DOI: doi.org/10.47363/JAICC/ICMLAIDS2026/2026(5)7

International Conference on Machine Learning, Artificial Intelligence and Data Science (ICMLAIDS 2026)

Conference Proceedings March 20 - 21, 2026 - Virtual

Privacy-Preserving Mental Health Analysis on Social Media Using Federated Deep Learning and Named Entity Recognition

Ujunwa Madububa Mbachu

Assistant Professor of Computer & Cyber Engineering, Mississippi, USA

Abstract

Background
The increasing prevalence of mental health disorders, alongside the ubiquity of social media platforms, offers a unique opportunity for early detection of psychological distress through the analysis of user-generated content. However, such analysis raises serious ethical and privacy concerns due to the sensitive nature of mental health data and the potential exposure of personally identifiable information (PII).

Objective
This study proposes a privacy-preserving framework that combines Federated Learning (FL), transformer-based language models, and Named Entity Recognition (NER) to analyze social media content for mental health indicators specifically depression, anxiety, and stress without compromising user privacy.

Methods
We developed a federated learning architecture that integrates pre-trained transformer models RoBERTa, BERT, and DistilBERT alongside a custom NER module designed to detect and mask PII in text data. All model training was conducted locally on user devices, with only encrypted model updates shared via secure aggregation protocols. The system was tested on Reddit and Twitter datasets annotated for mental health-related posts. A rigorous evaluation pipeline was employed: the data were split into a 70/15/15 train/validation/test ratio, with 10-fold stratified cross-validation applied during model training. The test set remained untouched throughout the tuning phase. To enhance generalizability, we included external validation using a holdout secondary dataset from kaggle. Training optimization incorporated early stopping, dropout, and hyperparameter tuning to optimize the F1 score. Model diagnostics included learning curves and confusion matrices. Furthermore, development aligned with the U.S. FDA's Good Machine Learning Practice (GMLP) standards to ensure model safety, efficacy, and clinical relevance.

Results
The proposed framework achieved performance comparable to centralized learning methods while significantly enhancing user privacy. RoBERTa achieved the best overall results, with an F1- score of 0.87, accuracy of 89%, and AUC-ROC of 0.91. BERT followed with an F1-score of 0.84 and AUC-ROC of 0.88, while DistilBERT scored 0.80 F1 and 0.85 AUC-ROC. The NER module effectively masked over 92% of PII entities, contributing to ethical compliance and reducing re-identification risk. Learning curves showed stable training dynamics, and confusion matrices demonstrated strong class-specific separation, particularly in depression detection. The external dataset results further validated the model's robustness under data drift.

Conclusions
This study demonstrates the viability of combining federated learning, transformer-based NLP models, and NER to build an ethical, scalable, and privacy-preserving system for mental health analysis on social media. By maintaining strong predictive performance while minimizing privacy risks, this approach offers a practical pathway to deploying real-time mental health monitoring tools in both research and clinical support settings. Future work will explore real-time deployment, multimodal signal integration, and partnerships with mental health platforms for broader societal.