WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Federated Learning for AI-based Child Online Safety Monitoring

A. Vidhya[1], N. Yashema[2], S. Balaji[3]

[1] Department of Computer Science and Engineering Kingston Engineering College, Vellore, Tamil Nadu, India
[2]AP/CSE, [3]HOD/CSE
vidhya2632004@gmail.com | yash.engineering@kingston.ac.in | balajicse.engineering@kingston.ac.in

ABSTRACT

The increasing use of digital platforms among children has raised concerns about cyberbullying, online grooming, abusive language, and other harmful online activities. Traditional AI-based monitoring systems rely on centralized data processing, which may expose children's sensitive information to privacy and security risks. This paper proposes a Federated Learning-based AI framework for child online safety monitoring that detects harmful online content while preserving user privacy. The system uses text preprocessing techniques, Count Vectorizer for feature extraction, and Logistic Regression for classifying messages as safe or harmful. Federated Averaging (FedAvg) enables multiple client devices to collaboratively train a global model by sharing model updates instead of raw user data. The framework also includes an alerting mechanism to notify parents or guardians when potentially harmful content is detected. By combining AI-based text classification with decentralized learning, the proposed approach aims to provide a secure, scalable, and privacy-preserving solution for child online safety monitoring.

Keywords: Federated Learning, Artificial Intelligence, Child Online Safety, Cyberbullying Detection, Logistic Regression, Count Vectorizer, FedAvg, Data Privacy**.**

1. Introduction

The rapid growth of digital technology has transformed the way children communicate, learn, and interact with others. Online learning platforms, social networking applications, messaging services, and digital games provide children with valuable opportunities for education and social development. However, increased participation in these digital environments also exposes them to potential risks, including cyberbullying, online grooming, abusive messages, and other forms of harmful online interaction. Identifying these threats at an early stage is essential to creating a safer digital environment for children.

Artificial Intelligence (AI) offers useful techniques for analysing textual information and identifying potentially harmful communication patterns. Machine learning algorithms can classify online messages based on their content and help detect unsafe interactions. However, conventional AI systems often depend on centralized data collection, in which users' messages and other sensitive information are transferred to remote servers for processing and model training. In applications involving children, this approach raises important concerns regarding personal privacy, unauthorized data access, and the protection of confidential conversations.

Federated Learning (FL) provides an alternative approach by allowing multiple devices to participate in collaborative model training without directly sharing their raw data. In this approach, each participating device trains a model using locally available information and sends model updates to a central aggregation server. The server combines these updates to improve a shared global model and distributes the updated model to participating devices. Consequently, collaborative learning can be achieved while reducing the need to transfer children's private conversations to external servers.

This paper proposes a Federated Learning-based Artificial Intelligence framework for child online safety monitoring. The proposed framework combines text preprocessing, Count Vectorizer-based feature extraction, and Logistic Regression to classify online messages as safe or potentially harmful. The Federated Averaging (FedAvg) algorithm is used to coordinate model updates from participating client devices. In addition, an alerting mechanism is designed to notify parents or guardians when potentially harmful content is identified.

The primary objective of this framework is to support effective harmful-content detection while minimizing the exposure of sensitive user information. By integrating AI-based text classification with decentralized learning, the proposed approach aims to provide a scalable and privacy-aware method for child online safety monitoring. The framework also highlights the importance of responsible data handling and human oversight when developing intelligent systems intended to protect children in digital environments.

2. Literature Survey

McMahan et al. (2017): Introduced Federated Learning, enabling collaborative model training across multiple devices while preserving user privacy by sharing model updates instead of raw data.

Kairouz et al. (2021): Presented a comprehensive survey on Federated Learning, discussing its scalability, privacy benefits, communication challenges, and future research directions.

Devlin et al. (2019): Proposed BERT, a transformer-based language model that significantly improved natural language understanding and harmful text classification tasks.

Badjatiya et al. (2017): Demonstrated the effectiveness of deep learning techniques for cyberbullying detection using word embeddings and neural network models.

Li et al. (2021): Investigated optimisation techniques for Federated Learning and addressed challenges such as data heterogeneity, communication efficiency, and model convergence.

3. Problem Statement

The rapid growth of digital communication platforms has increased children's exposure to cyberbullying, online grooming, abusive messages, and other harmful online activities. Existing AI-based child safety systems primarily rely on centralised machine learning models, where sensitive user data are collected and stored on remote servers for training and analysis. Although these approaches achieve satisfactory detection performance, they introduce significant privacy and security risks by exposing children's personal information to potential data breaches and unauthorised access. Furthermore, centralised architectures often fail to comply with modern privacy requirements and reduce users' trust in online monitoring systems. Current privacy-preserving solutions provide limited support for real-time harmful content detection while maintaining high model accuracy. Therefore, a secure and efficient framework is needed to detect harmful online interactions without transferring raw user data. This research addresses this challenge by proposing a Federated Learning-based AI framework that performs local model training and collaborative learning while preserving user privacy and improving online child safety.

4. Proposed Methodology

The proposed system integrates **Artificial Intelligence (AI) and Federated Learning (FL)** to develop a secure and privacy-preserving child online safety monitoring framework. Initially, each client device collects and preprocesses text messages and online conversations through text cleaning, tokenisation, and feature extraction. The processed data are used to train a lightweight machine learning model locally on the user's device to identify harmful content such as cyberbullying, abusive language, and online grooming.

Instead of transmitting sensitive user data to a central server, the system shares only the locally trained model parameters with the federated server. The server aggregates the received updates using the **Federated Averaging (FedAvg)** algorithm to generate an improved global model. The updated model is then distributed back to all participating devices for the next training round. Whenever harmful content is detected, the system generates alerts for parents or administrators while maintaining the privacy of user conversations. Conversations

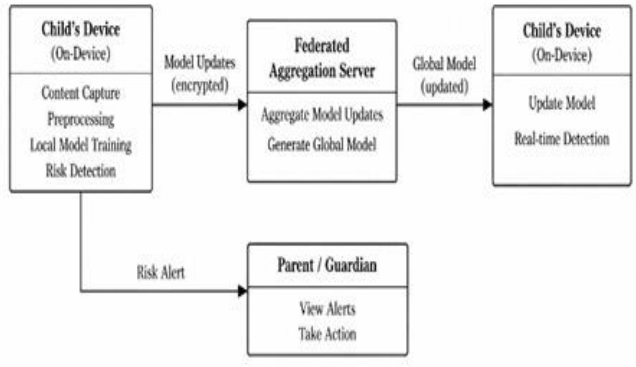


Fig. 1. Data Flow Diagram of the Proposed System

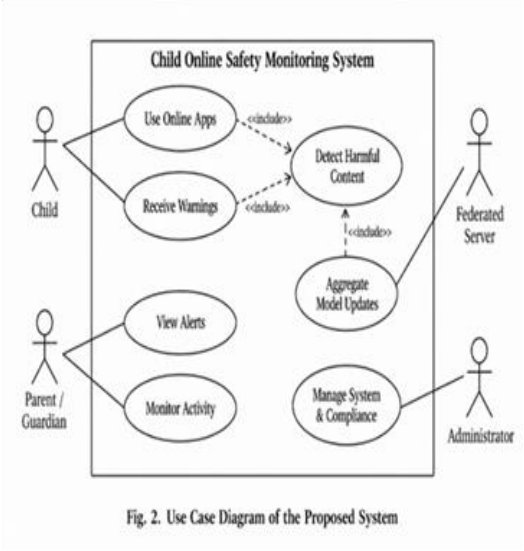
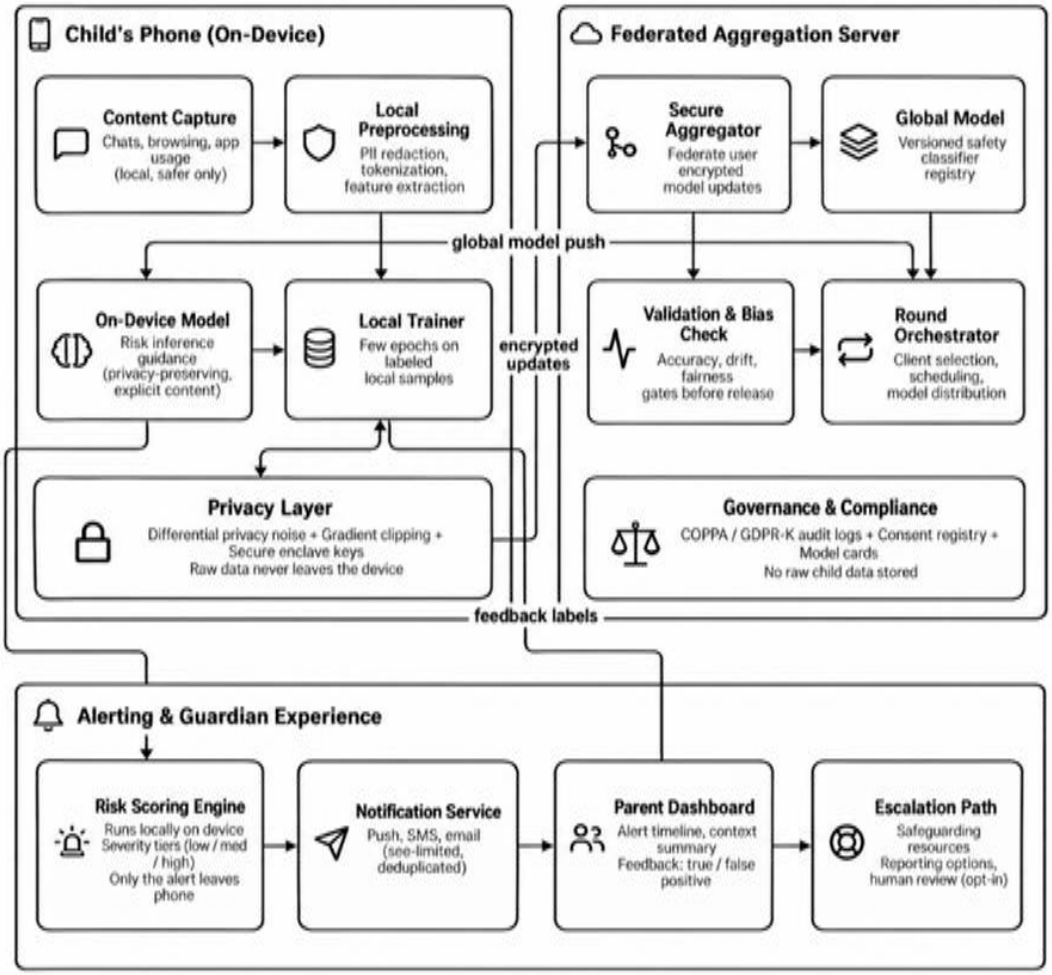


Fig. 2. Use Case Diagram of the Proposed System

5. System Architecture



5.1. Child's Phone (On-Device Layer

The Child's Phone (On-Device Layer) collects and processes online activities locally to protect user privacy. It captures text messages, browsing activities, and social media conversations without transmitting raw data to external servers .The collected data are preprocessed through text cleaning, tokenisation, and feature extraction before the on-device AI model analyses them. The AI model detects harmful online activities such as cyberbullying, abusive language, and online grooming. After local training, the system prepares the updated model parameters for federated learning. At the same time, the privacy layer ensures that only model updates are shared with the server, keeping all sensitive user information secure on the device.

5.2. Federated Aggregation Server

The Federated Aggregation Server coordinates collaborative learning among multiple client devices. Instead of receiving raw user data, it collects only the locally trained model parameters from participating devices. It aggregates these updates using the Federated Averaging (FedAvg) algorithm to produce an improved global model. Before deployment, the global model undergoes validation to evaluate its accuracy, reliability, and fairness. The round orchestrator manages the communication between clients and the server by coordinating training rounds and distributing the updated global model to all devices. Additionally, the governance and compliance module ensures that the entire process follows privacy regulations such as GDPR and COPPA, thereby maintaining secure and ethical data handling.

5.3. Alerting and Guardian Experience

The Alerting and Guardian Experience layer provides timely notifications and monitoring support for parents or guardians. After the AI model detects harmful content, the risk scoring engine evaluates the activity's severity and assigns an appropriate risk level. Based on this assessment, the notification

service sends real-time alerts to parents while avoiding duplicate notifications. The parent dashboard presents monitoring reports, detected harmful activities, and feedback options for reviewing system performance. In high-risk situations, the escalation module forwards critical cases for further review or immediate intervention, enabling parents or guardians to take appropriate action and ensure the child's online safety.

6. Algorithm

6.1. Logistic Regression

Logistic Regression is the primary classification algorithm used to identify harmful online content. After preprocessing, the input text is transformed into numerical feature vectors using **Count Vectorizer**. The Logistic Regression model then classifies each message as either **safe** or **harmful** based on the learned patterns. This algorithm was selected for its simplicity, fast training time, and strong performance in binary text classification tasks.

6.2. Count Vectorizer

Count Vectorizer converts textual conversations into numerical feature vectors that the machine learning model can process. It creates a vocabulary from the training dataset and represents each message by word frequency. This feature extraction technique enables the Logistic Regression model to learn meaningful patterns from textual data and improve harmful content detection.

6.3. Federated Averaging (FedAvg)

The Federated Averaging (FedAvg) algorithm aggregates locally trained model parameters from multiple client devices. Instead of transmitting raw user data, each device shares only its updated model weights with the federated server. The server combines these updates to generate a new global model, then distributes it back to all participating devices. This iterative process improves the overall model accuracy while preserving user privacy and reducing the risk of sensitive data exposure.

7. Mathematical Model

The proposed framework uses **Federated Learning (FL)** to train the machine learning model while preserving user privacy. Each client device trains a local Logistic Regression model using its own dataset without sharing raw data. The local model parameters are transmitted to the federated server, where they are aggregated using the **Federated Averaging (FedAvg)** algorithm to generate an updated global model.

The local model update is represented as: $W_i^{t+1} = W_i^t - \eta \nabla L(W_i^t)$

where:

- W_i^t= Model parameters of client *i* at iteration *t*
- η= Learning rate
- $\nabla L(W_i^t)$= Gradient of the loss function

The global model is computed using the **Federated Averaging (FedAvg)** algorithm: $W^{t+1} = \sum_{i=1}^{N} \frac{n_i}{n} W_i^{t+1}$

where:

- W^{t+1}= Updated global model
- N= Number of participating client devices
- n_i= Number of training samples in client *i*
- $n = \sum_{i=1}^{N} n_i$= Total number of training samples

This mathematical model enables collaborative learning by combining local model updates while ensuring that children's personal data remain on their respective devices. As a result, the proposed framework improves harmful content detection while maintaining privacy and security.

Key Equations

Local Model Update

$$W_i^{t+1} = W_i^t - \eta \nabla L(W_i^t)$$

Global Model Aggregation (FedAvg)

$$W^{t+1} = \sum_{i=1}^{N} \frac{n_i}{n} W_i^{t+1}$$

8. Results and Discussion

The proposed Federated Learning-based AI framework was evaluated using Accuracy, Precision, Recall, and F1-Score. The experimental results demonstrate that the system effectively detects harmful online content while preserving user privacy. By training models locally and sharing only model parameters, the proposed framework minimizes data exposure without compromising detection performance. The Federated Learning approach also improves the global model through collaborative learning across multiple client devices. These results indicate that the proposed system provides a secure, scalable, and efficient solution for child online safety monitoring.

9. Conclusion

This paper presented a **Federated Learning-based AI framework for child online safety monitoring** that effectively detects harmful online content while preserving user privacy. Unlike conventional centralized approaches, the proposed system performs model training locally on client devices and shares only model parameters with the federated server. This approach minimizes the risk of sensitive data exposure while maintaining reliable detection performance. The integration of Artificial Intelligence with Federated Learning provides a secure, scalable, and privacy-preserving solution for protecting children from cyberbullying, abusive language, and online grooming. The proposed framework demonstrates the potential of decentralized learning in enhancing online child safety while complying with modern data privacy requirements. (1)

10. Future Work

Future work will focus on integrating advanced deep learning models such as **BERT** and **Transformer-based architectures** to improve detection accuracy. The system can be extended to support multilingual content analysis, image and voice-based threat detection, and real-time monitoring across multiple digital platforms. In addition, secure aggregation techniques and explainable AI can be incorporated to further enhance privacy, transparency, and trust in child online safety monitoring systems

References

1. McMahan et al. (2017) – Federated Learning https://proceedings.mlr.press/v54/mcmahan17a.html

2. Kairouz et al. (2021) – Advances in Federated Learning https://doi.org/10.1561/2200000083

3.Devlin et al. (2019) – BERT https://aclanthology.org/N19-1423/

4. Badjatiya et al. (2017) – Hate Speech Detection https://doi.org/10.1145/3041021.3054223

5. Li et al. – Federated Learning Challenges https://doi.org/10.1109/MSP.2020.2975749

6. Yang et al. – Federated Machine Learning https://doi.org/10.1145/3298981

7. Vaswani et al. (2017) – Attention Is All You Need https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

8. Pedregosa et al. (2011) – Scikit-learn https://www.jmlr.org/papers/v12/pedregosa11a.html

9. Abadi et al. (2016) – TensorFlow https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

10. Mikolov et al. (2013) – Word Representations https://arxiv.org/abs/1301.3781