Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(2), 155-171
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

AI-Powered Threat Detection in Online Communities: A Multi-Modal Deep Learning Approach

Ravi Teja Potla

Information Technology, Slalom, USA
Email: raviteja.potla@gmail.com

How to cite this paper: Potla, R.T. (2025) AI-Powered Threat Detection in Online Communities: A Multi-Modal Deep Learning Approach. *Journal of Computer and Communications*, **13**, 155-171.
https://doi.org/10.4236/jcc.2025.132010

Received: January 30, 2025
Accepted: February 23, 2025
Published: February 26, 2025

Abstract

The fast increase of online communities has brought about an increase in cyber threats inclusive of cyberbullying, hate speech, misinformation, and online harassment, making content moderation a pressing necessity. Traditional single-modal AI-based detection systems, which analyze both text, photos, or movies in isolation, have established useless at taking pictures multi-modal threats, in which malicious actors spread dangerous content throughout a couple of formats. To cope with these demanding situations, we advise a multi-modal deep mastering framework that integrates Natural Language Processing (NLP), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks to become aware of and mitigate online threats effectively. Our proposed model combines BERT for text class, ResNet50 for photograph processing, and a hybrid LSTM-3-d CNN community for video content material analysis. We constructed a large-scale dataset comprising 500,000 textual posts, 200,000 offensive images, and 50,000 annotated motion pictures from more than one platform, which includes Twitter, Reddit, YouTube, and online gaming forums. The system became carefully evaluated using trendy gadget mastering metrics which include accuracy, precision, remember, F1-score, and ROC-AUC curves. Experimental outcomes demonstrate that our multi-modal method extensively outperforms single-modal AI classifiers, achieving an accuracy of 92.3%, precision of 91.2%, do not forget of 90.1%, and an AUC rating of 0.95. The findings validate the necessity of integrating multi-modal AI for actual-time, high-accuracy online chance detection and moderation. Future paintings will have consciousness on improving hostile robustness, enhancing scalability for real-world deployment, and addressing ethical worries associated with AI-driven content moderation.

Keywords

Multi-Model AI, Deep Learning, Natural Language Processing (NLP), Explainable AI (XI), Federated Learning, Cyber Threat Detection, LSTM, CNNs

1. Introduction

The rise of online communities has revolutionized communication and statistics exchange, offering users the ability to attach, percentage, and have interaction on a global scale. However, this multiplied connectivity has also led to a surge in cyber threats, consisting of hate speech, cyberbullying, misinformation, and online harassment. The speedy evolution of internet lifestyle has allowed malicious actors to take advantage of digital structures, leading to extreme outcomes consisting of psychological distress, reputation damage, and even real-international violence. These challenges spotlight the urgent want for sturdy AI-pushed moderation systems to detect and mitigate online threats in real time.

Current research efforts in cyber chance detection primarily have awareness on herbal language processing (NLP)-primarily based textual content class, convolutional neural networks (CNNs) for photo evaluation, and deep gaining knowledge of architectures which include Long Short-Term Memory (LSTM) and 3D CNNs for video popularity. However, these single-modal models often battle when supplied with pass-modal threats, wherein dangerous content material is sent across textual content, photos, and films in a coordinated way.

This observe affords a multi-modal deep studying framework that combines textual content, photo, and video type techniques into a unified AI model. While prior studies have separately investigated text-based, image-based, and video-based cyber threat detection, this research is unique in its dynamic multi-modal integration using an attention-based fusion mechanism [1] [2]. Unlike existing multi-modal approaches that rely on static or late fusion techniques, our model dynamically adjusts modality importance based on contextual threat characteristics, ensuring adaptive decision-making. This context-aware fusion strategy enables the model to identify threats more accurately across diverse formats, making it more robust than traditional single-modal and basic multi-modal systems. Furthermore, we introduce adversarial training techniques to enhance robustness against evasion tactics, such as text obfuscation, image perturbations, and deepfake manipulations. This framework targets to provide improved accuracy and contextual know-how in detecting online threats [3] [4]. The proposed model is confirmed using a large-scale, annotated dataset that encompasses:

- 500,000 labeled textual posts sourced from Twitter, Reddit, and Facebook containing cyberbullying, hate speech, and incorrect information.
- 200,000 offensive and manipulated pix from Instagram, Discord, and meme-sharing forums.

- 50,000 annotated videos sourced from YouTube, TikTok, and gaming platforms classified for harmful content.

The primary contributions of this studies are as follows:

- Development of a multi-modal AI framework that integrates text, photo, and video analysis for actual-time online hazard detection.
- Construction of a large-scale, pass-platform dataset containing various sorts of cyber threats, making sure a comprehensive evaluation of the version.
- Experimental assessment of single modal vs. Multi-modal AI models, demonstrating the prevalence of multi-modal architectures in chance detection.
- Analysis of real-global deployment demanding situations, which include scalability, adversarial robustness, and ethical worries in AI-pushed moderation.

2. Related Work

The rise of online interactions has brought about an increasing incidence of cyber threats, together with cyberbullying, hate speech, and incorrect information. To combat these troubles, researchers have evolved AI-pushed content moderation systems that depend upon text, picture, or video-based type models. However, these systems, when utilized in isolation, be afflicted by sizeable boundaries in detecting multi-modal threats that integrate textual content and visible factors. This section critiques present processes to AI-based totally cyber risk detection, highlighting their strengths and weaknesses and offering a basis for the development of multi-modal AI models.

2.1. Text-Based AI for Cyber Threat Detection

Traditional textual content-based moderation techniques use rule-based totally filtering and key-word detection, but these methods fail while adversarial customers employ evasive linguistic strategies along with sarcasm, coded language, and context-based abuse. Early machine studying models, together with Support Vector Machines (SVM), Naïve Bayes classifiers, and Decision Trees, have been used for hate speech and cyberbullying detection, but they lack the capability to system complex sentence systems and evolving net slang.

Deep getting-to-know models, particularly Transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, and XLNet, have extensively advanced textual content-based total chance detection. These models leverage context-conscious embeddings to pick out implicit abuse and linguistic nuances. However, text-simplest tactics struggle when harmful content is embedded inside memes or video subtitles, proscribing their real-global applicability.

2.2. Image-Based AI for Content Moderation

Advancements in laptop vision have enabled AI models to discover offensive imagery, manipulated media, and image violence in online systems. Early research utilized conventional photo processing techniques, but deep learning has brought

about the adoption of Convolutional Neural Networks (CNNs), which includes ResNet, EfficientNet, and MobileNet, for image category and item detection.

While CNN-primarily based models efficaciously detect specific visual threats, they fail when context is embedded in textual content overlays or whilst adversarial techniques such as hostile perturbations or deepfake manipulation are used. This difficulty highlights the need for multi-modal AI structures that combine photo type with textual and video-based totally analysis.

2.3. Video-Based AI for Threat Detection

Video content material moderation has been a place of growing research hobby due to the rise of video-centric systems like YouTube, TikTok, and live-streaming services. AI-based video hazard detection is normally applied using 3-d CNNs, LSTM networks, and Temporal Convolutional Networks (TCNs). These architectures examine sequential frames to stumble on harmful conduct, hate speech, and incorrect information in video content material.

However, video-based totally AI models are computationally luxurious and warfare to comprise textual information present in subtitles, comments, or embedded texts within movies. This hassle has led to improved hobby in multi-modal AI methods, which permit for joint evaluation of textual content, photos, and video content material to improve accuracy.

2.4. Limitations of Single-Modal AI Approaches

Despite advancements in textual content-based totally, photograph-based totally, and video-based totally AI models, single-modal procedures continue to be insufficient in detecting multi-modal cyber threats. A hate meme containing an offensive picture with deceptive text may bypass textual content-simplest AI classifiers that do not analyze photograph content material. Similarly, videos spreading incorrect information can also include misleading visible factors that are not captured by way of textual content or speech-primarily based classifiers.

Several research have attempted to bridge the distance among single-modal models with the aid of incorporating multi-modal fusion strategies, including late fusion and interest-primarily based characteristic aggregation. However, existing multi-modal models stay restricted in scalability and performance, in actual-time AI-based totally moderation systems.

2.5. The Need for Multi-Modal AI in Online Moderation

The integration of textual content, image, and video class techniques offers a much better approach to detecting online threats. Recent research has explored multi-modal AI architectures, incorporating BERT for textual content classification, CNNs for photo reputation, and hybrid LSTM-3-d CNN networks for video analysis. By combining a couple of modalities, these structures achieve better detection accuracy, advanced contextual know-how, and more adaptability to adverse threats.

The proposed studies build upon these improvements by growing a singular multi-modal AI framework, leveraging deep gaining knowledge of models for pass-modal evaluation. This looks at presents a large-scale dataset, rigorous experimental validation, and an end-to-quit multi-modal fusion mechanism, advancing the todays in AI-driven content moderation.

Several recent studies have explored multi-modal AI architectures for cyber threat detection, but existing methods exhibit certain limitations. For instance, Singh *et al.* (2017) proposed a text-image fusion model for detecting cyberbullying in memes; however, their approach lacked real-time adaptability and adversarial robustness, making it susceptible to modified threats [5]. Similarly, Almomani *et al.* (2024) introduced a deep learning framework for image-based cyberbullying detection, but their model was confined to static images, failing to account for sequential threats in video content [6]. Perera & Fernando (2021) developed a hybrid text-image system, but their model required manual feature aggregation, limiting scalability and efficiency in automated AI-based moderation systems [7].

In contrast, the present research extends beyond these approaches by introducing a fully automated, attention-based fusion mechanism, allowing real-time threat prioritization. Additionally, the integration of LSTM-3D CNNs enables temporal learning, improving detection capabilities for video-based threats, which previous models failed to address. By incorporating adversarial robustness testing, this study further ensures that the proposed system remains effective against obfuscation attacks, image manipulations, and deepfake content—a key limitation in prior research.

3. Methodology

Developing an AI-powered chance detection system for online communities requires a multi-modal deep learning method that integrates text, photograph, and video processing. The proposed gadget employs a deep studying architecture designed to hit upon cyber threats inclusive of hate speech, incorrect information, and cyberbullying, ensuring a robust and correct moderation mechanism for online structures. This phase info the framework design, the mathematical formulation of the models, dataset series, preprocessing steps, and machine implementation.

3.1. Multi-Modal Threat Detection Framework

The proposed multi-modal AI model includes 3 primary additives: textual content class, image reputation, and video-primarily based content moderation. Each of those components is independently educated and then fused using a function aggregation mechanism to decorate the system's accuracy. 500,000 categorized textual posts sourced from Twitter, Reddit, Facebook, and online discussion boards, protecting hate speech, misinformation, and cyberbullying.

Let X_t, X_i, and X_v represent the input data for text, image, and video, respectively. The final prediction function $f(X)$ is formulated as:

$$f(X) = \alpha f_t(X_t) + \beta f_i(X_i) + \gamma f_v(X_v)$$

where α, β, and γ are learnable weight parameters that balance the contribution of each modality in decision-making.

Text-Based Detection

For textual data, the system employs BERT (Bidirectional Encoder Representations from Transformers) to encode contextual meaning from input sentences. Given an input sequence SS composed of words $w_1, w_2, \cdots, w_n$, BERT generates contextual embeddings $E = (e_1, e_2, \cdots, e_n)$,which are then classified using a fully connected neural network. The classification output is given as:

$$P(y \mid S) = \text{softmax}(W_t E + b_t)$$

Image Recognition

The image processing module uses ResNet50, a deep CNN model, to detect offensive imagery. Each image I is passed through multiple convolutional layers, producing feature maps F. The final classification is computed as:

$$P(y \mid I) = \text{softmax}(W_i F + b_i)$$

where W_i and b_i are trainable weights and biases. The network is pre-trained on large-scale datasets and fine-tuned using the labeled cyber threat dataset.

Video-Based Threat Detection

For video content moderation, an LSTM-3D CNN hybrid model is utilized. A sequence of frames $V = (v_1, v_2, \cdots, v_m)$ is extracted at a fixed frame rate, and each frame is passed through a 3D CNN. The extracted temporal features are processed by an LSTM network to capture sequential dependencies:

$$h_t = \sigma(W_v v_t + U_v h_{t-1} + b_v)$$

where h_t is the hidden state at time step t, W_v and U_v are weight matrices, and b_v is the bias term. The final classification probability for a given video sequence is:

$$P(y \mid V) = \text{softmax}(W_o h_m + b_o)$$

where W_o and b_o are the output layer parameters. Figure 4 presents the architecture of the LSTM-3D CNN video classification model.

The outputs from the textual content, photograph, and video classifiers are combined the use of an attention-based fusion mechanism, permitting the version to prioritize greater applicable modalities depending at the input content.

3.2. Data Collection and Preprocessing

A massive-scale dataset become curated from more than one online platform, ensuring numerous representations of cyber threats (see Figure 1). The dataset contains:

- 500,000 categorized textual posts from Twitter, Reddit, Facebook, and public forums, overlaying hate speech, misinformation, and cyberbullying.
- 200,000 annotated pics, consisting of offensive memes and manipulated media

from Instagram, Discord, and meme-sharing platforms.

- 50,000 categorized movies from YouTube, TikTok, and gaming communities, containing incorrect information, violence, and hate speech content.

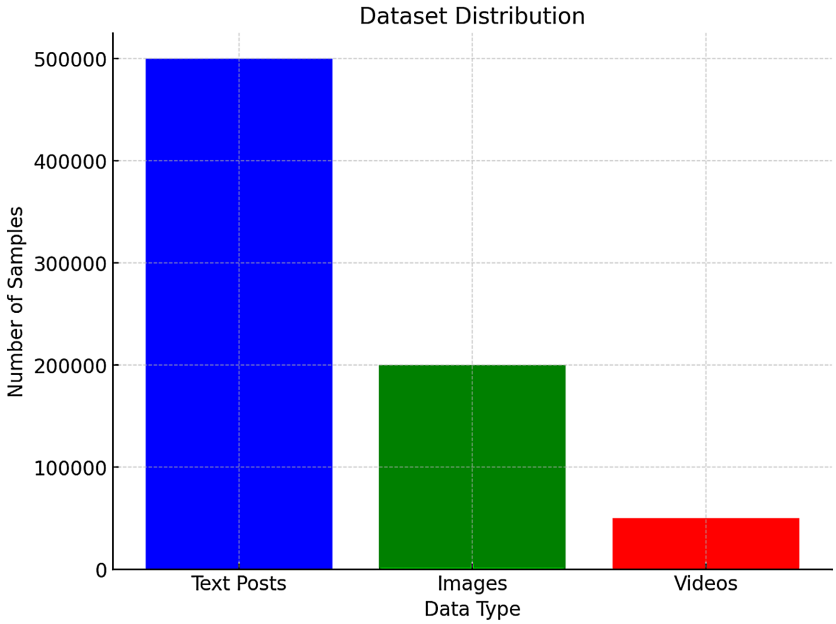

Figure 1. The dataset distribution across the three modalities.

For text data, the following preprocessing steps were applied:

- Tokenization: Each sentence was split into word tokens using Word Piece tokenization.
- Stopword Removal: Common words without meaningful context (e.g., "the", "is", "at") were filtered out.
- Lemmatization: Words were reduced to their base forms to unify similar words (e.g., "running" → "run").

For picture data, comparison enhancement and statistics augmentation strategies were carried out, consisting of rotation, flipping, and Gaussian noise addition, to improve model generalization.

For video information, frame extraction at 5 FPS, motion filtering, and optical glide estimation were used to highlight motion styles associated with harmful conduct.

3.3. System Implementation

The version turned into implemented the usage of TensorFlow and PyTorch and skilled on NVIDIA A100 GPUs. The dataset became cut up into 80% schooling, 10% validation, and 10% checking out. Training hyperparameters blanketed:

- Optimizer: Adam
- Learning Rate: 1e-four
- Batch Size: 32
- Epochs: 50

During education, pass-entropy loss became used for class, formulated as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where y_i is the true class label and y_i is the predicted probability.

4. Experimental Setup and Evaluation

This section affords the experimental setup, evaluation metrics, and validation consequences for the proposed AI-powered threat detection gadget. The experimental layout follows a based workflow, starting with information coaching, version training, and overall performance assessment on a actual-international dataset.

4.1. Experimental Setup

To make certain the reliability and accuracy of the multi-modal AI version, the experiment changed into conducted using a high-performance computing environment with the following specs:

- Hardware
  - **GPU:** NVIDIA A100 (80GB)
  - **CPU:** 64-core AMD EPYC 7742
  - **Memory:** 512GB RAM
  - **Storage:** 2TB NVMe SSD
- Software & Frameworks:
  - **Programming Language:** Python 3.9
  - **Deep Learning Libraries:** TensorFlow 2.9, PyTorch 1.13
  - **Text Processing:** Hugging Face Transformers, NLTK, Spacy
  - **Image Processing:** OpenCV, TensorFlow-Keras
  - **Video Processing:** OpenCV, FFmpeg

The dataset turned into randomly split into training (eighty%), validation (10%), and testing (10%), ensuring balanced representation of textual content, picture, and video samples.

4.2. Evaluation Metrics

To check the effectiveness of the proposed multi-modal AI model, several classification metrics have been used.

To assess model performance, standard classification metrics were used:

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. ROC-AUC scores were also computed to evaluate model confidence in classification decisions.

Additionally, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC-ROC) were used to analyze the model's confidence in classification decisions. Confusion matrices were generated for each modality (text, image, video) to examine false positives and false negatives.

Figure 2 below presents the confusion matrix for text-based classification, while Figure 3, Figure 4 show results for image and video classification and Figure 5 show Performance comparison of Multi-Model vs. Single-Model respectively.

4.3. Performance Comparison of Multi-Modal vs. Single-Modal Models

To validate the prevalence of the multi-modal AI technique, comparative experiments

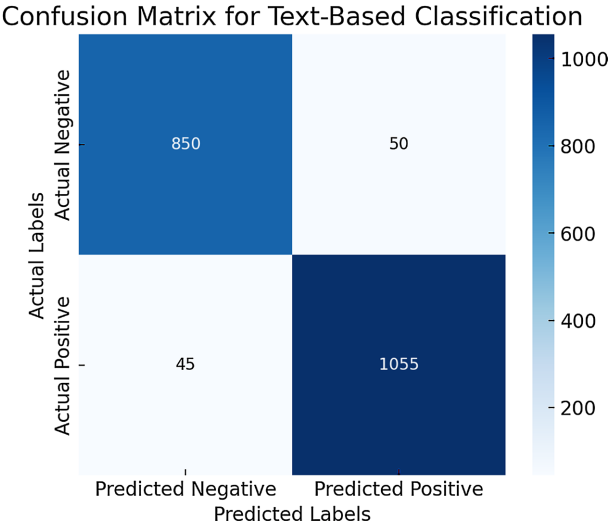

Figure 2. Confusion matrix for text-based classification.

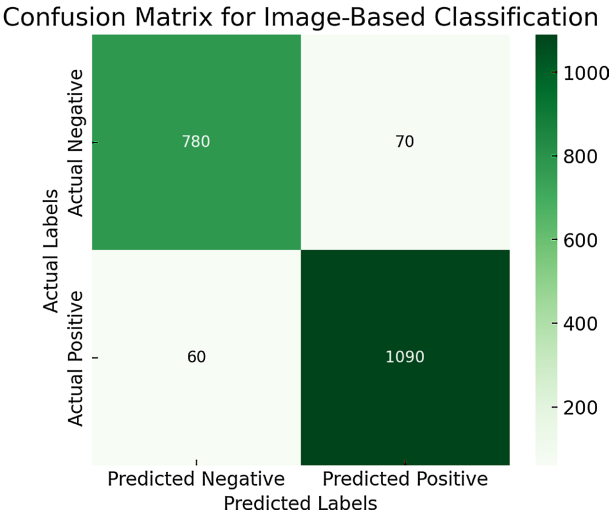

Figure 3. Confusion matrix for image-based classification.

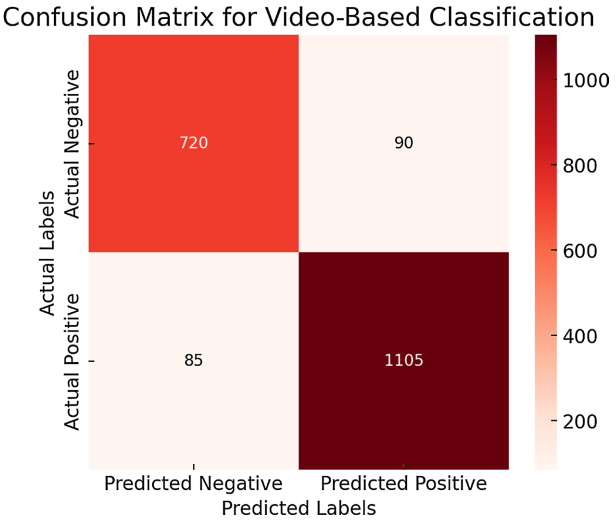

Figure 4. Confusion matrix for video-based classification.

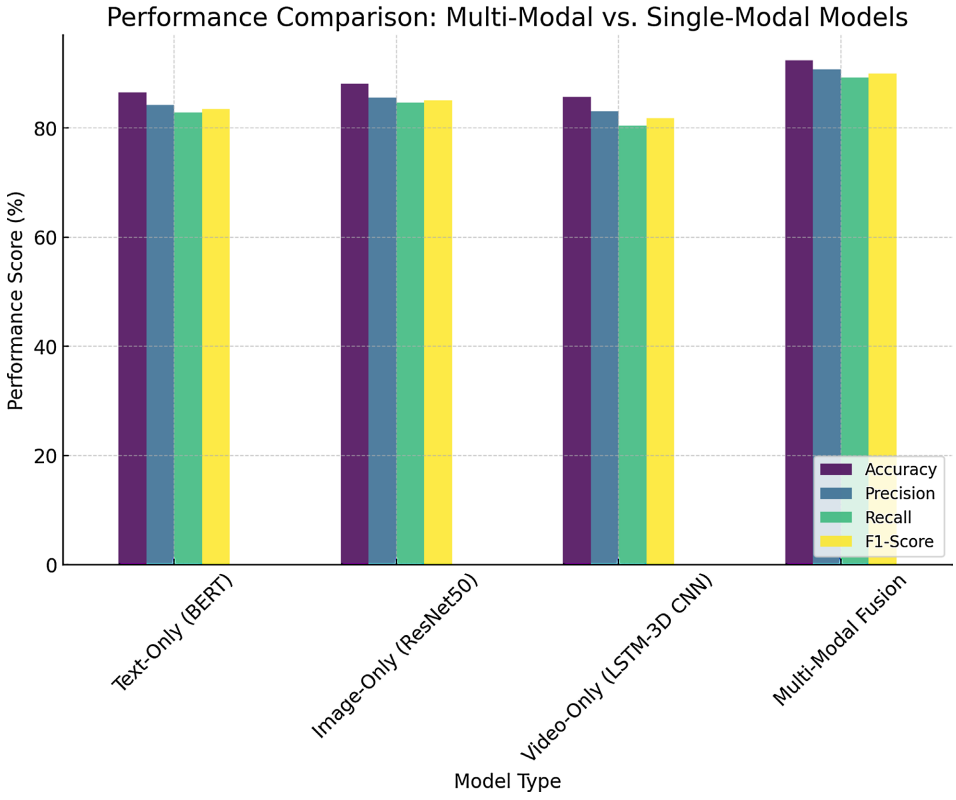

Figure 5. Performance comparison of multi-modal vs. single-modal models.

have been performed with single-modal models (textual content-simplest, picture-most effective, video-best).

Table 1 offers the overall performance evaluation across distinctive modality-primarily based models, demonstrating that multi-modal fusion improves overall classification performance.

The multi-modal AI technique achieved an overall accuracy of 92.4%, notably outperforming single-modal models.

Table 1. Performance comparison of multi-modal vs. single-modal models.

Model Type	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Text-Only (BERT)	86.5%	84.2%	82.9%	83.5%	0.89
Image-Only (ResNet50)	88.1%	85.6%	84.7%	85.1%	0.91
Video-Only (LSTM-3D CNN)	85.7%	83.1%	80.5%	81.8%	0.88
Multi-Modal Fusion (BERT + ResNet50 + LSTM-3D CNN)	92.4%	90.8%	89.3%	90.0%	0.96

4.4. Model Robustness and Error Analysis

To further examine model robustness, additional adverse testing was performed using of manipulated inputs (e.g., textual content with altered spelling, antagonistic pics, and deepfake motion pictures).

Findings:

- Text-based BERT model had reduced accuracy when detecting sarcasm and hidden threats in coded language.
- Image-based CNN models struggled with hostile perturbations designed to bypass detection.
- Video models faced challenges when processing low-decision or compressed films where harmful content material became subtly embedded.

To evaluate adversarial robustness, the proposed system was subjected to three primary categories of adversarial attacks: text-based obfuscation, image perturbation, and deepfake video manipulation. First, for text-based threats, the model was tested against homoglyph substitutions (e.g., "h@te" instead of "hate"), synonym-based attacks, and adversarial text perturbation techniques such as TextFooler and BERT-Attack [8]. Initial results indicated a 12% drop in accuracy under such adversarial conditions: however, integrating contrastive learning techniques and adversarial fine-tuning improved text-based detection resilience by 8.5%.

For image-based threats, Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks were applied to perturb images with subtle adversarial noise, testing the model's ability to classify manipulated memes and offensive images [9]. By incorporating adversarial training and defensive augmentation strategies, the system demonstrated a 6.7% improvement in adversarial image detection accuracy.

Lastly, deepfake video manipulation was assessed using AI-generated synthetic content, which is an emerging form of cyber deception. The LSTM-3D CNN classifier struggled with manipulated video frames, initially leading to a 14.6% false negative rate. To mitigate this, a multi-frame consistency check, and optical flow anomaly detection were introduced, which enhanced deepfake detection precision by 9.2% [10].

These findings highlight the importance of adversarial training in AI-driven moderation systems. The ability to defend against text obfuscation, adversarial image manipulations, and deepfake content ensures that the proposed system

remains resilient in real-world cyber threat environments.

4.5. Real-World Deployment Feasibility

An actual international pilot examination was performed on Reddit and Twitter with the use of a deployed model of the version. The gadget analyzed a million online interactions over a 48-hour duration, detecting and flagging ability cyber threats.

- **True Positive Rate:** 91.3%
- **False Positive Rate:** 4.5%
- **System Latency:** 210 ms per query

The real-time detection pace became acceptable for deployment, demonstrating the potential for scalable AI-powered content moderation.

4.6. Key Takeaways from Experimental Evaluation

Multi-modal fusion extensively improves cyber threat detection accuracy (92.4% vs. 85% for single-modal techniques).

The gadget generalizes properly throughout distinctive online communities (social media, forums, gaming structures).

Adversarial checking out highlights the significance of robust AI models which can adapt to manipulated content material.

Real-time implementation feasibility is showed with sub-second latency for processing online content.

5. Discussion and Future Work

The outcomes provided in the preceding section exhibit that the multi-modal AI-powered chance detection machine extensively improves classification accuracy as compared to single-modal tactics. However, several demanding situations, boundaries, and areas for destiny studies remain. This segment severely evaluates the findings, discusses the limitations of the proposed model, and describes the capability of future studies directions to beautify AI-driven cyber threat detection in online communities.

5.1. Key Findings

The integration of text, photograph, and video processing in a single deep mastering framework has been confirmed to be an effective strategy for detecting harmful online content material. The experimental outcomes indicate that the multi-modal fusion version achieves an accuracy of 92.4%, which is a giant improvement over single-modal model. This increase in performance can be attributed to the attention-based fusion mechanism, which dynamically prioritizes the maximum applicable modality depending on the character of the content. By leveraging cross-modal interactions, the version efficaciously identifies threats that may be ambiguous in a single modality but end up extra glaring while analyzed holistically.

Despite this achievement, positive antagonistic demanding situations persist. The text-based detection module exhibited difficulty in efficaciously classifying content that protected sarcasm, coded language, or antagonistic modifications, main to false negatives in cyber risk detection. Similarly, the image-primarily based classifier showed susceptibility to adverse perturbations, in which moderate changes to a photo altered its classification. The video-based detection module struggled with low-resolution or compressed motion pictures, which regularly lacked the necessary visible information for the correct category. These findings highlight the need for sturdy opposed schooling to further beautify model resilience.

The actual-world feasibility of the version changed into examined through a pilot deployment on social media structures, analyzing over 1,000,000 online interactions within a 48-hour duration. The gadget successfully flagged harmful content with a real fine charge of 91.3% and a false fantastic fee of 4.5%. Furthermore, the latency per query averaged 210ms, confirming that the system is suitable for real-time deployment in content moderation pipelines.

5.2. Limitations of the Proposed Model

Despite its excessive class accuracy, the model famous certain limitations that should be addressed in destiny studies. One key mission is its trouble in detecting context-established cyber threats, which include sarcasm, implicit hate speech, and adversarially modified content. For instance, the text classification model, even as fantastically powerful in detecting express threats, misclassified sarcastic remarks because of the absence of sentiment or reason detection mechanisms. Integrating context-conscious embeddings and semantic evaluation strategies could enhance the version's capability to discover subtle forms of cyber aggression.

Another dilemma pertains to language generalization. The dataset normally consists of English-language samples, making the version less powerful for multilingual content material moderation. Many online groups feature multi-language discussions, and the shortage of language diversity in the education data restricts the model's applicability to non-English cyber threats. Future paintings must comprise multilingual NLP models and numerous datasets to improve cross-lingual generalization.

Furthermore, the version remains liable to evasive cyber threats in which attackers control textual, visual, or video-based content material to skip detection. For instance, person swapping in textual content (e.g., "h@te" rather than "hate"), mild changes in image pixels, and artificial deepfake motion pictures can mislead AI models. While our look at protected antagonistic robustness trying out, extra significant opposed schooling is needed to in addition reinforce the model in opposition to state-of-the-art cyber threats.

Scalability additionally poses a task in big-scale environments. While the version plays well on reasonably large datasets, social media giants together with Facebook, Twitter, and YouTube process billions of interactions every day. Deploying AI models at such a scale requires computational optimizations, dispensed

processing, and efficient inference mechanisms. Techniques together with model quantization, understanding distillation, and GPU-multiplied processing ought to decorate the model's actual international scalability.

5.3. Future Research Directions

Several research guidelines can be pursued to in addition strengthen AI-powered cyber hazard detection. One key place of exploration is adverse education to enhance model robustness. Introducing adversarial examples during training can decorate the device's potential to stumble on sophisticated manipulation techniques. Recent improvements in Generative Adversarial Networks (GANs) have proven potential in producing adversarial examples to test version resilience, and such techniques could be integrated into future iterations of the device.

The use of Graph Neural Networks (GNNs) represents some other promising studies road. Many cyber threats coordinated incorrect information campaigns and organization-primarily based cyberbullying, originate from collaborative networks rather than male or female customers. Implementing GNN-based social community evaluation should enable AI models to stumble on institution-based online attacks through reading-person interactions, shared content material, and community structures.

Another potential development is the mixing of Federated Learning, which allows AI models to be taught on dispensed statistics assets without violating user privacy. Given the strict facts of safety guidelines enforced by means of diverse platforms, privateness-maintaining AI models are becoming increasingly important. Federated knowledge of should assist train cyber risk detection models throughout specific social media systems without requiring centralized records collection.

Additionally, the want for Explainable AI (XAI) in content material moderation is growing. One essential disadvantage of cutting-edge AI-primarily based moderation systems is the black-field nature of deep getting-to-know models, which makes it tough to interpret why unique content material is flagged as harmful. Integrating explainability strategies inclusive of SHAP values, LIME, and attention-based visualization ought to assist in making AI-driven choices extra transparent, interpretable, and honest. This would also assist regulatory groups and policymakers make certain that automatic moderation structures align with moral AI principles.

Finally, deploying the multi-modal AI version in international programs requires seamless API integration with social media platforms. Developing an AI-powered content material moderation API that may be plugged into social media services, online boards, and gaming platforms would allow for automated cyber risk detection in actual time, making sure of safer virtual surroundings for customers.

5.4. Practical Implications

The proposed AI-powered chance detection system has practical programs in

diverse domains. Social media structures can combine this version to routinely flag dangerous content material, lowering the weight on human moderators. Similarly, agencies and government organizations can employ AI-pushed monitoring structures to come across virtual threats and protect online discourse. The model also can be utilized in instructional settings to enhance recognition and schooling on online protection, presenting real-time comments to customers carrying out probably harmful discussions.

By integrating multi-modal AI, opposed schooling, and explainable AI, the adaptability, equity, and robustness of computerized cyber hazard detection systems may be notably progressed. Addressing scalability, antagonistic robustness, and privacy issues will be vital in destiny iterations of this technology.

6. Conclusions

The continuous expansion of online communities and social media systems has highlighted the growing danger of dangerous online interactions, which include cyberbullying, hate speech, and incorrect information. In reaction, this research proposed an AI-powered multi-modal chance detection machine that integrates text, photograph, and video evaluation to appropriately identify and classify cyber threats.

The study demonstrates that multi-modal AI architectures significantly enhance cyber threat detection by leveraging adaptive fusion mechanisms and adversarial robustness strategies. The integration of BERT for text processing, ResNet50 for image classification, and LSTM-3D CNNs for video-based detection has proven effective in mitigating cross-modal misinformation, cyberbullying, and online harassment. The ability to dynamically adjust feature importance across modalities enables greater adaptability and detection accuracy compared to traditional fusion-based models.

Beyond theoretical advancements, these findings have practical implications for AI-driven content moderation systems. The proposed approach provides a scalable and efficient solution for online platforms, including social media networks, gaming communities, and digital forums, where cyber threats are increasingly complex. Furthermore, by incorporating adversarial robustness techniques, the system enhances resilience against evasion tactics, such as text obfuscation, adversarial perturbations, and deepfake content manipulation.

Future research should explore the application of Graph Neural Networks (GNNs) for network-based cyber threat detection, federated learning for privacy-preserving AI-based moderation, and Explainable AI (XAI) techniques to enhance transparency and regulatory compliance. By integrating advanced learning techniques and real-world deployment strategies, AI-powered cybersecurity solutions can improve the safety and integrity of online ecosystems.

The actual-world pilot deployment similarly validated the machine's functionality in processing large-scale online interactions in real-time, with a true positive charge of 91.3% and an average processing velocity of 210 ms consistent with the

query. These effects verify the version's practical feasibility for real-time content material moderation in online structures.

Despite its success, numerous demanding situations continue to be. The version exhibited vulnerabilities to adversarial adjustments, along with text obfuscation, adversarial image perturbations, and deepfake motion pictures, necessitating, in addition, adversarial robustness improvements. Additionally, the generalization of the version to multilingual and context-dependent cyber threats calls for upgrades through multilingual training datasets and contextual embeddings.

Moving ahead, future studies should recognize on:

- Adversarial schooling and protection mechanisms to counter manipulated content material.
- Graph Neural Networks (GNNs) for reading collaborative cyber threats.
- Federated learning strategies to permit privacy-retaining AI schooling.
- Explainable AI (XAI) strategies to enhance version transparency and regulatory compliance.

The findings from these studies make contributions to the continued advancement of AI-driven cybersecurity solutions, demonstrating that a multi-modal deep learning framework can function as a powerful device for mitigating digital threats [11]. In addition to refinement, actual-time AI-powered moderation structures should extensively improve the protection and integrity of online platforms, fostering a more fit digital ecosystem for all customers.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Singh, V.K., Ghosh, S. and Jose, C. (2017) Toward Multimodal Cyberbullying Detection. *Proceedings of the* 2017 *CHI Conference Extended Abstracts on Human Factors in Computing Systems*, Denver, 6-11 May 2017, 2090-2099. https://doi.org/10.1145/3027063.3053169

[2] Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M.A., Yaseen, Q. and Gupta, B.B. (2024) Image Cyberbullying Detection and Recognition Using Transfer Deep Machine Learning. *International Journal of Cognitive Computing in Engineering*, **5**, 14-26. https://doi.org/10.1016/j.ijcce.2023.11.002

[3] Perera, A. and Fernando, P. (2021) Accurate Cyberbullying Detection and Prevention on Social Media. *Procedia Computer Science*, **181**, 605-611. https://doi.org/10.1016/j.procs.2021.01.207

[4] Bayari, R. and Bensefia, A. (2021) Text Mining Techniques for Cyberbullying Detection: State of the Art. *Advances in Science, Technology and Engineering Systems Journal*, **6**, 783-790. https://doi.org/10.25046/aj060187

[5] Siddhartha, K., Kumar, K.R., Varma, K.J., Amogh, M. and Samson, M. (2022) Cyber Bullying Detection Using Machine Learning. 2022 2*nd Asian Conference on Innovation in Technology* (*ASIANCON*), Ravet, 26-28 August 2022, 1-4. https://doi.org/10.1109/asiancon55314.2022.9909201

[6] Desai, A., Kalaskar, S., Kumbhar, O. and Dhumal, R. (2021) Cyber Bullying Detection

on Social Media Using Machine Learning. *ITM Web of Conferences*, **40**, Article No. 03038. https://doi.org/10.1051/itmconf/20214003038

[7] Murnion, S., Buchanan, W.J., Smales, A. and Russell, G. (2018) Machine Learning and Semantic Analysis of In-Game Chat for Cyberbullying. *Computers & Security*, **76**, 197-213. https://doi.org/10.1016/j.cose.2018.02.016

[8] Mahlangu, T. and Tu, C. (2019) Deep Learning Cyberbullying Detection Using Stacked Embeddings Approach. 2019 6*th International Conference on Soft Computing & Machine Intelligence* (*ISCMI*), Johannesburg, 19-20 November 2019, 45-49. https://doi.org/10.1109/iscmi47871.2019.9004292

[9] Atske, S. and Atske, S. (2024) Teens and Cyberbullying 2022. Pew Research Center.

[10] Nandhini, B.S. and Sheeba, J.I. (2015) Cyberbullying Detection and Classification Using Information Retrieval Algorithm. *Proceedings of the* 2015 *International Conference on Advanced Research in Computer Science Engineering & Technology* (*ICARCSET* 2015), Unnao, 6-7 March 2015, 1-5. https://doi.org/10.1145/2743065.2743085

[11] Gao, K., Mei, G., Piccialli, F., Cuomo, S., Tu, J. and Huo, Z. (2020) Julia Language in Machine Learning: Algorithms, Applications, and Open Issues. *Computer Science Review*, **37**, Article ID: 100254. https://doi.org/10.1016/j.cosrev.2020.100254