PARSHVANATH CHARITABLE TRUST

A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Information Technology
NBA Accredited

A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI
Academic Year 2022-2023

A Project Report on

Content Sanitization using CV & NLP

Submitted in partial fulfilment of the degree of
Bachelor of Engineering(Sem-8)
in

INFORMATION TECHNOLOGY

By
Parth Bhoir(19104050)
Pranav Mayekar(19104040)
Anjali Singh(20204006)

Under the Guidance of
Ms. Rujata Chaudhari
Mrs. Shital Agrawal

1.Project Conception and Initiation

1.1 Abstract

- This platform named coinplanet is useful for new projects and businesses to give a head start to their projects with the help of surveys with support for multimedia data.
- However with increasing number of people using the internet they might find toxic content on platform.
- To deal with this, our platform comes into picture. It handles toxic data and prevent them from spreading further to keep the online space safe.

1.2 Objectives

- To manage toxic text using Machine Learning.
- To detect and manage inappropriate images and videos using OpenCV by making use of CNN to detect features necessary for inappropriate image detection.
- To detect inappropriate audio by converting language to text using Natural Language Processing.

1.3 Literature Review

Sr. No.	Title	Key Findings	Year
1.	Nudity detection based on image zoning.	They used image zoning and skin filtering for nudity detection in their architecture.	2021
2.	Transfer learning based object detection.	Comparison of different pretrained CNN models.	2020
3.	Audio based toxic language detection and classification.	Proposed 2 models, first will process the words and calculate relevance toxicity and second will summarize meaning of audio.	2021

1.4 Problem Definition

- In our platform a user can create these survey forms and add contents like text, image, video and audio in those survey forms.
- The only issue is that if a user adds some inappropriate contents like obscene images and toxic language in the survey forms and then publishes the survey.
- Users of varying age and from different backgrounds visit the platform everyday, it is our duty to make sure that the content on our platform is safe for everyone.

1.5 Scope

- Can be used for censorship of videos.
- Can be used for audio censorship.
- Can be used by organization to reduce toxicity.
- Can be used by companies to automatically sanitize the data.

1.6 Technology stack

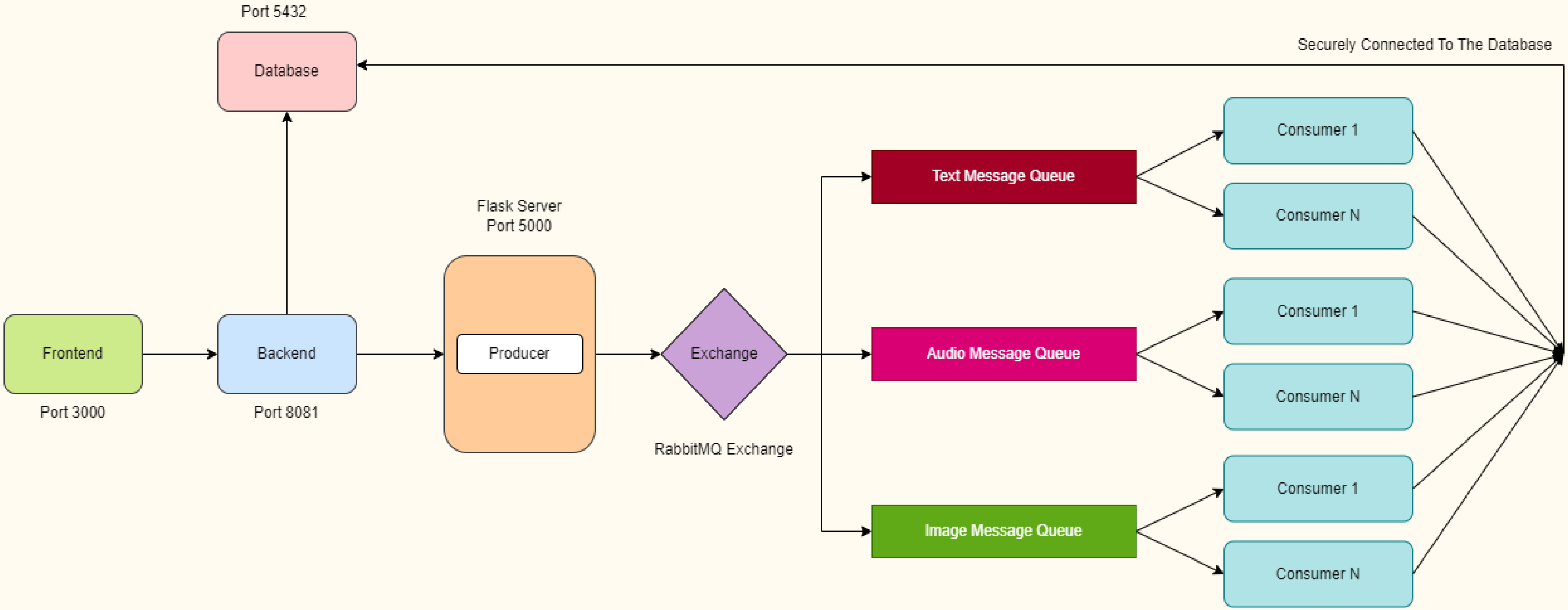
- ReactJS 18.2.0
- Node v14.1.0
- Flask 1.9.1
- RabbitMQ v3.9.0
- OpenCV 4.6.0
- Python v3.0
- CNN Algorithm
- Natural Language Processing

1.7 Benefits for environment & Society

-
-
-

2. Project Design

2.1 Proposed System



2.2 Design(Flow Of Modules)

•

•

•

2.3 Description Of Use Case

2.4 Activity diagram

3. Implementation

4. Testing

5. Result

6. Conclusion and Future Scope

- We achieved text, image, audio and video sanitization and exclude inappropriate multimedia for our website environment.
- We also achieved requests processing in bulk without dropping the requests in a Competing Consumer Pattern .
- This can be applied to large applications as well but the number of Consumers needs to be increased in accordance with the average number of request.
- In future, we are planning to develop an ecosystem on our platform so that functionalities implemented on platform can be integrated with other platforms in real time.

References

- Clayton Santos, Eulanda M. dos Santos, Eduardo Souto, "NUDITY DETECTION BASED ON IMAGE ZONING", "https://sci-hub.se/10.1109/ISSPA.2012.6310454" The 11th International Conference on Information Sciences, Signal Processing and their Applications: Special Sessions,2020
- Rahat Shahriar Islam, Raisa Siddiqui, Dipta Roy,"Blurring of Inappropriate Scenes in a Video Using Image Processing".
- Shoji Kido, Yasusi Hirano, Noriaki Hashimoto "Detection and Classification of Lung Abnormalities by Use of Convolutional Neural Network (CNN) and Regions with CNN Features (R-CNN)"

Thank You