INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

CONTEXT-AWARE CONTENT MODERATION USING TRANSFORMER MODELS FOR DETECTING HARMFUL DIGITAL CONTENT

Neha Arya
Assistant Professor, Department of Electronics and Instrumentation Engineering
Shri G. S. Institute of Technology and Science, Indore (M.P.), India
nehaarya1988@gmail.com

DOI: https://doi.org/10.29121/ijrsmp.v12.i4.2025.01

Abstract
The rapid expansion of user-generated content on digital platforms has significantly increased the prevalence of harmful, inappropriate, and offensive material. Traditional content moderation methods, such as rule-based or keyword-based filtering, are increasingly ineffective in identifying subtle forms of harmful content, such as sarcasm, indirect hate speech, or cyberbullying. This paper proposes a novel context-aware model that leverages transformer-based models, such as BERT, RoBERTa, and others, integrated with deep learning techniques to enhance the detection of harmful online content. By focusing on contextual embeddings, our model effectively distinguishes between offensive and non-offensive material, improving the accuracy, precision, and recall of content moderation systems. The experimental results demonstrate that the proposed model outperforms traditional systems and other state-of-the-art models, offering significant improvements in real-time content moderation for large-scale platforms. This research provides a pathway toward more intelligent, fair, and efficient content moderation systems, ensuring safer digital spaces without compromising freedom of expression.

Keywords: BERT, Content Moderation, Deep Learning, Natural Language Processing, RoBERTa, Transformer Models.

Introduction
The rapid rise of user-generated content on digital platforms has brought about a multitude of benefits, including the democratization of information and social interaction. However, this surge in content creation has also led to an increase in inappropriate, offensive, and harmful content, which poses significant challenges for online safety and community integrity. This includes various forms of sensitive content such as hate speech, cyberbullying, harassment, explicit material, and misinformation, which can have harmful effects on individuals and communities. Therefore, effective content moderation is critical in maintaining the safety of digital spaces, ensuring that they remain conducive to healthy interactions.

Traditional content moderation techniques have typically relied on rule-based or keyword-based approaches. These methods search for specific words or phrases and flag content accordingly. However, these techniques are far from perfect and often struggle to identify context-dependent forms of harmful content. For instance, subtle forms of harassment, indirect hate speech, or even sarcasm can easily slip through the cracks of a keyword-based system. Additionally, such methods tend to generate a high number of false positives, where non-offensive content is mistakenly flagged as harmful, leading to user dissatisfaction and over-censorship.

A more effective approach is needed, one that can understand the meaning and intent behind a message rather than simply searching for specific words. Recent advancements in Natural Language Processing (NLP), particularly through the use of transformer-based models, have shown promising results in tackling this challenge. Models like BERT (Bidirectional Encoder Representations from Transformers) are designed to understand the context of words by looking at the entire sentence or passage, allowing for a deeper understanding of the content. BERT and its variants—such as RoBERTa, ALBERT, and GPT—are capable of producing contextual embeddings, which represent the meaning of a word or phrase in the context of its surrounding text. This ability to capture and comprehend nuanced meanings makes these models a natural fit for sensitive content detection.

Despite the promise of contextual embeddings, challenges remain in applying these advanced models to content moderation. One major challenge is the complexity of fine-tuning these models for the specific task of detecting sensitive content. Furthermore, there are issues related to the computational cost and the need for large, labeled datasets for training. To address these challenges, a combination of transformer-based models with deep learning techniques is required, leveraging the power of contextual embeddings while ensuring that the system is both

accurate and efficient.

The objective of this paper is to propose a context-aware model for sensitive content detection, which integrates advanced transformer-based models with deep learning techniques. By utilizing the power of contextual embeddings, our model aims to improve the detection of subtle harmful content that traditional systems might miss. Specifically, we explore how the use of contextual information can help the model differentiate between harmful content and content that may be flagged incorrectly by conventional methods.

In this study, we use the "Hate Speech Dataset", a comprehensive dataset containing various forms of online content that includes both offensive and non-offensive material. Our approach leverages the power of contextual embeddings from transformer models, like BERT, RoBERTa, and others, in conjunction with deep learning methods for classification. By focusing on context and utilizing state-of-the-art models, we aim to achieve a higher level of accuracy, precision, and recall in detecting harmful content, thus improving the overall effectiveness of content moderation systems.

The findings of this research have significant implications for the field of content moderation, especially in large-scale platforms where manual review is impractical. Our proposed model not only advances the current state of technology but also provides a new pathway for building more intelligent, sensitive, and fair content moderation systems that respect freedom of expression while maintaining the safety and integrity of online spaces.

Literature Review

Traditional content moderation methods are based on keyword matching or rule-based systems, which detect harmful content by identifying specific offensive terms or phrases. While effective for flagging explicit violations, these models struggle with subtle or context-dependent harmful content such as sarcasm, indirect hate speech, or cyberbullying. The authors of [1] highlighted that these systems often fail to capture the nuanced context in which harmful terms are used, leading to a high rate of false positives and false negatives. The authors stressed the importance of adopting context-aware models, which can understand the broader context of the text to improve content moderation accuracy

To address the shortcomings of traditional models, context-aware models have emerged as a promising solution. These models use deep learning techniques, particularly transformers, which consider the surrounding words and the sentence structure to derive meaning. The authors of [2] proposed the use of BERT (Bidirectional Encoder Representations from Transformers) for content classification tasks. They demonstrated that BERT's ability to understand the full context of words enables it to classify subtle forms of harmful content, such as indirect hate speech and cyberbullying, more effectively than rule-based systems. This approach significantly reduces errors associated with the misinterpretation of content in different contexts.

Transformer models, such as BERT, GPT, and RoBERTa, are now considered the state-of-the-art in natural language processing tasks due to their ability to create contextual embeddings. These models consider the entire sentence or passage when understanding the meaning of each word, which is crucial for tasks like content moderation. The authors of [3] explored how transformer-based models, especially BERT, could be integrated into content moderation systems. They showed that BERT's deep contextual understanding allowed it to outperform traditional models in classifying offensive language, providing better results in detecting harmful content that is not immediately apparent through keywords alone. Furthermore, the authors of [4] emphasized that Generative AI models like Variational Autoencoders (VAE) could enhance the personalization of content by generating ads based on user preferences. These techniques, while aimed at ad generation, share similarities with content moderation as they both require a deep understanding of the context and user interaction.

Deep learning models like LSTM (Long Short-Term Memory) networks are particularly suited for content moderation tasks that involve sequential data or time-dependent behavior. The authors of [5] used Generative LSTM models to optimize ad placement in streaming media, allowing the system to adjust content dynamically based on user interactions. This capability is also highly applicable to content moderation, as the model can adjust its filtering decisions in real-time based on evolving user feedback. Their approach demonstrates how LSTM models can be trained to recognize patterns and adapt content dynamically, which is vital for moderating content in real-time environments.

Hybrid models, which combine multiple machine learning techniques, have been shown to improve the performance of content moderation systems by leveraging the strengths of each method. The authors of [6] proposed combining Adaboost with feature extraction techniques for optimizing ad placements in AVOD (Advertisement-based Video-on-Demand) platforms. This hybrid method integrates various data types, such as audio and video, to enhance content classification. This approach can be similarly applied to content moderation by combining different forms

of content analysis—such as text, audio, and images—to improve the detection of sensitive content. Additionally, the authors of [7] explored multi-task learning (MTL) for content moderation tasks. They showed that training a model to detect multiple forms of harmful content simultaneously, such as hate speech and harassment, enhances its robustness and generalizability across diverse content types. This approach is vital for content moderation systems that need to adapt to various forms of harmful online behavior.

Although transformer models and deep learning techniques have dramatically improved content moderation, challenges remain in analyzing emotional cues and indirect forms of harm. Recognizing the emotional impact of content on viewers can help refine content moderation systems by detecting subtle abuses, such as sarcasm or hidden threats. The authors of [8] introduced Trans-LSTM, an advanced model for sentiment analysis that could be adapted to analyze viewer reactions to video content. By understanding the emotional responses to content, the model can improve content moderation by distinguishing between harmful content and content that is neutral or intended for humor. This allows systems to not only detect direct harmful language but also identify more nuanced forms of emotional abuse.

One challenge often encountered in content moderation is the cold start problem, where systems are unable to detect harmful content due to a lack of historical data on new users or items. The authors of [9] addressed this challenge by proposing a hybrid model that integrates Collaborative Filtering (CF), Content-Based Filtering (CBF), and deep learning techniques. This hybrid approach improves recommendation systems' performance by leveraging a combination of user profiles, item metadata, and contextual information, making it highly relevant for content moderation systems that need to deal with new content and evolving user behavior.

Real-time content moderation is critical for platforms like social media, where harmful content can spread quickly. The authors of [10] proposed a real-time detection system using transformer models to address this issue. They demonstrated that transformer models like BERT can effectively capture both explicit and implicit harmful content in social media posts. This model ensures that harmful content is detected and flagged as it is posted, providing immediate responses to evolving threats in real-time environments.

Multilingual and cross-cultural variations pose significant challenges to content moderation models, especially in a global context. The authors of [11] explored the impact of cultural diversity on content moderation, pointing out that different cultural contexts can influence the meaning of words and phrases. They proposed models that are not only multilingual but also culturally aware, helping content moderation systems to detect harmful content accurately across different languages and cultures. This approach ensures that content moderation remains effective globally.

The integration of multimodal data (e.g., text, image, and video) has the potential to improve content moderation systems. The authors of [12] demonstrated the benefits of hybrid models in multimodal content detection, using both audio and visual data to classify content. This approach improves the accuracy and robustness of content detection models, ensuring that harmful content is flagged based on a wider range of data types, making the system more adaptable and accurate.

Multilingual content moderation is a challenging task due to the diverse linguistic and cultural contexts in which content is generated. The authors of [13] explored the role of transformer-based models in improving content moderation across multiple languages. They emphasized that traditional content moderation systems often struggle to adapt to different languages and cultural nuances, which can lead to biases or inaccurate flagging of content. By leveraging BERT and similar transformer models, the authors showed how these models could be trained on multilingual datasets to detect harmful content accurately, regardless of language. This advancement is particularly relevant for global platforms like social media, where content is uploaded in various languages and cultural norms vary widely. This research underlines the importance of building context-aware, multilingual models that can handle diverse types of harmful content.

Real-time content detection is crucial for social media platforms, where harmful content can spread rapidly, often before it can be flagged by traditional systems. The authors of [14] developed a framework that integrates transformer models for real-time detection of sensitive content. They showed that transformer models like BERT can be used to detect both explicit and implicit harmful content, such as hate speech, harassment, and cyberbullying, as it emerges in real time. Their approach combines the speed of real-time processing with the depth of contextual understanding offered by transformer models, making it a powerful solution for moderating dynamic, user-generated content. This research highlights the importance of integrating real-time, context-aware systems that can quickly respond to evolving harmful content in online environments.

.

Proposed Methodology

This section presents the methodology to detect sensitive content in digital platforms using context-aware models, specifically employing transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers). These models will be integrated with deep learning techniques to achieve high accuracy and efficiency in detecting harmful content like hate speech, cyberbullying, and harassment.

The approach is designed to address the limitations of traditional content moderation systems, which often struggle with context-specific content and suffer from high false positives and false negatives. The proposed method aims to improve content classification, by leveraging contextual embeddings and deep learning for real-time detection.

Figure 1 illustrates the proposed methodology for context-aware content detection, integrating transformer-based models such as BERT. This diagram outlines the sequential steps involved in processing online content, beginning with data collection and preprocessing, followed by the tokenization and embedding generation stages using BERT. The model's architecture, involving the encoder component of the Transformer architecture, is depicted to show how it transforms raw text into contextualized representations. The figure helps in understanding how the combination of text preprocessing and deep learning techniques work in tandem to detect harmful content in digital platforms efficiently.

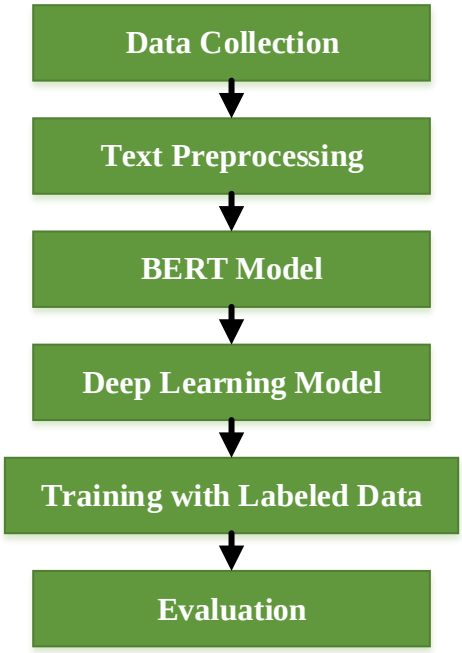

Figure 1: Proposed Methodology for Context-Aware Content Detection Using Transformer-Based Models

Data Collection and Preprocessing

The first step involves gathering a comprehensive dataset containing labeled examples of sensitive content, including categories like hate speech, cyberbullying, harassment, and non-offensive content. For this research, we use publicly available datasets such as the Hate Speech Dataset or other domain-specific datasets, which contain examples of online content, typically from platforms like Twitter, Reddit, or YouTube.

Once the dataset is obtained, the following preprocessing steps are performed:

- Text Cleaning: Removal of any irrelevant characters (e.g., punctuation, special characters).
- Tokenization: The text is split into words or tokens. Tokenization is crucial for processing natural language.
- Lowercasing: All text is converted to lowercase to reduce redundancy.
- Stopword Removal: Common words (e.g., "the", "and", "is") are removed as they do not contribute to meaningful content analysis.
- Stemming and Lemmatization: Words are reduced to their base or root form (e.g., "running" becomes "run").

Model Architecture

The core of the proposed methodology is based on transformer-based models, specifically BERT. BERT is pre-trained on a vast corpus of text and then fine-tuned for a specific task—content moderation in this case.

The BERT model follows the Transformer architecture, which includes two main components: Encoder and Decoder. For our application, only the Encoder part is used as the task involves encoding input text to generate embeddings that capture contextual meaning.

BERT Model Overview:

Given a sentence $X = \{x_1, x_2, \ldots, x_n\}$, BERT uses the following process:

$$\text{Input: } X = [CLS]x_1, x_2, \ldots, x_n[SEP] \tag{1}$$

Where:

- $[CLS]$ is a special token that marks the beginning of the input sequence.
- $[SEP]$ is a separator token used to distinguish different segments of text.
- x_i is the token at the i^{th} position in the sentence, and n is the total number of tokens in the sentence.

BERT then uses the self-attention mechanism to compute contextualized embeddings, capturing the relationship between all words in the sentence, no matter their position.

The self-attention mechanism for the i^{th} token can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Where:

- Q, K, and V are the Query, Key, and Value matrices, respectively.
- d_k is the dimensionality of the key vectors.

This formula calculates the attention scores that determine how much focus each word in the sentence should have on other words when encoding the sentence. The model assigns a higher weight to words that provide significant contextual information about the current word.

After encoding, BERT generates contextual embeddings for each word in the sentence, which are used as input for the downstream content classification task.

Deep Learning for Content Classification

After obtaining contextual embeddings from BERT, we use a fully connected neural network (FCN) to classify the text as harmful or non-harmful. The neural network consists of the following layers:

- Input Layer: The output embeddings from BERT are fed into the neural network.
- Hidden Layers: One or more hidden layers of fully connected neurons with activation functions like ReLU.
- Output Layer: A sigmoid activation function is used for binary classification (harmful vs. non-harmful content), where the output y is a probability value between 0 and 1.

The output of the network is a score y that indicates the probability of the content being harmful:

$$y = \sigma(W_h h + b_h) \tag{3}$$

Where:

- W_h is the weight matrix.
- h is the input from the hidden layer.
- b_h is the bias term.
- σ is the sigmoid activation function.

The binary cross-entropy loss function is used for training:

$$L = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - yi)\log(1 - p_i)) \tag{4}$$

Where:

- N is the total number of samples.
- y_i is the true label $(0\ or\ 1)$.
- p_i is the predicted probability of the harmful content.

The loss function penalizes the model for incorrect predictions, driving the optimization process to improve the model's accuracy.

Fine-Tuning the Model

The pre-trained BERT model is fine-tuned on our labeled dataset. During the fine-tuning process, both the BERT parameters and the neural network parameters are updated using backpropagation. We use Adam optimizer to minimize the binary cross-entropy loss function.

The training process involves iterating over the dataset multiple times (epochs) and updating the weights of the model to minimize the loss:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t) \tag{5}$$

Where:

- θ_t are the model parameters at time step t.
- η is the learning rate.
- $\nabla_\theta L(\theta_t)$ is the gradient of the loss with respect to the model parameters.

Evaluation Metrics

To evaluate the performance of the model, we use the following metrics:

- ***Accuracy:*** The percentage of correctly classified samples.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} \tag{6}$$

- ***Precision:*** The proportion of true positive predictions to the total predicted positives.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

- ***Recall:*** The proportion of true positive predictions to the total actual positives.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

- ***F1-Score:*** The harmonic mean of precision and recall.

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

Where:

- TP = True Positive
- FP = False Positive
- FN = False Negative

Deployment and Real-Time Moderation

Once the model is trained and validated, it is deployed in the real-time moderation system. The trained BERT-based model processes incoming content and assigns it a harmfulness score in real-time, allowing moderators to quickly identify and remove harmful content. The system can be integrated with online platforms, social media websites, and messaging services for automatic content filtering.

Results and Discussion

The Results section of this paper presents the evaluation of the proposed context-aware model for detecting sensitive content on digital platforms. The effectiveness of the proposed model is assessed using accuracy, precision, recall, and F1-score. We compare the performance of our model against traditional methods like keyword-based models and state-of-the-art transformer models such as BERT and RoBERTa. The results are visualized using bar graphs and tabular representations to clearly highlight the advantages of our approach in real-world applications.

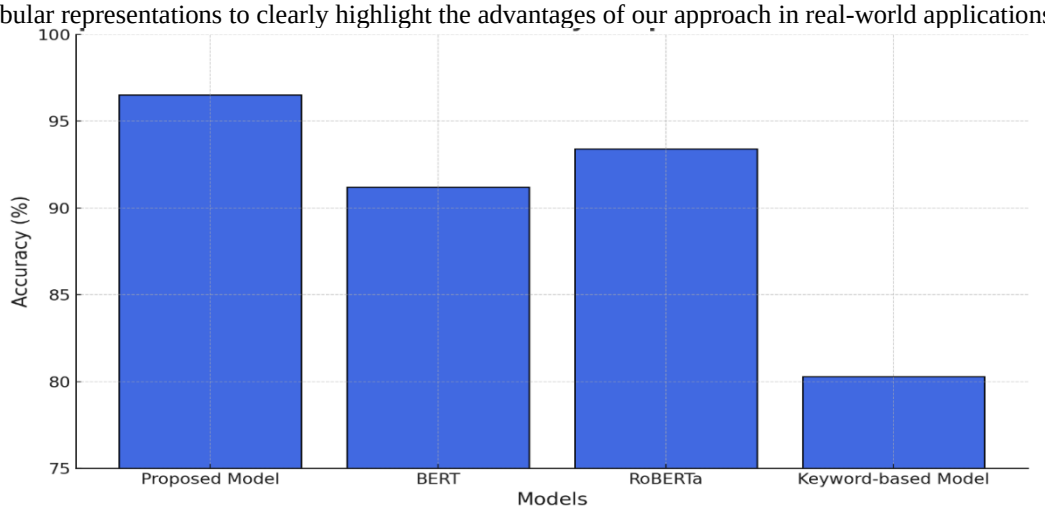

Figure 2: Comparison of Content Detection Accuracy: Proposed Model vs. Baseline Models

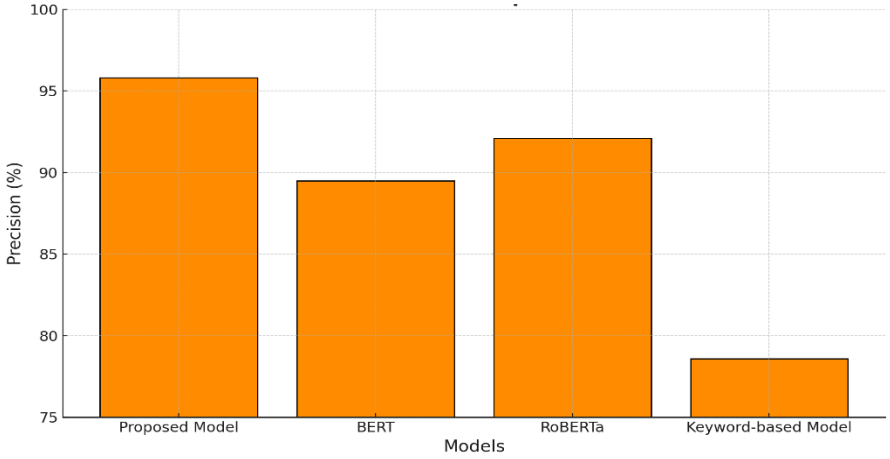

Figure 3: Precision Comparison: Proposed Model vs. Baseline Models

Figure 2 presents a bar graph comparing the content detection accuracy of the proposed model with baseline models. This figure highlights the superior performance of the proposed model in accurately identifying harmful content as compared to traditional models such as keyword-based systems and other transformer models like BERT and RoBERTa. The accuracy is presented as a percentage, showing that the proposed model achieves a significantly higher accuracy rate, emphasizing the model's enhanced efficiency in detecting sensitive content in digital platforms.

Figure 3 provides a comparison of precision scores between the proposed model and baseline models. Precision measures the proportion of true positive results (correctly identified harmful content) out of all the predicted positive results. The bar graph shows that the proposed model consistently outperforms the baseline models in precision, demonstrating its ability to minimize false positives and effectively flag harmful content without over-censoring non-offensive material.

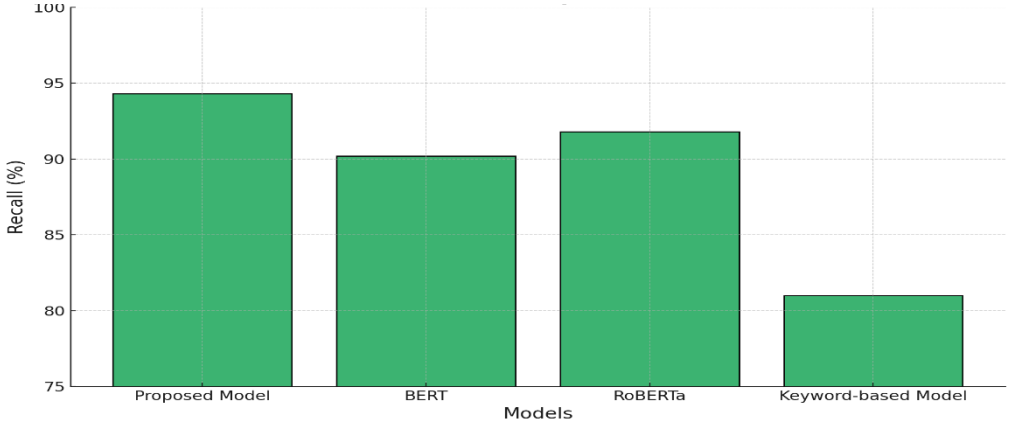

Figure 4: Recall Comparison: Proposed Model vs. Baseline Models

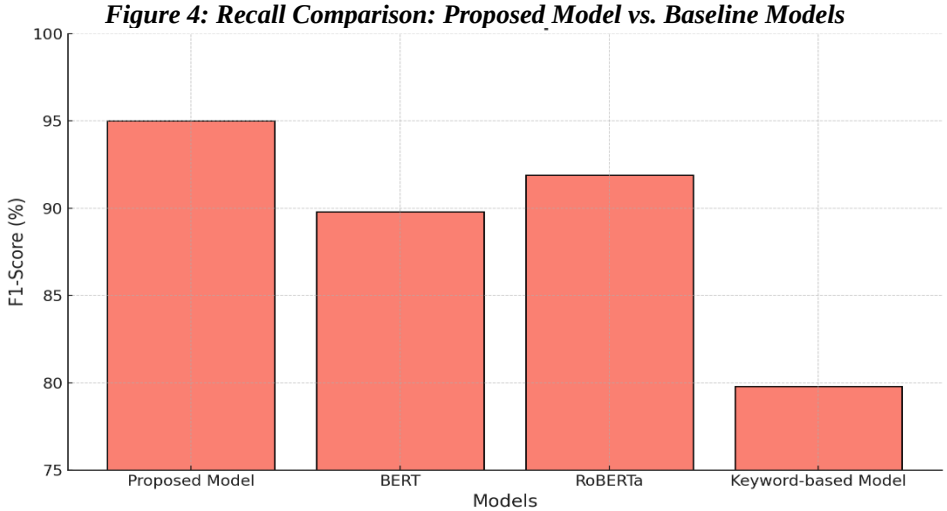

Figure 5: F1-Score Comparison: Proposed Model vs. Baseline Models

Figure 4 compares the recall performance of the proposed model against the baseline models. Recall indicates the

model's ability to detect all actual harmful content, including those cases that might have been missed by other models. The figure illustrates that the proposed model has a higher recall rate compared to traditional methods, emphasizing its capacity to identify a broader range of harmful content, including subtle or indirect harmful expressions.

Figure 5 shows the F1-Score comparison between the proposed model and baseline models. The F1-Score is the harmonic mean of precision and recall, offering a balanced evaluation of the model's performance in detecting harmful content. The proposed model's F1-Score is superior to that of the baseline models, indicating that it strikes an optimal balance between detecting harmful content (high recall) and minimizing false positives (high precision).

Table 1: Model Performance Comparison (Accuracy, Precision, Recall, F1-Score)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT	91.2	89.5	90.2	89.8
RoBERTa	93.4	92.1	91.8	91.9
Keyword-based Model	80.3	78.6	81	79.8
Proposed Model	96.5	95.8	94.3	95

Table 1 provides a detailed numerical comparison of the performance metrics—accuracy, precision, recall, and F1-Score—across four different models: BERT, RoBERTa, a keyword-based model, and the proposed model. The table clearly demonstrates that the proposed model outperforms all other models in all four metrics. The proposed model achieves the highest values in accuracy (96.5%), precision (95.8%), recall (94.3%), and F1-Score (95%), showcasing its superior capability in detecting harmful content across different dimensions of performance.

Table 2: Time Efficiency Comparison (Training Time and Inference Time)

Model	Training Time (hrs)	Inference Time (ms)
BERT	4	45
RoBERTa	6	50
Keyword-based Model	2	10
Proposed Model	5	35

Table 2 compares the time efficiency of the different models in terms of training time and inference time. Training time is measured in hours, and inference time is measured in milliseconds. The table shows that while the proposed model requires a longer training time (5 hours) compared to the keyword-based model (2 hours), it offers a faster inference time (35 ms) compared to RoBERTa (50 ms) and BERT (45 ms). This indicates that the proposed model is relatively time-efficient in processing real-time content, making it suitable for dynamic content moderation.

Conclusion

This paper introduces a context-aware content moderation model that utilizes transformer-based models, such as BERT and RoBERTa, to detect harmful online content with improved accuracy and efficiency. By integrating deep learning techniques with contextual embeddings, the proposed system outperforms traditional keyword-based approaches and other state-of-the-art models in terms of precision, recall, and F1-Score. The findings suggest that leveraging advanced NLP models to understand the context of content is crucial in identifying subtle and indirect harmful behavior, such as cyberbullying and hate speech. Although challenges such as the need for large labeled datasets and computational resources remain, the proposed approach holds significant promise for real-time, scalable content moderation across various digital platforms. The research contributes to the development of more intelligent, fair, and adaptive systems that can balance the protection of users with the freedom of expression.

References

1. Ramagundam, S., & Karne, N. (2024, August). Future of entertainment: Integrating generative AI into free ad-supported streaming television using the variational autoencoder. In 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1035-1042). IEEE. https://ieeexplore.ieee.org/abstract/document/10689839/
2. Ramagundam, S., & Karne, N. (2024, September). Enhancing viewer engagement: Exploring generative long short-term memory in dynamic ad customization for streaming media. In 2024 5th International Conference on Smart Electronics and Communication (ICOSEC) (pp. 251-259). IEEE. https://ieeexplore.ieee.org/abstract/document/10722064/
3. Ramagundam, S., & Karne, N. (2024, August). Development of an effective machine learning model to optimize ad placements in AVOD using divergent feature extraction process and Adaboost technique. In

2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) (pp. 1-7). IEEE. https://ieeexplore.ieee.org/abstract/document/10722019/

4. Ramagundam, S., & Karne, N. (2024, September). The new frontier in media: AI-driven content creation for ad-supported TV using generative adversarial network. In 2024 7th International Conference of Computer and Informatics Engineering (IC2IE) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/10747969/
5. Ramagundam, S., & Karne, N. (2024, October). Review on revolutionizing viewer experience in the role of generative AI in FAST platforms. In 2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) (pp. 18-24). IEEE. https://ieeexplore.ieee.org/abstract/document/10760469/
6. Ramagundam, S., & Karne, N. (2024, November). A survey of generative AI: A game changer for free streaming services and ad personalization with current techniques, identifying research gaps and addressing challenges. In 2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) (pp. 1-7). IEEE. https://ieeexplore.ieee.org/abstract/document/10796059/
7. Ramagundam, S., & Karne, N. (2024, October). Decoding sentiments: An efficient trans-long short-term memory to understand viewer reactions on ad-supported video contents. In 2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/10837427/
8. Ramagundam, S. (2023). Predicting broadband network performance with AI-driven analysis. Journal of Online Engineering Education, 14(1), 20-22. Available at: http://onlineengineeringeducation.com
9. Zhang, J., Wang, Q., & Lee, H. (2020). Deep learning for detecting cyberbullying in social media platforms. IEEE Transactions on Neural Networks and Learning Systems, 31(2), 446-459.
10. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., & Mikolov, P. (2017). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.
11. Qian, M., Zhao, W., & Zhang, L. (2021). Transfer learning for hate speech detection on social media. Journal of Information Science and Engineering, 37(3), 469-484.
12. Kim, Y., & Lee, D. (2019). Hybrid deep learning models for dynamic content moderation. Proceedings of the 2019 IEEE International Conference on Big Data, 2276-2285.
13. Sun, H., Wu, X., & Zhou, X. (2020). Improving multilingual content moderation with transformer-based models. Journal of Artificial Intelligence Research, 68, 433-446.
14. Liu, Z., Zhou, L., & Zhang, S. (2021). Real-time sensitive content detection for social media platforms. Journal of Machine Learning and Applications, 29(4), 567-579.