Content Moderation on Social Media Platforms

Sprucing Up Social Spaces with NLP Magic

Eileen Cai

Data Scientist

The problem

Problem

With the widespread of Social Media influence comes the risk of exposure to harmful content that can negatively impact teen's mental health.

Solution

Leveraging the power of machine learning to:

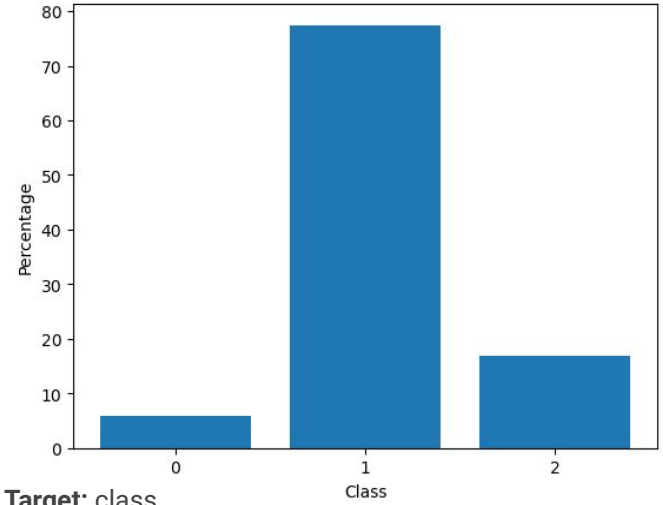
- Automatically detect and flag texts containing cyberbullying, hate speech & other harmful material.

Impact

Make meaningful impact by promoting online safety and well-being for the younger generation.

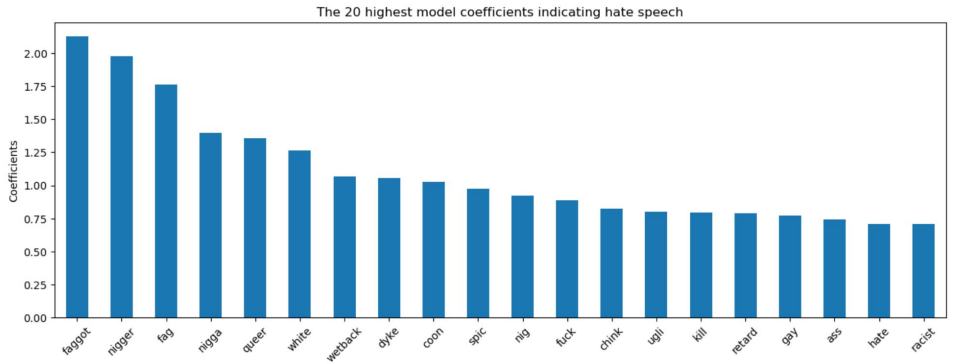
EDA

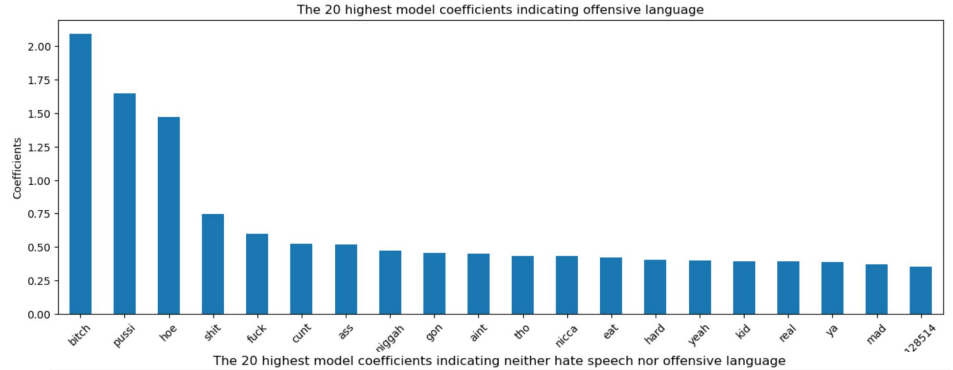
- 24783 rows & 6 columns
- 0 null values, 0 duplicated rows
- class 0 (6%) - hate speech
- class 1 (77%) - offensive language
- class 2 (17%) - neither

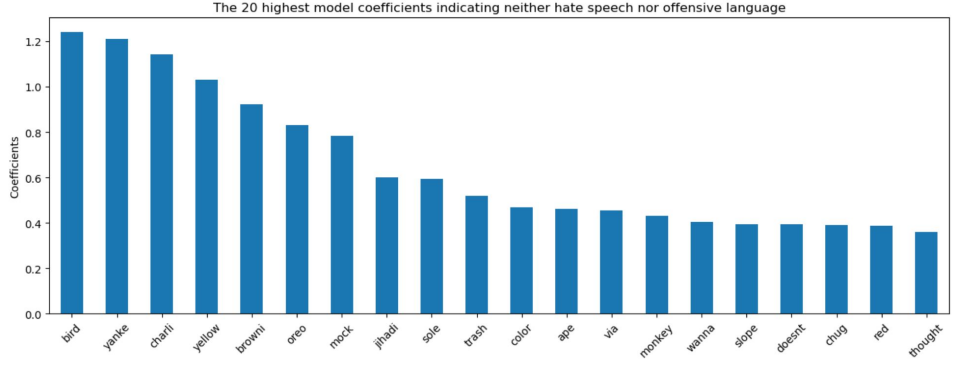


Target: class
Feature: tweet

Data source: Huggingface hate_speech_offensive







Preprocessing Procedures

Train / Test Split

Split data into train (80%) & test (20%) for modeling.

- The TRAIN set has 19826 data points.
- The TEST set has 4957 data points.

Re-sampling

Upsample class 0 and downsample class 1.

- Class 0 before: 1144
- Class 0 after: 3330
- Class 1 before: 15352
- Class 1 after: 3330
- The TRAIN set has 9990 data points.
- The TEST set has 4957 data points.

Vectorization

Transform text into numbers.

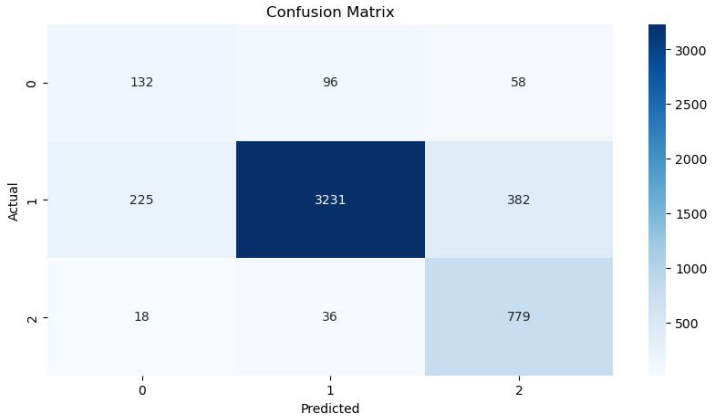
- Custom tokenizer
- Count Vectorization using Bag-of-Words
- Word Embeddings using Word2Vec

Model Selection & Evaluations

Here is our model summary:

Model	Vectorizer	Parameters	Train Accuracy (%)	Test Accuracy (%)	F1-Score (weighted avg)	Recall (weighted avg)
Logistic Regression	CountVectorizer	C=1	83.2	79.3	82	79
Random Forest	Sentence2vecTransformer	max_depth=20, n_estimators=900	99.8	82.1	83	82
XGBoost	Sentence2vecTransformer	max_depth=4, n_estimators=481	99.8	83.6	85	84

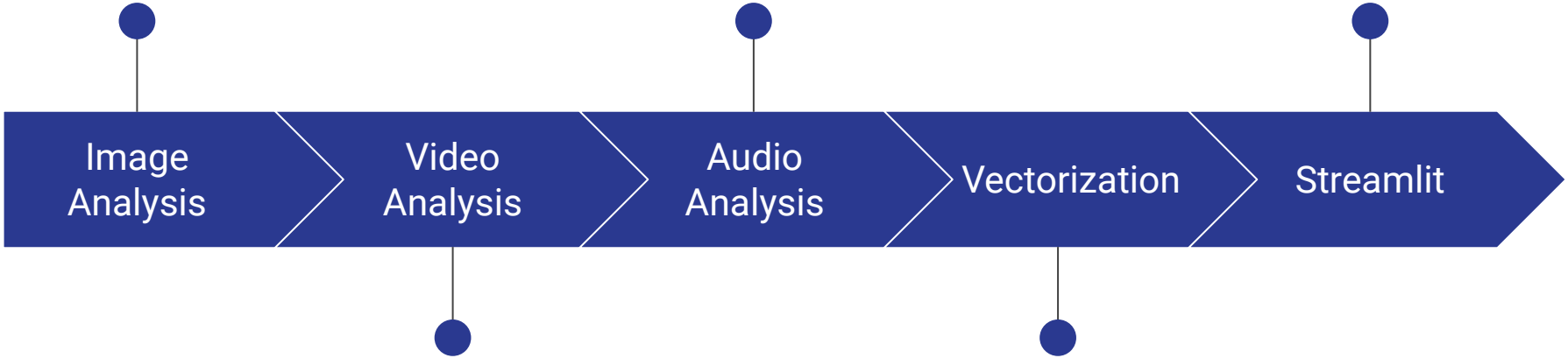
	precision	recall	f1-score	support
0	0.35	0.46	0.40	286
1	0.96	0.84	0.90	3838
2	0.64	0.94	0.76	833
accuracy			0.84	4957
macro avg	0.65	0.75	0.69	4957
weighted avg	0.87	0.84	0.85	4957



Detect harmful images/memes

Detect harmful audios

Create app to deploy ML model



Detect harmful videos

Explore alternative vectorization techniques

Questions