# Obscenity Blocker Extension

Anirudh Arcot- RA2211026010449
Ayushi Roy-RA2211026010439

# Table of contents

# Introduction

The internet serves as a vast repository of information, but it also hosts content that may be inappropriate or harmful, especially for vulnerable users such as children or employees. To address this challenge, we present an advanced content filtering and obscenity detection system that aims to safeguard users from accessing objectionable material online.

Traditional methods of content filtering and obscenity detection do not effectively mitigate these risks. Static blacklists can be circumvented, while keyword-based filters may fail to accurately identify nuanced forms of objectionable content. By combining machine learning algorithms, image processing techniques, and natural language understanding, our system offers a multi-faceted approach to content filtering and obscenity detection.

# Background

Traditional approaches to internet security have relied on static blacklists and keyword-based filters to identify and block inappropriate content. While effective to some extent, these methods often struggle to keep pace with the rapidly evolving landscape of online threats. Moreover, the widespread adoption of multimedia content formats, including images, videos, and audio clips, has necessitated more sophisticated techniques for content analysis and classification.

Against this backdrop, there has been a growing emphasis on the development of advanced technologies and methodologies for enhancing internet security. Machine learning algorithms, such as Convolutional Neural Networks (CNNs) and Natural Language Processing (NLP) models, offer promising avenues for automating the detection of objectionable content across diverse formats and contexts

# Motivation

The motivation behind our endeavor stems from a profound commitment to fostering a safer and more secure online environment for users of all ages and backgrounds. In an era marked by the exponential growth of internet usage, ensuring the integrity of online spaces is paramount to safeguarding individuals from exposure to harmful or inappropriate content. By developing an advanced content filtering and obscenity detection system, we seek to empower users with the tools and resources necessary to navigate the digital landscape with confidence and peace of mind. Our motivation is driven by a recognition of the pressing need for innovative solutions that can adapt to the dynamic nature of online threats, thereby promoting responsible digital citizenship and upholding fundamental principles of decency and respect in the virtual realm.

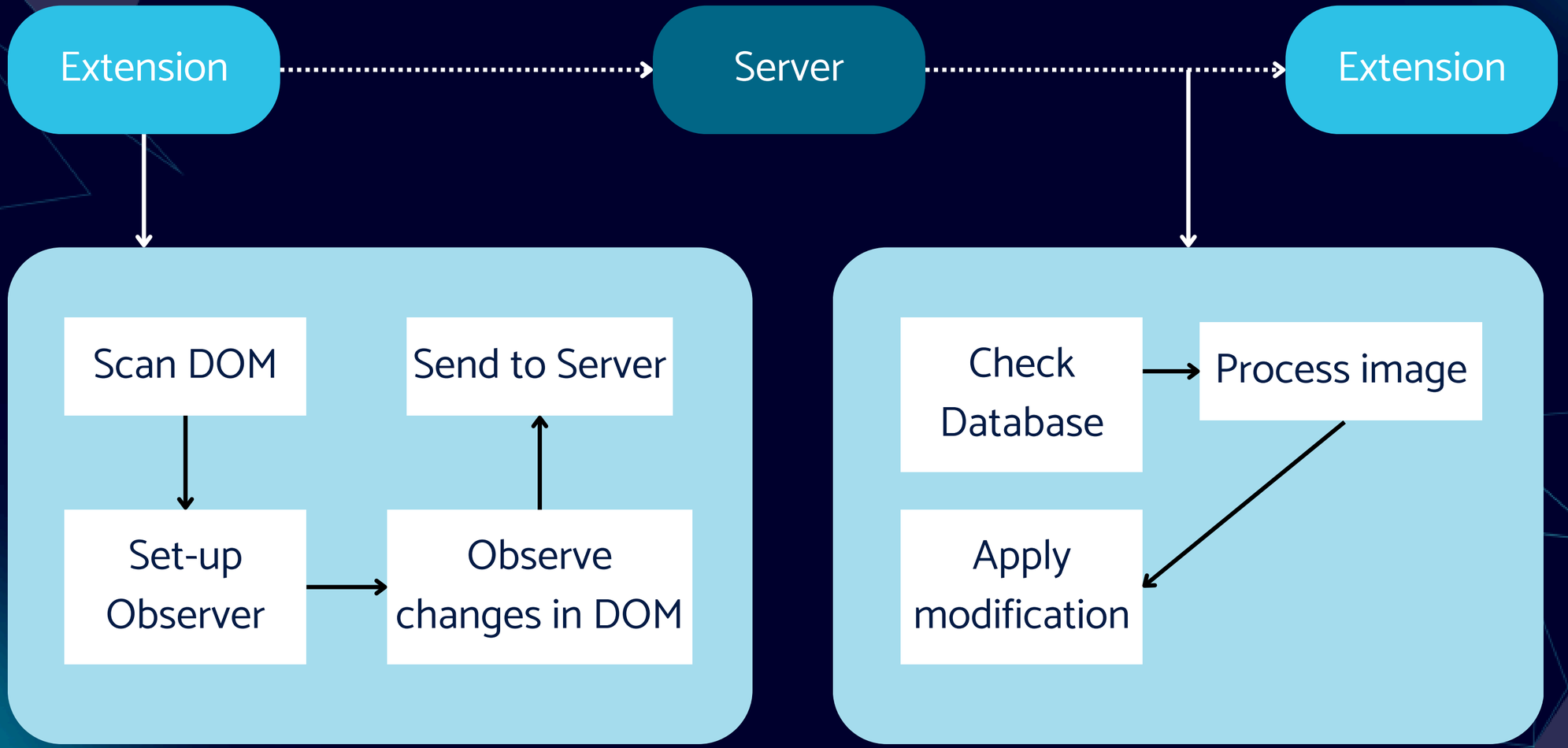
# Methodology

The provided code establishes a convolutional neural network (CNN) for image classification using TensorFlow and Keras. Initially, the dataset is prepared, loading images from a directory structure and splitting them into training and validation sets. Data augmentation techniques are then applied to enrich the training dataset with variations of the original images. Subsequently, the CNN model architecture is defined, comprising convolutional layers for feature extraction, dropout for regularization, and fully connected layers for classification. During training, checkpoints are saved at regular intervals, and early stopping is employed to prevent overfitting. After training, both the model weights and the entire model are saved for future use. Finally, the training history is visualized through plots showing the accuracy and loss over epochs, offering insights into the model's learning progress and performance.

# Machine Learning Model

The machine learning methodology employed here revolves around constructing a robust image classification model using convolutional neural networks (CNNs). The process begins with data preparation, where the dataset is organized and preprocessed, ensuring compatibility with the model architecture. Data augmentation is then applied to augment the dataset, enhancing its diversity and aiding the model's generalization capability. Subsequently, the CNN model is defined, featuring convolutional layers to extract hierarchical features from images, along with fully connected layers for classification. Throughout training, the model's parameters are optimized using the Adam optimizer, and its performance is evaluated using metrics such as accuracy and loss. Techniques like early stopping and checkpointing are implemented to prevent overfitting and ensure model stability. After training, the model's effectiveness is further assessed through validation data, and the trained model is saved for deployment in real-world scenarios. This methodology encapsulates a systematic approach to building and training CNN models for image classification tasks, underlining the iterative process of experimentation and refinement in machine learning workflows.

# Workflow


Extension
Server
Extension
Scan DOM
Send to Server
Set-up Observer
Observe changes in DOM
Check Database
Process image
Apply modification

# Technology Stack

## Backend

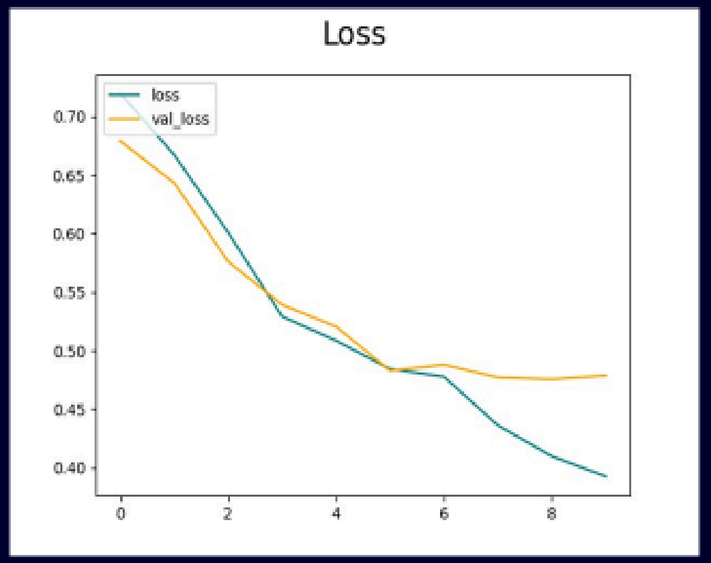




## Frontend

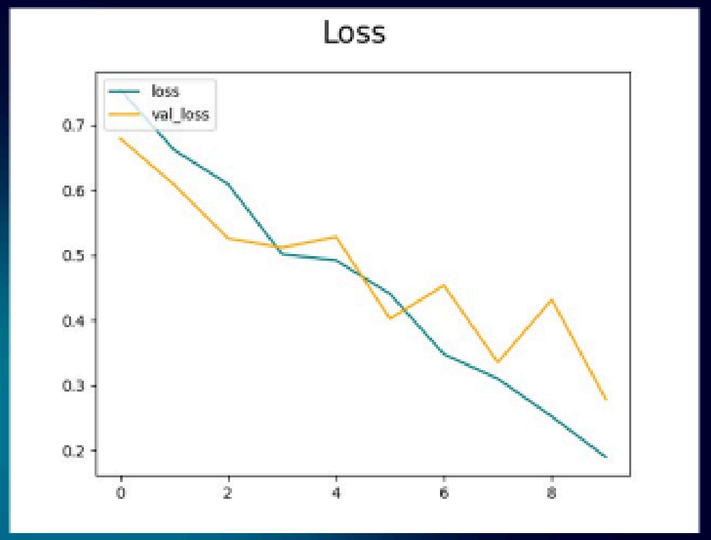

## Database

# Analysis




1

The model employs a Convolutional Neural Network (CNN) architecture, a cutting-edge technology used to train machines by analyzing images in a manner akin to human perception.




2

For the model to achieve accurate predictions, a meticulously curated custom dataset consisting of approximately 30,000 images has been meticulously annotated into three distinct categories: "Gore," "Adult," and "SFW"

| | | |
|---|---|---|
| conv2d_2 (Conv2D) | (None, 60, 60, 32) | 4,640 |
| batch_normalization_2 (BatchNormalization) | (None, 60, 60, 32) | 128 |
| conv2d_3 (Conv2D) | (None, 58, 58, 32) | 9,248 |
| batch_normalization_3 (BatchNormalization) | (None, 58, 58, 32) | 128 |
| max_pooling2d_1 (MaxPooling2D) | (None, 29, 29, 32) | 0 |
| conv2d_4 (Conv2D) | (None, 27, 27, 64) | 18,496 |
| batch_normalization_4 (BatchNormalization) | (None, 27, 27, 64) | 256 |
| conv2d_5 (Conv2D) | (None, 25, 25, 128) | 73,856 |
| batch_normalization_5 (BatchNormalization) | (None, 25, 25, 128) | 512 |
| max_pooling2d_2 (MaxPooling2D) | (None, 12, 12, 128) | 0 |
| conv2d_6 (Conv2D) | (None, 10, 10, 128) | 147,584 |
| batch_normalization_6 (BatchNormalization) | (None, 10, 10, 128) | 512 |
| conv2d_7 (Conv2D) | (None, 8, 8, 256) | 295,168 |
| batch_normalization_7 (BatchNormalization) | (None, 8, 8, 256) | 1,024 |
| max_pooling2d_3 (MaxPooling2D) | (None, 4, 4, 256) | 0 |
| conv2d_8 (Conv2D) | (None, 2, 2, 1024) | 2,360,320 |
| batch_normalization_8 (BatchNormalization) | (None, 2, 2, 1024) | 4,096 |
| max_pooling2d_4 (MaxPooling2D) | (None, 1, 1, 1024) | 0 |
| flatten (Flatten) | (None, 1024) | 0 |
| dropout (Dropout) | (None, 1024) | 0 |
| dense (Dense) | (None, 512) | 524,800 |
| dense_1 (Dense) | (None, 3) | 1,539 |

Total params: 3,445,203 (13.14 MB)

```
WARNING:absl:Compiled the loaded model, bu
model loaded
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 380ms/step
False
predict:  https://www.goremedical.com/site
WARNING:absl:Compiled the loaded model, bu
model loaded
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 249ms/step
42.081546783447266 gore
True
```

```
models, prefer using an `Input(shape)` object as the first layer in the model instead.
  super().__init__(
Model: "sequential_1"
```

| Layer (type) | Output Shape | Param # |
|---|---|---|
| sequential (Sequential) | (None, 128, 128, 3) | 0 |
| conv2d (Conv2D) | (None, 126, 126, 16) | 448 |
| batch_normalization (BatchNormalization) | (None, 126, 126, 16) | 64 |
| conv2d_1 (Conv2D) | (None, 124, 124, 16) | 2,320 |
| batch_normalization_1 (BatchNormalization) | (None, 124, 124, 16) | 64 |
| max_pooling2d (MaxPooling2D) | (None, 62, 62, 16) | 0 |
| conv2d_2 (Conv2D) | (None, 60, 60, 32) | 4,640 |
| batch_normalization_2 | (None, 60, 60, 32) | 128 |

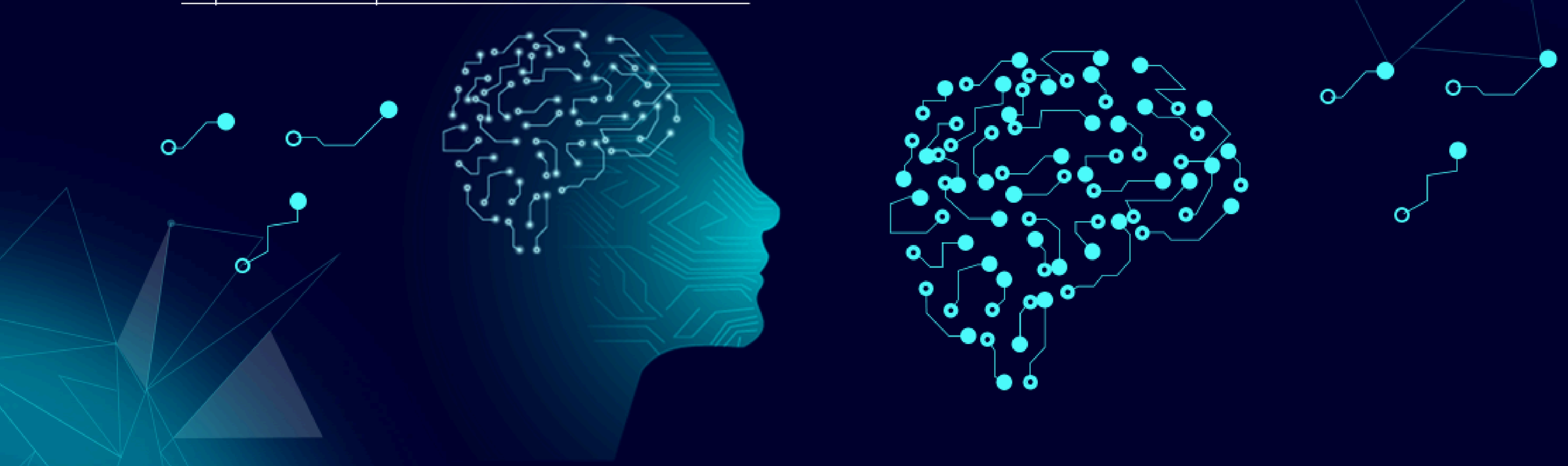
# Conclusion

Evaluation of the system's performance demonstrates its effectiveness in filtering objectionable content and detecting obscenity across various formats. Quantitative metrics and qualitative assessments provide insights into the system's accuracy and efficiency.

The proposed content filtering and obscenity detection system represents a significant advancement in internet security technology. By leveraging cutting-edge techniques, it offers users enhanced control over their online environments while promoting a safer browsing experience. Further research and development efforts are warranted to continuously improve the system's capabilities and adaptability to evolving online threats.

# References


- https://datagen.tech/guides/computer-vision/cnn-convolutional-neural-network/#:~:text=A%20Convolutional%20Neural%20Network%20(CNN,representations%20from%20raw%20input%20images.
- https://d2l.ai/chapter_recurrent-modern/lstm.html

# Thank you

Anirudh Arcot- RA2211026010449

Ayushi Roy- RA2211026010439