Federated Learning for Early Mental-Health Detection Using Social Media Text

MD Mazharul Islam Nabil
Department of Computer Science
American International University-Bangladesh
Email: 23-50025-1@student.aiub.edu

Abed Rahman Bhuiyan
Department of Computer Science
American International University-Bangladesh
Email: 23-50144-1@student.aiub.edu

MD. Mehedi Hasan Anik
Department of Computer Science
American International University-Bangladesh
Email: 22-48937-3@student.aiub.edu

Wasimul Bari Rahul
Department of Computer Science
American International University-Bangladesh
Email: 23-50913-1@student.aiub.edu

I. Literature Review

Artificial Intelligence (AI) is becoming increasingly important in mental healthcare, and the field is moving away from storing all data in one central place toward using decentralized methods that protect user privacy. Wajid et al. [1] reviewed current AI practices in mental health and observed that although centralized models usually achieve strong accuracy, they also create serious risks because all private data is kept in one location. This problem becomes even more serious in academic environments. Ebrahimi et al. [2] explained that placing sensitive student mental health information on a single server can easily lead to privacy violations. Because of these issues, Federated Learning (FL) has gained attention as a safer alternative. Grataloup and Kurpicz-Briki [3] noted that FL is increasingly replacing older centralized methods because it allows training to happen directly on users' devices without sharing raw data.

Recent research shows many successful uses of FL for structured and sensor-based data. Khan et al. [4] proposed *FedEmo*, which uses IoMT devices to recognize emotions from handwriting, proving that biometric data can be processed in a federated way. Dubey et al. [5] extended this idea by combining physiological signals with clinical text, demonstrating that using several types of data together improves results compared to using only one. Huremagic et al. [6] developed *MINDDS-connect*, a system that supports FL on large genetic datasets. However, these studies focus mostly on structured or stable data types and do not address the rapidly changing and unstructured nature of social media text.

Using FL for Natural Language Processing (NLP) introduces several unique challenges, especially because people write in different styles and languages. Khan et al. [7] showed that non-IID data—where each user writes very differently—makes it harder for the shared global model to learn properly. Even so, Khalil et al. [8] proved that multilingual models can still detect depression patterns federatedly across different user groups. Ahmed et al. [9] took this further by creating an on-device FL method for Reddit posts, making the system run quickly on mobile devices. Nevertheless, strong centralized models such as *DABLNet* (Saeed et al. [10]) and *SMILE* (Pesqueira et al. [11]) still perform better because they use large global attention mechanisms that are hard to reproduce in fully private FL setups. Alonge et al. [12] attempted to solve this issue using secure federated BERT models, but these models still require heavy computation, which limits their use in real-time systems.

FL also does not guarantee full security. Latif et al. [13] pointed out that federated networks can be attacked by harmful users, and they proposed neural network–based Intrusion Detection Systems (IDS) to detect such threats. For text-generation tasks, Wong et al. [14] introduced a defense method that protects next-word prediction systems from backdoor attacks, which is especially important for mental health chatbots. Privacy leakage remains another concern. Murala et al. [15] created *MedShieldFL*, which reduces leakage by grouping similar clients for aggregation. However, Vu et al. [16] argued that common Differential Privacy (DP) techniques are not detailed enough for text data because important emotional clues may disappear. Shenoy et al. [17] supported this claim, noting that normal privacy metrics do not fully capture risks related to writing style.

Personalization has become an essential focus because mental health data varies widely between individuals. Pirmani et al. [18] showed that personalized FL methods work better than universal global models for predicting clinical outcomes. Mateus et al. [19] extended this idea by showing that FL can handle data from different hospital groups. For practical situations such as student mental health monitoring, Bakirova et al. [20] highlighted the need for lightweight models, and Salman et al. [21] explained how Knowledge Distillation (KD) can shrink large models so they work more efficiently on local devices. Another key concern is model transparency. Ibrahimov et al. [22] and Hameed et al. [23] emphasized that Explainable AI (XAI) is necessary for earning clinical trust. Yet, according to Ducange et al. [24], XAI methods for

federated NLP tasks are still rare.

Overall, the literature shows a clear research gap. Although strong FL systems exist for sensor data (such as *FedEmo*) and structured medical records (such as *MedShieldFL*), there is still no complete framework designed specifically for social media text that can provide: (1) fine-grained privacy to protect writing style, (2) protection against backdoor attacks, (3) personalized learning using distillation for different writing habits, and (4) explainability for clinical evaluation. Zenk et al. [25] also pointed out that fair benchmarking for decentralized text classification remains missing. This research aims to fill these gaps by designing a secure, personalized, and interpretable FL system focused on early mental health detection from social media posts.

II. Methodology

A. Overview of the Proposed Framework

This research introduces a privacy-preserving Federated Learning (FL) framework designed specifically for early depression detection on resource-constrained edge devices. The system architecture is decentralized, consisting of a central aggregation server and a network of K distributed clients. Unlike traditional centralized learning, where sensitive user data is uploaded to a cloud server, our framework retains all personal social media data locally on the client's device.

We formally define the problem as a decentralized optimization task. The global objective is to minimize a loss function $\mathcal{L}(w)$ without exposing the local datasets $\mathcal{D}_k$ residing on client devices. The system iterates between local training on private data and global aggregation of model updates Δw, addressing the critical trade-off between privacy preservation and diagnostic accuracy.

The complete system architecture, illustrating the interaction between the data preparation phase, local client training, and server aggregation, is visualized in Fig. 1.

B. Data Acquisition and Partitioning

To validate the framework within a controlled experimental environment, we utilize the **Twitter Depression Dataset** [26], a widely recognized benchmark for linguistic analysis in mental health. The dataset comprises archival microblogging posts labeled for depressive and non-depressive content.

For this study, we curate a balanced corpus of $N = 10,000$ distinct samples (5,000 Depressive, 5,000 Control), denoted as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents the textual input and $y_i \in \{0, 1\}$ is the binary label. To simulate a non-IID (Independent and Identically Distributed) federated environment, the dataset is partitioned into $K = 10$ disjoint subsets $\mathcal{D}_1, \ldots, \mathcal{D}_K$, such that $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$. Each client represents a hypothetical user with unique linguistic patterns, ensuring the model is tested against realistic data heterogeneity.

C. Data Preprocessing Pipeline

Raw social media text contains significant noise that can degrade model performance. We implement a four-stage preprocessing pipeline strictly following the protocols outlined in recent NLP studies [27]:

1) **Text Cleaning:** We remove platform-specific artifacts including HTML tags, URLs, and user mentions (e.g., "@username"), as these tokens often contain no sentiment value and increase feature sparsity.
2) **Normalization:** All characters are converted to lowercase. English contractions are expanded (e.g., "I'm" $\rightarrow$ "I am") to ensure consistent token representation.
3) **Stopword Retention:** Unlike general topic modeling, we retain stopwords (e.g., "I", "myself"), as previous psychological research indicates that depressed individuals use self-referential pronouns at a significantly higher frequency.
4) **Sequence Formatting:** We employ a Keras-based Tokenizer with a fixed vocabulary size of $V = 10,000$. Each post x_i is converted into a sequence of integer indices $S_i = [s_1, s_2, \ldots, s_L]$, where $L = 100$ is the fixed sequence length. Padding is applied where $|S_i| < L$ to ensure compatibility with the fixed-size tensor inputs required by the neural network.

D. Feature Engineering Strategy

We employ a hybrid feature extraction mechanism that combines deep semantic learning with explicit emotional quantification.

1) Learnable Semantic Embeddings: The integer-encoded tokens are passed through a trainable **Embedding Layer**. This layer projects each token index $s_t \in S_i$ into a dense vector space $\mathbb{R}^{d_{emb}}$, where $d_{emb} = 100$. During training, this layer learns to map semantically similar words (e.g., "sad" and "hopeless") to proximal points in the vector space.

2) Lexical Emotion Vectors: To guide the model towards clinically relevant signals, we augment the embeddings with the **NRC Emotion Lexicon** [28]. For every input post x_i, we compute a normalized 8-dimensional vector $v_{emo} \in \mathbb{R}^8$ representing the intensity of basic emotions:

$$v_{emo} = \frac{1}{|x_i|} \sum_{w \in x_i} \mathbf{1}_{emo}(w) \tag{1}$$

where $\mathbf{1}_{emo}(w)$ is a binary indicator vector for emotions (anger, fear, anticipation, trust, surprise, sadness, joy, disgust). This auxiliary feature vector is concatenated with the deep learning representation in the final stages of the network.

E. Deep Learning Model Architecture

To ensure the system allows for deployment on mobile and edge devices, we implement a lightweight **Bidirectional Long Short-Term Memory (BiLSTM)** network. A 2024 comparative study by Zhang et al. [29] demonstrated that BiLSTM architectures achieve detection accuracy comparable to Transformer models (e.g., BERT) while requiring less than 10% of the computational resources.

The architecture processes the input sequence $X = (x_1, \ldots, x_T)$ in two directions. For a given time step t, the forward LSTM unit computes the hidden state $\vec{h}_t$ using the following gate equations:

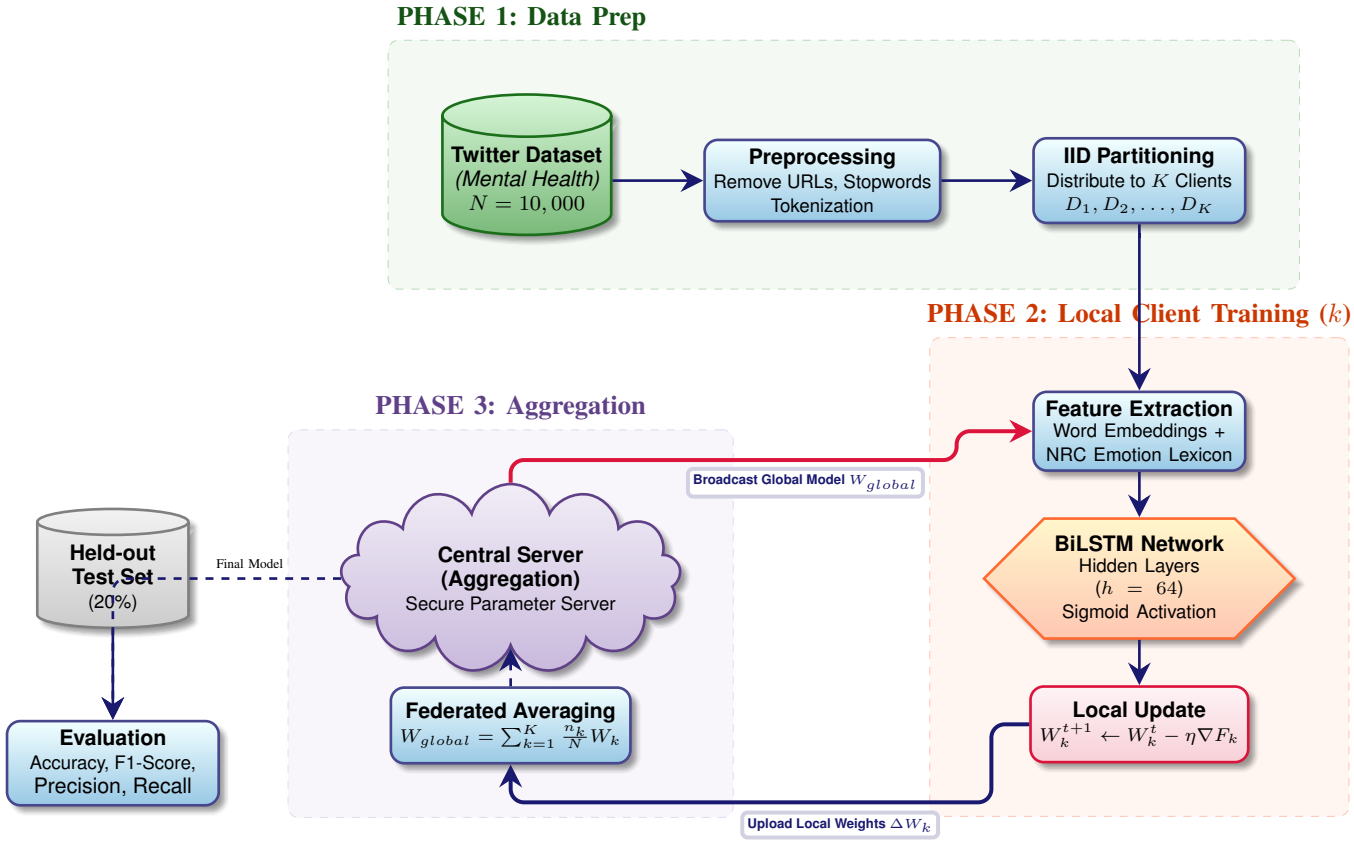

Fig. 1. Proposed Federated Learning Architecture for Mental Health Detection. The system partitions user data locally, trains BiLSTM models on embeddings and emotion features, and securely aggregates weights via Federated Averaging.

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [\vec{h}_{t-1}, e_t] + b_f) \\
i_t &= \sigma(W_i \cdot [\vec{h}_{t-1}, e_t] + b_i) \\
\tilde{C}_t &= \tanh(W_C \cdot [\vec{h}_{t-1}, e_t] + b_C) \\
C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\
\vec{h}_t &= \sigma(W_o \cdot [\vec{h}_{t-1}, e_t] + b_o) \odot \tanh(C_t)
\end{aligned} \tag{2}
$$

where σ is the sigmoid function, $\odot$ denotes element-wise multiplication, and f_t, i_t, o_t represent the forget, input, and output gates.

A simplified backward pass computes $\overleftarrow{h}_t$ by processing the sequence from $t = L$ to 1. The final representation is the concatenation of the last hidden states from both directions:

$$h_{final} = [\vec{h}_L \oplus \overleftarrow{h}_1] \tag{3}$$

This vector is passed through a Dropout layer ($p = 0.5$) to prevent overfitting and a final Dense layer with a Sigmoid activation function to output the probability of depression $\hat{y} \in [0, 1]$.

F. Federated Learning Implementation

We implement the standard **Federated Averaging (FedAvg)** algorithm [30]. The training procedure iterates through the following steps for T communication rounds:

1) **Global Broadcast:** In round t, the central server broadcasts the global model weights W_G^t to all active clients.
2) **Local Training:** Each client k loads W_G as their starting point and trains on their local partition $\mathcal{D}_k$ for $E = 1$ epoch using the Adam optimizer. This produces a local weight update W_k.
3) **Secure Aggregation:** Clients upload their encrypted updates to the server. The server aggregates these updates using a weighted average based on the number of samples n_k held by each client:

$$W_G^{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{N} W_k^t \tag{4}$$

This cycle repeats until the global loss function converges or the maximum number of rounds is reached.

G. Experimental Evaluation Metrics

The system's performance is rigorously evaluated on a held-out test set comprising 20% of the data, which is never seen during the training phase. We utilize four standard metrics defined as follows, where TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives:

- **Precision:** $P = \frac{TP}{TP+FP}$
- **Recall:** $R = \frac{TP}{TP+FN}$

- **F1-Score:** The harmonic mean of Precision and Recall, prioritized for its robustness in medical diagnosis:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{5}$$

- **Accuracy:** $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

REFERENCES

[1] A. Wajid, F. Azam, and M. S. Anwar, "Applications of artificial intelligence in mental health: a systematic literature review," *Discover Artificial Intelligence*, vol. 5, p. 332, 2025.

[2] M. Ebrahimi, R. Sahay, S. Hosseinalipour, and B. Akram, "The transition from centralized machine learning to federated learning for mental health in education: A survey," *IEEE Access*, vol. 12, pp. 45 120–45 145, 2024.

[3] A. Grataloup and M. Kurpicz-Briki, "A systematic survey on the application of federated learning in mental state detection and human activity recognition," *Frontiers in Digital Health*, vol. 6, p. 135, 2024.

[4] Z. A. Khan, Y. Xia, and W. Jiang, "FedEmo: A federated learning framework for privacy-preserving emotion detection from handwriting on consumer IoMT devices," *IEEE Internet of Things Journal*, vol. 12, no. 4, pp. 3400–3412, 2025.

[5] P. Dubey, P. Dubey, and P. N. Bokoro, "Federated learning for privacy-enhanced mental health prediction with multimodal data integration," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 13, no. 1, p. 19822, 2025.

[6] S. Huremagic, D. Patel, and S. O'Connor, "MINDDS-connect: a federated data platform integrating biobanks for meta cohort building and analysis," *European Journal of Human Genetics*, vol. 33, pp. 1539–1546, 2025.

[7] Y. Khan *et al.*, "Federated learning-based natural language processing: a systematic literature review," *Artificial Intelligence Review*, vol. 57, no. 12, p. 320, 2024.

[8] S. S. Khalil, N. S. Tawfik, and M. Spruit, "Federated learning for privacy-preserving depression detection with multilingual language models in social media posts," *Patterns*, vol. 5, no. 7, p. 100990, 2024.

[9] M. Ahmed, A. Muntakim, N. Tabassum, and M. A. Rahim, "On-device federated learning in smartphones for detecting depression from reddit posts," *arXiv preprint*, 2024, arXiv:2410.xxxxx.

[10] Q. B. Saeed, S. Ram, and E. Park, "Multi-modal deep-attention-BiLSTM based early detection of mental health issues using social media posts," *Scientific Reports*, vol. 15, no. 1, p. 35152, 2025.

[11] A. Pesqueira, M. J. Sousa, and R. Pereira, "Designing and implementing SMILE: An AI-driven platform for enhancing clinical decision-making in mental health," *Computational and Structural Biotechnology Journal*, vol. 24, pp. 102–115, 2025.

[12] M. Alonge *et al.*, "Secure federated sentiment analysis for mental health support systems using BERT," *ResearchGate Preprint*, 2025, available at ResearchGate.

[13] N. Latif, W. Ma, and H. B. Ahmad, "Advancements in securing federated learning with IDS: a comprehensive review of neural networks," *Discover Computing*, vol. 22, no. 3, pp. 45–62, 2025.

[14] J. K. W. Wong, K. K. Chung, and Y. W. Lo, "Practical implementation of federated learning for detecting backdoor attacks in a next-word prediction model," *Scientific Reports*, vol. 15, no. 1, 2025.

[15] D. K. Murala *et al.*, "MedShieldFL: A privacy-preserving hybrid federated learning framework for intelligent healthcare systems," *Scientific Reports*, vol. 15, no. 1, p. 43144, 2025.

[16] D. N. L. Vu *et al.*, "Granularity is crucial when applying differential privacy to text," *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1200–1215, 2024.

[17] D. Shenoy, T. Nguyen, and P. Wright, "Exploring privacy mechanisms and metrics in federated learning: A survey," *Artificial Intelligence Review*, vol. 58, no. 4, pp. 11 170–11 205, 2025.

[18] A. Pirmani, E. De Brouwer, and Á. Arany, "Personalized federated learning for predicting disability progression in multiple sclerosis using real-world routine clinical data," *npj Digital Medicine*, vol. 8, no. 1, p. 478, 2025.

[19] P. Mateus *et al.*, "Multi-cohort federated learning shows synergy in health outcomes," *Artificial Intelligence in Medicine*, vol. 148, p. 102755, 2025.

[20] G. Bakirova, G. Bektemyssova, and N. B. Ali, "Federated machine learning for monitoring student mental health in kazakhstan," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 16, no. 10, pp. 212–220, 2025.

[21] H. Salman, C. Zaki, and A. Nasser, "Knowledge distillation in federated learning: a comprehensive survey," *Discover Computing*, vol. 24, p. 112, 2025.

[22] Y. Ibrahimov, T. Anwar, and T. Yuan, "Explainable AI for mental disorder detection via social media: a survey and outlook," *arXiv preprint arXiv:2401.09876*, 2024.

[23] S. Hameed *et al.*, "Explainable AI-driven depression detection from social media using NLP and black-box models," *Frontiers in Artificial Intelligence*, vol. 8, p. 123, 2025.

[24] P. Ducange *et al.*, "Federated learning of XAI models in healthcare: a case for interpretable models for disease progression," *Machine Learning and Knowledge Extraction*, vol. 6, pp. 220–240, 2024.

[25] M. Zenk, L. Ochoa, and I. Svensson, "Towards fair decentralized benchmarking of healthcare AI algorithms with the federated tumor segmentation (FeTS) challenge," *Nature Communications*, vol. 16, no. 1, p. 60466, 2025.

[26] G. Shen *et al.*, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 3838–3844.

[27] C. Sweeney *et al.*, "Text-based depression prediction on social media using machine learning: Systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 27, p. e59002, 2025.

[28] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[29] Y. Zhang *et al.*, "A novel improved bilstm method for depression detection on social media," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 1066–1072.

[30] B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," *Artificial intelligence and statistics*, pp. 1273–1282, 2017.