International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

AI Powered Content Moderation System

Mr. Kyatham Sai Charan, M.Tech
Assistant Professor
Department of Information Technology,JBIET Autonomous ,Hyderabad, Telangana,India
saicharan155@gmail.com

P.Samaya
Student,
Department of Information Technology,JBIET Autonomous ,Hyderabad, Telangana,India
poturisamaya1811@gmail.com

M.Tanay
Student,
Department of Information Technology,JBIET Autonomous ,Hyderabad, Telangana,India
tanaoymongolla@gmail.com

K.Mahalaxmi
Student,
Department of Information Technology,JBIET Autonomous ,Hyderabad, Telangana,India
kummarikuntlamahalaxmi921@gmail.com

B.Amol
Student,
Department of Information Technology,JBIET Autonomous ,Hyderabad, Telangana,India
amolbassi9515@gmail.com

***Abstract*— With the rapid increase in user-generated content on social media and other online platforms, the need to have intelligent and scalable content moderation systems is even more crucial. The automated moderation systems based on AI should not only be utilized to locate poisonous, dangerous, and inappropriate multimodal content but also ensure that the process is transparent and credible. We outline a content moderation model, based on AI analysis of images, text, and audio. Picture moderation is done using deep transfer learning architectures such as VGG16, Xception, and ResNet50. The system is also made more resilient by using a hybrid ensemble model that consists of Xception and ResNet50. Image preparation includes changing the size, normalizing, converting to NumPy, and utilizing pretrained convolutional neural networks to get features. Transformer-based language models BERT and RoBERTa are moderated to use with the right preprocessing, label encoding, and embedding techniques to moderate text. The process involves transcribing audio with openAI whisper, followed by classifying the text with trained transformer models to moderate audio. Explainable AI methods like LIME and SHAP make guarantee that models can be understood. In the case of real-time moderation services, the system is configured using Flask framework and SQLite-based authentication. Experimental evaluation indicates that BERT is more effective than the Hybrid Ensemble model when it comes to text classification (99.20% accuracy) and picture classification (93.10% accuracy). This implies that both the models are very dependable and can be applied in practical scenarios.**

"Keywords— Artificial Intelligence, Content Moderation, BERT, RoBERTa, Transfer Learning, Hybrid Ensemble, Explainable AI, Flask Deployment."

I. INTRODUCTION

The emergence of the digital space and social media has precipitated an unprecedented increase in user-created content, including text, images, and audio. This enormous stream of material has numerous opportunities of communication, collaboration, and knowledge sharing, but there are also enormous issues with ensuring that web spaces are safe and healthy. Harmful, abusive, or poisonous information may spread quickly, affecting users' mental health, breaking up online groups, and perhaps breaking the law and moral rules [1]. Sites are increasingly struggling to monitor much multimedia simultaneously. Manual moderation is time-consuming, not always accurate and may produce errors [2].

AI can provide a scalable mechanism of filtering this kind of information as it makes it possible to analyze and filter large amounts of information with high accuracy and do it automatically. ML and DL algorithms are capable of locating the inappropriate content, identifying potential threats and applying the rules of the platform effectively [3]. The issues with current content can be addressed by multimodal moderation, which considers text, graphics, and audio simultaneously, as these issues can manifest themselves in a vast number of different ways simultaneously [4]. These AI methods can also be used by platforms to rapidly recognize or remove negative content, reducing the chances of exposing vulnerable individuals and rendering the internet a safer space [5].

Moderation besides the automatic analysis also allows moderators to intervene instantly to prevent the rapid dissemination of inflammatory or fake news. The users can easily interact with the moderation system, submit content to be reviewed, and receive comments with clear insights due to integration with web-based frameworks [6]. Two examples of explainable AI strategies are LIME and SHAP, which further open and trustworthy the process of highlighting content by providing a clear explanation of why it was highlighted [7]. This approach ensures moderation is not just right, but also responsible, which fosters the trust of users and promotes responsible online dialogue [8].

This system aims to offer an AI-based system capable of swiftly and precisely assessing and filtering text, pictures, and audio content posted by users. The objectives of the system are to deliver real-time moderation, simplify things by offering interpretable insights and to make the internet a safer place overall by restricting access to dangerous or

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

inappropriate information [9]. The system aims to ensure the integrity of digital platforms and user trust and engagement through scalable automation and interfaces that are easy to use [10].

II. Related Work

Due to the rapid proliferation of social media and online platforms, there has been a need to devise effective mechanisms of curtailing content to cope with the increasing number of abusive, offensive, or destructive content. Parker discusses the tension between privacy and freedom in AI-based content filtering, highlighting the challenges of balancing user rights and the protection of the platform, and emphasizing the ethical consequences of automated interventions [11]. Malec and Lešetický consider how social media sites filter content, such as filtering techniques and how people might attempt to bypass AI detection. This illustrates the fact that content moderation is evolving at all times [12]. Wiesner, Schafer, and Lecheler discuss the perspectives of professional moderators, highlighting the difficulties in decoding uncivil comments and the role of AI solutions in supporting the human decision-making process, indicating that moderation is not only a technical issue but a human-related problem [13].

Divya, Samprakash, Yazhini, Kesavan, Saravanakumar, and Lakshmi present a ML and DL-based system that detects offensive content in multimedia data and demonstrate how AI-based automated moderation can be improved by the models, particularly the classification of contextually inappropriate content [14]. Jha, Pandey, Srivastava, Rajput, Pant, and Mukherjee discuss how AI could be integrated into the social media applications to enhance the user experience and make it more personal and, at the same time, monitor the content. This demonstrates that intelligent systems are able to enhance safety and engagement [15]. Parmar and Murari discuss the ways in which AI moderation systems must evolve with the changing trends online. They emphasize the fact that constant learning is required to address new kinds of hazardous material and that dynamic models, which evolve depending on the behavior of the user, are needed [16].

Fahrudin, Tiwari, and Rahmi discuss the psychological impact on human moderators within AI-based mechanisms, emphasizing the significance of wellness strategies to reduce stress and burnout, depicting the relationship between human and AI factors in moderation processes [17]. Pardhi discusses some of the challenges of generative AI regulation, and how automated systems have to cope with more complex material generated by AI models, which poses new threats to the security of the platform [18]. Nitheeshwari and Malar propose a hybrid AI-based content moderation structure, where various learning techniques are employed to enhance its capability to identify issues in text, images, and audio content. This indicates the effectiveness of multimodal solutions [19]. Proebsting, Anigboro, Crawford, Metaxa and Friedler discuss identity-related speech suppression in generative AI moderation, exploring biases and fairness issues that arise when moderating sensitive topics, and highlighting the ethical responsibility of automated systems [20].

Overall, these works demonstrate how moderation of the content has evolved in the past to be transformed by human and rule-based systems to intelligent, AI-driven mechanisms that can process multiple forms of information. They emphasize the importance of being capable of comprehending, believing, and adhering to ethical regulations, and the human control is required. They also emphasize the necessity of robust, mobile and explicit solutions to ensure that digital communication is secure and accountable. Such studies are the basis of scalable frameworks that are real time and explainable AI moderation of text, graphics, and audio. They also pave the way to future advancements of automated content regulation.

III. Materials And Methods

The system relies on AI to develop a multimodal content moderation, which examines text, pictures, and audio and identifies threatening, abusive, or inappropriate material. Deep transfer learning models such as VGG16, Xception, and ResNet50 are employed to extract hierarchical features which exhibit both local and global patterns to perform image moderation. Xception + ResNet50 gives better classification accuracy and strength [21]. Transformer-based models such as BERT and RoBERTa are used in text moderation to determine the relationship between words in meaning and context. This makes it easier to find harmful or poisonous language [22]. To start with, audio is transcribed by speech-to-text algorithms. Then, trained text models are used to find abusive speech, making sure that all types of speech are covered [23]. The system is configured with the help of a web interface developed on the basis of Flask that does not complicate the registration and the entry of the system (safe) and the delivery of material to the moderators (safe). The interface performs pre-processing and passes inputs to the appropriate models to be classified [24]. Models can be made more transparent and understandable with explainable AI methods such as LIME and SHAP. They provide specific details on judgment of models [25]. This hybrid approach ensures moderation is performed in real time, can scale with the platform and is equitable, which safeguards the users and the platform.

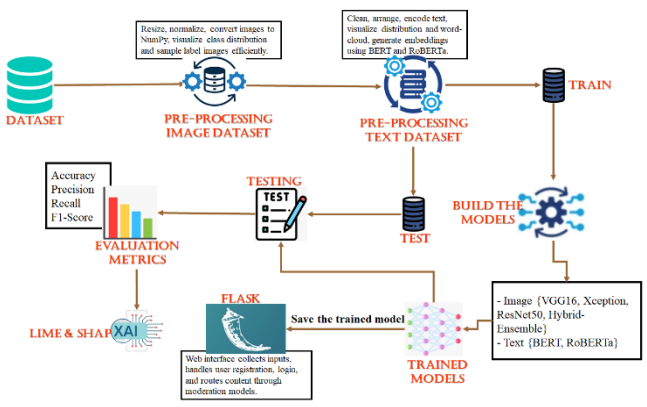

Fig.1 System Architecture

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

The method shows a complete ML pipeline for datasets of images and text. VGG16, Xception, ResNet50, and BERT models have been trained and tested after undergoing certain pre-processing and are evaluated based on their metrics such as F1-score and accuracy. The Hybrid-Ensemble has the best performance, with an accuracy of 0.931, as seen in Fig. 1. Xception has positive results as well.

A) Dataset Collection:

The system is used to facilitate the multimodal content moderation process that involves textual and visual content with the help of two different datasets. Picture moderation is done on the SIMAS dataset. It contains numerous images which are tagged differently and display various types of content which may not be appropriate. To ensure that the models are trained to differentiate between safe and hazardous visual images, this dataset was carefully collected to encompass a diverse array of situations. To collect and organize the photos, you will have to check the labels, remove duplicates and place them in training and testing folders which apply machine learning methods. The pictures are categorized to enhance the collection even more. This assists in making sure that there is a balance in training and that the efficacy of the model is sufficiently tested.

The Kaggled toxic-comments dataset is used to filter out text. It includes the user comments that have been marked as offensive, abusive, or poisonous. Numerous other social media site samples that show how people actually communicate are included in the data. Text preprocessing eliminates noise, homogenizes the content and converts category labels into numbers which can be utilized to train models. Since the DL models are trained in both datasets, the system has the capability to manage material in any media effectively and properly to ensure that dangerous information is never forgotten in real-time.

B) Pre-Processing:

Before you can use raw data in ML and DL models, you need to do some pre-processing. It facilitates in making sure that the inputs are clean, uniform and suitable to train models and obtain features. Pre-processing improves the accuracy, decreases the noise and enhances the uniformity of pictures and text in their formats. This enables models to learn helpful patterns in a short time, retaining their accuracy over a broad array of datasets.

Image Pre-processing: Pre-processing of the images is of importance before DL models can utilize the visual input. At first, the dataset is shown using bar graphs to see how the class labels are spread out. This makes sure that the model training is balanced. We compare sample photographs to discover variations in quality, light and content. This assists us in making decisions of what to do next in the preprocessing stages. Then, every image is resized to a standard size that is compatible with the selected models, like VGG16, Xception, and ResNet50, to ensure that all works together in the network. The pixel values are converted to floating-point values and rescaled to the range of 0-1. This assists the model train to accelerate faster and makes the calculations not to be unstable. NumPy arrays can also be converted to images, allowing it to be simple to perform math and interoperate with DL models. Additionally, the features that reflect hierarchical representations of visual patterns are extracted using pre-trained models of transfer learning. These features encode both local information and global patterns, making it easier to differentiate between safe and unsafe information by the models. This rigorous pre-processing pipeline ensures that the dataset is clean, standardized and packed with useful data, which results in high and viable image moderation results.

Text Pre-processing: Text pre-processing is highly significant to prepare unstructured text data to DL and other transformer-based algorithms such as BERT and RoBERTa. The initial measure is to visualize data to examine the distributions of classes, comment size and common language patterns. Word clouds can also help you locate words that appear frequently, which could possibly assist you in determining how to eliminate noise. Then, the raw text is stripped of special letters, unnecessary spaces and URLs since they are not relevant to semantic understanding. Splitting sentences into words or subwords that are good to embed models is called tokenization. Label encoding converts categorical target labels into numbers and categorical conversion ensures that the model can be trained on them. Two normalization techniques are lowercasing and lemmatization that reduce repetition and make the text representation more uniform, thus rendering models more generic. The methods of embedding, based on BERT and RoBERTa, identify semantic and contextual links. They encode raw tokens into dense vectors which encode local and global language information. This process aids models to interpret nuances such as irony, slang, and phrases consisting of more than one word. Then, the model is trained using pre-processed embeddings, which make sure that the text input is clean, consistent, and full of meaning. Combining these steps, the system is capable of moderating text efficiently, accurately and on a massive scale, by detecting poisonous, abusive, or inappropriate content in a large selection of user-generated text.

C) Train & Test:

The post-processing dataset is further divided into training and testing to test the ability of the system to generalize with new data. The ratio of 80:20 is mostly applied, with 80 per cent of the data to be utilized in training the models and 20 per cent to be utilized in testing. The training set is used to educate the models to locate patterns and relationships in the input data that may be pictures, words or sounds. This ensures that the models are familiar with the characteristics of various classes such as safe, dangerous, or unsuitable information, and become more proficient at making accurate predictions.

The testing set, contrarily, is not utilized in training at all. It provides an objective evaluation of the quality of the

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

system work, and it is more convenient to detect that there is overfitting, underfitting, or generalization. The train-test split ensures that the two subsets of picture data contain instances of the various classes and image conditions, e.g., illumination, orientation, and scale. The divide also ensures that the heterogeneity in the language, style and context are maintained in both text and audio instruction and evaluation.

Before dividing the data, it is important to shuffle it to avoid any sequential bias and keep it random. This makes sure that the subsets are a good representation of the whole dataset. This chasm allows one to assess things in a just and systematic manner, which assists in refining the models to difficult, real world content moderation tasks.

D) Algorithms:

VGG16: VGG16 is a kind of deep convolutional neural network, which consists of 16 layers, most layers are convolutional, and fully connected. The network employs small 3x3 convolution filters throughout and max-pooling layers to reduce the size of the spatial dimensions gradually retaining important information. VGG16 is extensively used by people when it comes to activities such as recognizing and classifying images since it is able to effectively extract hierarchical visual information. It transforms raw pixel data into valuable feature maps, which enable models to distinguish between classes in a correct manner. VGG16 trains on the input photos to moderate their contents by learning patterns, textures, and structures in the image. This lets it filter out improper or harmful visual content automatically with a high level of accuracy.

Xception: Xception is a deep convolutional neural network with depthwise separable convolutions to space and channel segregated operations. The design simplifies the calculations and yet achieves high quality feature extraction. It is also effective in memorizing intricate and complex picture patterns thus suitable in occupations that require the classification of high-resolution images. Xception scans images uploaded in content moderation systems to identify obscene, violent or otherwise inappropriate images. It is able to make proper predictions about moderation decisions to consider both local and local factors. Its profound design allows it to acquire new features that enhance the system in distinguishing between safe and dangerous pictures and yet remain efficient enough to work in real time.

ResNet50: ResNet50 is a 50-layer convolutional neural network that employs residual connections to correct the issue of the gradients vanishing. These skip connections enable the directness of the gradients across layers and this aspect facilitates very deep networks to learn well. ResNet50 gets hierarchical characteristics from photos, such as edges, textures, and more abstract semantic information. It evaluates photographs that users send in to put them into safe or improper categories as part of content moderation. Its residual form allows it to remain accurate when dealing with large, complex data. The network is able to moderately cope with moderation due to its ability to generalize with various forms of visual material. This renders it a reliable component of automated image analysis systems which are applied in real-time applications.

Hybrid Ensemble (Xception + ResNet50): The hybrid ensemble takes the predictions of both Xception and ResNet50 to obtain optimal results of both networks. Ensemble learning pools together the results of multiple models to make them broader, less changeable, and more robust. The ensemble consults both networks to assess photos during content moderation and finally synthesizes their forecasts to come up with a final decision. This method makes it easier to find hazardous or unsuitable visual information. The hybrid model is effective as Xception is able to detect features in a fast manner and ResNet50 is able to learn profoundly on the residuals, thus it becomes feasible to find complex patterns in images. This makes individuals become more persuasive regarding the outcome of moderation and reduce the possibility of errors, thus aiding in the dependable real-time content filtering of pictures.

BERT: BERT is a language model based on transformers that considers both the meaning of the words preceding and succeeding a particular word in a sentence. It employs attention processes to figure out how words in a text are related to each other and how they depend on each other over vast distances. BERT converts text inputs into dense representations of vectors that represent meaning on a highly contextual level. In the context of content moderation, BERT examines text posted by users to detect toxic or abusive or inappropriate content. The fact that it is bidirectional means that it is able to capture small contextual details that would be missed by rule-based systems. BERT can ensure that communication is safe on online platforms, providing good embeddings and contextual sensitivity, making it simpler to accurately categorize content into moderation labels.

RoBERTa: RoBERTa is a better version of BERT that uses bigger datasets, dynamic masking, and longer training times to enhance pretraining. It eliminates the following sentence prediction task and is more effective in modelling complex language patterns. RoBERTa develops high-quality contextual embeddings that can pick up on subtle meanings, irony, and indirect toxicity in text. RoBERTa classifies user inputs into three (safe, offensive, and harmful). This is the way it operates in case of content moderation. It is able to learn more, therefore it can deal with a broad variety of words, such as the casual speech, slang, and lengthy, poisonous sentences. The deep sensitivity to context allows RoBERTa to detect the wrong data with a large degree of accuracy, which contributes to the security of online spaces on automatic moderation systems.

IV. Experimental Results

Accuracy: The accuracy of the test is its ability to distinguish sick and healthy patients. So to determine the accuracy of a test we have to determine the number of true positives and

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

true negatives in all the instances we considered. In mathematics, it can be expressed as:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} (1)$$

Precision: Precision is a measure of the number of the samples or instances called positives which were correct. The equation to calculate the accuracy is, therefore,:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} (2)$$

Recall: In machine learning, recall is a measure of how well a model can find all the instances of a certain class that are important. It is calculated as the proportion of the number of those that are actually predicted positively to the total number of actual positives. This tells you how well a model is able to capture examples of a certain class.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} (3)$$

F1-Score: To measure the accuracy of a ML model, the F1 score can be used. It combines accuracy and recall scores of the model. Accuracy statistic is the number of times a model made a valid prediction on the entire dataset.

$$F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision} * 100 (1)$$

Table.1 Performance Evaluation Table – Image Dataset

Model	Accuracy	Precision	Recall	F1_score
VGG16	0.824	0.840	0.823	0.828
Xception	0.926	0.931	0.925	0.927
ResNet50	0.917	0.914	0.918	0.914
Hybrid-Ensemble	**0.931**	**0.929**	**0.934**	**0.931**

The successfulness of the picture models is shown in Table 1, and VGG16, Xception, ResNet50, and Hybrid-Ensemble have high accuracy, precision, recall, and F1-score.

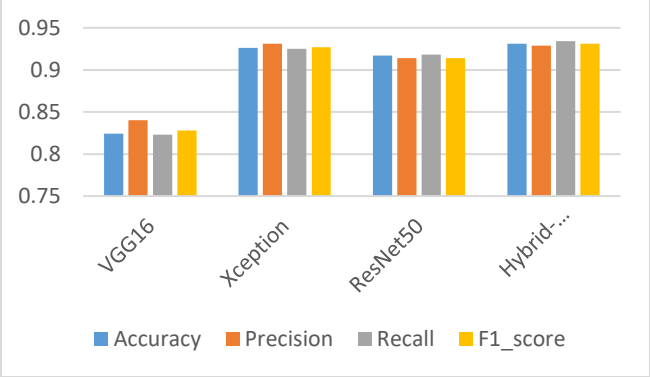

Fig.2 Comparison Graph – Image Dataset

The Comparison Graph of Image Dataset in Figure2 indicates that performance metrics and the Hybrid-Ensemble is the most accurate.

Table.2 Performance Evaluation – Text Dataset

Model	Accuracy	Precision	Recall	F1 Score
BERT	**0.9920**	**0.9921**	**0.9919**	**0.9920**
RoBERTa	0.9867	0.9870	0.9866	0.9867

The BERT and RoBERTa text models performed well, as indicated in Table 2, in terms of their accuracy, precision, recall, and F1-score.

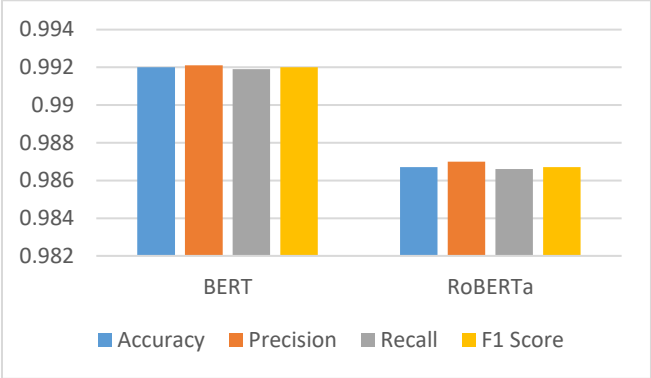

Fig.3 Comparison Graph – Text Dataset

Fig. 3 graph of text dataset comparison indicates that BERT is superior to RoBERTa as both models have a score of above 0.98.

V. Conclusion

We have designed and developed a smart multimodal content moderation platform that can effectively identify harmful and toxic text, visual, and audio content. The integration of deep transfer learning models in analyzing the images with transformer-based language models in interpreting the text makes it possible to extract robust semantic and visual features. The highest picture classification accuracy of 93.10 indicated the Hybrid-Ensemble (Xception + ResNet50) model has the potential to be a useful one in detecting improper visual content. As it has the highest accuracy rating of 99.20, meaning that it is relatively good at detecting toxic and abusive words, BERT is the most suitable model to control content. The trained model of BERT will process the audio input and enable the speech-based content to be controlled, which will make OpenAI Whisper a reliable audio-to-text translation service. Explainable AI such as SHAP and LIME can also be helpful in clarification by providing a comprehensible insight into predictions. This instills trust and accountability. The Flask framework with SQLite based authentication to support deployment enables safe interaction with users, real time prediction and online integration which can be expanded as required. The ultimate outcome is that the AI-based moderation system is effective, user-friendly and can be applied in practice to filter information in various types of data.

The next round of enhancements could be more multimodal moderation capabilities by exploiting spatiotemporal DL networks such as 3D CNNs and Vision

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Transformers to process video data on-the-fly. It would also be possible to moderate in more regional languages and therefore useful all around the world since it would add multilingual transformer models. Also, you can add constant learning mechanisms to assist them to learn new terminologies, hazardous patterns, and enemy attacks. Federated learning methods can be explored to be trained on decentralized data and contribute to the protection of user privacy. Quantization and lightweight model compression can be used to assist with edge deployment as they can be used to enhance performance. In addition, complex Explainable AI systems that operate interactive visual dashboards may enable automated decision-making systems on large online platforms to be easier to comprehend, less rule-breaking or less prone to uncertainties.

REFERENCES

[1] Ananthajothi, K., Meenakshi, R., & Monica, S. (2024, May). Promoting Positive Discourse: Advancing AI-Powered Content Moderation with Explainability and User Rephrasing. In 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-6). IEEE.

[2] Ami, R. (2021). AI in automated content moderation on social media. International Journal of Artificial Intelligence and Machine Learning, 4(3).

[3] Gowda, C., Charishma, G., Darshan, B., & Yadav, L. P. D. (2025, June). AI Powered Multimodal Content Moderation for Online Safety in Social Media Platforms. In 2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1748-1758). IEEE.

[4] Sun, H., & Ni, W. (2022). Design and Application of an AI-Based Text Content Moderation System. Scientific Programming, 2022(1), 2576535.

[5] Vidhya, K., Abhishek, V. S., Akash, G., Karti, S. A., & Harshini, R. (2023, December). AI-Powered content moderation. In *AIP Conference Proceedings* (Vol. 2914, No. 1, p. 050027). AIP Publishing LLC.

[6] Banchio, P. R. (2024). Legal, ethical and practical challenges of AI-driven content moderation. Ethical and Practical Challenges of AI-Driven Content Moderation (September 26, 2024).

[7] Raza, H. (2024). AI-Driven Content Moderation: Assessing the Privacy Implications and Safeguarding Free Speech. Research Gate

[8] Vamsikeshwaran, M. (2024, December). AI Powered Video Content Moderation Governed by Intensity Based Custom Rules with Remedial Pipelines. In International Conference on Computer Vision and Image Processing (pp. 390-403). Cham: Springer Nature Switzerland.

[9] Mengade, S., Chopade, P., Tate, P., & Patil, S. (2025). Facilitating Anonymous Communication on Social Networks via AI-Driven Content Moderation. Journal of the ACS Advances in Computer Science, 16(1).

[10] Bortnyk, K., Bahniuk, N., Kondius, I., Melnyk, K., Melnychuk, Y., & Kondius, K. (2024, October). Effective Content Moderation Using Modern AI Tools. In 2024 14th International Conference on Dependable Systems, Services and Technologies (DESSERT) (pp. 1-8). IEEE.

[11] Parker, O. (2024). Navigating the privacy-freedom dilemma: The impact of ai on content moderation and free speech. Journal of Digital Ethics and Policy, 12(3), 201-219.

[12] Malec, L., & Lešetický, J. (2024). Social media content moderation, censorship and ai detection evasion techniques. IDIMT-2024: Changes to ICT, Management, and Business Processes through AI.

[13] Wiesner, A., Schäfer, S., & Lecheler, S. (2025). Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools. new media & society, 27(3), 1215-1234.

[14] Divya, P., Samprakash, G., Yazhini, B., Kesavan, R., Saravanakumar, R., & Lakshmi, S. J. (2025, July). AI-based Content Moderation System for Offensive Data Detection. In 2025 8th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1803-1809). IEEE.

[15] Jha, K., Pandey, S. K., Srivastava, A., Rajput, A., Pant, S., & Mukherjee, P. (2025, October). Design and development of a social media application integrated with artificial intelligence capabilities: Enhancing user experience, content moderation, and personalization. In AIP Conference Proceedings (Vol. 3297, No. 1, p. 050002). AIP Publishing LLC.

[16] Parmar, H., & Murari, U. K. (2025). Adapting AI Moderation: Navigating Emerging Online Trends and Mitigating Evolving Harmful Content. In Content Moderation in the Age of AI (pp. 29-54). IGI Global Scientific Publishing.

[17] Fahrudin, A., Tiwari, S. P., & Rahmi, K. H. (2025). Psychological Well-Being of Human Content Moderators and Wellness Strategies in AI-Driven Content Moderation Organizations. In Content Moderation in the Age of AI (pp. 221-250). IGI Global Scientific Publishing.

[18] Pardhi, P. (2025). Content moderation of generative AI prompts. SN Computer Science, 6(4), 329.

[19] Nitheeshwari, S., & Malar, T. (2025, December). Hybrid AI-Driven Content Moderation. In 2025 IEEE 9th International Conference on Information and Communication Technology (CICT) (pp. 1-5). IEEE.

[20] Proebsting, G., Anigboro, O. I., Crawford, C. M., Metaxa, D., & Friedler, S. A. (2025, November). Identity-related speech suppression in generative AI content moderation. In Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (pp. 185-217).

[21] Wang, S. (2023). Factors related to user perceptions of artificial intelligence (AI)-based content moderation on social media. Computers in Human Behavior, 149, 107971.

[22] Parmar, H., & Murari, U. K. (2025). Human-AI Synergy in Ethical Content Moderation: Navigating Fairness, Accountability, and Transparency Challenges. In Ethical AI Solutions for Addressing Social Media Influence and Hate Speech (pp. 191-212). IGI Global Scientific Publishing.

[23] Rajput, R. S., Shah, S., & Neema, S. (2023). Content moderation framework for the LLM-based recommendation systems. Journal of Computer Engineering and Technology (IJCET), 14(3), 104-17.

[24] Khare, P., & Raghuwanshi, V. (2025). Legal frameworks surrounding the use of AI in online content moderation. In Ethical AI Solutions for Addressing Social Media Influence and Hate Speech (pp. 235-260). IGI Global Scientific Publishing.

[25] Oh, D., & Downey, J. (2025). Does algorithmic content moderation promote democratic discourse? Radical democratic critique of toxic language AI. Information, Communication & Society, 28(7), 1157-1176.