Foundational Models and Generative AI

Team SafeLens
Nakkina Vinay (B21AI023)
Renu Sankhla (B21AI028)
Nitish Bhardwaj(B21AI056)

Project Final Evaluation

Abstract

Our project focuses on fine-tuning foundational models across different modalities to evaluate the safety of online content. As part of our downstream task, the system will identify and flag unsafe material on social media platforms, including content that may violate community guidelines, pose harm, or create trauma to users.

1. Introduction

We are working on three modalities: text, image, and audio. Till now, we have done data collection for all modalities. We have done a detailed analysis of our dataset collected for text and image. Also trained our model for image modality.

Dataset: SafeLens Dataset

Github Repository:
https://github.com/Niticodersh/SafeLens

2. Data Collection Technique

We collected 900 samples in total.

300 samples of each modality with a balanced number of safe and unsafe samples.

For each modality, we collected content from:

a. Social media like LinkedIn, Instagram, and Youtube
b. Did ground-level data collection
c. From our created SafeLens website https://niticodersh.github.io/SafeLens/
d. Some content from pre-existing available datasets as well. Available datasets ranged from basic sentiment analysis to adult content datasets. We picked those samples that aligned with the motive of dataset creation.

3. Datasets

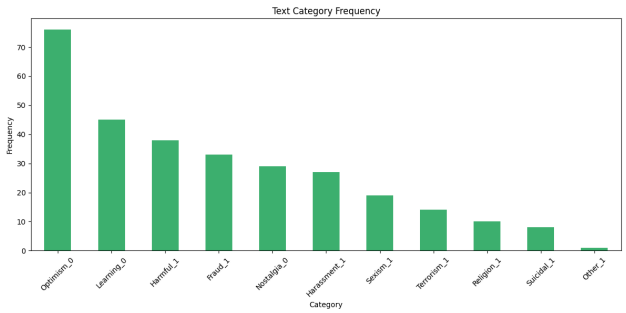
3.1. Text

We have collected text samples from different categories. We hereby classify a broader range in which samples were collected.

Safe texts are categorized under three categories:

'Optimism' if anything is related to happiness, gratitude, etc., 'Learning' if there are some comments related to Education, skills, etc., and 'Nostalgia' if some content describes past memories, etc.

Unsafe texts are categorized as follows: 'Sexism' for targeting vulnerabilities like body shaming, 'Suicidal' for self-harm indicators, 'Harassment' for targeted threats, 'Religion' for sensitive religious topics, 'Terrorism' for promoting violence or extremist ideas, 'Fraud' for deceitful content, 'Harmful' for potential risks, and 'Other' for miscellaneous unsafe content.



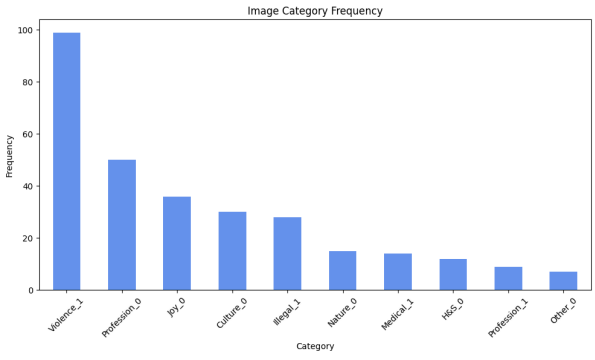
3.2. Image Dataset

We have collected image samples from various categories. We hereby classify a broader range in which samples were collected.

Safe images are categorized under five categories: ‘Culture’ if they reflect cultural practices, traditions, or heritage; ‘H&S’ (Health & Safety) if they promote health awareness, safety measures, and protective practices; ‘Nature’ if they showcase natural scenes, landscapes, or wildlife; ‘Joy’ if they evoke happiness, celebration, or positive emotions; and ‘Profession’ if they represent various occupations or professional activities under safe guidelines.

Unsafe images mainly include ‘Violence’ for those depicting or promoting physical harm or aggression,

‘Medical’ if they relate to health care, medical practices, or healthcare environments, ‘Illegal’ if they depict or suggest illegal activities or behaviors, ‘Profession’ if it includes risky activities as well as ‘Other’ for images that do not fit into the predefined categories and encompass miscellaneous content.



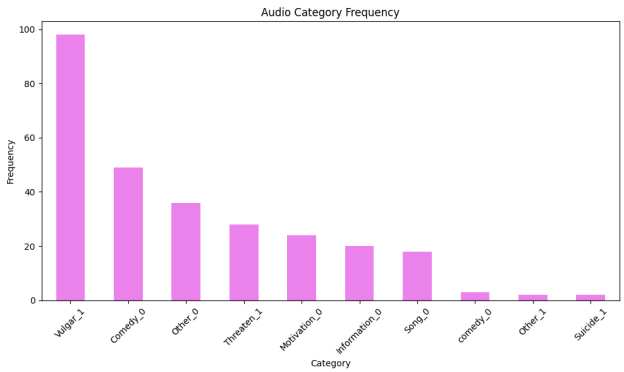
3.3. Audio Dataset

We collected samples by ground-level data collection from voice samples of our batchmates, with their permission. Some audio samples were collected from Instagram and YouTube videos. All samples are of nearly 15 sec.

We have collected these samples from various categories. We hereby classify a broader range in which samples were collected.

Safe audios are categorized under three categories: ‘Motivation’ for audios that share inspiring or encouraging messages, speeches, or quotes; ‘Comedy’ for fun and entertaining audios like jokes or funny dialogues that make people laugh; and ‘Songs’ for music or tunes from different genres that are enjoyable and emotional.

Unsafe audios mainly include ‘Vulgar’ for audios with rude or offensive language; ‘Threaten’ for audios that include scary or harmful messages meant to frighten someone; and ‘Suicide’ for audios that talk about self-harm or feelings of giving up, which can be harmful to listeners.



4. Model Fine-Tuning

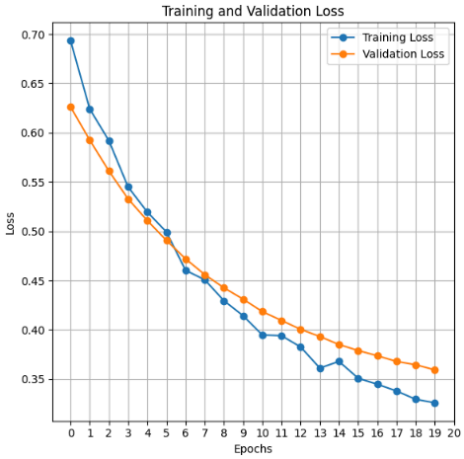
4.1. Images

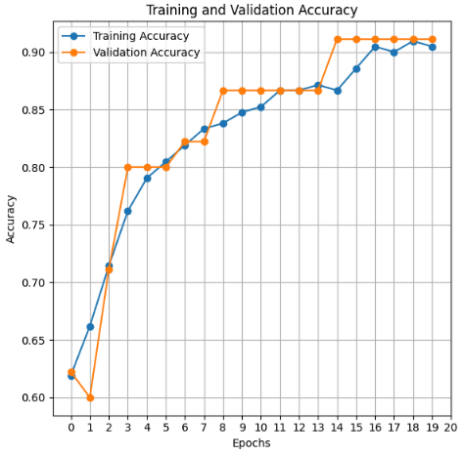
Colab : [Image_Codebase]

Fine-tuned three models for the Image dataset:

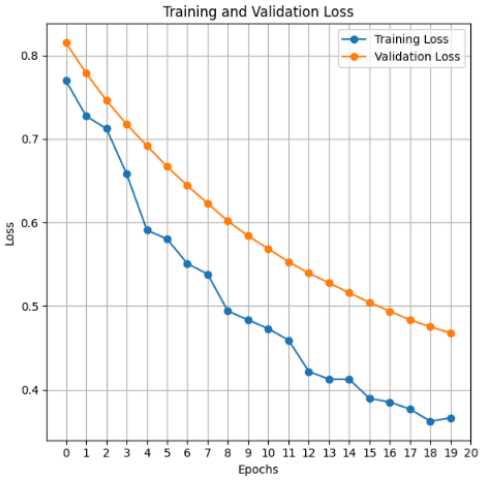
- **Google Vision Transformer** (google/vit-base-patch16-224)
  Preprocessing: Resized images to 224x224 and normalized images to mean of [0.485, 0.456, 0.406] and standard deviation of [0.229, 0.224, 0.225].

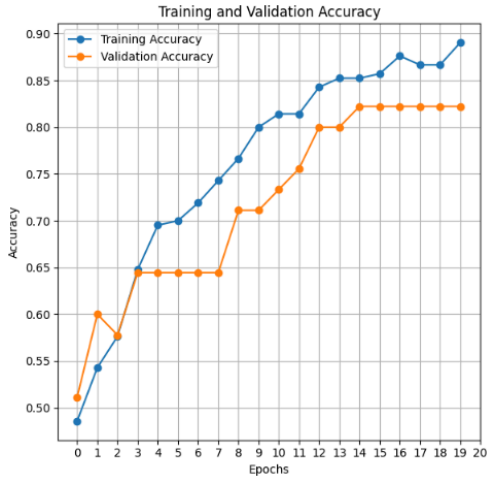
Extracted the 1000-dimensional latent representations (logits) from a pre-trained Vision Transformer (ViT) model and passed them through an MLP for binary classification. The ViT model's weights were frozen to retain its pre-trained features, while only the MLP was fine-tuned for the classification task.



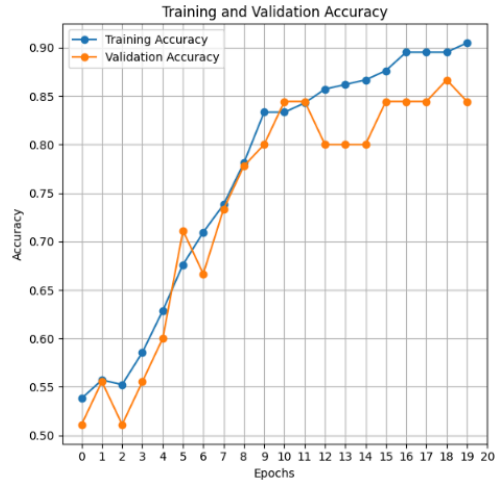


- **Microsoft Swin Transformer** (microsoft/swin-tiny-patch4-window7-224)
  Preprocessing: Resized images to 224x224 and normalized images to mean of [0.485, 0.456, 0.406] and standard deviation of [0.229, 0.224, 0.225].
  The Swin Transformer also produces 1,000-dimensional logits, similar to Google ViT. The processing pipeline and approach remain consistent with that of the ViT model.





- **OpenAI Clip Transformer** (openai/clip-vit-base-patch16)
  Preprocessing: Used pre-trained CLIP processor of the same model.
  In this model, extracted last hidden state features (dimension of 768) followed by global average pooling to reduce sequence dimension, further integrated with MLP for binary classification. Rest pipeline remains consistent as above models.



Batch size = 32, Optimizer = Adam, lr = 1e-4, loss function = Cross Entropy

Overall, all models have fine-tuned successfully over 20 epochs, demonstrating that they were already well-pretrained on large datasets and have now been effectively fine-tuned for the specific task of binary classification between safe and unsafe categories.

4.2. Audio

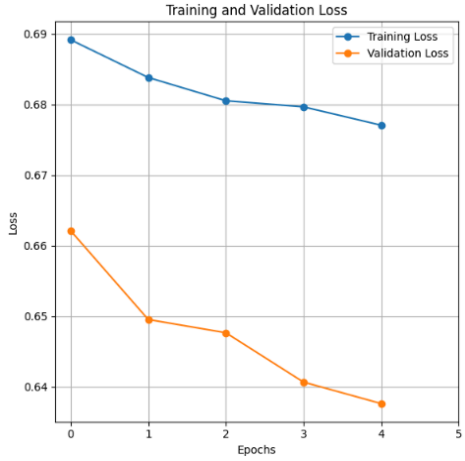
Colab : [Audio_Codebase]

Preprocessing: Audio files were clipped to a maximum length of 15 seconds, and if they were shorter than 15 seconds, they were padded with zeros on the right side. All models accepted audio with a sampling rate of 16 kHz.

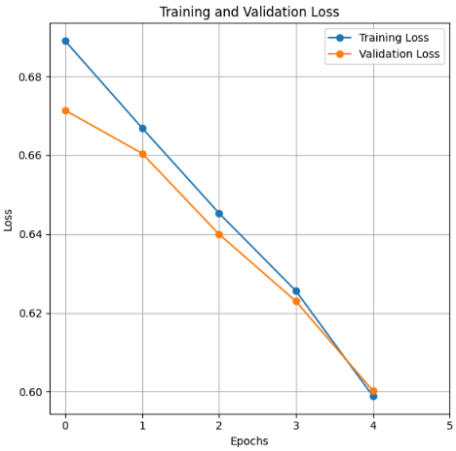
Further pretrained processors of each model were used for final preprocessing as per model requirements.

Fine-tuned three models for the audio dataset:

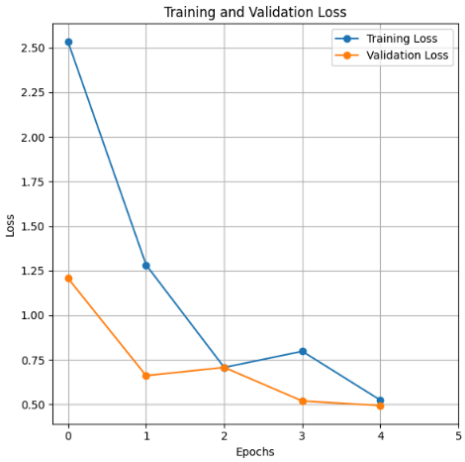
- **Facebook Wave2Vec2**
  (facebook/wav2vec2-base-960h)
  Extracted the 768 dimensional feature vector from the last hidden state of the model, applied average global pooling and passed through the MLP layers for binary classification.



- **OpenAI Whisper**
  (openai/whisper-large-v3-turbo)
  Extracted the 1280 dimensional feature vector from the last hidden layer from the encoder block of the model, applied average pooling and passed through the MLP layers for binary classification.



- **Facebook Hubert**
  (facebook/hubert-large-ls960-ft)
  Extracted the 1024 dimensional feature vector from the second last layer and passed through MLP layers for binary classification.



Batch size = 8, Optimizer = Adam, lr = 1e-4, loss function = Cross Entropy
All three models were fine-tuned for 5 epochs with Whisper model fine-tuned better smoothly. The pre-trained model weights were kept frozen while the MLP layers were fine-tuned.

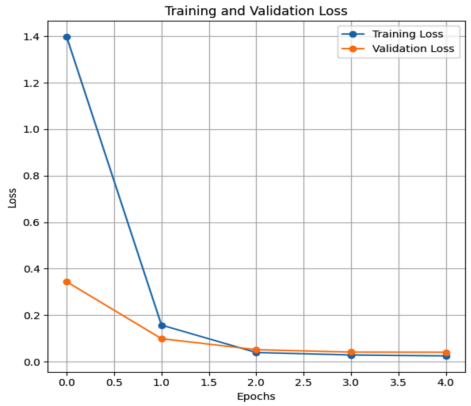
4.3. Text

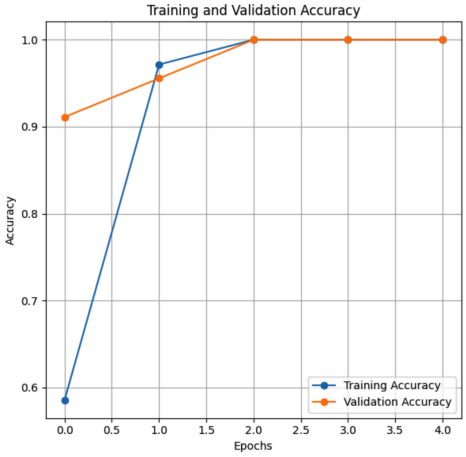
Colab: Text_Codebase

Preprocessing: Truncated the text to max sequence length of 128 and applied padding if smaller than that. Further, used the pretrained tokenizer of each model to tokenize as per the model requirements.

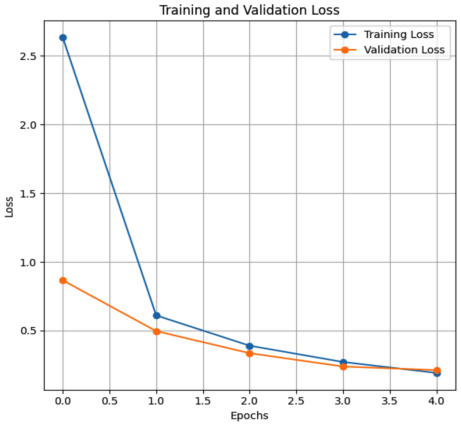
We fine-tuned the pre-trained model weights. Fine-tuned three models for the audio datasets.

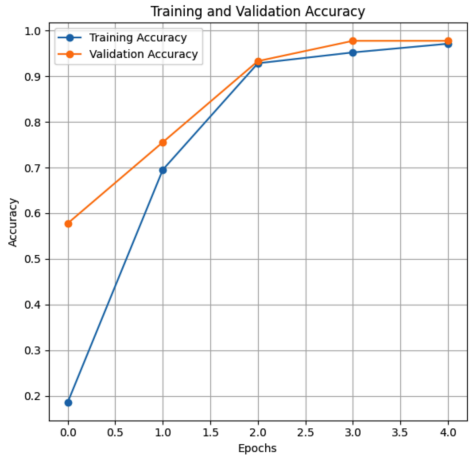
- **Finite Automata BERTweet**
  (finiteautomata/bertweet-base-sentiment-analysis)
  This model is based on the BERTweet architecture, which is a variant of RoBERTa specifically pre-trained on English tweets.



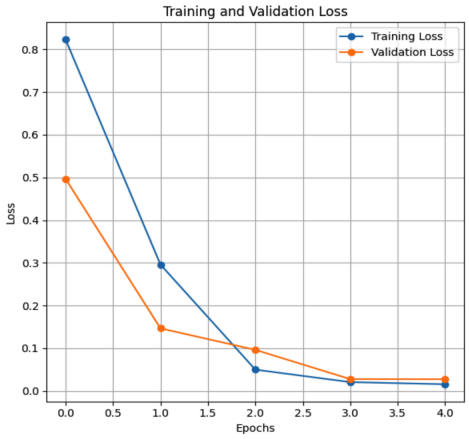


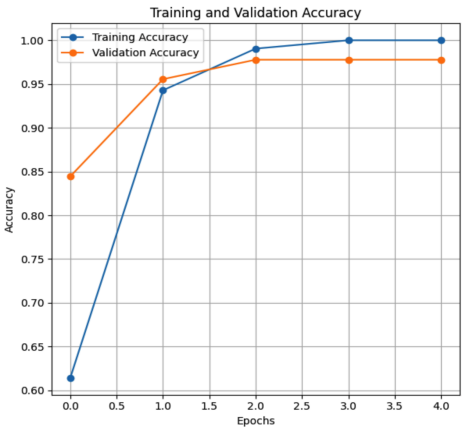
- **Distilbert Base Model**
  (distilbert/distilbert-base-uncased-finetuned-sst-2-english)
  This model is a distilled version of BERT that retains 97% of BERT's language understanding while being faster and smaller. Using this model we fine-tuned on a diverse dataset to enhance its performance in text classification tasks.





- **Cardiffnlp Twitter Roberta**
  (cardiffnlp/twitter-roberta-base-sentiment-latest)
  This model is another sentiment analysis model specifically trained on Twitter data





Batch size = 8, Optimizer = AdamW, lr = 2e-5, loss function = Cross Entropy, Scheduler = Linear

Overall all three models were good at fine-tuning even with just 5 epochs, demonstrating that they were already well-pre trained on large datasets and have now been effectively fine-tuned for the specific task of binary classification between safe and unsafe categories.

5. Results

After fine-tuning all the models of all modalities, we tested them on our test datasets and calculated the loss and test accuracy of the predicted labels by comparing them with the original ones. We have also done our observations category-wise, like calculating the evaluation metrics (Accuracy, Precision, Recall, and F1-Score) for each category present in the data individually for all the models of all modalities.

While benchmarking, we will focus on Recall more as we want that unsafe content should not be classified as safe, although we can tolerate misclassification of safe as unsafe.

5.1. Images

- **Google Vision Transformer Model Results**

Test Loss: 0.3498, Test Accuracy: 0.8444

	Category	Accuracy	Precision	Recall	F1 Score
0	Joy_0	0.833333	1.0	0.833333	0.909091
1	Violence_1	0.866667	1.0	0.866667	0.928571
2	H&S_0	1.000000	1.0	1.000000	1.000000
3	Culture_0	0.750000	1.0	0.750000	0.857143
4	Profession_1	0.000000	1.0	0.000000	0.000000
5	Profession_0	1.000000	1.0	1.000000	1.000000
6	Medical_1	1.000000	1.0	1.000000	1.000000
7	Illegal_1	1.000000	1.0	1.000000	1.000000
8	Other_0	0.000000	1.0	0.000000	0.000000

Overall Classification Report:

	precision	recall	f1-score	support
0	0.83	0.87	0.85	23
1	0.86	0.82	0.84	22
accuracy			0.84	45
macro avg	0.85	0.84	0.84	45
weighted avg	0.84	0.84	0.84	45

The model performs well in categories such as H&S_0, Medical_1, and Violence_0 but struggles with rare or nuanced categories like Profession_1, where it shows zero recall.

- **Microsoft Swin Transformer Model Results**

Test Loss: 0.4830, Test Accuracy: 0.7333

	Category	Accuracy	Precision	Recall	F1 Score
0	Joy_0	1.0	1.0	1.0	1.000000
1	Violence_1	1.0	1.0	1.0	1.000000
2	H&S_0	0.0	1.0	0.0	0.000000
3	Culture_0	1.0	1.0	1.0	1.000000
4	Profession_1	0.5	1.0	0.5	0.666667
5	Profession_0	0.2	1.0	0.2	0.333333
6	Medical_1	1.0	1.0	1.0	1.000000
7	Illegal_1	1.0	1.0	1.0	1.000000
8	Other_0	0.0	1.0	0.0	0.000000

Overall Classification Report:

	precision	recall	f1-score	support
0	0.92	0.52	0.67	23
1	0.66	0.95	0.78	22
accuracy			0.73	45
macro avg	0.79	0.74	0.72	45
weighted avg	0.79	0.73	0.72	45

H&S_0, Other_0 have 0% recall, indicating these categories are highly prone to misclassifying unsafe as safe.
Profession_1 and Profession_0 also show significant issues with recall (50% and 20% respectively), so these categories are quite vulnerable.

- **OpenAI Clip Transformer Model Results**

Test Loss: 0.5469, Test Accuracy: 0.8667

	Category	Accuracy	Precision	Recall	F1 Score
0	Joy_0	1.000000	1.0	1.000000	1.000000
1	Violence_1	0.666667	1.0	0.666667	0.800000
2	H&S_0	1.000000	1.0	1.000000	1.000000
3	Culture_0	1.000000	1.0	1.000000	1.000000
4	Profession_1	1.000000	1.0	1.000000	1.000000
5	Profession_0	1.000000	1.0	1.000000	1.000000
6	Medical_1	1.000000	1.0	1.000000	1.000000
7	Illegal_1	0.500000	1.0	0.500000	0.666667
8	Other_0	1.000000	1.0	1.000000	1.000000

Overall Classification Report:

	precision	recall	f1-score	support
0	0.79	1.00	0.88	23
1	1.00	0.73	0.84	22
accuracy			0.87	45
macro avg	0.90	0.86	0.86	45
weighted avg	0.89	0.87	0.86	45

Violence_1 and Illegal_1 categories have slightly lower recall, indicating they are more prone to classifying unsafe content as safe (especially Illegal_1, with only 50% recall).
Most categories have perfect recall (H&S_0, Culture_0, Profession_1, Profession_0, Medical_1, Other_0, and Violence_0), meaning unsafe instances in these categories are identified correctly.
Violence_1 is the primary category where the model is at risk of missing unsafe instances.

Overall Analysis:
By comparing the results of all three models, we can see that the test accuracy of the **OpenAI Clip Transformer** model is high compared to other models.
The models that are most prone to misclassifying unsafe content as safe are Swin Transformer
OpenAI CLIP Transformer showed the least proneness to misclassifying unsafe as safe, though it still has some issues with Violence_1 and Illegal_1.

5.2. Audio

- **Facebook Wave2Vec2 Model Results**

```
Test Loss: 0.6946, Test Accuracy: 0.5238
```

Category-wise Metrics:

	Category	Accuracy	Precision	Recall	F1 Score
0	Information_0	0.625000	1.0	0.625000	0.769231
1	Threaten_1	0.000000	1.0	0.000000	0.000000
2	Vulgar_1	0.166667	1.0	0.166667	0.285714
3	Other_0	1.000000	1.0	1.000000	1.000000
4	Other_1	0.000000	1.0	0.000000	0.000000
5	Song_0	1.000000	1.0	1.000000	1.000000
6	Comedy_0	1.000000	1.0	1.000000	1.000000
7	Motivation_0	1.000000	1.0	1.000000	1.000000

Overall Metrics:
Accuracy: 0.5238
Precision: 0.4770
Recall: 0.5238
F1 Score: 0.4405

Overall Classification Report:

	precision	recall	f1-score	support
0	0.54	0.87	0.67	23
1	0.40	0.11	0.17	19
accuracy			0.52	42
macro avg	0.47	0.49	0.42	42
weighted avg	0.48	0.52	0.44	42

Categories like Other_0, Song_0, Comedy_0, and Motivation_0 show perfect accuracy, precision, recall, and F1 scores (all 1.0). These categories seem to be easier to classify accurately, representing more distinct or clear audio content.
The Threaten_1 and Other_1 categories show poor performance with a recall of 0.0. The model fails to correctly identify any instances of these categories, which may suggest they are either underrepresented or harder to distinguish based on the audio features used by Wave2Vec2.

- **OpenAI Whisper Model Results**

```
Test Loss: 0.6033, Test Accuracy: 0.7857
```

Category-wise Metrics:

	Category	Accuracy	Precision	Recall	F1 Score
0	Information_0	1.000000	1.0	1.000000	1.000000
1	Threaten_1	0.666667	1.0	0.666667	0.800000
2	Vulgar_1	0.583333	1.0	0.583333	0.736842
3	Other_0	1.000000	1.0	1.000000	1.000000
4	Other_1	0.000000	1.0	0.000000	0.000000
5	Song_0	0.666667	1.0	0.666667	0.800000
6	Comedy_0	1.000000	1.0	1.000000	1.000000
7	Motivation_0	1.000000	1.0	1.000000	1.000000

Overall Metrics:
Accuracy: 0.7857
Precision: 0.8163
Recall: 0.7857
F1 Score: 0.7757

Overall Classification Report:

	precision	recall	f1-score	support
0	0.73	0.96	0.83	23
1	0.92	0.58	0.71	19
accuracy			0.79	42
macro avg	0.82	0.77	0.77	42
weighted avg	0.82	0.79	0.78	42

The Whisper model excels in the Information_0, Other_0, Comedy_0, and Motivation_0 categories with perfect classification scores.
The Other_1 category again performs poorly with a recall of 0.0, similar to Wave2Vec2. Threaten_1 and Vulgar_1 are better than Wave2Vec2 but still have room for improvement.

- **Facebook Hubert Model Results**

```
Test Loss: 0.5158, Test Accuracy: 0.7143
```

```
Category-wise Metrics:
        Category  Accuracy  Precision    Recall  F1 Score
0  Information_0  0.875000        1.0  0.875000  0.933333
1     Threaten_1  0.500000        1.0  0.500000  0.666667
2       Vulgar_1  0.666667        1.0  0.666667  0.800000
3        Other_0  0.833333        1.0  0.833333  0.909091
4        Other_1  0.000000        1.0  0.000000  0.000000
5         Song_0  0.666667        1.0  0.666667  0.800000
6       Comedy_0  0.666667        1.0  0.666667  0.800000
7   Motivation_0  1.000000        1.0  1.000000  1.000000

Overall Metrics:
Accuracy: 0.7143
Precision: 0.7171
Recall: 0.7143
F1 Score: 0.7089

Overall Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.83      0.76        23
           1       0.73      0.58      0.65        19

    accuracy                           0.71        42
   macro avg       0.72      0.70      0.70        42
weighted avg       0.72      0.71      0.71        42
```

Categories like Information_0, Other_0, Comedy_0, and Motivation_0 show decent to high accuracy, similar to Whisper.
Other_1 shows no predictions with a recall of 0.0, similar to the other models, highlighting that some categories may be inherently harder for all models.
Unlike the other two models, Hubert's scores are more evenly distributed across all categories, reflecting a somewhat more consistent performance.

Overall Analysis:
By comparing the results of all three models, we can see that the test accuracy of the **OpenAI Whisper Model** is high compared to other models.

Whisper Model stands out as the best overall performer. With high accuracy, precision, recall, and F1 score, it is a strong choice for tasks where recall is a critical factor, especially in minimizing the risk of classifying unsafe content as safe.
Wave2Vec2, despite excelling in certain categories, has low recall and F1 scores, especially in harder categories like Threaten_1 and Other_1. It may need further fine-tuning to handle such categories better.
Hubert offers a balanced approach but lags slightly behind Whisper in terms of recall and precision.

5.3. Text

- **Bertweet-base-sentiment-analysis Model Results**

```
Test Set Accuracy: 0.9556

Test Set Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        23
           1       1.00      0.91      0.95        22

    accuracy                           0.96        45
   macro avg       0.96      0.95      0.96        45
weighted avg       0.96      0.96      0.96        45

Test Set Classification Report (Category-wise):
              precision    recall  f1-score   support

     Fraud_1       1.00      1.00      1.00         6
Harassment_1       1.00      0.67      0.80         3
   Harmful_1       1.00      1.00      1.00         6
  Learning_0       1.00      1.00      1.00         7
 Nostalgia_0       1.00      1.00      1.00         7
  Optimism_0       1.00      1.00      1.00         9
  Religion_1       1.00      0.50      0.67         2
    Sexism_1       1.00      1.00      1.00         2
  Suicidal_1       1.00      1.00      1.00         1
 Terrorism_1       1.00      1.00      1.00         2

    accuracy                           0.96        45
   macro avg       0.83      0.76      0.79        45
weighted avg       1.00      0.96      0.97        45
```

The model performs excellently in categories like Fraud_1, Harmful_1, Learning_0, Nostalgia_0, and Optimism_0, etc. with perfect precision, recall, and F1 scores (1.00). These categories seem to represent well-defined or easily distinguishable content.
Categories like Harassment_1, and Religion_1 are problematic, with the model not making all predictions predictions for Harassment_1 and Religion_1 indicating these categories are harder for the model to classify.

- **Distilbert-base-uncased-fine-tuned Model Results**

```
Test Set Accuracy: 0.8889

Test Set Classification Report:
              precision    recall  f1-score   support

           0       0.82      1.00      0.90        23
           1       1.00      0.77      0.87        22

    accuracy                           0.89        45
   macro avg       0.91      0.89      0.89        45
weighted avg       0.91      0.89      0.89        45

Test Set Classification Report (Category-wise):
              precision    recall  f1-score   support

     Fraud_1       1.00      0.83      0.91         6
Harassment_1       1.00      0.67      0.80         3
   Harmful_1       1.00      0.83      0.91         6
  Learning_0       1.00      1.00      1.00         7
 Nostalgia_0       1.00      1.00      1.00         7
  Optimism_0       1.00      1.00      1.00         9
  Religion_1       1.00      0.50      0.67         2
    Sexism_1       1.00      0.50      0.67         2
  Suicidal_1       1.00      1.00      1.00         1
 Terrorism_1       1.00      1.00      1.00         2

    accuracy                           0.89        45
   macro avg       0.67      0.56      0.60        45
weighted avg       1.00      0.89      0.93        45
```

The model performs very well in categories such as Learning_0, Nostalgia_0, Optimism_0, and Suicidal_1 with perfect scores in terms of precision, recall, and F1.
While it faces problem in categories unsafe categories well like Religion_1, Sexism_1, Fraud_1, Harassment_1 and Harmful_1

- **Twitter-roberta-base-sentiment-latest Model Results**

```
Test Set Accuracy: 0.9778

Test Set Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        23
           1       1.00      0.95      0.98        22

    accuracy                           0.98        45
   macro avg       0.98      0.98      0.98        45
weighted avg       0.98      0.98      0.98        45

Test Set Classification Report (Category-wise):
              precision    recall  f1-score   support

     Fraud_1       1.00      1.00      1.00         6
Harassment_1       1.00      1.00      1.00         3
   Harmful_1       1.00      0.83      0.91         6
  Learning_0       1.00      1.00      1.00         7
 Nostalgia_0       1.00      1.00      1.00         7
  Optimism_0       1.00      1.00      1.00         9
  Religion_1       1.00      1.00      1.00         2
    Sexism_1       1.00      1.00      1.00         2
  Suicidal_1       1.00      1.00      1.00         1
 Terrorism_1       1.00      1.00      1.00         2

    accuracy                           0.98        45
   macro avg       0.91      0.89      0.90        45
weighted avg       1.00      0.98      0.99        45
```

CardiffNLP Twitter RoBERTa performs excellently across all categories, achieving 1.00 precision, recall, and F1 scores for several categories, including Fraud_1, Harassment_1, Learning_0. It only faced problems in Harmful_1 category.

Overall Analysis:

By Comparing the results of all three models, we can see that the test accuracy of the **Twitter-roberta-base-sentiment model** is high compared to other models.

Future Work

Future efforts will focus on expanding the dataset size and implementing robust fine-tuning techniques to address the critical issue of misclassifying unsafe content as safe. This misclassification is a major concern, as the opposite case of misclassifying safe content as unsafe is yet tolerable.

References

[1] An image is worth 16x16 words: Transformers for image recognition at scale, 2021. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl- vain Gelly, Jakob Uszkoreit, and Neil Houlsby.

[2] https://www.kaggle.com/datasets/harsh03/sexually-explicit-comments/data

[3] Suicide and Depression Detection

[4] https://huggingface.co/datasets/valurank/Adult-content-dataset

[5] https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset

[6] https://huggingface.co/datasets/JailbreakBench/JBB-Behaviours