DATASCI 207 - Final Presentation
Content Moderation Classifier for LLMs

Kevin Coppa, Rick Pereira, and Ryan Schaefer

Research Questions

- **Question:** How effectively can multi-label classifiers, trained on a dataset of prompts sent to LLM, accurately identify and categorize multiple unsafe content labels?

Importance & Interest

- **Scale of the Problem:**
    1. LLMs are getting adopted widely and rapidly and are already being pushed to generate unsafe content.
    2. Malicious actors continuously develop new techniques to bypass safety filters, creating a constant arms race. Effective multi-label classifiers are crucial for staying ahead of these evolving threats.
- **Ethics and Societal Implications**
    1. Preventing Harm: as generative models become popular, we should minimize its harmful effects by any means necessary.
    2. Maintaining Trust: effective content moderation is necessary for building trustworthy machine learning systems.
    3. Legal and Regulatory Compliance: Governments and regulators are increasingly becoming concerned about LLMs and their potential for misuse.

Data Source

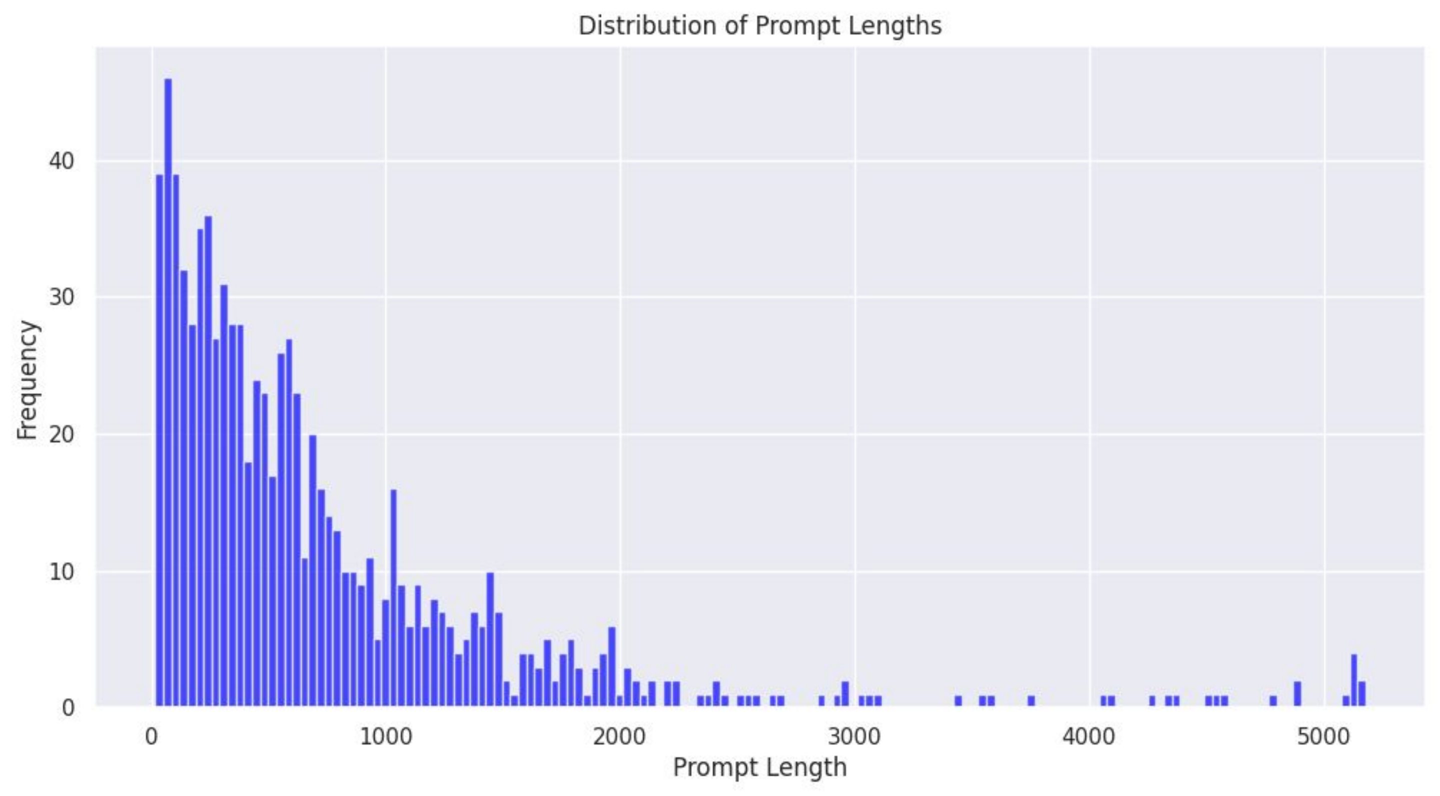
Data Source: OpenAI content moderation dataset provided form their research paper "A Holistic Approach to Undesired Content Detection."

Dataset Size: 1,680 text prompts

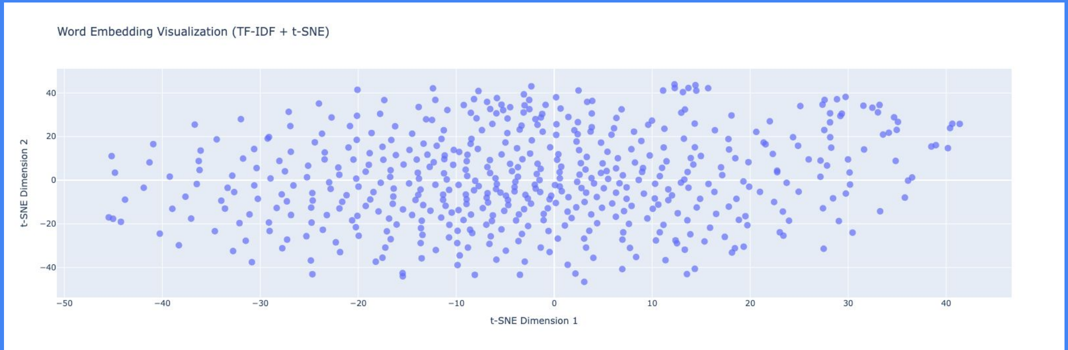
Labels: Binary content moderation flags for 8 categories of unsafe content. The category labels are defined according to the following taxonomy:

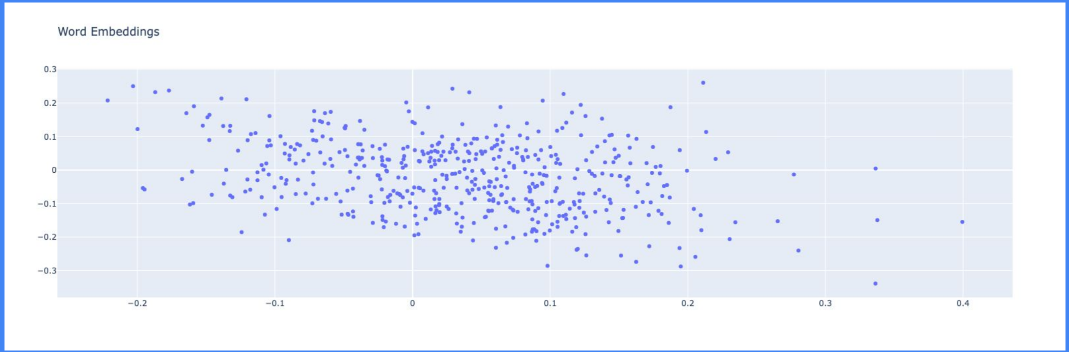
- **sexual (S)**: Content meant to arouse sexual excitement.
- **hate (H)**: Content that expresses, incites, or promotes hate.
- **violence (V)**: Content that promotes or glorifies violence.
- **harassment (HR):** Content that may be used to torment or annoy individuals.
- **self-harm (SH)**: Content that promotes, encourages, or depicts acts of self-harm.
- **sexual/minors (S3)**: Sexual content that includes an individual who is under 18 years old.
- **hate/threatening (H2)**: Hateful content that also includes violence or serious harm.
- **violence/graphic (V2)**: Violent content that depicts death, violence, or serious physical injury in extreme graphic detail.

Data Summary - Text



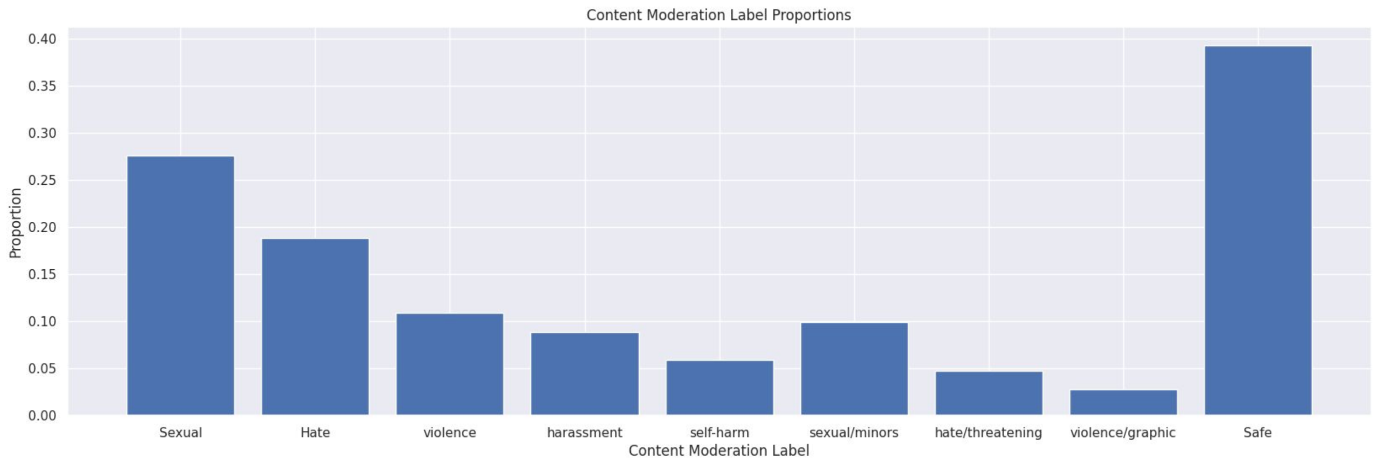
Embeddings


Word Embedding Visualization (TF-IDF + t-SNE)
t-SNE Dimension 2
40
20
0
−20
−40
−50
−40
−30
−20
−10
0
10
20
30
40
t-SNE Dimension 1


Word Embeddings
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4

Models

Baseline



Baseline accuracy, training: 0.416
Baseline accuracy, validation: 0.378
Baseline accuracy, testing: 0.337

Multi-Label Logistic Regression

Overview of multi-label LR:

- Predicts multiple labels
- Applies logistic function to predict class probabilities for each label

Key parameters:

- Learning rate, regularization

Model Architecture:

- Input layer
- Each label predicted independently using sigmoid activation

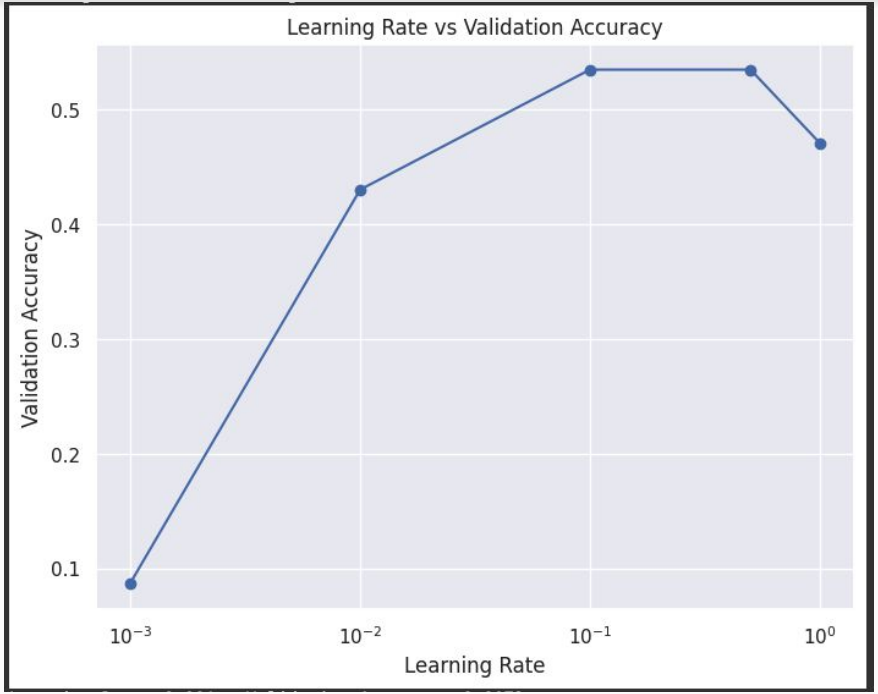
Loss:

Binary_crossentropy

Multi-Label Logistic Regression

Hyper-parameter tuning:

- For such a simple model, we tuned the learning rate only



Feed Forward Neural Network

Overview of feed forward NN:

- Predicts multiple labels
- Contains one or more hidden layers with non-linear activations

Key parameters:

- Layers, dropout, units

Model Architecture:

- Input layer
- Hidden layer with ReLU activation
- Dropout layer
- Each label predicted independently using sigmoid activation

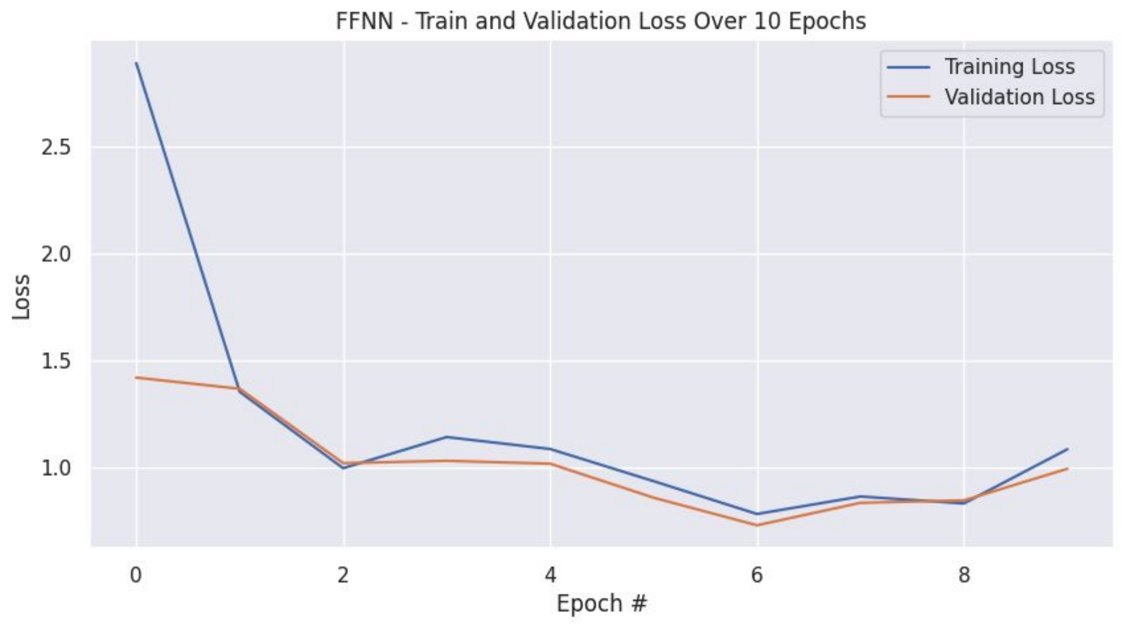
Loss:

Binary_crossentropy

Feed Forward Neural Network

Learning rate tuning:

learning_rate	val_loss
0.000214	0.587
0.000200	0.430
0.003150	0.262



1D CNN

Learned Keras Embeddings

conv_rounds	num_filters_0	hidden_layers	neurons_0	val_loss
2	40	2	384.0	0.275349
1	64	1	256.0	0.277306
2	40	2	384.0	0.278203
2	64	2	384.0	0.278498
2	64	2	384.0	0.282622
2	24	2	64.0	0.283209
2	40	0	320.0	0.283320
1	32	0	448.0	0.283749
2	48	1	384.0	0.284442
2	16	0	384.0	0.285944

TF-IDF Embeddings

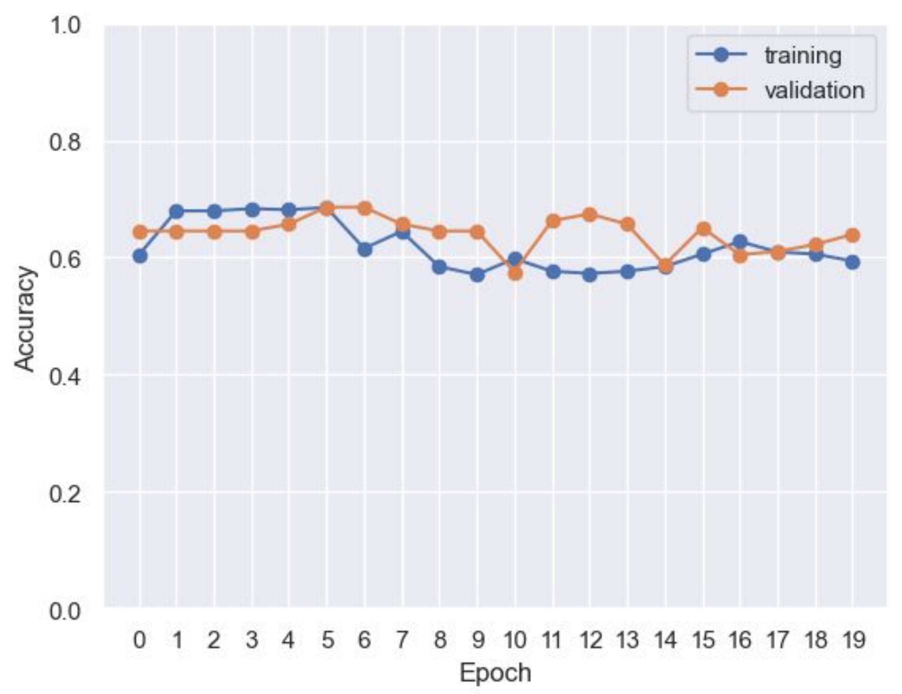
conv_rounds	num_filters_0	hidden_layers	neurons_0	val_loss
1	40	2	320.0	0.246201
1	56	3	384.0	0.249464
1	16	0	256.0	0.254305
1	24	2	384.0	0.256814
1	32	2	512.0	0.260282
2	56	1	256.0	0.271895
1	16	0	256.0	0.275356
1	24	2	384.0	0.276608
1	56	3	384.0	0.284873
2	56	1	448.0	0.294575

1D CNN

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 997, 40)	200
max_pooling1d_1 (MaxPooling1D)	(None, 332, 40)	0
dropout_1 (Dropout)	(None, 332, 40)	0
flatten_1 (Flatten)	(None, 13280)	0
dense_3 (Dense)	(None, 320)	4249920
dense_4 (Dense)	(None, 512)	164352
dense_5 (Dense)	(None, 8)	4104

```
Total params: 4418576 (16.86 MB)
Trainable params: 4418576 (16.86 MB)
Non-trainable params: 0 (0.00 Byte)
```

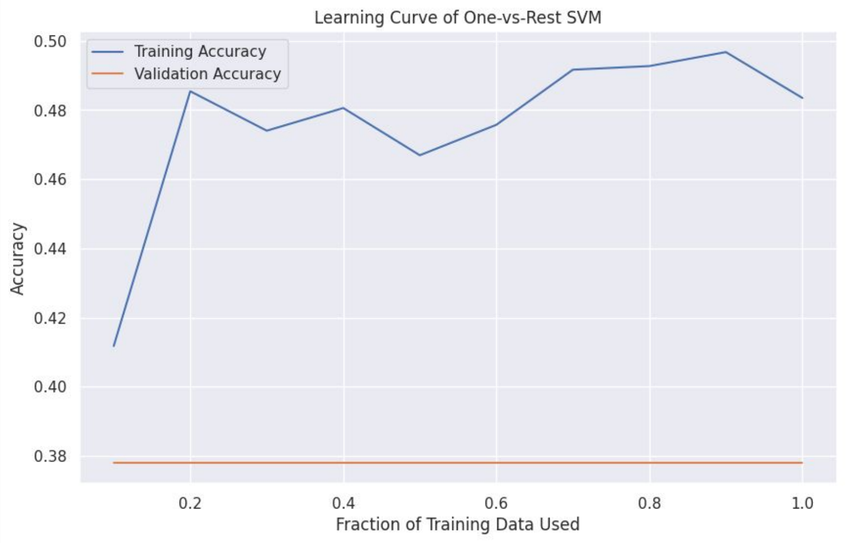
1D CNN



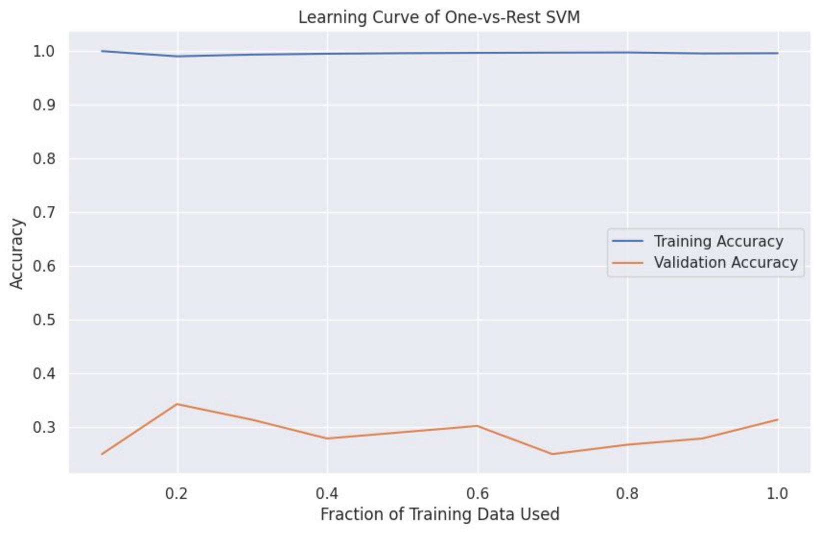
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Epoch
training
validation

Multi-Label Support Vector Machines

IF-IDF



Learned Embeddings



Multi-Label Support Vector Machines

C	Kernel	Gamma	Training Accuracy	Validation Accuracy
0.1	linear	scale	0.9320	0.3605
0.1	linear	auto	0.9320	0.3605
0.1	rbf	scale	0.4155	0.3779
0.1	rbf	auto	0.4155	0.3779
0.1	sigmoid	scale	0.4155	0.3779
0.1	sigmoid	auto	0.4155	0.3779
1	linear	scale	0.9961	0.2965
1	linear	auto	0.9961	0.2965
1	rbf	scale	0.4194	0.3837
1	rbf	auto	0.4194	0.3837
1	sigmoid	scale	0.4194	0.3895
1	sigmoid	auto	0.4194	0.3895

Multi-Label Support Vector Machines

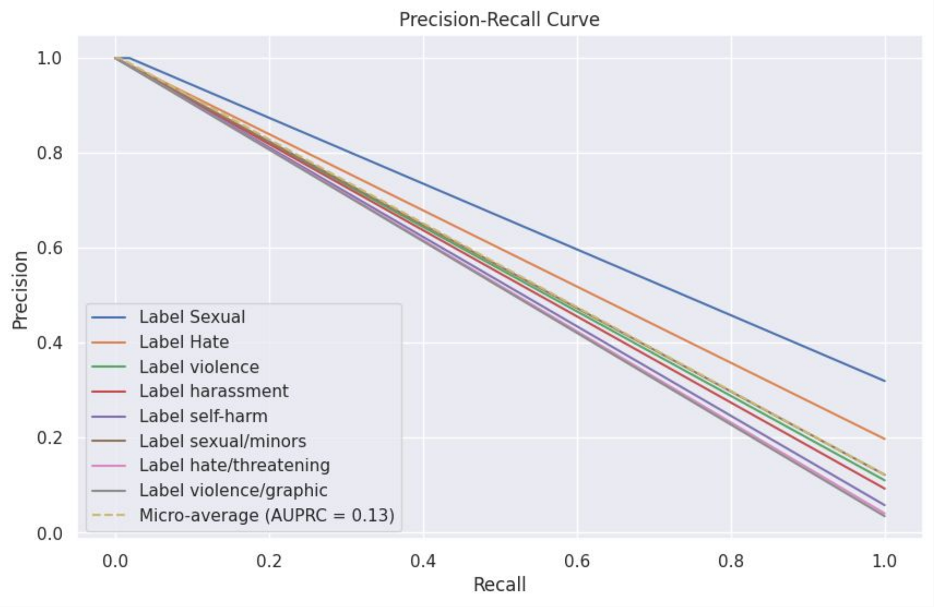
Micro-averaged Precision: 1.0000

Micro-averaged Recall: 0.0060

Micro-averaged F1-Score: 0.0118

Micro-averaged AUPRC: 0.1273

Hamming Loss: 0.1214



Test Data Results

Model Type	Accuracy	Recall	AUPRC
Baseline	0.337	0.000	0.122
Logistic Regression	0.314	0.145	0.146
FF Neural Network	0.430	0.444	0.304
Convolutional Neural Network	0.453	0.510	0.359
Support Vector Machine	0.385	0.006	0.127

Fairness

- Our models are likely prone to bias!

Future Work

- LSTM & Transformer Model
- Sophisticated Learned Embeddings (e.g BERT)
- More Data!!

Thank You!

GitHub

GitHub Codebase:
https://github.com/rickypereira/Content-Moderation-Classifier-for-LLMs

References

- Dataset:
  - https://huggingface.co/datasets/mmathys/openai-moderation-api-evaluation
- Code in Colab:
  - https://colab.research.google.com/drive/18XJEazwQVdBYHtFa0vf0KmcrTYzENNsl
- Evaluation paper for dataset:
  - Markov, Todor, et al. "A holistic approach to undesired content detection in the real world." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 12. 2023.
  - https://arxiv.org/abs/2208.03274