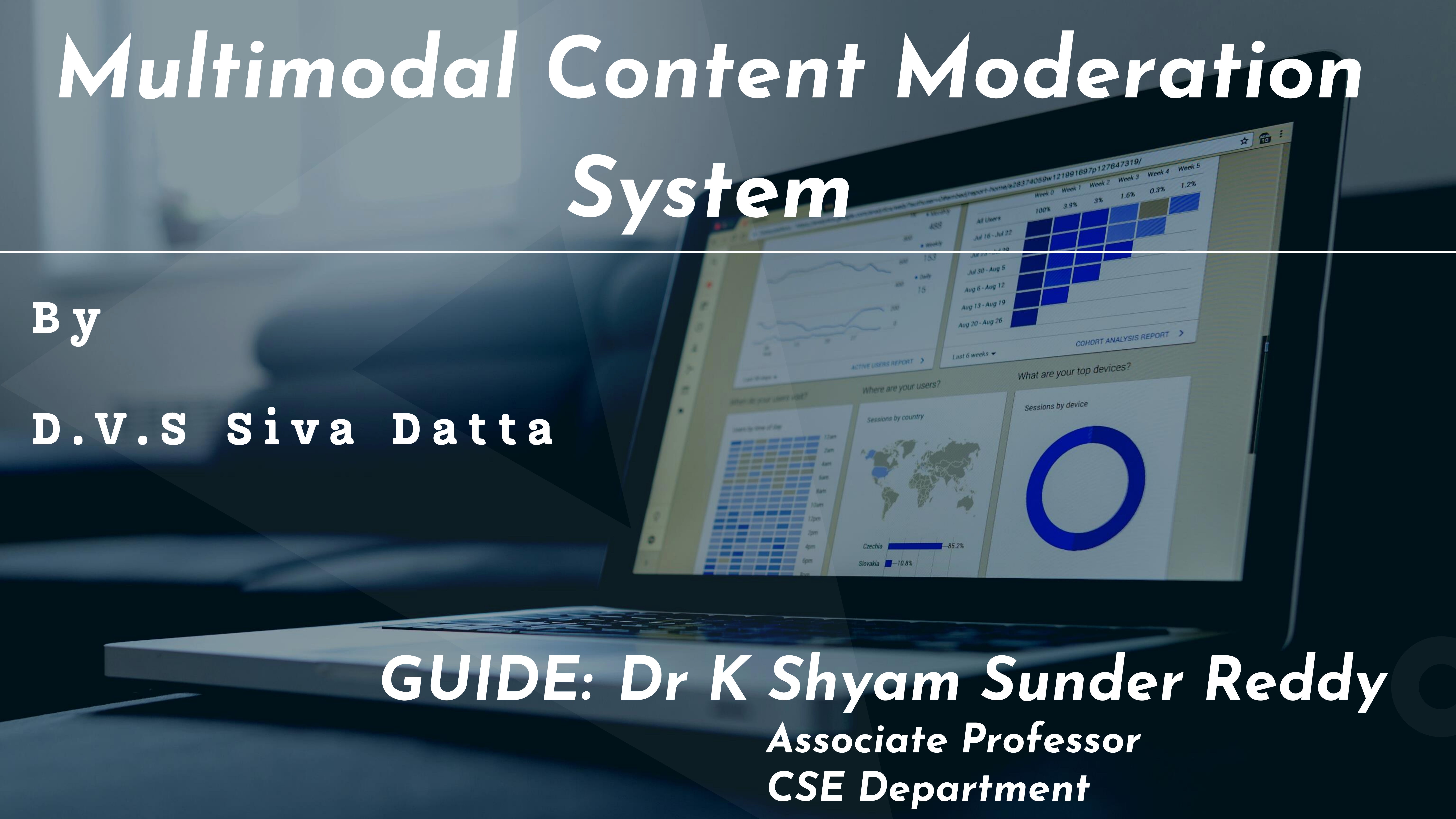
Multimodal Content Moderation System

By

D.V.S Siva Datta

GUIDE: Dr K Shyam Sunder Reddy

Associate Professor
CSE Department



Problem Definition

- In an increasingly digital world, the need for effective content moderation has become paramount to maintain safe and inclusive online spaces.

- Our project presents a comprehensive, real-time content moderation framework that combines natural language processing (NLP) and computer vision techniques to ensure the safety and integrity of text and images shared on online platforms.

- The framework includes a user-friendly dashboard that provides real-time monitoring of content moderation activities like visualizing statistics & trends related to the flagged content.

- This framework addresses various aspects, including
  - ➢ Text Classification
  - ➢ Sentiment Analysis
  - ➢ Profanity Detection
  - ➢ Named Entity Recognition (NER) for abusive words
  - ➢ Masking for Profanity
  - ➢ Image classification for NSFW, violence, and nudity

Literature Survey

Ref no	PAPER TITLE	AUTHOR	DESCRIPTION
1)	**Censored, suspended, shadowbanned**: User interpretations of content moderation on social media platforms	Sarah Mayers West	It explores how users perceive moderation systems, their emotional connection to platforms, and actions taken to address concerns. The study suggests shifting towards an educational model for content moderation, going beyond debates on freedom of expression.
2)	**STATE-OF-THE-ART IN NUDITY CLASSIFICATION:** A COMPARATIVE ANALYSIS	Fatih Cagatay Akyon , Alptekin Temizel	The paper compares existing nudity classification techniques, focusing on CNN-based models, vision transformers, and safety checkers from Stable Diffusion and LAION for content moderation. It highlights limitations in current evaluation datasets, advocating for more diverse datasets. The study underscores the importance of ongoing improvements in image classification models for the safety of online platform users.
3)	**Content Moderation on Social Media in the EU:** Insights From the DSA Transparency Database	Chiara Drolsbach, Nicolas Pröllochs	The Digital Services Act (DSA) mandates large EU social media platforms to disclose content removal details in "Statements of Reasons" (SoRs). Analyzing 156 million SoRs over two months, the study finds variations in moderation frequency, with TikTok leading. It suggests inconsistencies in DSA implementation, highlighting the need for clearer guidelines to ensure common standards for handling rule-breaking content on social media platforms.
4)	**Toxicity Detection is NOT all you Need:** Measuring the Gaps to Supporting Volunteer Content Moderators	Yang Trista Cao, Lovely-Frances Domingo, Sarah Ann Gilbert,Michelle Mazurek, Katie Shilton, Hal Daumé III	Ongoing efforts to automate content moderation focus on identifying toxic content but may not fully address the needs of volunteer moderators. The paper reveals a gap between past research on automation and the requirements of moderators by reviewing models on Hugging Face and testing state-of-the-art LLMs (GPT-4 and Llama-2), showing a significant recall gap on platform rules.

DEVELOPMENT ENVIRONMENT

Software Requirements

- Google Colab/Jupyter Notebook
- Python
- Flask/Streamlit
- NLP and Computer Vision Libraries

Hardware Requirements

- Laptop/ Personal Computer (PC)
- Random Access Memory (RAM): 8 GB or above
- Central Processing Unit (CPU): 1.7 GHz Processor and above
- Operating System (OS): Windows 10 and above

Project Modules

Various features of the Content Moderation System Framework are:

- **Text Classification (Ham or Spam)**

We begin by classifying textual content into "ham" (safe) or "spam" categories using machine learning models. This is essential for weeding out unwanted and harmful content from user-generated text. This is done by training a model by using dataset from Kaggle. Algorithms used are Logistic Regression, Scikit-Learn.

- **Sentiment Analysis**

The framework further analyzes text to determine the sentiment as positive, negative, or neutral and gives its probability. This helps in understanding the emotional tone of messages, comments, and posts in real-time.This is done by using Vader(SentimentIntensityAnalyzer) and Flair which are pre-trained models.

Project Modules

- **Profanity Detection**

To identify explicit words(hate,abuse,toxicity) and the probability of toxicity in text(toxicity,severe toxicity,obscene,threat,insult,identity attack), we employ profanity classification models like detoxify and better_profanity & Logistic Regression. This aids in identifying and filtering out offensive language and abusive content.

- **Custom NER for Abusive Words**

In addition to general profanity detection, we use custom named entity recognition(ABUSIVE', 'SEXUALLY EXPLICIT', 'SUGGESTIVE', 'VIOLENCE', 'WARNING', 'INSULT) to identify and classify specific abusive words, slurs, or derogatory terms to provide a more tailored approach to moderation.We use custom annotations as a part of our data and the model used is Spacy Custom NER. Packages used are Spacy, en_core_web_lg.

Project Modules

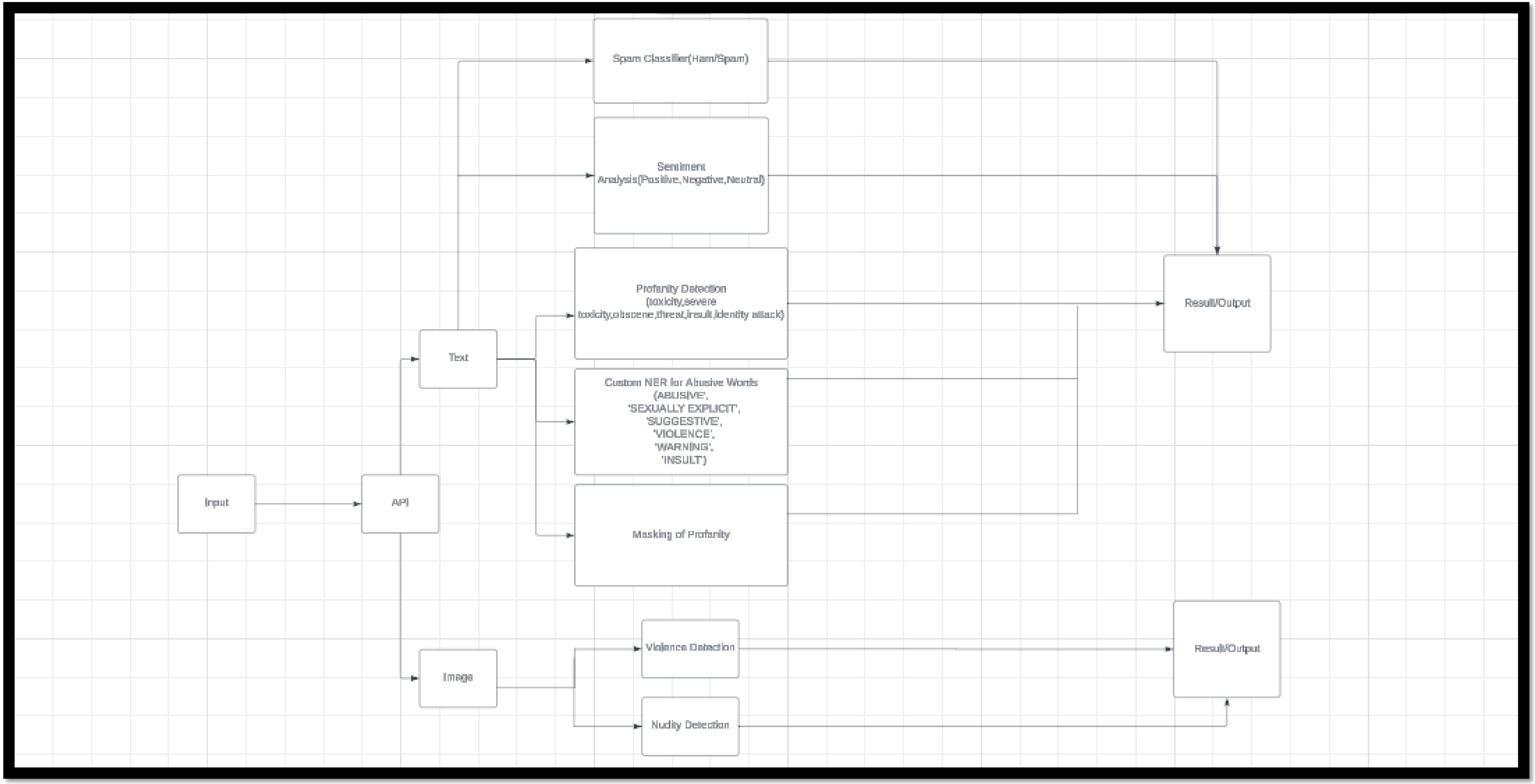
- **Masking of Profanity**

Detected profanity is then masked or replaced with appropriate placeholders to prevent it from being visible to users, ensuring a cleaner online environment.

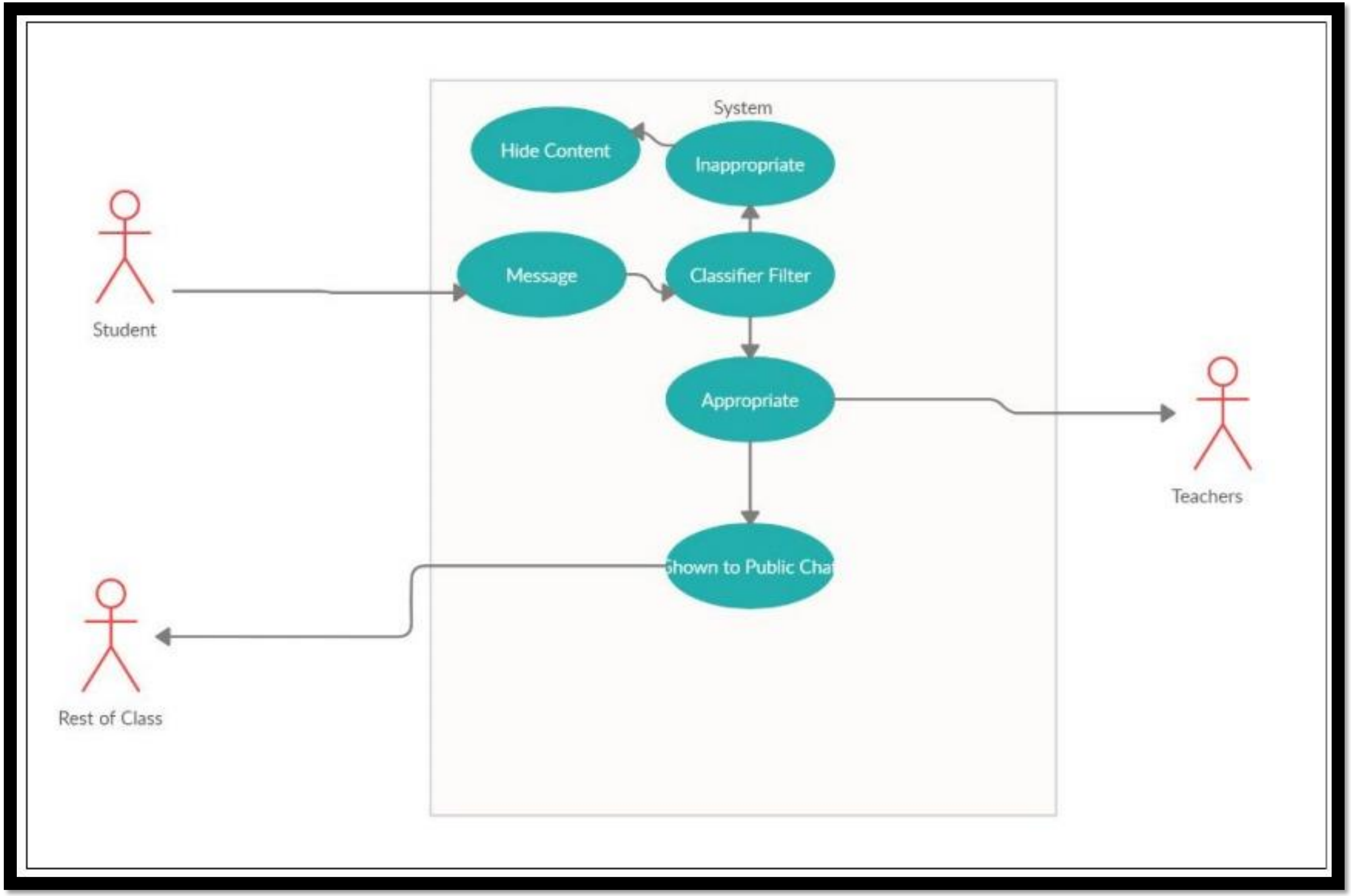
- **Image Classification (Safe vs. Unsafe)**

-Moving beyond text, the framework utilizes computer vision to classify images into safe or unsafe categories based on the presence of NSFW, violence, or nudity. This ensures that explicit and harmful images are filtered out.
-For NSFW Image Detection the packages used are nsfw_detector, nudenet and the model used is nsfw_mobilenet2.224x224(pre-trained).
-For Image Violence Detection the model used is Resnet50(fine tuned on custom data) and packages used are torch and torchvision.

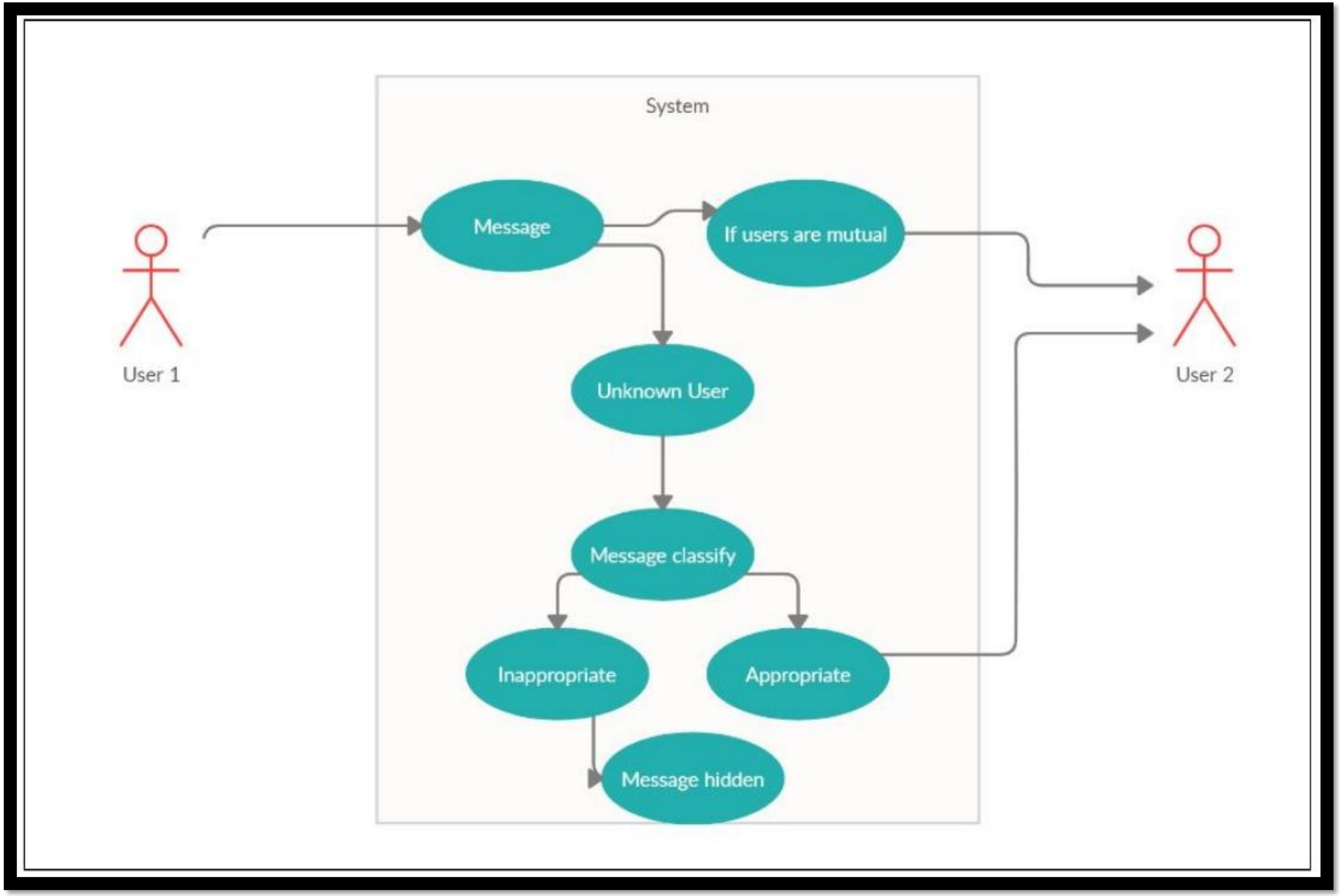
System Architecture


Spam Classifier(Ham/Spam)
Sentiment Analysis(Positive,Negative,Neutral)
Profanity Detection (toxicity,severe toxicity,obscene,threat,insult,identity attack)
Custom NER for Abusive Words (ABUSIVE', 'SEXUALLY EXPLICIT', 'SUGGESTIVE', 'VIOLENCE', 'WARNING', 'INSULT')
Masking of Profanity
Result/Output
Text
Input
API
Image
Violence Detection
Nudity Detection
Result/Output

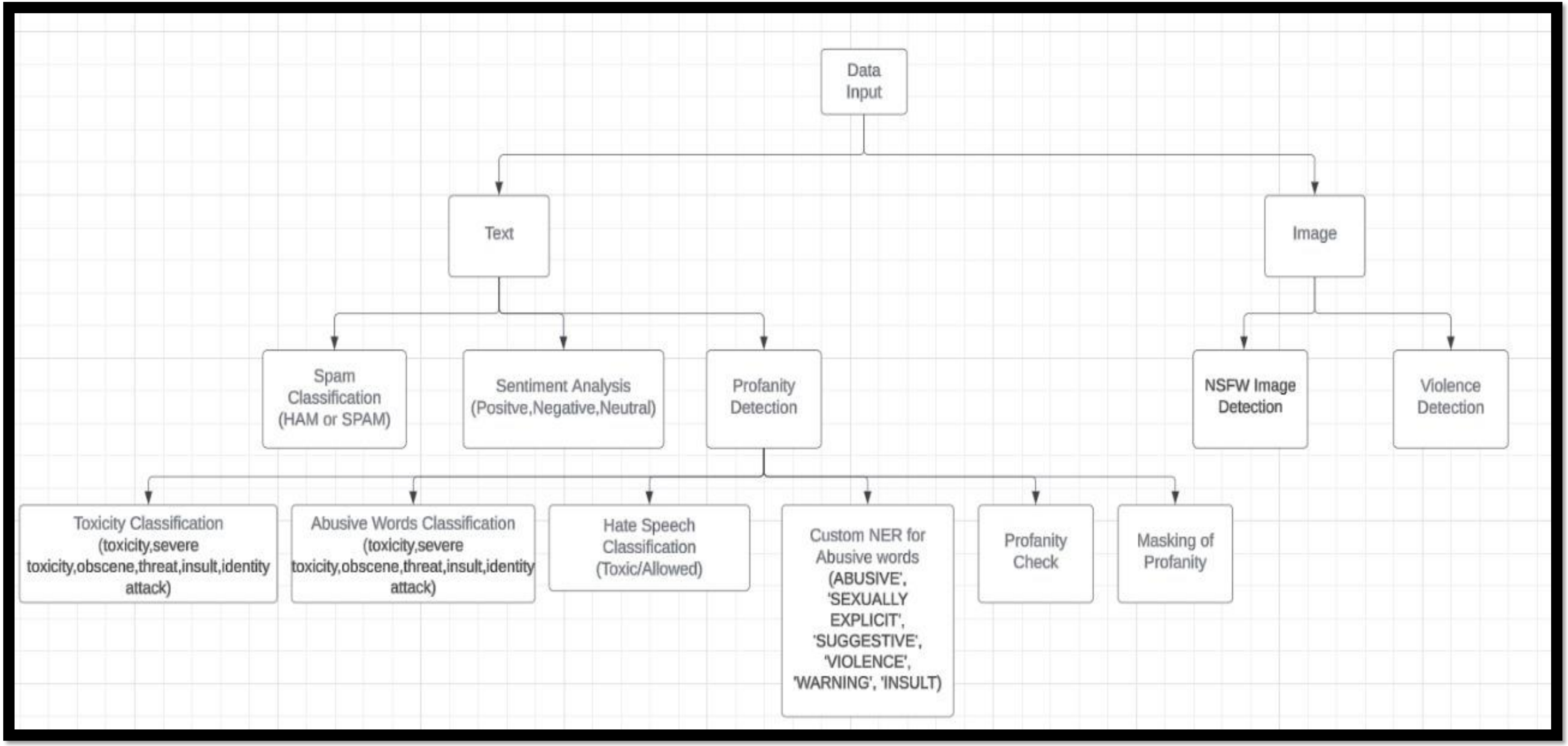
Use Case Diagram(Online Education)



Use Case Diagram(Social Media)



Flow Chart



Data Input

Text

Image

Spam Classification (HAM or SPAM)

Sentiment Analysis (Positve,Negative,Neutral)

Profanity Detection

NSFW Image Detection

Violence Detection

Toxicity Classification (toxicity,severe toxicity,obscene,threat,insult,identity attack)

Abusive Words Classification (toxicity,severe toxicity,obscene,threat,insult,identity attack)

Hate Speech Classification (Toxic/Allowed)

Custom NER for Abusive words ('ABUSIVE', 'SEXUALLY EXPLICIT', 'SUGGESTIVE', 'VIOLENCE', 'WARNING', 'INSULT)

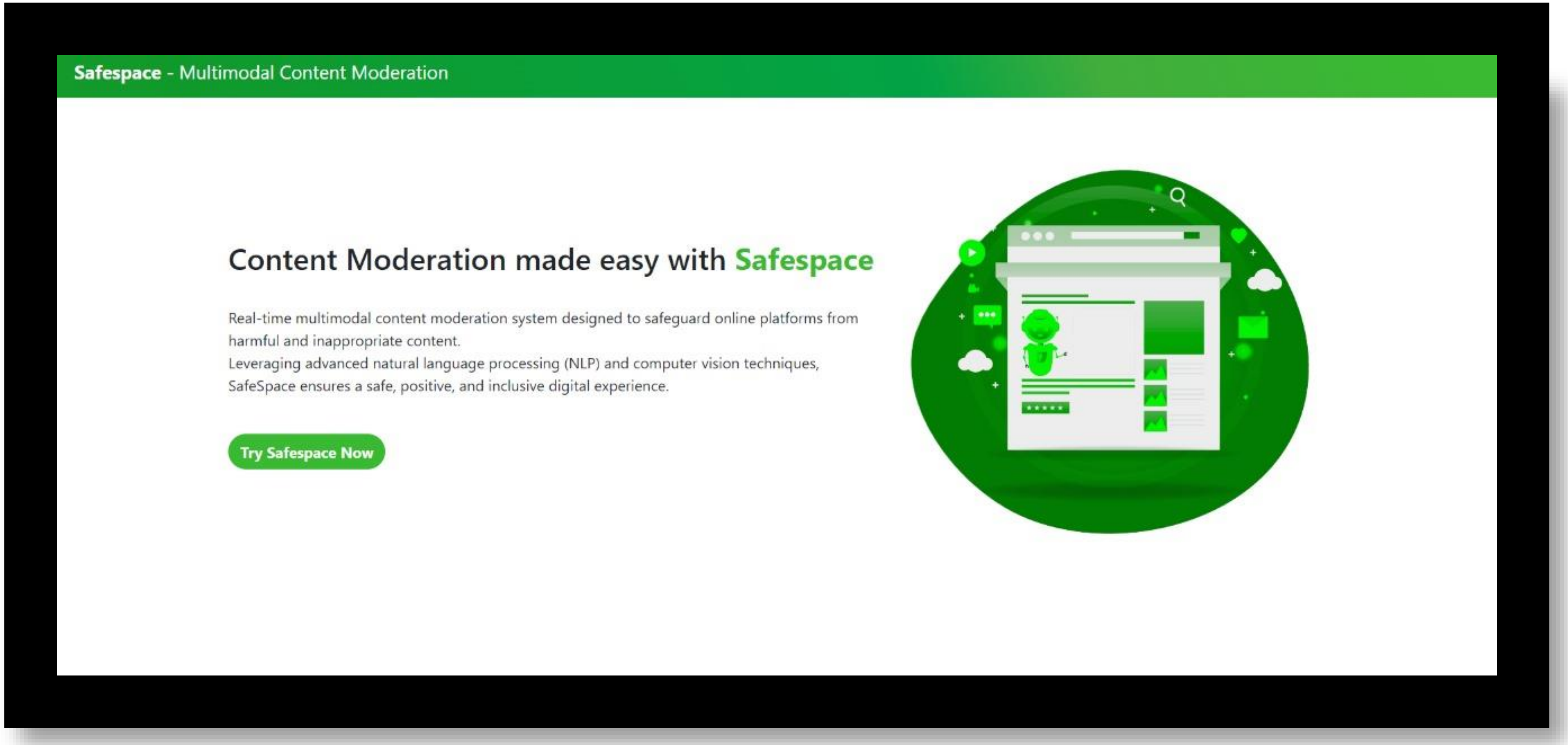
Profanity Check

Masking of Profanity

LIBRARIES USED

- numpy, pandas for text data manipulation
- matplotlib, seaborn for visualization
- re, nltk, spacy for text preprocessing and NER
- scikit-learn, Logistic Regression for hatespeech classification
- opencv-python, pillow for image preprocessing
- torch, torchvision.transforms, resnet50 for violence detection
- better-profanity for profanity check
- Detoxify for toxicity classification
- vader for sentiment analysis
- nudenet for NSFW detection
- Flask/Streamlit for UI

User Interface

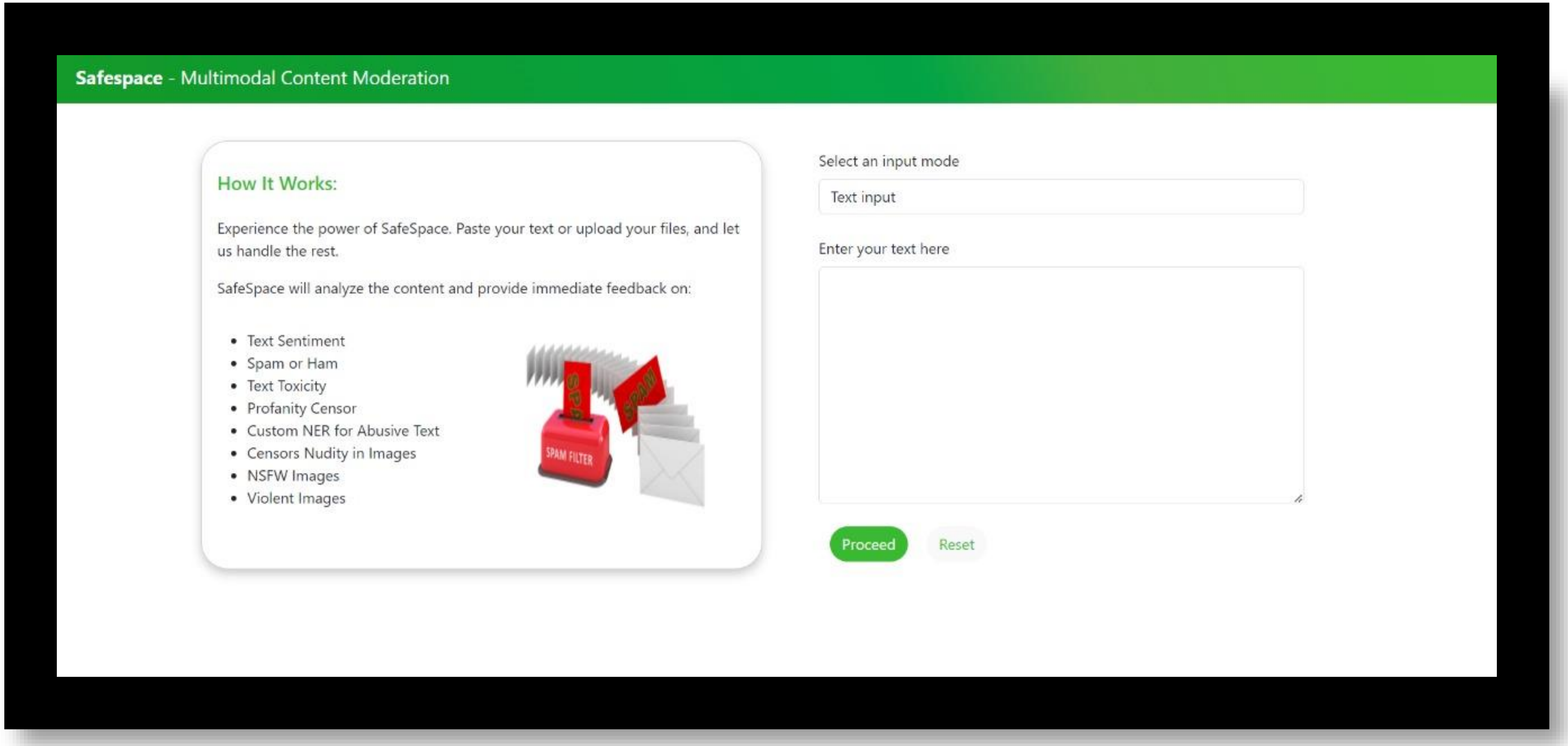


Safespace - Multimodal Content Moderation

Content Moderation made easy with Safespace

Real-time multimodal content moderation system designed to safeguard online platforms from harmful and inappropriate content.
Leveraging advanced natural language processing (NLP) and computer vision techniques, SafeSpace ensures a safe, positive, and inclusive digital experience.

User Interface



Safespace - Multimodal Content Moderation

How It Works:

Experience the power of SafeSpace. Paste your text or upload your files, and let us handle the rest.

SafeSpace will analyze the content and provide immediate feedback on:

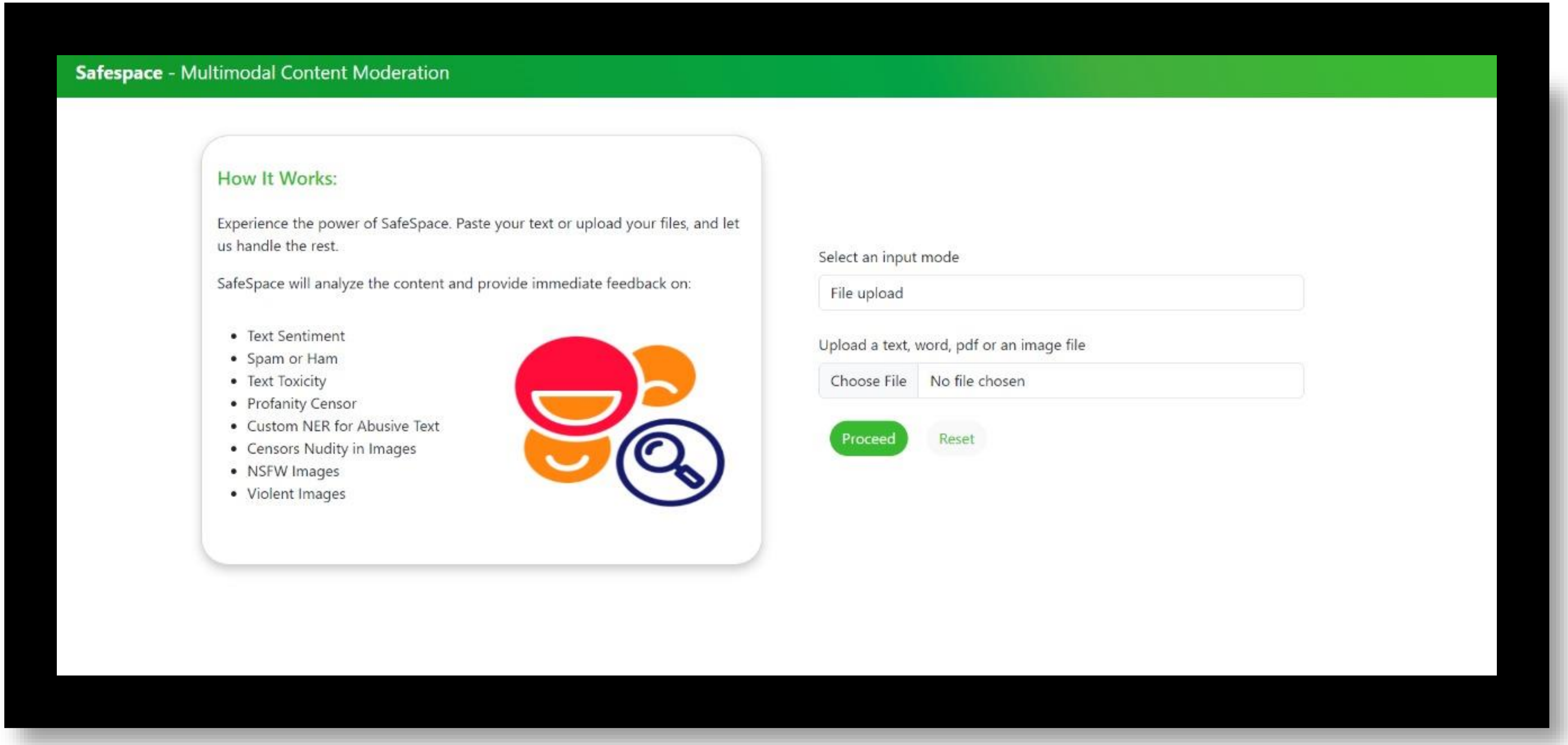
- Text Sentiment
- Spam or Ham
- Text Toxicity
- Profanity Censor
- Custom NER for Abusive Text
- Censors Nudity in Images
- NSFW Images
- Violent Images

Select an input mode

Text input

Enter your text here

User Interface



Safespace - Multimodal Content Moderation

How It Works:

Experience the power of SafeSpace. Paste your text or upload your files, and let us handle the rest.

SafeSpace will analyze the content and provide immediate feedback on:

- Text Sentiment
- Spam or Ham
- Text Toxicity
- Profanity Censor
- Custom NER for Abusive Text
- Censors Nudity in Images
- NSFW Images
- Violent Images

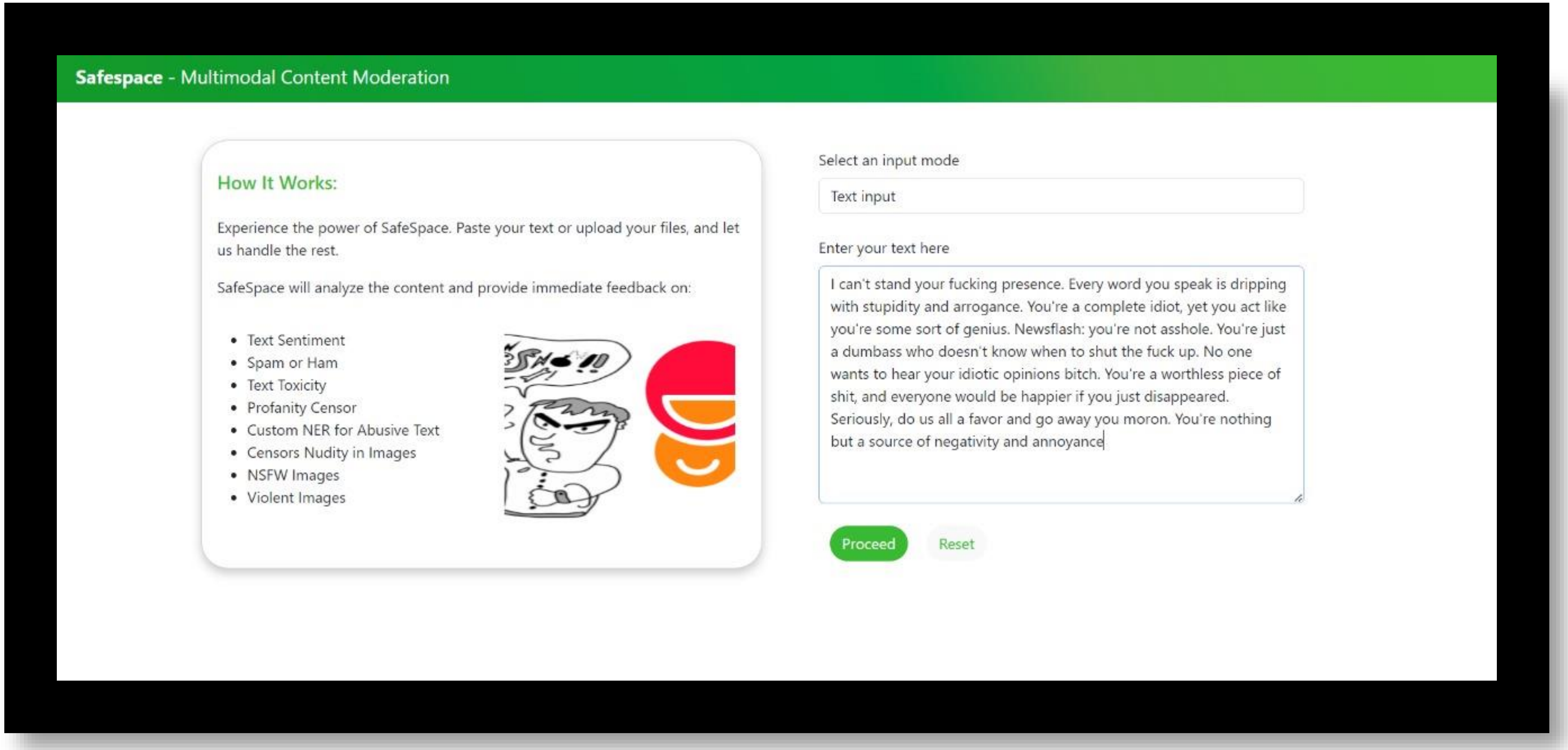
Select an input mode

File upload

Upload a text, word, pdf or an image file

Choose File No file chosen

User Interface



Safespace - Multimodal Content Moderation

How It Works:

Experience the power of SafeSpace. Paste your text or upload your files, and let us handle the rest.

SafeSpace will analyze the content and provide immediate feedback on:

- Text Sentiment
- Spam or Ham
- Text Toxicity
- Profanity Censor
- Custom NER for Abusive Text
- Censors Nudity in Images
- NSFW Images
- Violent Images

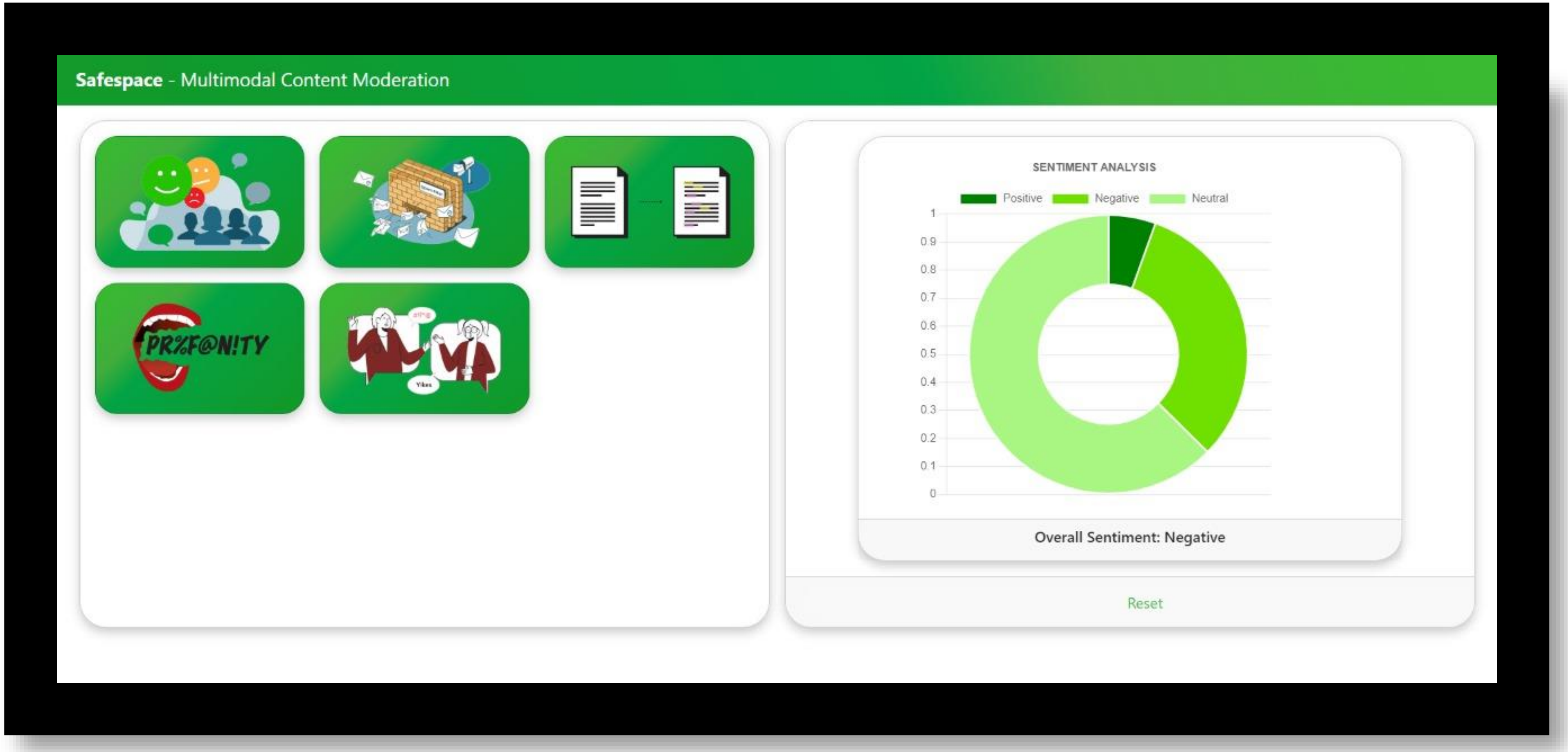
Select an input mode

Text input

Enter your text here

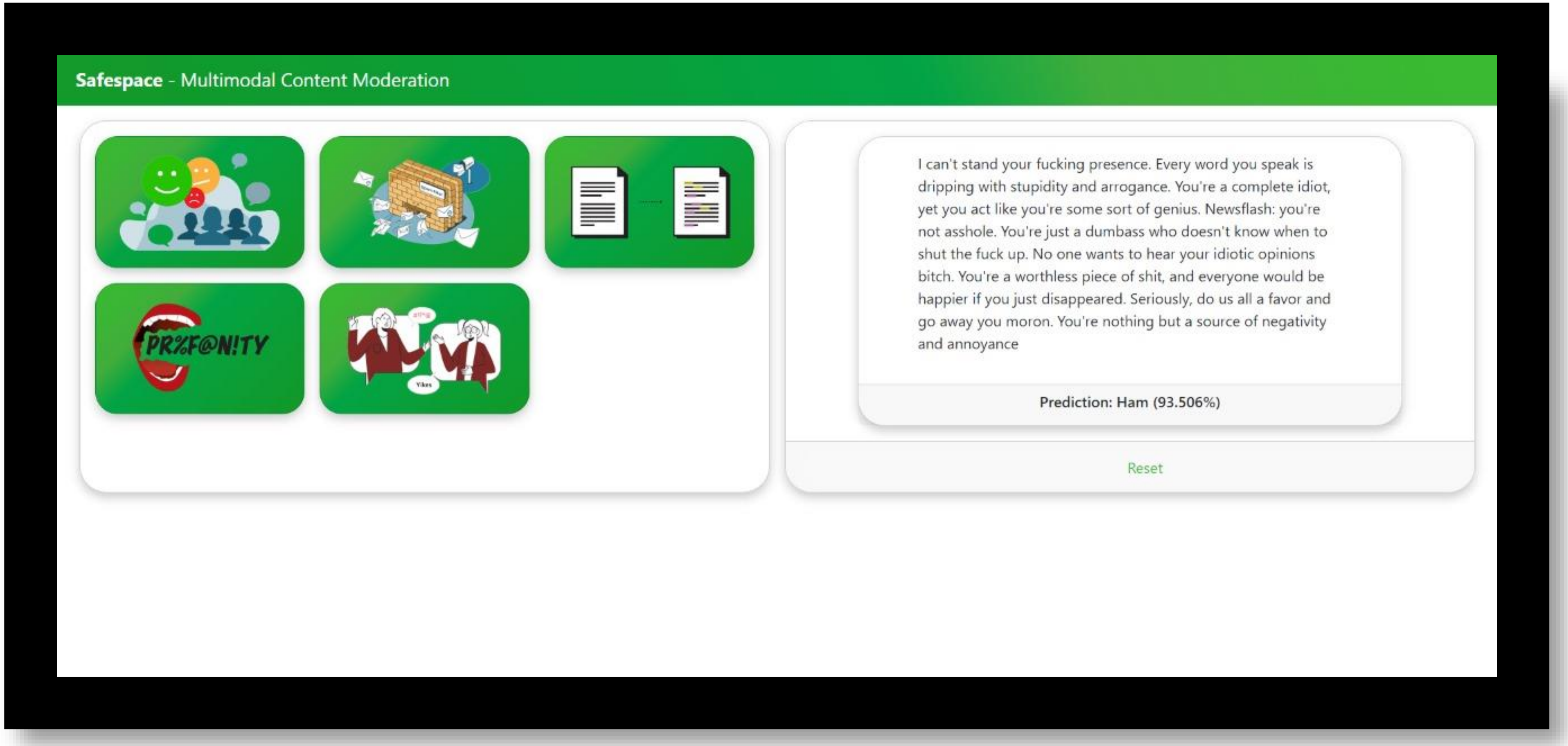
I can't stand your fucking presence. Every word you speak is dripping with stupidity and arrogance. You're a complete idiot, yet you act like you're some sort of genius. Newsflash: you're not asshole. You're just a dumbass who doesn't know when to shut the fuck up. No one wants to hear your idiotic opinions bitch. You're a worthless piece of shit, and everyone would be happier if you just disappeared. Seriously, do us all a favor and go away you moron. You're nothing but a source of negativity and annoyance

Text Input Results



Overall Sentiment: Negative

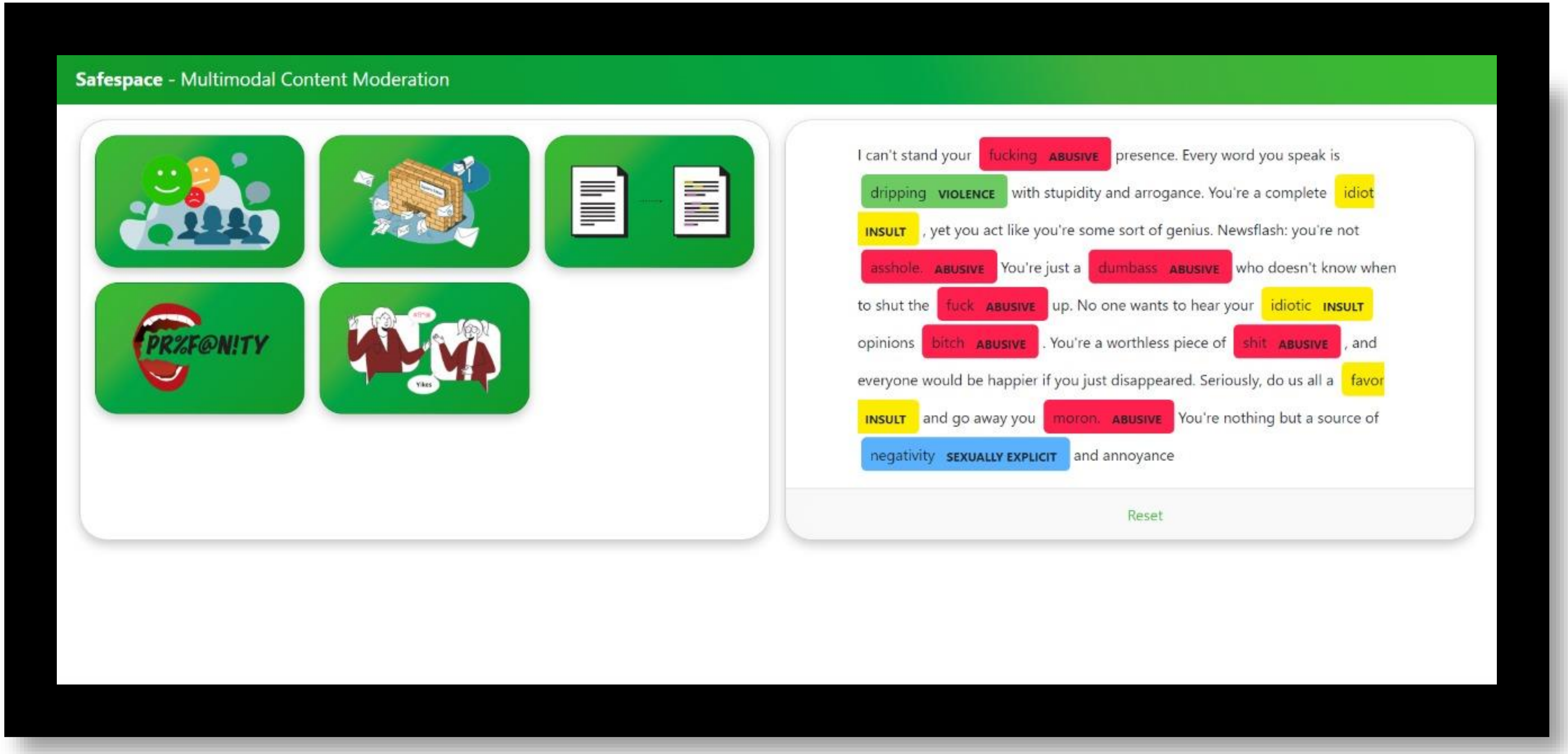
Text Input Results



I can't stand your fucking presence. Every word you speak is dripping with stupidity and arrogance. You're a complete idiot, yet you act like you're some sort of genius. Newsflash: you're not asshole. You're just a dumbass who doesn't know when to shut the fuck up. No one wants to hear your idiotic opinions bitch. You're a worthless piece of shit, and everyone would be happier if you just disappeared. Seriously, do us all a favor and go away you moron. You're nothing but a source of negativity and annoyance

Prediction: Ham (93.506%)

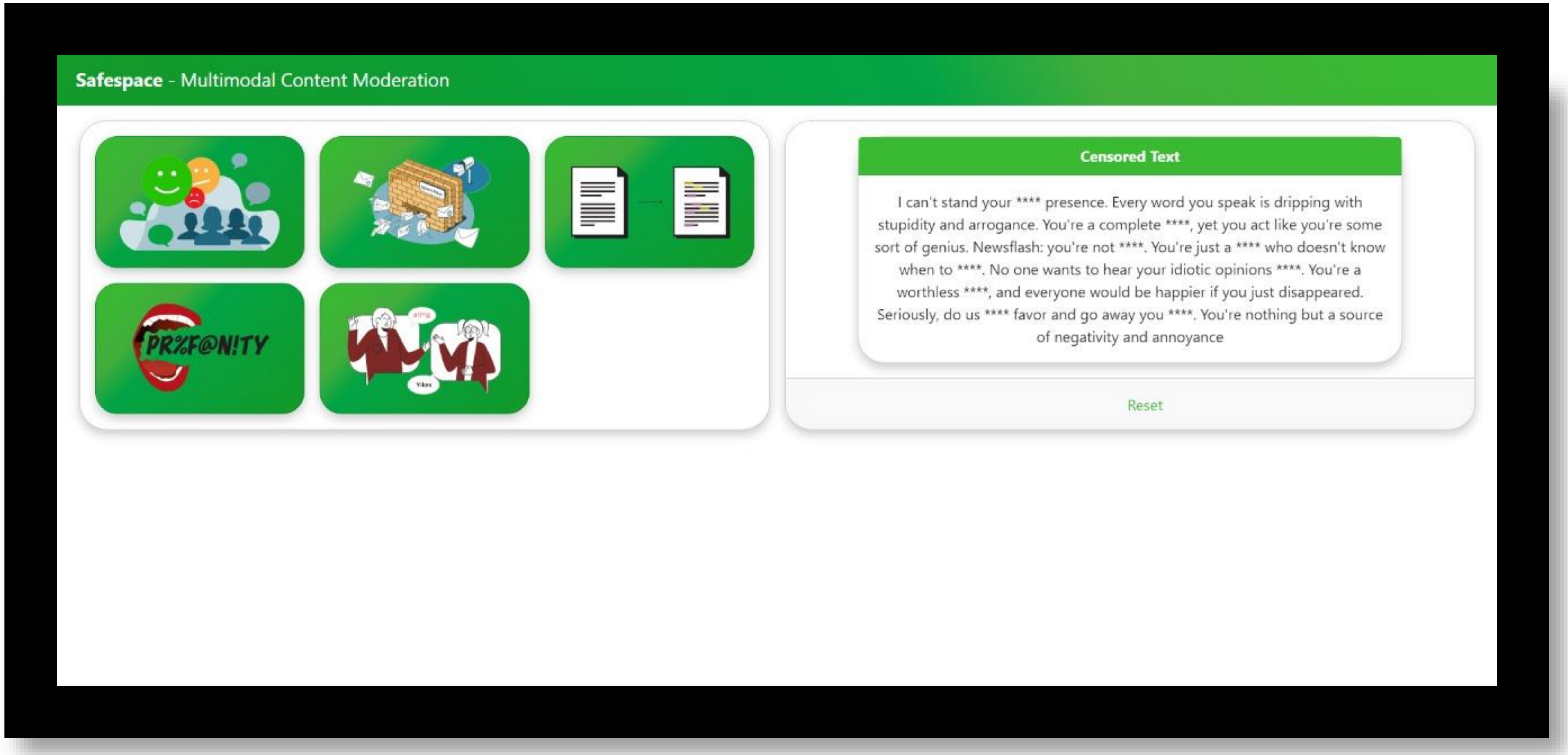
Text Input Results



Safespace - Multimodal Content Moderation

I can't stand your fucking ABUSIVE presence. Every word you speak is dripping VIOLENCE with stupidity and arrogance. You're a complete idiot INSULT , yet you act like you're some sort of genius. Newsflash: you're not asshole. ABUSIVE You're just a dumbass ABUSIVE who doesn't know when to shut the fuck ABUSIVE up. No one wants to hear your idiotic INSULT opinions bitch ABUSIVE . You're a worthless piece of shit ABUSIVE , and everyone would be happier if you just disappeared. Seriously, do us all a favor INSULT and go away you moron. ABUSIVE You're nothing but a source of negativity SEXUALLY EXPLICIT and annoyance

Text Input Results



Safespace - Multimodal Content Moderation

Censored Text

I can't stand your **** presence. Every word you speak is dripping with stupidity and arrogance. You're a complete ****, yet you act like you're some sort of genius. Newsflash: you're not ****. You're just a **** who doesn't know when to ****. No one wants to hear your idiotic opinions ****. You're a worthless ****, and everyone would be happier if you just disappeared. Seriously, do us **** favor and go away you ****. You're nothing but a source of negativity and annoyance

Text Input Results

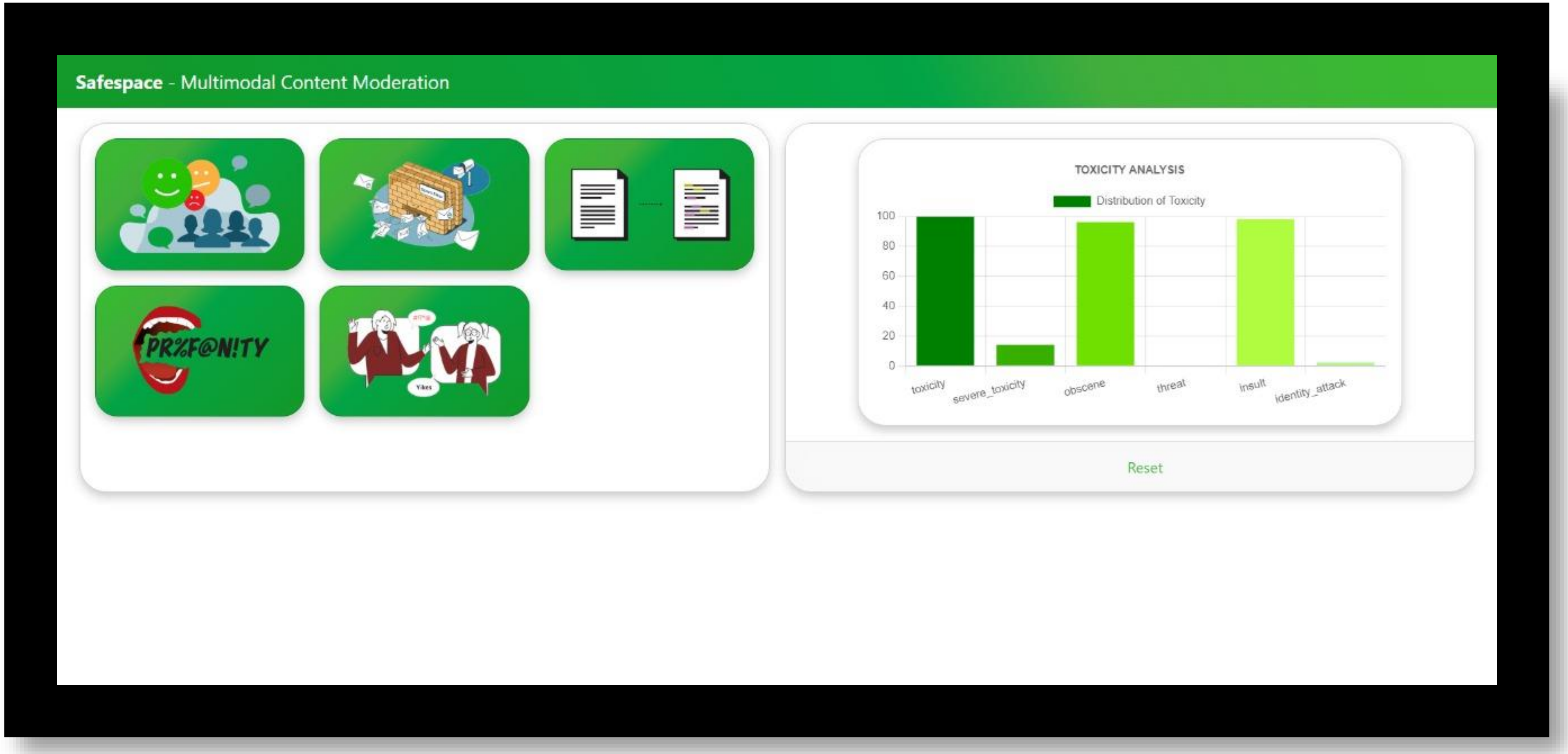


Image Input Results

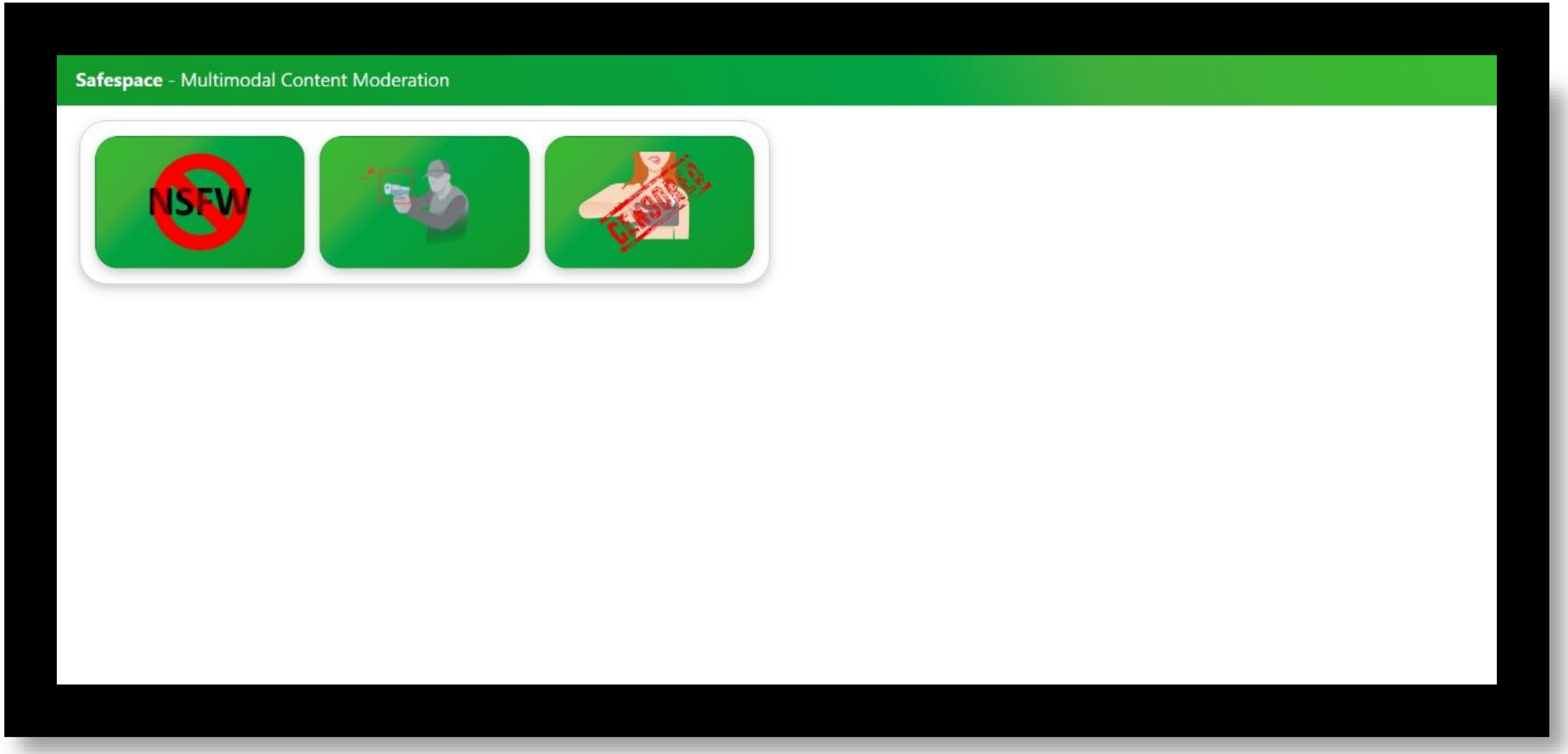
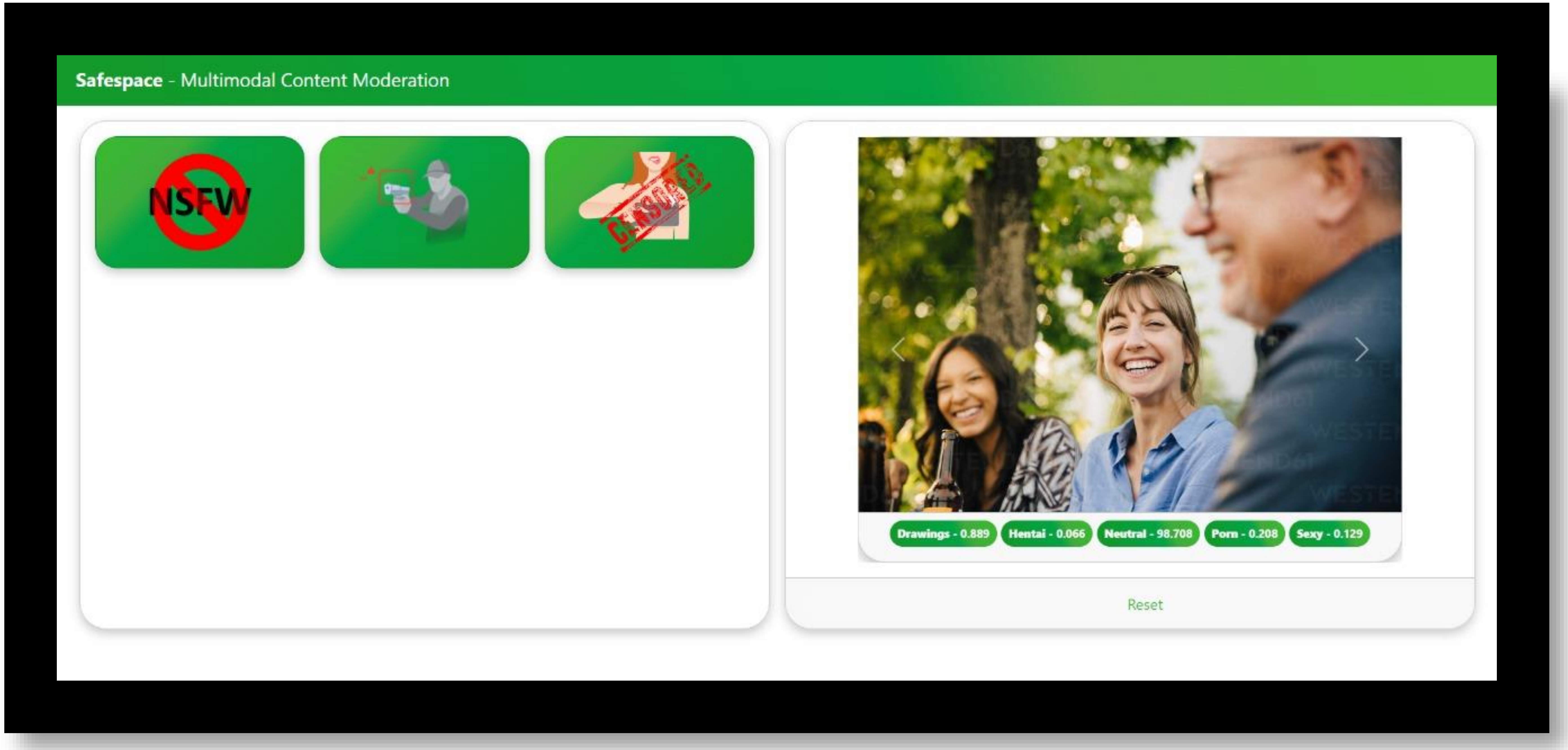


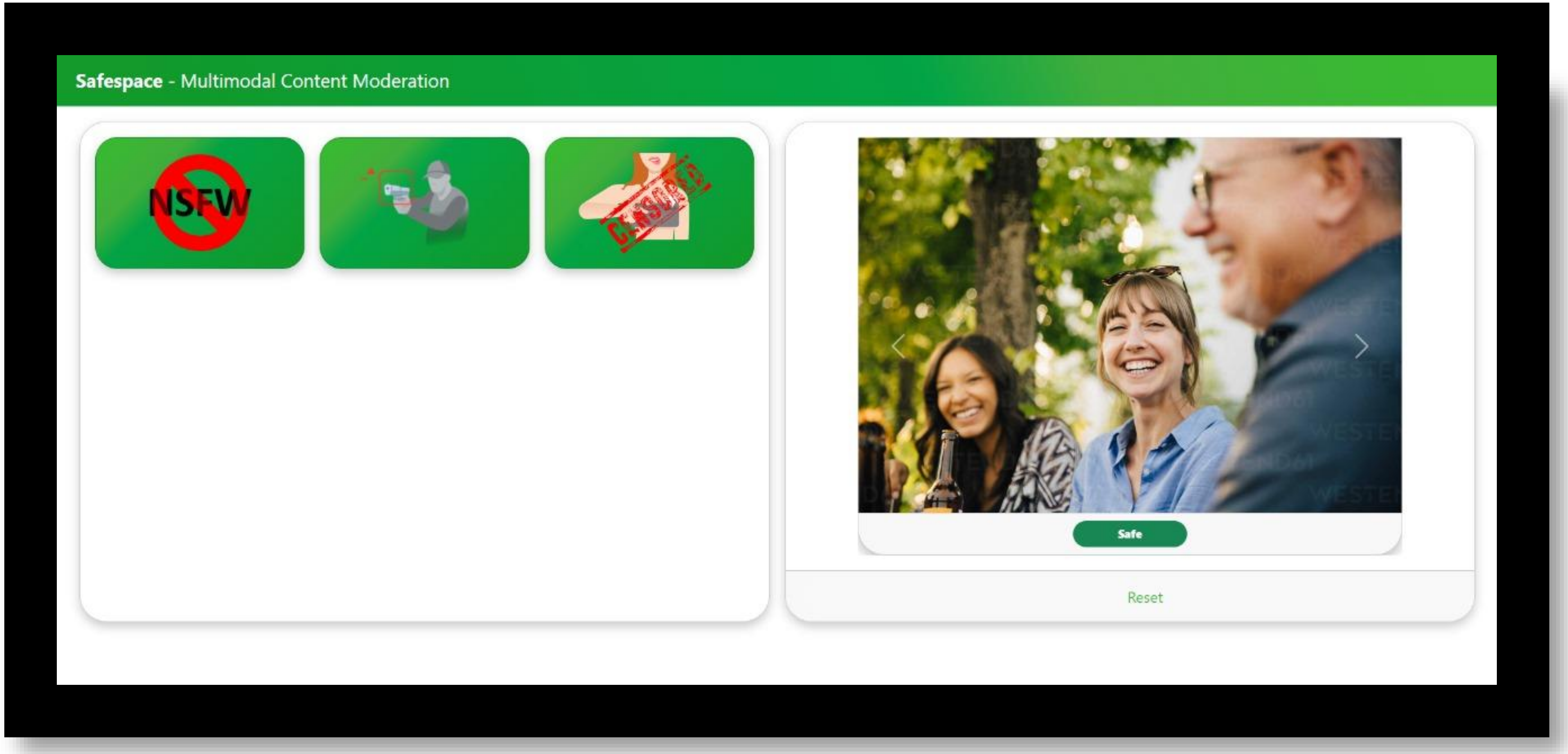
Image Input Results



Safespace - Multimodal Content Moderation

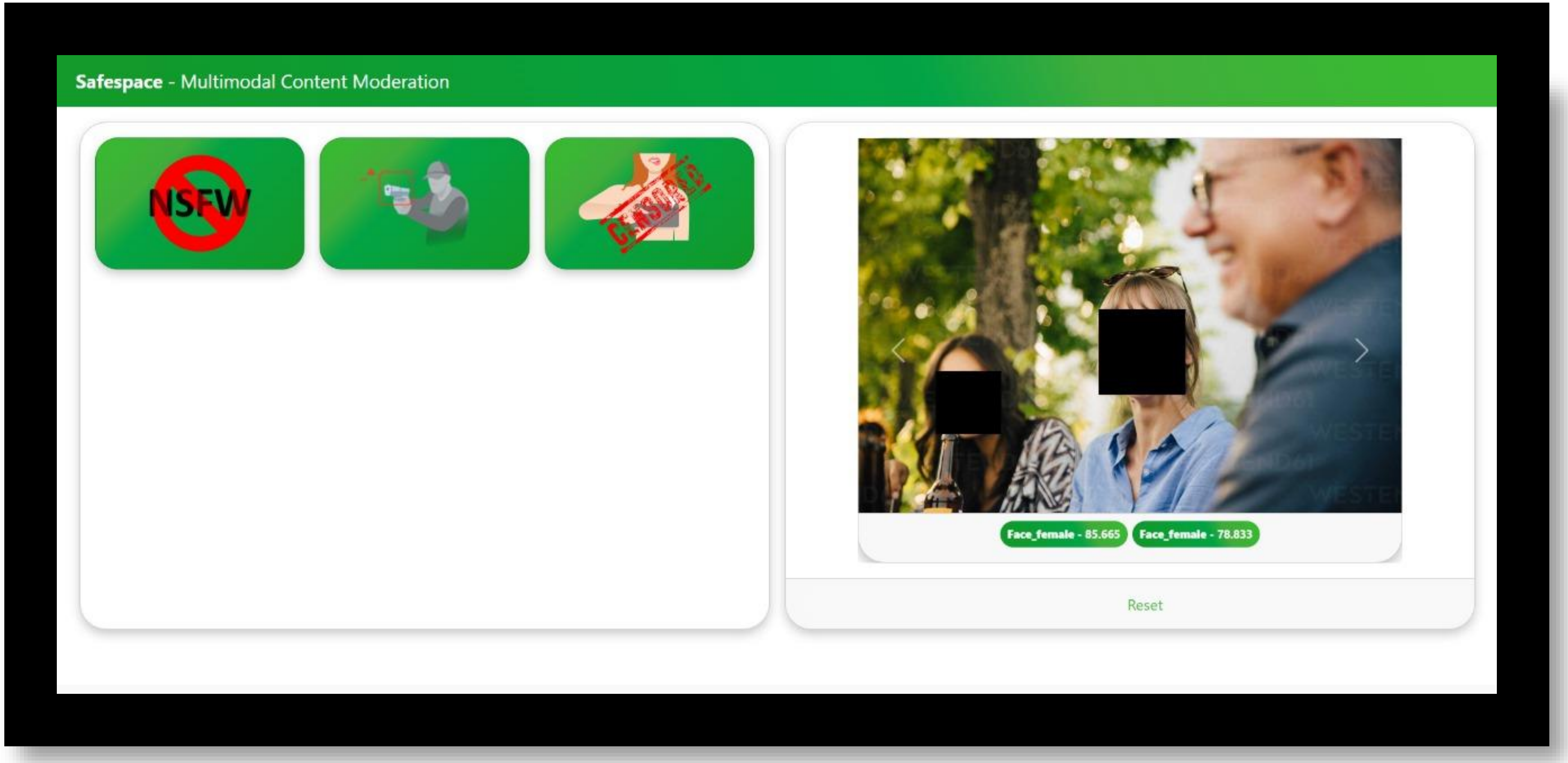
Drawings - 0.889 Hentai - 0.066 Neutral - 98.708 Porn - 0.208 Sexy - 0.129

Image Input Results



Safespace - Multimodal Content Moderation

Image Input Results



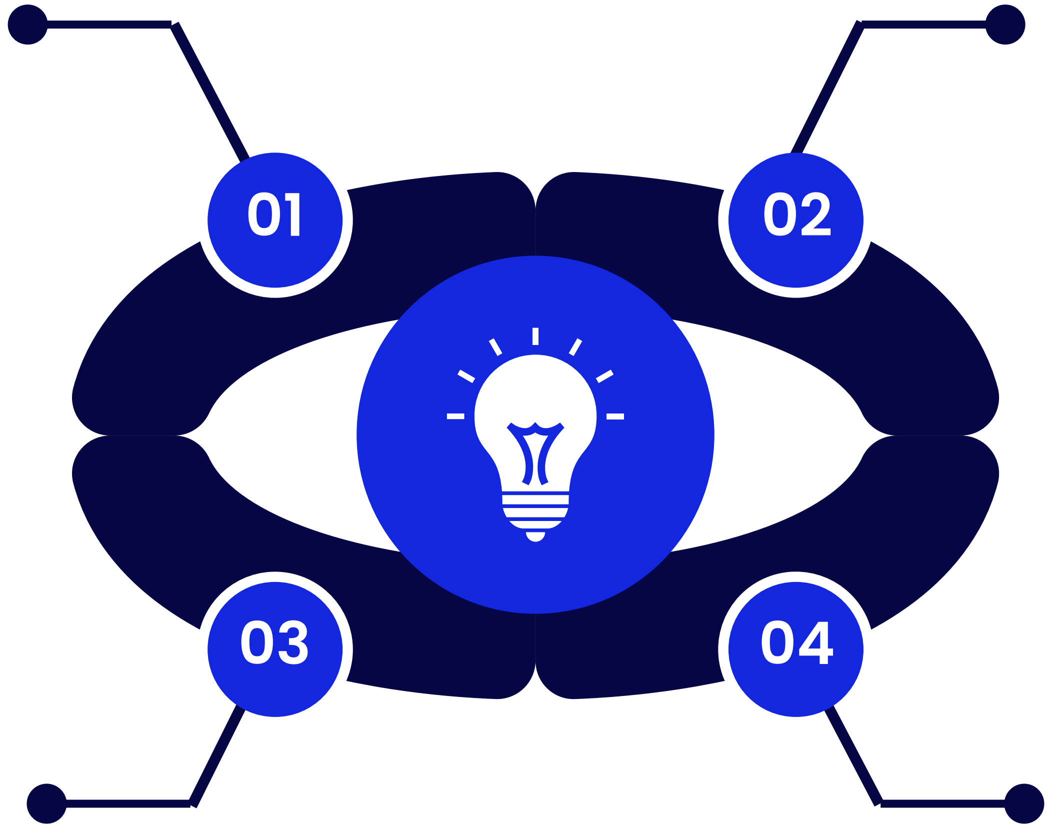
Applications

Social Media Platforms

This framework is invaluable for social media networks, ensuring that user-generated content aligns with community guidelines, creating a safer and more positive user experience.

Online Forums and Communities

Online forums and community platforms can use this solution to automatically moderate discussions, maintaining respectful and constructive conversations. For example Gaming Communities. In the gaming industry, where interactions can be intense and competitive, the framework helps maintain a respectful environment by filtering out abusive language and inappropriate images.



E-commerce Websites

E-commerce platforms can safeguard their product listings and reviews against spam, profanity, and inappropriate images, enhancing the overall shopping experience.

Educational Platforms

Online learning environments benefit from content moderation to ensure a safe and respectful space for both students and instructors.

Thank You