Social Media Content Moderation Using AI

[1]N. Hari Krishna, [2]Nagendrababu Kallem, [3]Ansar Katrapadu, [4]Srinivas Naik Ketavath, [5]Balu Orsu

1 Assistant Professor, Department of CSE, KKR & KSR Institute of Technology and Sciences, A.P, India.

2,3,4,5 B.Tech Student, Department of CSE, KKR & KSR Institute of Technology and Sciences, A.P, India.

Abstract

Social media platforms have become a part in our daily lives for communication, sharing information, ideas and expressions with other peoples. There is no guarantee that secure information is posted while conversation in social media platforms, the users are increasing day to day, due to the increase of the users the quality of the posted content also varied because different peoples have various mindsets that leads to posting of inappropriate content that makes insecure or harmful to the user. So, this project aims to detect such type of inappropriate content and prevent spreading of the inappropriate information through development of an advancement algorithms, However Here we develop the Machine Learning models to recognize the inappropriate content, initially we use text as input to the model. The text processed to detect the inappropriate information if it finds then it stops the post and warns to the sender for responsible misinformation.

Keywords: *Inappropriate information, vulgar words, Machine Learning, Text Analysis, words expansion, User Feedback*

I. INTRODUCTION

In modern days, the usage of the social media increased exponentially. The rise of mobile connectivity and technology impacts on the social media platforms that results widespread of the social media. Before Online media platforms bullying is happen by face to face, but after introduce of the Online platforms it happens by virtually. Any Random person can post in social media irrespective of relationship if he knows the ID or any public social media access. In social media any one can posts what they thought and there was no regulation and rights to stop them without posting. The reports are showing that the increase usage of social media platforms in corona pandemic time went to 80%. At that time the vast number of peoples are use the mobiles and advance technologies to get relax and work from home facilities, Due to this increase usage of mobiles challenges are also face in social media platforms. Most of the peoples are taking the social media platform to communicate with peoples in global wide that leads to posts inappropriate content because various peoples have various mindsets and intentionally posts the inappropriate information. The usage of the social media leads to various types of content posts in the social media platforms, in that posts most of the information is inappropriate content and that makes the user insecure.

Day by day the bad content posts are increasing rapidly and there is no central controller to detect and prevent to post the content. Due to rapid posts in the social media will leads to the negative impression on the online platforms and peoples get distressed. The social media has a big impact on the human mental health. Expose of the bad content on social media will it leads to the negative mind sets to the peoples. It impacts mainly on the teenage persons. The teenage peoples are very flexible to adopt new things, if at the time bad content surroundings that person it leads to bad impact on him and unable to change his mindsets if once his mindset changes into negative.

By taking all this into consideration we are going develop the AI based algorithms to detect all this type of inappropriate content and prevent the user from spreading of the inappropriate information. The AI system which acts as the mediator in the message travelling, It ensures to understand whether the content is inappropriate or not which is send by the user. Here it mainly uses the machine learning algorithms to detect the vulgar words or inappropriate content in the posts for that detection it uses the Natural Language Processing (NLP) models.

The remainder of this paper is as follows. Section 1 Discussion on the previous literature. Section 2 identified the challenges and problem identification by based in the previous existing content moderations. Section 3 Mention how my project overcame all mentioned drawbacks in the previous existing and what features we add in my project all are placed in the proposed systems. Section 4 Describes the Architecture of the AI system and methodologies. Section 5 It explains about the code execution and conclusion.

II. LITERATURE SURVEY

[1] Shobha Tyagi, Adarsh Pai, Jeson Pegado· Ajinkya Kamath. Published in the year 2019 in IEEE. "A Proposed Model for Preventing the spread of misinformation on Online Social Media using Machine Learning." This paper describes a system for real-time tweet content classification in the twitter. For development of this project its uses Tweepy, NLTK, and Naïve Bayes/Decision Tree algorithms. The AI model classifies the content based on providing datasets. It detects the content when the inappropriate tweets are occurred.

However, it has drawback that lies in the assumption of the statistical independence in the Naïve Bayes classifier, which may lead to low accuracy in scenarios where this assumption doesn't hold.

[2] Prof. Aarti Burghate, Poonam Bramhane, Pranjal Kuhikar, Pranjali Kanhekar, Ruchika Parate, Anjali Dongre. Published in the year 2023 in international journal of advanced research in science, communication and Technology. "Implementation of AI-Based Social Media for Vulgar Content Detector and Remover". This paper mainly focuses on the web application with an admin allowing administrators to control messages and activities. This paper uses the Python Django-based web application with a CNN algorithm to reduce vulgar content and negative comments. It has more complexity to develop and to train the model and it has more cost because it connected layers while in training.

[3] Eugene Yang, David D. Lewis and Ophir Frieder. Published in 2021 "TAR on Social Media: A Framework for Online Content Moderation". TAR is Text Analysis and Retrieval, It uses the logistic regression models to get more accuracy. It has a features random sampling, uncertainty sample and feedback. It uses the libact, it is an open-source framework for learning experiments and logistic regression models to train the data. The application design involves algorithms, and evaluation metrics, with a notable drawback being the computational cost in the development of a CNN algorithm for content moderation and it uses the oracle to recall the estimation which may not accurately in real-world.

[4] Mukul Anand, Dr.R. Eswari. published in 2019 "Classification of Abuse comments in social media using Deep Learning". It uses the deep learning algorithms to identify "trolls" and harmful information on social media platforms. For better prediction abilities, it's better to use machine learning approach for the identification of harmful pictures based on integrated text content. The drawback of this paper was there were no updating datasets and not dynamic updating.

[5] Prof. Giovanni Sartor, Prof. Giovanni Sartor and Dr. Andrea Loreggia. Published in 2020 "The impact of algorithms for online content filtering or moderation". The main objective of this project is to understand and address the challenges faced by the rapid growth of the data over internet. It focusses on the controlling of the content by providing the automated filter methods. The project uses the machine learning algorithms and it focus on the text, audio, images and video media. For development of this project, it uses the Machine learning algorithms and neural network, and deep learning. It is very complex and resource intensive, it needs to scalable computational resources and drawbacks include delays and persistence of harmful material.

[6] Asmita Poojari· Pallavi K. N, McEnroe Ryan Dsilva, Jagadevi N. Kalshetty. Published in 2023 in ijisae "A Novel Deep Learning Technique for Detection of Violent Content in Videos". The objective of this project is to use Deep learning models to detect the violent content in videos. It uses the manual detection it is time consuming and subjective. This classifying of video is the main motto of this project. It collects the data and audio processing and video processing and model evolution. The project developed by Python Django web app employs CNNs to identify vulgar content on social media. The deep learning uses the quality of the training data if the training data not have much quality, then it is difficult to train the machine, if it done then it confirms gives the biased data. It has the drawbacks in algorithmic complexity and evaluation metrics. It faces a computational cost drawback, especially in fully connected layers while training and needs to enhance computational to ensures practical.

[7] paolo sernan, Nicola Falcionell . Taylor & Francis in Applied Artificial Intelligence on the 6th of February 2020. "Violence Detection in Videos by Combining 3DConvolutional Neural Networks and Support Vector Machines". It is designed to check the effectiveness of pre-training conventional 3D network. But basically, it designed to check the actions in the sports videos for classification of violent videos. It extracts the violent content from the videos. This project uses the 3D conventional and 3D pooling for processing frames. It consists of the linear support vector Machine (SVM) classifiers. It has limitation that it only works for the sports videos, if the videos other than sports it doesn't give much accurate and data bias that nothing but it should train by the quality of data if not it gives unrelated assumptions.

[8] Barnes, Michael Randall. 2022. *Feminist Philosophy Quarterly* 8 (3/4). Article 6. "Online Extremism, AI, and (Human) Content Moderation". It describes context of the content moderation on social media platforms. It extracts the bad content which spreads the false news in social media platforms. It conducts different literature surveys to detect the violent extremism. It identifies the technical and ethical challenges in the content posted. The article says that how online platforms connect extremists, and the "individual route," focusing on recommendation algorithms fostering radicalization. However, limitations in attributing causality to online platforms and there no dynamic updation of datasets.

[9] TAÍS FERNANDA BLAUTH, OSKAR JOSEF GSTREIN, AND ANDREJ ZWITTER. Published in the year 2022 "Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI". The project ensures the

malicious use and abuse content by using AI system. It detects the malicious AI actions activities. Which are leads to the cyber enabled and cyber dependent and adopting a suppressive approach. The study identifies vulnerabilities, unintended outcomes, and concerns in algorithmic trading, the need for policy responses in the face of evolving technology. It had the limitations in limited scope of analysis and also it doesn't support the different languages and also it depends on the news articles reports.

[10] MAYUR GAIKWAD SWATI AHIRRAO, SHRADDHA PHANSALKAR. IEEE Access in 2021. "Online Extremism Detection: A Systematic Literature Review with Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools". The main objective of this project is to provide the different ways to detect and classifies the content based on the effectiveness and applicability. This uses the PRISMA instructions adapted from kitchen ham and charters. This study focuses on the works related to the inappropriate content detection. The limitation of the project is scope limitation, that it valid for few years. There also a risk in find the extraction process. The methodologies in extraction of the content was not much accurate.

III. PROBLEM IDENTIFICATION & CHALLENGES IDENTIFIED

Now a days the spread of the inappropriate information grows rapidly, especially real-world sensitive incidents, it can harm to the person. Most of the public actively participate in the social media platforms to generate inappropriate content that includes vulgar, harassment, and other against to the guidelines.

The reason for generation of the inappropriate content is absence of the central moderation in flow of content in social media platform. Manual moderation is time consuming and accuracy in detection also low. It is difficult to detect the inappropriate content through manual content moderation. There are some automated detections are also existing in some social media platforms like Facebook, Instagram, and Twitter etc.., but it is not dynamic, it is impossible to detect the new content and it doesn't update the datasets.

The following challenges were identified:

A. Time taken in Manual Moderation

In few online platforms like Twitter uses the Manual moderation that takes more time and also have an idea on the words that leads to mislead the words. The more persons should present in the moderation if the posts are rapidly increases and more complex in manual moderation. The person should expose to the offensive content daily due to this the persons can get metal health. The manual moderation should take more time to understand the context of the conversation.

B. Content Flagging

The Social media platforms were not providing the flagging to the user which seems to be vulgar or bad content. Due to this it's impossible to find the content which harms to the user. The user flags can help to train the AI model more accurately. It is impossible to find all the vulgar or inappropriate words directly those are generated continuously, if we provide the user flagging it helps to get more accurate.

C. Contextual Misinterpretation

Contextual misinterpretation in moderation cited that a situation where the meaning of the post is misunderstood due to lack of context of communication. It is possible to use the bad words or content in friends' communication but the AI system not focus on the content of communication it's leads the suppression of the content. Using the bad words in the friend's conversation common thing but it is identified as bad or vulgar word is not correct.

D. Multilingual Challenges

Multiple languages are existing in the world due to the cultures and contexts. Due to widespread of the social media the users are increases while increasing this user the content posts are also increases. But all are not able to understand global languages so they communicate in their region languages only but existing social media platforms are not focus on the region languages content moderation, they were uses only global language that is English but most of the of the vulgar content is generated in regional languages its leads inconvenience to the sensitive users, so it should focus on regional languages also.

IV. PROPOSED SYSTEMS

- **Advanced Moderation System:**
  Introduction of advanced moderation system that uses NLP-based machine learning models (state of the art) for the accurate identification of vulgar or inappropriate content in the communication and stops the message without sending to the opposite user to whom he wants to send.
- **User Feedback Integration:**
  User Feedback method is the better to enhance the content understanding. By the User Feedback methods we can train the machine learning more effectively and it most possible to detect content accurately. It can update to the datasets without repeating for next time.

- **Ambiguous Content Handling:**
  There are some words which gives leads bad content but it is difficult to detect whether the content is vulgar or not, For this we proposed the AI model which checks the deeper meanings of the content and if it is vulgar then it update to the datasets. These all are done by AI system there no intervention of the humans. The datasets dynamically updated.
- **Multilingual Moderation:**
  The proposed AI System not only detect one language content here we facilitate to detect the different regional language content detection.
  It helps to detect the content even it is in the user's native language. This leads to more advancement and regional peoples can also use. The AI system easily detect the content without depending on the single language.
- **Cross-Lingual Word and Phrase Mapping:**
  Basically, the users are uses the words which are cross lingual that means using different words from different languages. So, this system has a capability to detect languages dynamically and allowing for efficient word and mapping across the different languages contexts.

- **Real-time Updates and Continuous Learning:**
  The incorporation of user feedback, word expansion, and dataset updates contribute to a dynamic system that learns and evolves in real-time. Continuous learning ensures the adaptation of moderation models to evolving language patterns and user behavior.
  The integration of the User Feedback and Deeper expansion of the words and updating of the datasets contribute to dynamic system that learns real-time. Continuous learning of the machine learning leads to accurate and ensures user behavior.

ARCHITECTURE

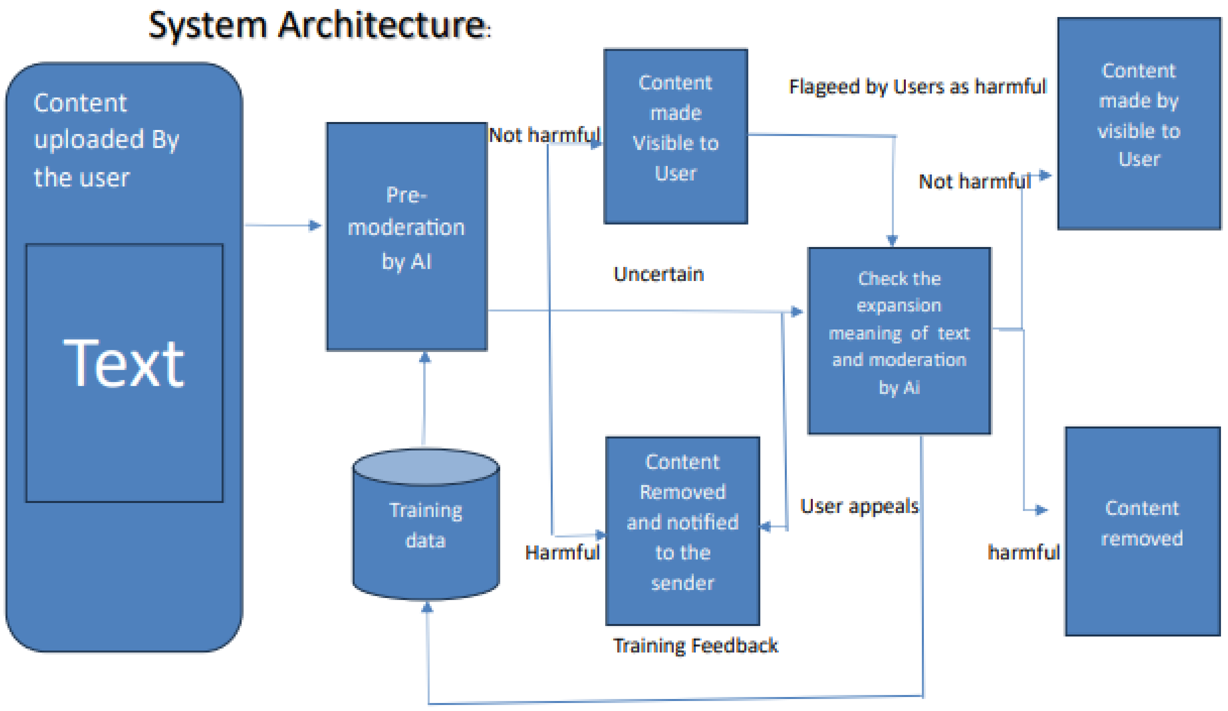

System Architecture
Content uploaded By the user
Text
Pre-moderation by AI
Training data
Not harmful
Content made Visible to User
Flageed by Users as harmful
Content made by visible to User
Not harmful
Uncertain
Check the expansion meaning of text and moderation by AI
Content Removed and notified to the sender
Harmful
User appeals
harmful
Content removed
Training Feedback

Figure:1 System Architecture

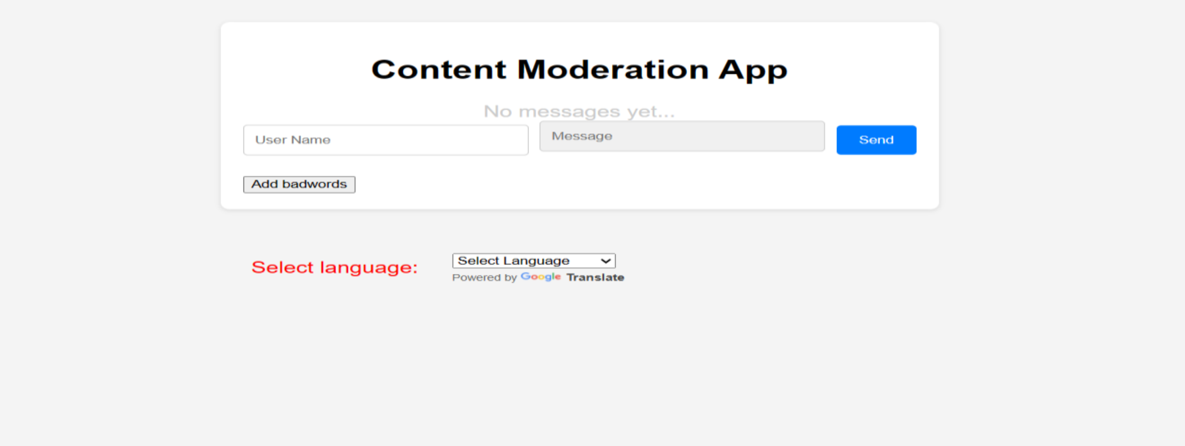
The message which sends by the sender is undergoes into the process to validate whether the message contains any vulgar or any inappropriate content exists or not. If the message contains any inappropriate content, then it removed and notified to the sender that sending bad words otherwise it goes to the next process that is expansion of the meaning.

After enter into the block check the expansion meaning text it will process and it checks whether the meaning of the sentence is related to the inappropriate content or not, if yes then it will automatically update to the dataset for training machine learning model otherwise it displays on the screen of the opposite user.

Here, we also facilitate the User Feedback to recommend the bad words or vulgar words if they find any while communication.

The User Feedback content will check by the admin and verify whether the user flagging data is vulgar or not. If the words are the vulgar then it sends to the dataset.

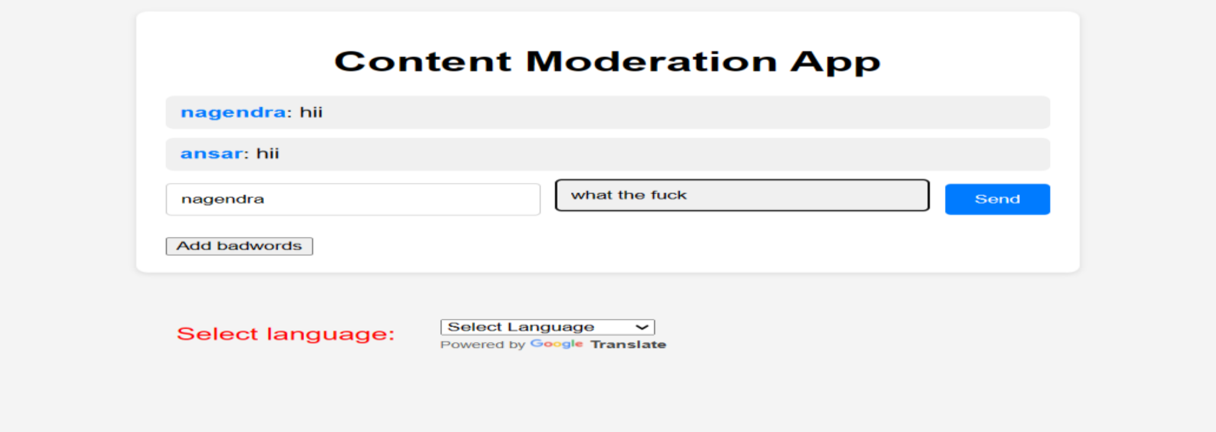
V. RESULTS


Content Moderation App

No messages yet...

Figure:2 Interface of the application

Figure 2 shows the interface of the application, It have fields user name, message block, submit button, and provided the link to the page where users can add the bad words to the datasets of dynamic updating and provided language selection button.


Content Moderation App

nagendra: hii

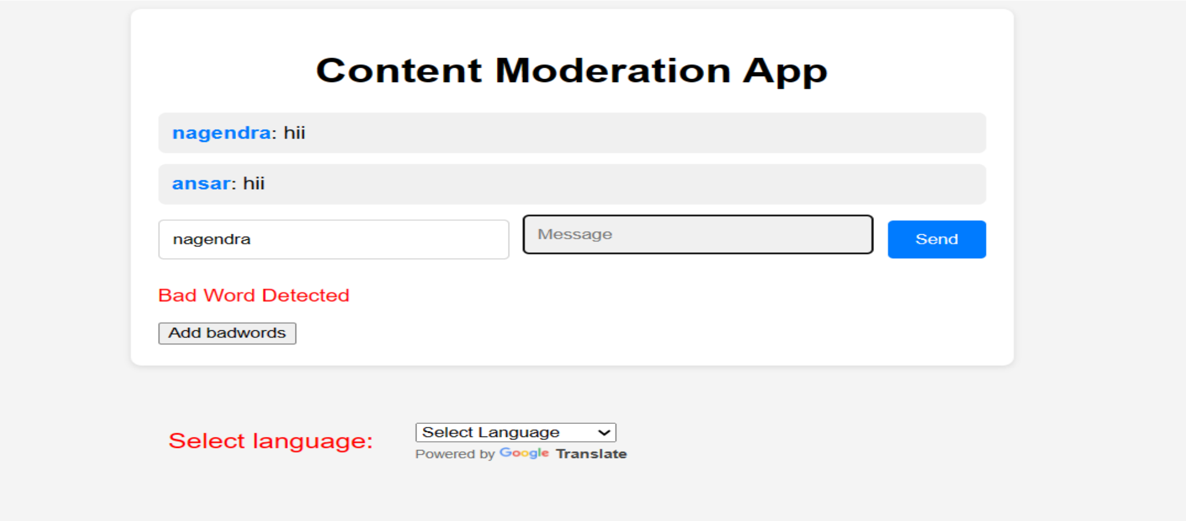
ansar: hii

nagendra

what the fuck

Figure:3 Input blocks

In figure 3 we can see that the messages are displaying on the screen just above the input blocks after click submit button. After click the submit button the message went to processing and check whether the content is appropriate or not. If the message is not vulgar or inappropriate then it displays on the screen otherwise it stops to send the message and notify the sender with prompt “Bad word detected” refer below figure number 4.



Content Moderation App

nagendra: hii

ansar: hii

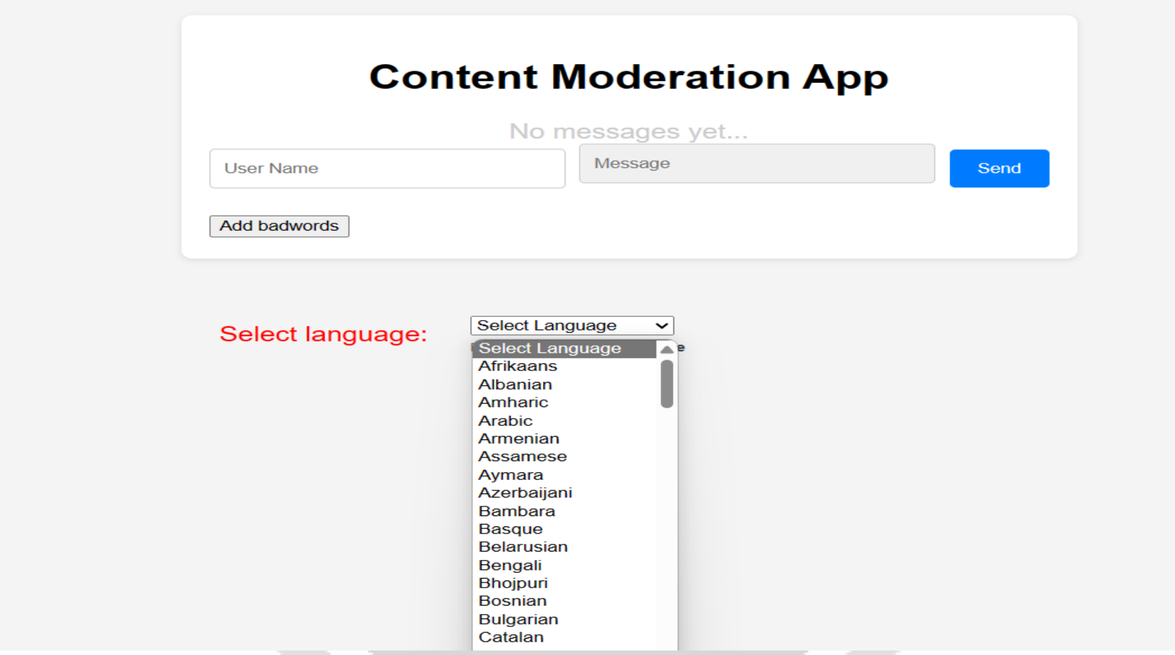
nagendra

Bad Word Detected

Select language: Select Language

Figure:4 Bad Word Detected

In figure 4 discussed about the sender can send text data to the receiver can reaches the message we have not select the language so it shows the bad word detected.



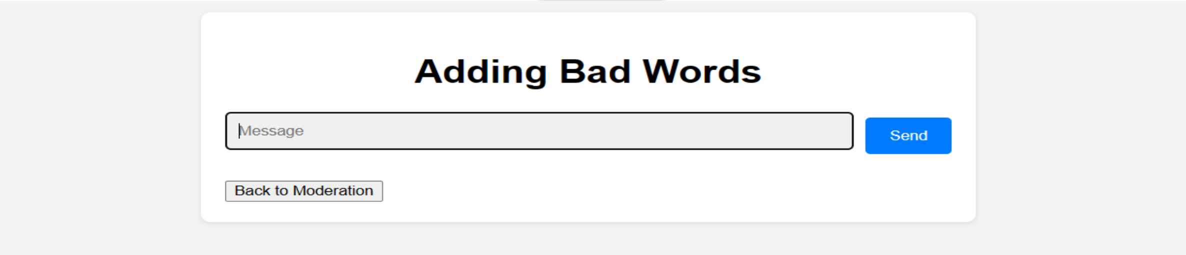
Content Moderation App

No messages yet...

Select language: Select Language

Figure:5 Language selection

There is an option to select the language in our application, we provided the drag list to select the language what users can prefer.



Adding Bad Words

Figure:6 Adding Bad words

The application provided a option to enter the bad words into the datasets. If the user find any bad word in the post then he can add by entering in the above figure 6 input field and also provide navigate button to main page.

VI. CONCLUSION

The impact of the Online platforms on individuals and society can't be ignored. There are some content moderation platforms are existing even though those not much accurate in detection of the vulgar content and take action or stop the content, to resolve we came with new idea that dynamic updating of datasets, words expansion, User Feedback, and multilingual languages. My AI model detects the inappropriate message or vulgar message when the sender sends if it is vulgar then it stops the sending and notified to the sender with prompt you sending bad words detected. If the message not detected in the preprocessing, then it goes to next processing expansion meaning block where we implement an advance code for ensures whether the meaning of the word is related to the vulgar or not, if it vulgar then it automatically update to the datasets otherwise displays to the user. It also facilitates the User Feedback to complain the messages which harms them. Those user feedback messages are undergone for validation to the Admin, the admin will verify the user feedback message if it vulgar then he sends to the datasets.

REFERENCES

[1] Shobha Tyagi , Adarsh Pai , Jeson Pegado , Ajinkya Kamath Published in 2019 "A Proposed Model for Preventing the spread of misinformation on Online Social Media using Machine Learning", published in the year 2019 in IEEE.
[2] Prof. Aarti Burghate, Poonam Bramhane, Pranjal Kuhikar, Pranjali Kanhekar, Ruchika Parate, Anjali Dongre in year 2023 "Implementation of AI-Based Social Media for Vulgar Content Detector and Remover".
[3] Eugene Yang, David D. Lewis and Ophir Frieder in the year 2021 "TAR on Social Media: A Framework for Online Content Moderation".
[4] Mukul Anand , Dr.R. Eswari in 2019 "Classification of Abuse comments in social media using Deep Learning".
[5] Prof. Giovanni Sartor, Prof. Giovanni Sartor and Dr. Andrea Loreggia published in year 2020"The impact of algorithms for online content filtering or moderation".
[6] Asmita Poojari , Pallavi K. N, McEnroe Ryan Dsilva, Jagadevi N. Kalshetty published in year 2023"A Novel Deep Learning Technique for Detection of Violent Content in Videos".
[7] paolo sernan, Nicola Falcionell . Taylor & Francis in Applied Artificial Intelligence on the 6th of February 2020. "Violence Detection in Videos by Combining 3DConvolutional Neural Networks and Support Vector Machines".
[8] Barnes, Michael Randall in year 2022. *Feminist Philosophy Quarterly* 8 (3/4). Article 6. "Online Extremism, AI, and (Human) Content Moderation".
[9] TAÍS FERNANDA BLAUTH, OSKAR JOSEF GSTREIN, AND ANDREJ ZWITTER in 2022 "Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI".
[10] MAYUR GAIKWAD SWATI AHIRRAO, SHRADDHA PHANSALKAR .IEEE Access in year 2021"Online Extremism Detection: A Systematic Literature Review With Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools".