IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Ai-Driven Multimodal Cyberbullying Detection: Analyzing text, Video, Image, And Audio For Digital Safety

"A Unified Deep Learning Frameworks for Real-Time Online Abuse Detection"

[1]Nagaraju Vassey, [2]Manne Naga VJ Manikanth, [3]Manmai Vimandi

[1]Assistant Professor, [2]Guest Faculty, [3]Student

Department of IT & CA, AU College Of Engineering Department of Information Technology & Computer Applications

Andhra University, Visakhapatnam, India

Abstract: This paper addresses this critical need by proposing an AI-driven multimodal cyberbullying detection system that integrates Natural Language Processing, Computer Vision, and Deep Learning techniques. By leveraging large-scale, diverse datasets and advanced feature fusion methods, the system aims to significantly improve detection accuracy and reduce false alarms. The proposed approach supports safer digital environments, empowering platforms to proactively identify and mitigate cyberbullying across multiple media formats, thus protecting vulnerable users more effectively than existing unimodal solutions.

Index Terms - Cyberbullying, Multimodal Machine Learning, Natural Language Processing, Computer Vision, Deep Learning, Social Media Analytics

I. INTRODUCTION

Cyberbullying is a pervasive issue in online social communities, often involving harmful content across text, images, and videos. Traditional detection methods mainly focus on text, leaving multimedia cyberbullying inadequately addressed. To bridge this gap, this paper proposes an AI-driven multimodal cyberbullying detection system that integrates Natural Language Processing, Computer Vision, and Deep Learning techniques for comprehensive analysis.

Research Objectives:

- To design and develop an AI-powered multimodal system capable of detecting cyberbullying across text, images, and videos in real time.
- To evaluate the effectiveness of feature fusion techniques in improving detection accuracy and reducing false positives.
- To validate system performance on diverse datasets representing real-world social media content.

Research-Hypothesis:

The study hypothesizes that a multimodal AI-based cyberbullying detection system, combining text, image, and video analysis, can achieve higher accuracy and robustness compared to unimodal approaches, thereby enhancing digital safety in online social environments.

II. ABBREVIATIONS AND ACRONYMS

AI - ARTIFICIAL INTELLIGENCE

NLP - NATURAL LANGUAGE PROCESSING CV - COMPUTER VISION

DL - DEEP LEARNING

API - APPLICATION PROGRAMMING INTERFACE LLM-LARGE LANGUAGE MODEL

UI-USER INTERFACE UX-USER EXPERIENCE DB-DATABASE

HTTP-HYPERTEXT TRANSFER PROTOCOL GPU-GRAPHICS PROCESSING UNIT

ML-MACHINE LEARNING

III. PROPOSED METHADOLOGY

A. SYSTEM OVERVIEW

The proposed system is an ai-driven multimodal cyberbullying detection platform capable of analyzing text, images, and videos. It utilizes a modular architecture that integrates natural language processing (NLP), computer vision (CV), and deep learning (DL) models via restful APIS to detect bullying content across different modalities. Each module communicates through APIS to ensure scalability, maintainability, and integration with external applications like social media platforms or content moderation tools.

B. DATA PREPARATION AND PREPROCESSING

- Text data: cleaned using tokenization, stop-word removal, and lemmatization. Labeled text samples are derived from dataasets like hate x plain and cyberbullying detection dataset.
- Image data: preprocessed using open cv (resizing, normalization), and annotated based on visual indicators of cyberbullying.
- Video data: extracted into frames and analyzed using 3d-cnns or frame-level CNN+ RNN combinations to assess bullying context.
- Data storage: preprocessed data is stored in a mongo DB/CSV format, accessible by APIS for real-time inference.

C. MODEL ARCHITECTURE

Each modality is handled by a separate deep learning model, accessible via APIS:

- Text module: fine-tuned transformer model (e.g., BERT or ROBERT A), accessed through /API/analyze-text.
- Image module: CNN-based model (e.g., or efficient net), accessible via /API/analyze-image.
- Video module: 3d CNN or CNN-RNN hybrid, exposed via /API/analyze-video.

These models are trained individually and deployed using flask or fast API as microservices.

D. API INTEGRATION AND FLOW

The detection system operates as a suite of rest APIS. The high-level flow is:

1. Client uploads text/image/video via a unified /API/submit-content endpoint.
2. Content is routed to the respective analysis service:
   - /API/analyze-text for NLP
   - /API/analyze-image for cv
   - /API/analyze-video for video content
3. Each service returns:
   - Classification: bullied / non-bullied
   - Confidence score: 0–100%
   - Threat indicator: BOOLEAN (true/false)
4. API gateway aggregates responses and forwards results to the UI or external app.

All APIs are secured via JWT tokens and https encryption to prevent misuse.

A. AGGREGATION AND DECISION LOGIC

An intelligent fusion layer combines the outputs from each modality using a weighted score mechanism or ensemble learning techniques (e.g., majority voting, stacked model). This step ensures robustness and reduces false positives.

Score=α(text)+β(image)+γ(video)score = α(text) + β(image) + γ(video) score=α(text)+β(image)+γ(video)

Where $\alpha + \beta + \gamma = 1$ (customized based on modality importance).

E. DEPLOYMENT ENVIRONMENT

- Backend: python 3.10+, flask/ fast API

- Model serving: tensor flow/on n x / PY-torch models served via torch serve or tensor flow serving
- Database: MONGODB or POSTGRESQL for user data and logs
- API testing: postman, swagger UI
- Deployment: docker containers deployed on AWS ec2 or local servers
- Security: aes-256 encryption for data, JWT for authentication

F. ETHICAL AND PRIVACY CONSIDERATIONS

- All data is anonymized before training.
- APIs are rate-limited to prevent abuse.
- System logs and predictions are stored securely and periodically reviewed for fairness and bias.

IV. LITERATURE REVIEW

Cyberbullying has become a significant social concern with the rise of social media platforms and user-generated content. Traditional detection approaches have primarily focused on **text-based analysis** using Natural Language Processing (NLP) techniques. These methods include **keyword spotting**, **sentiment analysis**, and **machine learning classifiers** such as SVMs and decision trees to detect abusive or toxic language [1][2]. While these approaches have demonstrated moderate success, they are limited in scope, particularly in capturing **implicit abuse**, **sarcasm**, and **non-verbal cues**.

Recent research has begun to explore the use of **Computer Vision (CV)** and **Deep Learning (DL)** to analyze non-textual content such as images and videos. Techniques like **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)** have been employed to identify offensive visual elements and behavioral cues in multimedia data [3][4]. However, these approaches often treat each modality in isolation, leading to fragmented insights and reduced detection accuracy.

To overcome these limitations, **multimodal learning** has emerged as a promising paradigm. It enables systems to **fuse features** from multiple modalities—text, images, and videos—providing a more **comprehensive understanding** of the content. Studies have shown that **feature fusion techniques** and **ensemble models** improve the robustness and contextual interpretation of cyberbullying detection systems [5].

Additionally, the integration of **cloud-based APIs** like Google Cloud Vision and Speech-to-Text has streamlined real-time multimedia analysis, reducing the complexity of developing end-to-end detection systems. However, concerns remain regarding **latency**, **scalability**, and **privacy** when relying heavily on external services.

Despite these advances, significant challenges persist:

- High **false positive** and **false negative** rates due to a lack of contextual awareness,
- Limited datasets that fail to represent **cultural, linguistic, and behavioral diversity**,
- Difficulty in handling **real-time, multimodal data streams** at scale.

This research addresses these gaps by developing a **unified, AI-driven multimodal framework** that

incorporates NLP, CV, and DL, leveraging **pre-trained models**, **feature fusion**, and **modular APIs** to detect cyberbullying across different media types. By focusing on real-time responsiveness, ethical data handling, and improved accuracy, this approach aims to create a more effective and scalable cyberbullying prevention tool for digital safety.

V. METHADOLOGY

This section outlines the research methods, data collection procedures, analysis techniques, and ethical considerations followed in the development and evaluation of the cyber bullying detection.

5.1 Research Methods

This study follows an experimental research methodology, where a multimodal AI system is designed, developed, and evaluated for cyberbullying detection across text, images, and videos. The system integrates **Natural Language Processing (NLP)** for textual content, **Computer Vision (CV)** for image analysis, and **Deep Learning (DL)** techniques for video processing. A modular architecture is adopted to enable real-time classification using pre-trained models and APIs for processing multimodal inputs.

5.2 Data Collection Procedure

The system is trained and tested using publicly available and ethically sourced datasets that include a wide range of cyberbullying examples:

- **Hate-X-plain Dataset** (annotated for hate and offensive speech in text),
- **Instagram Multimodal Cyberbullying Dataset** (containing images and text pairs),
- **Cyberbullying Detection Dataset** (for textual bullying patterns),
- **UCF-101** and similar video datasets annotated for aggressive behavior..

3.3 Analysis Techniques

Multimodal feature fusion is employed to combine insights from different modalities. The following techniques are used:

Text Analysis: Tokenization, embedding (via BERT/GPT models), and classification using fine-tuned transformers.

Image Analysis: Feature extraction using CNNs (e.g., Res Net, Mobile Net) and object detection for context.

Video Analysis: Frame extraction with temporal modeling using 3D CNNs or LSTM layers.

Fusion & Classification: Late fusion (combining outputs of individual models) and meta-classifiers (e.g., Random Forest, Logistic Regression) to finalize bullying classification. APIs like **Google Vision API**, **Speech-to-Text API**, and **Hugging Face transformers** are used to enhance modularity, performance, and scalability.

3.4 Ethical Considerations

This research was conducted with strict adherence to ethical standards:

- No personal data was collected; all datasets were publicly available and anonymized.
- No content was published or stored without ensuring privacy protection.
- All third-party APIs and datasets used comply with their respective licensing agreements.
- The system processes content locally when possible and avoids cloud-based storage for user data.

VI. RESULTS AND DISCUSSION

This section presents the findings from evaluating the proposed AI-driven multimodal cyberbullying detection system. The analysis covers precision, recall, and F1-score across different content types, response efficiency, and comparison with unimodal approaches. Results are interpreted both quantitatively and qualitatively, and validated against existing literature.

6.1 Evaluation Setup

The system was tested using a hybrid dataset composed of real-world and synthetic examples from social media platforms. The dataset included:

- **Text** posts (from Twitter, Reddit, etc.)
- **Image**-text pairs (e.g., Instagram posts)
- **Video** clips (short form content from YouTube, TikTok) Each data sample was processed through:
- **Text-only model** using NLP (BERT/GPT)
- **Image-only model** using CNNs (ResNet/MobileNet)
- **Video-only model** using temporal models (3D CNN/LSTM)
- **Multimodal fusion model** combining all three Performance was evaluated on:
- **Cyberbullying classification accuracy**
- **Detection robustness (F1-Score)**
- **Processing time per sample**

Modality	Precision	Recall	F1 Score	Avg. Time (sec)	Modality	Precision
Text (NLP)	88.2%	85.4%	86.8%	0.9	Text (NLP)	88.2%
Image (CV)	84.5%	82.1%	83.3%	1.2	Image (CV)	84.5%
Video (DL)	81.6%	78.9%	80.2%	2.5	Video (DL)	81.6%
Multimodal Fusion	**91.3%**	**89.5%**	**90.4%**	**2.7**	**Multimodal Fusion**	**91.3%**

Table 6.1: 1 **Comparison of Detection Performance**

Table 6.1 The performance table compares different modalities—Text (NLP), Image (CV), Video (DL), and Multimodal Fusion— for cyberbullying detection based on precision, recall, F1 score, and processing time. Text-based analysis (NLP) shows strong results with an F1 score of 86.8% and the fastest processing time (0.9 seconds), making it highly effective for analyzing written content like comments or messages. Image analysis (CV) follows with an F1 score of 83.3% and moderate speed (1.2 seconds), useful for detecting visual abuse such as harmful memes. Video analysis (DL) performs the lowest (F1 score of 80.2%) and is relatively slow (2.5 seconds), reflecting the complexity of processing video data. However, Multimodal Fusion—combining text, image, and video inputs—delivers the best overall performance with an F1 score of 90.4%, albeit with the highest processing time (2.7 seconds). This demonstrates that while single-modality systems can be efficient, combining modalities yields superior accuracy for comprehensive cyberbullying detection.

FiguresandTables

Fig6.2:Detection Accuracy Comparison Across Modalities

The bar chart titled "F1-Score Comparison Across Modalities" visually compares the effectiveness of different data types—Text (NLP), Image (CV), Video (DL), and Multimodal Fusion—for cyberbullying detection. Among the modalities, Multimodal Fusion achieves the highest F1-score at 90.4%, indicating superior accuracy when combining text, image, and video inputs. Text (NLP) follows closely with an F1-score of 86.8%, demonstrating strong performance and efficiency in detecting harmful language. Image (CV) performs moderately with an F1-score of 83.3%, suitable for identifying offensive visual content. Video (DL) has the lowest F1-score at 80.2%, reflecting the challenges in accurately analyzing temporal and visual cues in videos. Overall, the chart highlights that while each individual modality contributes valuable insights, fusing them yields the most accurate and robust results for detecting cyberbullying.

6.3 **Sample Detection Output**

Example 1: Text with Visual Sarcasm

- **Input:** “Nice pic. Hope you get more likes next time 😂” (Image shows bullying meme)
- **Text Model Output:** Non-bullying
- **Image Model Output:** Bullying
- **Fusion Output:** Bullying (Correct)

6.4 **Sample Detection Output**

Example 2: Video Clip of Subtle Harassment

- **Input:** Short video showing a student being mocked with gestures
- **Video Model Output:** Bullying
- **Text Model Output (from captions):** Non-bullying
- **Fusion Output:** Bullying (Correct)

Interpretation

These examples highlight the AI system's ability to:

- **Detect sarcasm and contextual bullying using combined cues**
- **Understand non-verbal aggression in video format**
- **Resolve contradictions between modalities through late fusion**

Such capabilities confirm that the proposed AI model goes beyond traditional single-modality tools, offering deeper semantic-level understanding and improved digital safety monitoring.

VII. CONCLUSION

The AI-Driven Multimodal Cyberbullying Detection system effectively leverages advanced machine learning techniques to analyze text, images, and videos for identifying cyberbullying behavior. By integrating multiple data modalities, the system enhances detection accuracy and provides a comprehensive safety solution for digital platforms. This approach not only helps in timely identification and mitigation of harmful content but also supports creating safer online environments. Future work can focus on improving real-time detection capabilities, expanding the range of abusive behaviors detected, and enhancing user privacy during analysis. Overall, this system represents a significant step toward empowering users and organizations to combat cyberbullying with intelligent, scalable technology.

A. Summary of Key Findings

This study successfully developed and evaluated a multimodal AI-driven cyberbullying detection system capable of analyzing text, images, and videos. By integrating Natural Language Processing (NLP), Computer Vision (CV), and Deep Learning (DL) through feature fusion, the system achieved a high F1-score of 90.4%, outperforming single-modality models. It demonstrates enhanced accuracy, reduced false positives, and real-time responsiveness using pre-trained APIs and threading for simultaneous input processing.

B. Implications for Theory and Practice

The results contribute to the growing field of AI ethics and digital safety, showcasing how multimodal analysis can bridge gaps in content moderation across social platforms. The system's design highlights the potential of combining language and visual understanding to detect nuanced cyberbullying instances that may be overlooked in traditional systems. Practically, this can be integrated into social media moderation pipelines or educational tools for early detection and intervention.

C. Limitations of the Study

Despite promising results, the system faces limitations:

- **Dataset Constraints**: Training was limited to publicly available datasets, which may not represent all forms of cyberbullying across cultures and languages.
- **Real-Time Limitations**: While effective, real-time video analysis can increase computational overhead, especially on low- end hardware.
- **Dependence on Pre Trained APIs**: The system relies on external APIs for certain tasks, raising concerns over latency and long-term availability.

VIII. REFERENCES

[1] M. A. Sazzed and M. U. Mahmud, "Cyberbullying detection on social media using multimodal deep learning," IEEE Access, vol. 10, pp. 45781–45792, 2022. DOI: 10.1109/ACCESS.2022.3169442

[2] K. Zhang, C. Xu, and M. Xu, "Multimodal learning for cyberbullying detection," in Proc. 26th ACM Int. Conf. Multimedia, Seoul,
South Korea, 2018, pp. 1483–1491. DOI: 10.1145/3240508.3240649

[3] S. Vidgen, B. Derczynski, and D. Nguyen, "Challenges and frontiers in abusive content detection," in Proc. 28th Int. Conf. Computational Linguistics, Barcelona, Spain, 2020, pp. 5030–5043.

[4] H. Hossein mar di, S. A. Mattson, R. Han, Q. LY, and S. Mishra, "Detection of cyberbullying incidents on the Instagram social network," ARXIV preprint, arXiv:1503.03909, 2015.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[6] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[7] Meta AI, "Hateful Memes Dataset," Facebook AI, 2020. [Online]. Available: https://ai.facebook.com/datasets/hateful- memes/

[8] C. Wang and W. Wang, "Video classification with convolutional neural networks," in Proc. Int. Conf. Image Processing (ICIP), Paris, France, 2014, pp. 1125–1129.

- Incorporate transformer-based fusion models (e.g., CLIP or Flamingo) for better cross-modal understanding.
- Expand datasets to include multilingual, region-specific, and anonymized real-world samples.
- Optimize the system for mobile or edge devices to broaden accessibility.
- Conduct user studies and expert evaluations to assess the system's social and psychological impact.