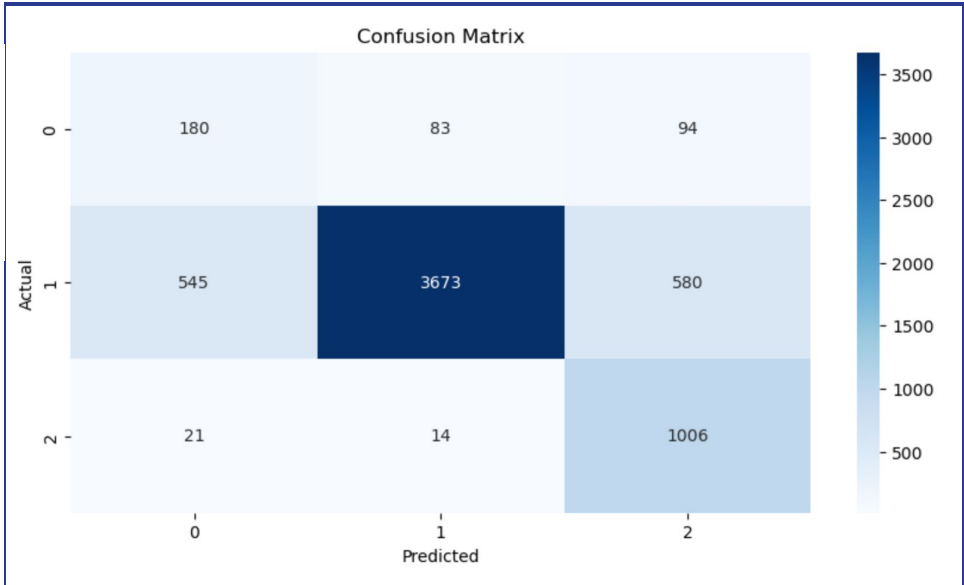


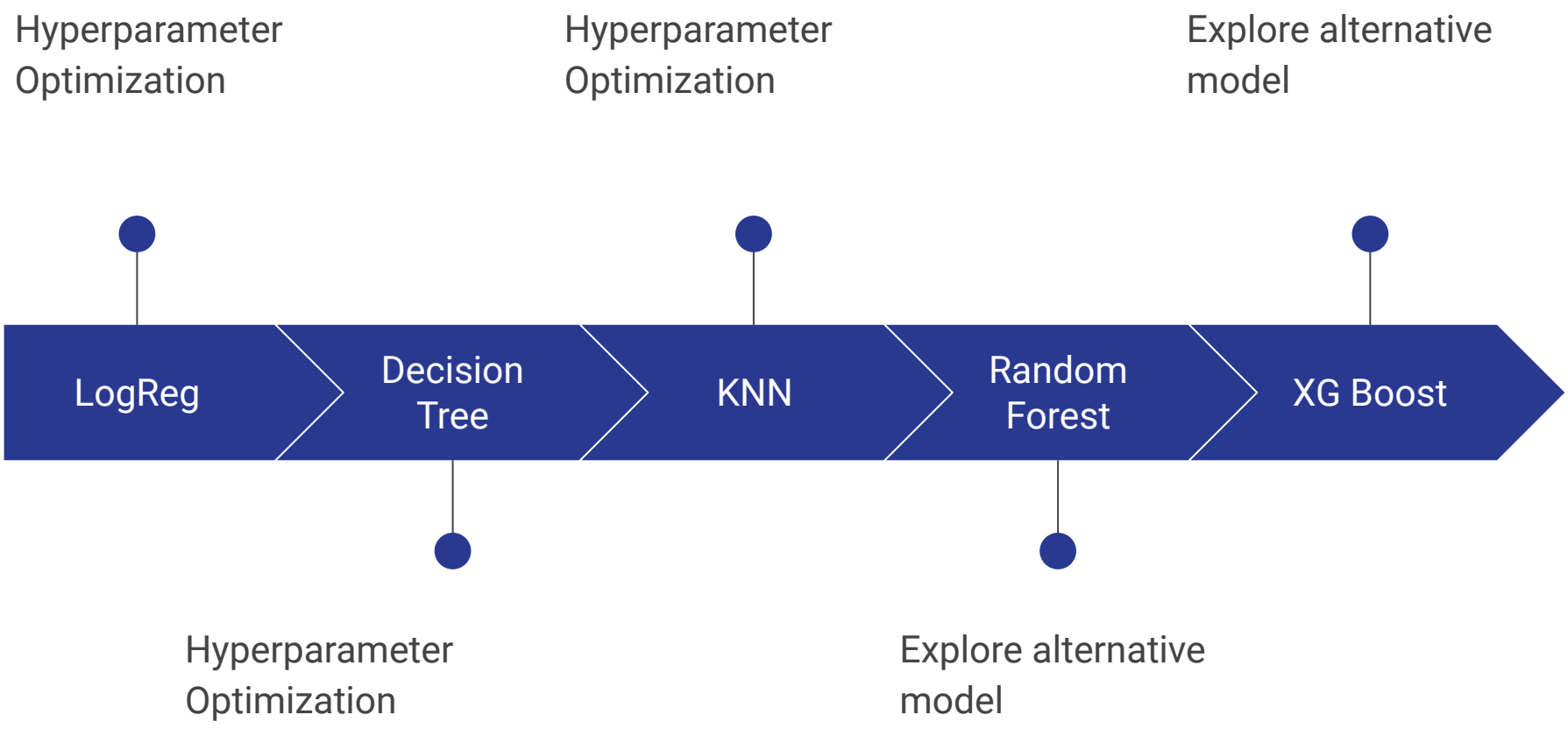
Baseline models

Model	Parameters	Train Accuracy (%)	Test Accuracy (%)	Notes
Logistic Regression	C=0.1	84.4	81.3	No overfitting
Decision Tree	max_depth=10	80	78.4	No overfitting
KNN	n/a	89.6	80	Slight overfitting

Evaluation Metrics

	precision	recall	f1-score	support
0	0.24	0.50	0.33	357
1	0.97	0.77	0.86	4798
2	0.60	0.97	0.74	1041
accuracy			0.78	6196
macro avg	0.60	0.75	0.64	6196
weighted avg	0.87	0.78	0.81	6196




Hyperparameter Optimization
Hyperparameter Optimization
Explore alternative model
LogReg
Decision Tree
KNN
Random Forest
XG Boost
Hyperparameter Optimization
Explore alternative model

Questions