# FedMental: Evaluating Federated Learning for Mental Health Detection from Social Media Data

**Nuredin Ali Abdelkadir**[1], **Anjali Ratnam**[1], **Zeerak Talat**[2], **Stevie Chancellor**[1]

[1]University of Minnesota
{ali00530,ratna038,steviec}@umn.edu
[2]University of Edinburgh
ztalat@ed.ac.uk

## Abstract

Social media text data are often used to train Machine Learning (ML) models to identify users exhibiting high-risk mental health behaviors. However, sharing this sensitive data poses privacy risks and limits the growth of benchmark datasets. We comprehensively evaluate whether privacy-preserving ML techniques can enable safer data sharing while preserving performance. Specifically, we apply federated learning (FL) and Differentially Private FL for two widely-studied mental health prediction tasks: depression detection on X (Twitter) and suicide crisis detection on Reddit. We simulate realistic data-sharing scenarios by treating each user as a client in a non-IID setting, evaluating across different client fractions, aggregation strategies, and privacy budgets. While FL achieves comparable performance to centralized training (centralized $F1 = 85.63$; best FL model $F1 = 83.16$) on depression identification, we find that Differentially Private FL has a large performance-privacy trade-off (up to $F1 = 27.01$ drop) even with low levels of noise ($\epsilon = 50$). This is due to the distortion of highly informative yet sparse mental health linguistic markers related to mental health, like health topics and emotion words. This research empirically demonstrates the potential and limitations of current privacy preservation techniques for mental health inference tasks.

## 1 Introduction

In the last decade, social media has emerged as an important source of data for predicting mental health conditions (De Choudhury and De, 2014; Yang et al., 2024) such as early detection of many mental illnesses (Coppersmith et al., 2016; Benton et al., 2017), inference about support in online communities (Pendse et al., 2019), and crisis interventions (Teague et al., 2022). Approaches to these tasks have explored the use of data across modalities, including text, visual data, and interaction/engagement metadata (Lin et al., 2020, 2017). Stigma and risk management around mental health disclosures inhibit honest disclosure online, and data can be difficult to source (Chancellor et al., 2023; Fiesler and Proferes, 2018). Yet, data which contains personal information, particularly as it pertains to healthcare is protected data, and is regulated in the European Union and United Kingdom (European Parliament and Council of the European Union; ICO, 2024). Such privacy concerns apply in research as well: a vast majority of users in public mental health datasets are not properly de-identified (Ayers et al., 2018).

Data protection protocols like Ethics Review Boards and Data Use Agreements help protect user data, but shift responsibility onto data controllers, and present institutional barriers to researchers who do not have access to institutional review (Zirikly et al., 2019; Shing et al., 2018a; Coppersmith et al., 2015). This also makes lapses in enforcement risks, exposing sensitive user data. Thus, benchmark datasets in this area are hard to find and can be small (Chancellor and De Choudhury, 2020), limiting the growth of benchmark datasets. Yet methods for ensuring the privacy of users' data in this domain are vital but remain under-explored.

In this work, we comprehensively evaluate two data privacy strategies for the most common tasks in mental health inference in English: depression detection on X (Twitter) and suicidal thought and behavior (STB) detection on Reddit. First, we evaluate Federated Learning (FL), where a global model is trained by distributing computation onto client devices, which train the model locally and send parameter updates, rather than raw data (i.e., text) to the global model. However, while FL can provide additional layers of privacy compared to centralized learning, it does not provide *guarantees* of privacy. Thus, we evaluate the combination of FL with differential privacy (DP) to study the utility-privacy tradeoffs. DP adds calibrated noise to limit inferences about individual users.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40929–40944

We find that the Federated setting is promising for these prediction tasks in both performance loss and efficiency; however, adding DP to FL creates an untenable utility-privacy tradeoff due to distortions of important yet sparse linguistic features. For FL, there are no statistically significant differences in performance between FL and centralized model; FL models also require less data to obtain comparable performance. However, in comparison to the federated setting, the DP-FL models result in a large performance-privacy trade-off (F1=27.01 drop) even with low levels of noise ($\epsilon = 50$).

Our results suggest that FL is a viable option; however, mathematically ensuring privacy using DP may not be practical due to significant performance drops and the negative impact of errors in mental health inference.

## 2 Related Work

**Mental Health Prediction.** ML has been applied to mental health prediction from social media to identifyhigh-risk mental health behaviors and disorders. ML techniques have been used to predict behaviors and disorders at a post or user level (AlSagri and Ykhlef, 2020; Aldarwish and Ahmad, 2017; Islam et al., 2018; Govindasamy and Palanichamy, 2021). These methods rely on data across different modalities such as text, visual data, and interaction/engagement metadata (Lin et al., 2020, 2017), and have recently explored language models pre-trained on mental health corpus (Ji et al., 2021, 2023). These works rely on a centralized training approach, where users' mental health disclosure data is shared with researchers or scraped and labeled from social media.

**Federated Learning.** Federated learning is a decentralized training mechanism for machine learning where client (edge) devices collaboratively train a shared global model (Li et al., 2020a; McMahan et al., 2016) while avoiding the transfer of raw data (e.g., text). Within the FL training paradigm, clients contribute to training the global model by first receiving a copy of the model, which is then trained on locally on the client device on locally held data. Finally, clients share the local model updates with the global server (McMahan et al., 2016). The collected updates are then aggregated using algorithms such as Federated Averaging (FedAvg, McMahan et al., 2017), FedProx (Li et al., 2020b), FedOPT (Reddi et al., 2020).

Federated learning has been applied in mental health domains with similar risks from different data sources. Khalil et al. (2024) surveyed 16 papers on federated learning within psychiatric tasks, the most common being depression detection, specifically used to predict durations of hospitalization using Electronic Health Records (Pfohl et al., 2019), detect depression using wearable sensors (Aminifar et al.; Wang et al., 2024; Gupta et al., 2024), or self-reported assessment data sources (Kuang et al., 2025). Others used mobile device data (keystrokes and accelerometer values) and clinical surveys to predict mood, and found that training on the IID setting yields better performance (Xu et al., 2021).

Using social media as a data source, federated approaches to mental health detection have also been explored (Vasconcelos et al., 2023; Basu et al., 2021; Liu, 2024; Ji et al., 2019). However, prior work has either focused on either IID settings or classifying individual posts. For instance, Vasconcelos et al. (2023) applied a federated approach to the eRisk depression dataset to predict whether a post is labeled as depression or control. Basu et al. (2021) explored federated settings in both IID and non-IID data distribution using BERT-based models to predict depression and sexual harassment posts. Liu (2024) uses federated learning on cross-platform and multilingual social media data to predict depression. Finally, Ji et al. (2019) trains a federated CNN and LSTM classifiers and proposes an advanced optimization scheme for data protection learning framework (AvgDiffLDP) for predicting suicidal ideation on Reddit in an IID setting. Data privacy concerns are underexplored in these domains in a few ways. First, these works lack evaluating data privacy considerations holistically (e.g., modeling the tradeoffs between privacy preservation and performance across various client fractions and aggregation algorithms). Second, most of these approaches do not use a naturalistic non-IID setting of FL representing each user with their post history as a separate client.

**Differential Privacy.** While federated learning affords an additional layer of data protection compared to centralized training, which requires sharing raw data, differential privacy is a mechanism for obtaining privacy through the addition of calibrated noise to the training process (Shan et al., 2024). The addition of the calibrated noise seeks to mitigate the identification of any individual data point while maintaining global aggregates and pat-

terns, which can be used for training models under mathematical guarantees of privacy. Thus, where federated learning affords privacy of data by not sharing raw data, differential privacy maintains privacy by adding noise to each data point, and can be used independently of federated settings.

Prior work at the intersection of federated learning and differential privacy has investigated different aspects of machine learning for mental health. Basu et al. (2021) have examined the identification of sexual harassment and depression in individual posts; they found that utility degradation is higher in a non-IID than an IID setting, and noise addition has more effect when training on a small dataset size. Recently, Sarwar and Dipta (2025) applied DP fine-tuning exclusively to Low-Rank Adaptation (Hu et al., 2021) to reduce communication and memory consumption.

## 3 Experimental Setup

### 3.1 Datasets

We selected benchmark datasets for high-risk behaviors and disorders, specifically depression and suicidality, from the most popular tasks in MH inference, from X and Reddit. We focus on user-level prediction from users' historical posts from five relevant datasets. These disorders, datasets, and platforms have been examined in prior work (Chancellor and De Choudhury, 2020).

| Disorder | Classes | Train | Validation | Test |
|---|---|---|---|---|
| Depression | Treatment | 1,844 | 263 | 528 |
| | Control | 2,123 | 303 | 608 |
| Suicide | Treatment | 597 | 85 | 171 |
| | Control | 534 | 76 | 153 |

Table 1: Datasets used in our experiments.

#### 3.1.1 Depression

We used three datasets from X, as about 50% of datasets for detecting depression rely on X (Aldkheel and Zhou, 2024). These datasets provide user post histories that enable user-level predictions and are collected using keywords and phrases that disclose depression, such as *"(I'm / I was / I am / I've been) diagnosed with depression."*

**CLPsych.** CLPsych was released with the 2015 CLPsych Shared Task for Depression (Coppersmith et al., 2015). The data is split into treatment (i.e., users who self-disclose diagnoses) and a control group. It is manually annotated to verify the authenticity of the disclosures. The dataset contains up to 3,000 posts for each user, with the self-disclosure posts removed. We use 477 users labeled as Depression, and the 871 control users who do not disclose depression.

**MTL-D.** The Multitask Learning-Depression (MTL-D) dataset (Shen et al., 2017) contains one month of user post history for 1,840 users labeled as depressed and 1,840 labeled as control users. The posts in the dataset contain images and textual data. In our work, we make use of the textual data.

**CCD.** The Cross-Cultural Depression (CCD) dataset (Abdelkadir et al., 2024) contains up to 3,200 posts for 267 users labeled as treatment (i.e., depressed) and 264 labeled as control users. This dataset is manually annotated and samples from seven English-speaking countries that are culturally and geographically diverse.

#### 3.1.2 Suicidal Thoughts and Behaviors (STB)

Reddit has emerged as the predominant data source for ML for identifying suicidal thoughts and behaviors. There are dedicated mental health-support subreddits about this topic, such as r/SuicideWatch, r/selfharm, and r/StopSelfHarm. These subreddits can offer insight into the language of users who display STB and of those who present a high risk of engaging in suicidal behaviors.

Following prior work (Chancellor and De Choudhury, 2020; Shing et al., 2018b), we split users into treatment (at-risk users) and control groups (very low or no risk users).

**C-SSRS.** The Columbia Suicide Severity Rating Scale dataset (C-SSRS) dataset (Gaur et al., 2019) was constructed for predicting suicide risk by categorizing users from mental health fora on Reddit into five groups following the C-SSRS, ranging from least to most concerning: supportive (110 users), indicator (100 users), ideation (170 users), behavior (75 users), and attempt (45 users). In our work, we group users labeled as ideation, behavior, and attempt as our treatment group of STB (290 users), and supportive and indicator (210 users) as our control group.

**UMD-RD.** The UMD Reddit Suicidality dataset (UMD-RD) dataset (Shing et al., 2018a, 2020) is manually labeled to verify that they exhibit genuine STB using four levels of suicide risk: No Risk, Low Risk, Moderate Risk, and Severe Risk. The control group consists of users who did not post

in any mental health-related subreddits while the treatment group are sampled from users who have posted to the *r/SuicideWatch* subreddit. We group users with Moderate (256) and Severe (302) Risk into our treatment group (358 users), and users labeled as No Risk (195) and Low Risk (113) into our control group (308 users).

### 3.2 Preprocessing

We combine the three datasets for identifying depression and the two datasets for identifying STB into a depression dataset and an STB dataset, respectively. We then perform a stratified split of the combined datasets into training (70%), validation (10%), and test (20%) sets.[1] See Table 1 for summaries of the dataset splits. We then preprocess the datasets to remove retweet tokens, username mentions, URLs, and numeric values, and expand English word contractions.

### 3.3 Models

We conducted our experiments with one linear and six transformer-based pre-trained models. Our linear model is a Logistic Regression, chosen for its quick training time, competitive performance, and interpretable predictions (Benton et al., 2017; Jiang et al., 2018; Harrigian et al., 2020). For the transformer-based models, we use the general-purpose architectures BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu, 2019), and their distilled counterparts: DistilBert and DistilRoBERTa (Sanh, 2019), as these models have been used for classification tasks (Aftan and Shah, 2023). Following recent work, we also experiment with two mental health-specific models: MentalBERT (Ji et al., 2021), which has been trained on ~13 million sentences from Reddit subreddits for discussing depression, anxiety, and suicide topics; and MentalLongformer (Ji et al., 2023), which is optimized for longer token sequences..

For all transformer-based models, we replace the pre-trained head of the transformer models with a randomly initialized classification head prior to conducting the training. We train the Logistic Regression model and fine-tune the transformer-based models to classify users as depression or control, and STB or control, respectively, based on their post history. We do not experiment with zero-shot settings for two reasons. First, a zero-shot setting requires transmitting data to a centralized server, thereby foregoing preservation of privacy. Second, prior work reports sub-par performance in zero-shot settings for the mental health domain (see e.g, Yang et al., 2023).

### 3.4 Training Schema

Next, we discuss the centralized and federated learning approach applied to our experiments, applied to the seven models above.

#### 3.4.1 Standardized/Centralized Approach

The centralized setting serves as a baseline. This standard ML approach requires datasets and models to be located in the same location, or on a single server. For all transformer-based models, we replace the pre-trained heads with randomly initialized classification heads. We train our models for 50 epochs and set early stopping to 5 epochs. Please refer to Appendix A.2 for further details on model setup and hyperparameters.

#### 3.4.2 Federated Learning Approach

For FL, we conceptualize client devices as individual users in the datasets. Our setup implies that the labels are not uniformly distributed across clients. However, the number of clients in the control and treatment groups is balanced. Moreover, the user histories that are available vary across clients (see depression (Figures 2 and 3) and STB (Figures 4 and 5) in Appendix A.1). Thus, our models are not trained on independent and identically distributed (IID) data. Prior work (e.g., Gandhi et al., 2022; Xu et al., 2021) has found that IID setups result in higher performances than non-IID setups for federated learning. This affords a realistic use case, in which a person accessing healthcare or using social media for well-being enrolls in a monitoring system, or a user consents to share their data with a research team.

**Training.** We train our models for 100 rounds on the server and train clients for 50 epochs per round. That is, we perform 100 rounds of retrieving client updates and aggregate them on the server, and we train each client model for 50 epochs in each round. We experiment with four different client fractions (c = 10, 30, 50, and 70), which represent the percentage of client devices/users included in the aggregation step on the server in each round (i.e., at $c = 10$, 10% of all clients are randomly sampled for the inclusion of their updates in the global model).[2]

[1] We include validation and test splits into the training set, for datasets in which they are separated.

[2] Selecting clients for inclusion in the aggregation step on the server is an open research area with different benefits

We set the client learning rate to $4e-5$ and the server learning rate to $1e-3$. See Appendix A.2 for full experimental details.

We experiment with three different aggregation algorithms: FedAvg, FedProx, and FedOPT. FedAvg is a common aggregation algorithm and serves as our baseline federated method, taking the weighted average of the received client updates. FedProx and FedOPT address issues of slow convergence with the FedAvg algorithm (Moshawrab et al., 2023). FedProx addresses the poor suitability of FedAvg for heterogeneous data situations by adding a proximal term, while FedOPT introduces separate optimizers for client and server models to introduce adaptive optimizers to federated learning.

In our experiments, we set the proximal constant $\mu = 0.01$ for FedProx and apply a server optimizer for FedOPT. All other hyperparameters are shared across the three algorithms.

### 3.5 Differentially Private FL

For our differentially private FL (DP-FL) setup, we implement a client-level DP mechanism. Each client trains a model locally and computes the model update in the traditional federated learning setup. Prior to transmitting the update vector from client to server, we apply an $\ell_2$-norm clipping to the update vector and add calibrated Gaussian noise. This is adaptively scaled as a factor of the training round. Next, on the server, we evaluate the resulting utility-privacy trade-off across different privacy budgets ($\epsilon$, $\delta$=$1e-5$), where the total accumulated privacy loss is tracked in each round (see Appendix A.2 for further experimental details). We used the best-performing models and hyperparameters from the traditional FL setting as our baseline. Specifically, we fine-tune a BERT model for the depression identification task using the FedProx aggregation methods. To investigate whether a change in the pre-trained model, aggregation algorithm, or the inference task affects the differentially private FL, we conduct additional experiments utilizing MenatlBERT as a model, FedAvg aggregation algorithm across both prediction tasks (depression and suicidal thoughts and behaviors). The primary goal of this investigation is to provide a comprehensive feasibility study of DP-FL under these conditions.

To ensure the reliability of our findings, we conducted stability analyses for representative models across all settings. We report 95% Confidence Intervals (CIs) calculated over five random seeds for these representative configurations in the centralized, standard FL, and DP-FL settings.

## 4 Results

We evaluate the performance of our models using the F1 score and recall. The best-performing model for both depression detection ($F1 = 85.63$, see Table 2) and suicidal ideation ($F1 = 85.44$, see Table 3) is the centralized MentalLongformer model, which has been pretrained on mental health subreddits. In the standard federated setting (without differential privacy), models achieved competitive performance: MentalLongformer using only 10% of client fractions reached $F1 = 83.16$ for depression (Table 2), and MentalBERT using 50% of client fractions reached $F1 = 84.09$ for suicidal ideation (Table 3). This suggests that beyond privacy preservation, federated learning offers a computationally efficient alternative to centralized training. We present detailed findings for depression identification (Section 4.1), suicidal thoughts and behaviors (Section 4.2), data efficiency (Section 4.3), the effect of differential privacy (Section 4.4), and an analysis of the resulting privacy-utility trade-off due to differential privacy (Section 5).

### 4.1 Depression Analysis

Federated approaches perform similarly to the centralized training approach for depression detection (see Table 2). While there is minimal drop in performance for federated models, the decrease is not statistically significant ($p = 0.21875$, $a = 0.05$ using the Wilcoxon signed-rank test). The largest performance difference between federated and centralized models is 4.61 point drop in F1 score by the Logistic Regression classifier. For transformer models, the largest performance drop of federated models is 2.47 point drop in F1 score. There are also models for which the federated settings outperformed the centralized setting. Specifically for MentalBERT, DistilBERT, and DistilRoBERTa we see small performance gains between 0.26 and 0.66 points in F1-score. See Tables 7 to 9 in Appendix B for full results.

### 4.2 STB Analysis

For identifying users with STB, federated models sometimes outperform their centralized counterparts (see Table 3). For instance, the federated Logistic Regression model (5.50 points in F1

and drawbacks to each aggregation method (Fu et al., 2023; Gouissem et al., 2024).

| Model | Centralized | | Federated | |
|---|---|---|---|---|
| | Recall | F1 | Recall | F1 |
| Logistic Regression | **73.59** | **73.67** | 69.03 | 69.06 |
| MentalBERT | 76.94 | 76.74 | **77.25** | **77.08** |
| MentalLongformer | **85.82** | **85.63** | 82.99 | 83.16 |
| BERT | **78.73** | **78.86** | 77.62 | 77.57 |
| RoBERTa | **79.26** | **79.03** | 76.85 | 76.83 |
| DistilBERT | 76.71 | 76.76 | **77.33** | **77.42** |
| DistilRoBERTa | 79.32 | 79.47 | **79.63** | **79.73** |

Table 2: Depression models comparing the centralized approach vs the best performing federated setting.

score) and federated MentalBERT (4.12 points in F1 score) outperform their centralized counterparts. Indeed, for three models for identifying STB, models trained using the federated approach outperform the centralized models. While this is encouraging for federated models for identifying STB, our analysis indicates no statistically significant differences between performances ($p = 0.6875$, $a = 0.05$ via Wilcoxon signed-rank test). See full results in Tables 10 to 12 in Appendix B.

| Model | Centralized | | Federated | |
|---|---|---|---|---|
| | Recall | F1 | Recall | F1 |
| Logistic Regression | 60.50 | 59.62 | **65.24** | **65.12** |
| MentalBERT | 79.84 | 79.97 | **83.95** | **84.09** |
| MentalLongformer | **85.43** | **85.44** | 81.56 | 81.73 |
| BERT | 81.48 | 81.60 | **82.65** | **82.80** |
| RoBERTa | **82.59** | **82.63** | 81.51 | 81.62 |
| DistilBERT | **79.22** | **79.35** | 77.81 | 77.75 |
| DistilRoBERTa | 81.15 | 81.27 | **81.63** | **81.79** |

Table 3: Suicide models comparing the centralized approach vs the best performing federated setting.

### 4.3 Efficiency Analysis for FL

We further find that in some instances, federated models compete with centralized models while using only $10-50\%$ of the total available data. Comparing the different aggregation algorithms trained on different client fractions with the centralized approach, we observe that training on 10% offers the best performance. For instance, see MentalBERT and DistilRoBERTa trained only on 10% of the data outperforms the centralized approach and other trainings on larger client fractions in the depression and STBs identification cases (see Table 8 and 11 in Appendix B). Increasing the client fraction does not correspond to a significant performance increase in either F1 or recall.

Similar to Gala et al. (2023), we argue that data efficiency gains may be due to the selected clients being a representative sample of the overall data. Client selection can, therefore, constitute an interesting area for future work.

### 4.4 Differentially Private FL

Our experiments with the DP-FL setting show a large utility drop of 27.01 F1 ($F1_{FL} = 77.57 \rightarrow F1_{FL-DP} = 50.56$) for depression identification, where $\epsilon$=50 and $c$=0.7 (see Table 4 and Figure 1). As the strength of the privacy increases with lower values of $\epsilon$, the performance further degrades. For a strong privacy budget ($\epsilon = 5$), the drop is even more significant (F1 = 34.86). In line with prior work, e.g., Tramer and Boneh (2020), we find that smaller client fractions result in larger drops in utility, thereby negating data efficiency benefits of the pure FL setup. Further, such large drops in utility challenge the viability of the DP-FL setting compared to centralized and federated settings. We found similar trends in the STB task (Table 5).

| Privacy Budget | Client Frac. 10% | | Client Frac. 70% | |
|---|---|---|---|---|
| ($\epsilon$) | Recall | F1 | Recall | F1 |
| 1 | 50.00 | 34.86 | 50.00 | 34.86 |
| 5 | 50.00 | 31.73 | 50.00 | 34.86 |
| 10 | 50.00 | 34.86 | 49.12 | 38.57 |
| 50 | 50.33 | 39.67 | 54.34 | 50.56 |
| 100 | 49.27 | 49.03 | 67.83 | 67.97 |

Table 4: DP-FL performance across different privacy budgets and client fractions. Larger client fractions improve the DP utility-privacy trade-off.

| Privacy Budget | Client Frac. 10% | | Client Frac. 70% | |
|---|---|---|---|---|
| ($\epsilon$) | Recall | F1 | Recall | F1 |
| 1 | 50.00 | 34.55 | 50.00 | 32.07 |
| 5 | 50.00 | 34.55 | 50.00 | 32.07 |
| 10 | 45.35 | 32.08 | 50.00 | 30.91 |
| 50 | 46.34 | 46.28 | 66.25 | 65.58 |
| 100 | 47.11 | 34.35 | 70.02 | 69.07 |

Table 5: Suicidal Thoughts and Behaviors–DP-FL performance across privacy budgets and client fractions.

To better understand whether this privacy-utility drop-off introduced when applying DP-FL is an artifact of the model, aggregation selection, or the task, we discuss our experimental findings on different settings on 95% confidence intervals over 5 random seeds. The results show a large utility-privacy

trade-off regardless of model or aggregation choice. For instance, for depression identification, MentalBERT with FedAvg achieves $F1 = 0.3361 \pm 0.0207$ ($\epsilon = 10$, $c = 0.1$) and $0.4599 \pm 0.0287$ ($\epsilon = 100$, $c = 0.1$), comparable to BERT with FedOpt $F1 = 0.3361 \pm 0.0172$ ($\epsilon = 10$, $c = 0.1$), despite changing both the model architecture and aggregation algorithm. Similarly, for STB identification, MentalBERT achieves $0.3798 \pm 0.0628$ ($\epsilon = 10$, $c = 0.1$) and $0.3915 \pm 0.0215$ ($\epsilon = 100$, $c = 100$). Both results remain far below standard FL ($F1 = 70.96$ STB; 76.85 depression). Those findings show the trade-off is not an artifact of the model, aggregation selection, or task.

Performance drops are generally expected when striving to achieve mathematical guarantees for privacy. Better convergence performance leads to lower levels of privacy protection (Wei et al., 2020). For example, Wei et al. (2020) find that there can be up to a 20% accuracy drop when privacy guarantees are increased. This significant drop in performance can be associated with the limited data, which is a common issue in the social media mental health detection domain. In settings with limited data, low levels of noise can distort the model training in DP settings (Tramer and Boneh, 2020; Jana and Biemann, 2021). Similarly, although our non-IID setup is more realistic, it may also be a factor in the utility drop for DP-FL (Basu et al., 2021).

### 4.5 Stability and Variance Analysis

To assess the reliability of our results, we conducted stability analyses across five random seeds for representative models in each experimental setting (see Table 6). From the sample of experiments conducted, we find that variance remains low in the centralized and standard FL settings (e.g., $SD \pm 0.0102$ for MentalBERT or $SD \pm 0.0298$ for DistilBERT). These variance intervals support the reliability of our single-run centralized and standard federated setting results. However, in the DP-FL setting, we observe increased instability ($SD \pm 0.0758$ for BERT at $c = 0.7$), highlighting the trade-off between privacy guarantees and model training stability in such sensitive mental health inference tasks from social media.

## 5 Privacy-Utility Trade-off

To understand the performance drops in the DP-FL setting, we conduct two mixed-method analyses for depression detection.

| Setting | Model | Task | F1 ($\mu \pm \sigma$) |
|---|---|---|---|
| *Centralized* | MentalBERT | STB | 0.8271 ± 0.0102 |
| *Centralized* | BERT | STB | 0.8113 ± 0.0190 |
| *Standard FL* | $\text{DistilBERT}_{c=0.7}$ | Depression | 0.7617 ± 0.0298 |
| *Standard FL*$^*$ | $\text{BERT}_{c=0.7}$ | STB | 0.6703 ± 0.0420 |
| *DP-FL*$^\dagger$ | $\text{BERT}_{c=0.1}$ | STB | 0.3310 ± 0.0140 |
| *DP-FL*$^\dagger$ | $\text{BERT}_{c=0.7}$ | STB | 0.4134 ± 0.0758 |
| *DP-FL*$^\ddagger$ | $\text{BERT}_{c=0.1}$ | STB | 0.4246 ± 0.0641 |
| *DP-FL*$^\ddagger$ | $\text{BERT}_{c=0.7}$ | STB | 0.6849 ± 0.0429 |

Table 6: Stability analysis across five random seeds. $c$ denotes client fraction. $^*$FedAvg. FedOPT aggregation is used for the rest. $^\dagger\epsilon = 10$, $^\ddagger\epsilon = 100$. $\mu \pm \sigma$ denotes Mean ± SD. Standard and Centralized settings show some stability, while DP-FL introduces higher variance.

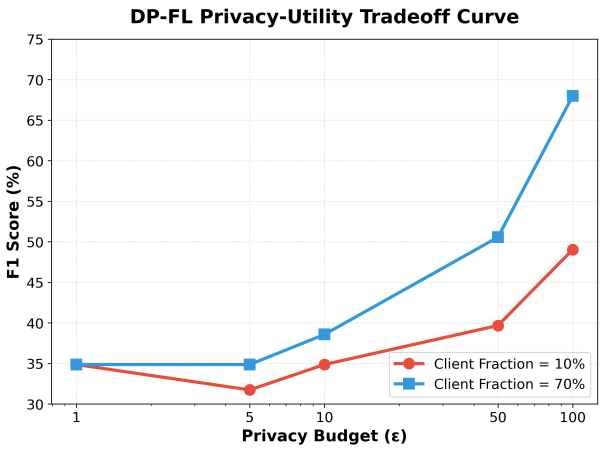


Figure 1: Utility-privacy trade-off for different client fractions and privacy budgets for depression detection.

### 5.1 Stronger Privacy Distorts Sparse Depression Linguistic Markers

In the context of DP-FL, model performance deteriorates privacy guarantees increase, i.e., by decreasing values of $\epsilon$ (see Tables 4 and 5). While we expect performance drops when applying privacy-preserving mechanisms, the performance drops we observe with DP-FL are larger than expected. One potential explanation is that the noise distorts important linguistic markers of mental illness.

**Method.** To evaluate why performance is dropping substantially, we compare three models trained at different privacy budgets while controlling all other hyperparameters. We compare ($\epsilon = 10$, 50, 100), where a higher epsilon indicates lower privacy protections. In the DP-FL setting, these three models vary in performance ($f1$=49.12, $f1$=54.34, $f1$=67.97), respectively. We extracted top features using SHAP (SHapley Additive exPlanations) and qualitatively analyzed important features. Given the importance of socio-linguistic cues in mental

health prediction, we also conducted LIWC analysis (Pennebaker et al., 2015) on the top 50 features to identify psychological and emotional markers learned at different privacy budgets.

**Result.** We find that increased noise does distort important markers for depression and relevant linguistic cues related to mental health. The model with the weakest privacy guarantees ($\epsilon$=100, F1=67.97), which is the best performing DP model, correctly learns 16% of the top 50 keywords as pertinent, half of which are health-related keywords (e.g., "tumor", "prescribed", "aching", "intermittent") and the other half related to negative emotions (e.g., "witnessed", "vanish"). The LIWC-based analysis further confirms this; we find health, social/work, and religion-related terms that are all potential indicators of depression.

In contrast, with stronger privacy guarantees ($\epsilon = 50$, $F1 = 54.34$), features learned by the model are irrelevant terms and patterns. For example, 22% are generic keywords (e.g., "premise", "traditionally", "confluence") and 8% are entertainment-related (e.g., "Eurovision", "maverick") keywords. In our LIWC analysis, we find that there is minimal inclusion of keywords related to affect, negative emotion, or keywords related to health.

The model trained with the strongest privacy guarantees ($\epsilon = 10$, $F1 = 49.12$) identifies keywords related to random topics (e.g., "Egypt", "penguins", "witchcraft") which is supported by our LIWC-based analysis, in which we find the occurrence of LWIC categories such as percept/see (e.g., term: graphics), affect (e.g., term: magnificent), drives/power/work (e.g., term: governments) – which may not be direct indicators of depression. Across our three settings, we thus find a clear trend: the stronger the privacy guarantees and the stronger the noise distortion, the more models rely on irrelevant words for prediction. This supports our hypothesis that the performance drops in DP-FL may be caused by models not learning important mental health-related keywords due to noise disrupting the sparse linguistic signals that are critical for this task.

### 5.2 Smaller Client Fractions Impacted with Low Noise

Our findings indicate that smaller client fractions reduce the DP-FL performance, the opposite of standard FL, which improved data efficiency when trained on lower client fractions. We therefore analyze the impact of noise on models with different client fractions to validate our hypothesis that small amounts of noise can produce large performance disparities by distorting sparse, yet critical cues.

**Method.** We compare models with different client fractions ($c = \{0.1, 0.7\}$) with fixed $\epsilon$ values, and use SHAP and LIWC to analyze how the relationship between client fraction and privacy guarantees impacts learning markers of depression.

**Result.** We find that smaller client fractions are particularly susceptible to noise, severely distorting markers of depression. At $c = 0.1$ and $\epsilon = 100$, the few depression markers are distorted in the training data, and the model relies primarily on irrelevant terms (e.g., "Egypt", "witchcraft", "penguins", "ethanol", "fresco") and a small set of genuine depression markers (e.g., "withdrawal", "restraint", "escalated") for classification. The identified LIWC categories in this low client fraction include percept/see (e.g., scans), relative/motion (e.g., removal), relative/space (e.g., eastwood, environments) – all of which are potentially irrelevant markers of depression. While at $c = 0.7$ and $\epsilon = 100$, the top features include more markers of mental health conditions (e.g., health or bio terms: tumor, prescribed, intermittent; and social/work: advising, and religion: pilgirm). Thus, we find that in considering the relationship between noise and classification for mental health conditions, we must take into account noise introduced, client fractions, and the density of relevant terms to our task.

## 6 Implications

We find that there is no statistical difference between models trained in centralized settings, which ensure no data privacy, and models trained using FL, which protect raw user data. However, our findings also show that differentially private FL results in a large utility drop when trying to mathematically guarantee privacy. Our findings indicate that FL can be a viable candidate for for the identification of high-risk behaviors, specifically the identification of people with depression and people who communicate suicidal ideation. In contrast, adding noise to mathematically ensure privacy may not be a viable option, as adding noise has a large impact on model utility, and the negative impact of errors in mental health inference. Our findings have several implications.

First, our setup approximates a realistic, real-

world setting by maintaining a non-identical data distribution, i.e., the token distributions (see Appendix A.1). Although IID data distribution across clients is common for FL research (e.g., Ji et al., 2019; Gala et al., 2023; Gandhi et al., 2022), creating IID settings would result in discarding relevant data, collecting multiple users on each client, or other bootstrapping approaches to ensure balanced datasets. Next, we consider two future implications of more private ML techniques for this tasks.

**Data Privacy and Dataset Sharing for Institutions:** Privacy concerns prevent researchers (especially those without access to institutional review boards) from developing, accessing, and sharing mental health datasets. Our results indicate space for privacy-preserving approaches to data sharing, which can help address the privacy risks of sharing datasets. We consider the implications of this.

Researchers might build infrastructures that afford federated training approaches for datasets that are usually not shared due to ethics board restrictions. This approach leverages the client-server model of FL – it allows institutions to retain control over local data and how it is shared while enabling other researchers to develop new predictive models for identifying high-risk behaviors and disorders. We are excited by the opportunity for future work.

**Promoting More Consentful and User-Driven Data Sharing:** Finally, we consider how researchers and practitioners may engineer more consentful and user-driven data sharing. Indeed, a major concern in ML and mental health research is the lack of explicit consent given by the data subjects for model training. Pendse et al. (2024) and Ajmani et al. (2024) discuss these concerns and suggest direct consent from users can be solicited to address this problem. However, centralized consent models are infeasible for a single institution to manage or gain for large datasets, and centralized ML infrastructure needs large datasets to demonstrate its prediction efficacy (Chancellor et al., 2019).

As we show in this work, FL could involve opt-in data sharing and training from consenting users, while not decreasing performance on common tasks. Individuals could enroll in ML ecosystems, where they share model updates but keep their data on their own devices. Moreover, users could opt into model training for specific goals (triage, crisis intervention) rather than others they do not support (advertisements). A downstream impact of this consentful data sharing is that it could improve users' trust in mental health and AI technologies, which could support more adoption and potential applications. This must be done in ways that center users' perspectives and avoid harming them (Pendse et al., 2024).

Combining with the aforementioned degradation in performance to guarantee privacy, we encourage the research community to develop infrastructures that allow users to opt in to research to identify high-risk behaviors and to create infrastructures that can safely share data with user consent.

## 7 Conclusion

This work provides a comprehensive analysis investigating the applicability of federated learning for the identification of high-risk behavior and disorders, spanning depressive disorder and suicidal ideation across two major platforms X and Reddit using benchmark datasets. We compare centralized approaches with federated learning settings and differentially private federated learning, where users are simulated as client devices. We demonstrate that the federated approach performs comparably to the centralized method, and that performance differences for both tasks are statistically insignificant. Whereas there is a major utility-privacy trade-off when applying differentially private FL, which results in a large utility degradation when trying to mathematically guarantee the privacy. Training in smaller client fractions results in a larger utility drop-off. These findings open new avenues for mental health detection researchers to leverage federated learning, lowering data-sharing barriers that limit access to restricted social media mental health datasets and engaging in more consentful practices while adhering to the principle of preserving sensitive user data about their health and well-being.

## Ethical Considerations

Mental health behavior identification from social media has several ethical challenges. Inferring the mental status of users from their social media is risky; different actors can target individuals from these predictions. This may result in different harms, including exacerbating their mental health conditions. At the same time, such inference can also be beneficial. Early identification of these behaviors can pave the way to intervention. However, we must be aware of risks and engage in ethical practices suggested in prior work (Benton et al.;

Chancellor et al., 2019) as we deal with sensitive user social media data.

The datasets used in our analysis are publicly available or accessed through IRB approval or data usage agreements to protect the users' privacy. The MTL and C-SSRS datasets were publicly available. The IRB institution at the University of Minnesota (Study ID: STUDY00022028) determined that our research does not involve human subjects, as we do not interact directly with the users. Hence, we accessed the other datasets through data use agreements, and shared this determination form through our requests for data use. Additionally, we ensured the data was only accessed by people included in the IRB determination form (the co-authors on this paper). Both the depression and STB datasets are anonymized and preprocessed to protect the identification of the users in future models. We applied further preprocessing techniques, such as cleaning any user mentions and URLs, to prevent these details from being included in training.

## Limitations

For our analysis, we merged labels in the STB task into binary classifications of whether a user has STBs vs. not (see Datasets). This collapsing of labels follows established prior work of both dataset creators: (Shing et al., 2018a), UMD-RD, collapse fine-grained risk levels to binary (at-risk vs control), and (Gaur et al., 2019), C-SSRS, who collapsed their five categories into 3 (where supportive and indicator classes are merged into one class: no-risk), for experimental purposes. However, it is important to note that binary labels may obscure clinically important distinctions, particularly whether DP noise disproportionately harms the detection of severe cases. Our goal in this work is to establish the viability of federated approaches before tackling fine-grained classifications.

Furthermore, we studied depression on X data and STBs on Reddit data, which are the most commonly studied behaviors and platforms in the domain (Chancellor and De Choudhury, 2020). Our results do not yet extend to other conditions, such as anxiety, eating disorders, PTSD, or other social media platforms, such as Facebook, Weibo, or TikTok. Future work should examine other disorders and platforms.

Although federated learning presents improved data privacy over centralized training paradigms in terms of not sharing raw data, it is not a "silver bullet" for privacy concerns (Jere et al., 2020), nor does it exempt researchers from making datasets to consider privacy as a core value in mental health and ML. Data could still be extracted from trained models (Lyu et al., 2020), and model performances can be degraded through model poisoning attacks (Jere et al., 2021). Our findings show that when ensuring privacy by applying DP on FL, a large degradation in performance. It is therefore necessary for the community to investigate further mechanisms to ensure the data is protected from attacks (Jere et al., 2021) such as homomorphic encryption (Wen et al., 2023) and further experiments with differential privacy to realize the secure implementation of the proposed setting in the social media mental health detection domain. Other techniques such as parameter-efficient federated fine-tuning (LoRA/PEFT), personalized federated approaches, and analysis of the communication costs present future directions in this domain, building on the baselines presented by this work. Furthermore, the relationship between federated learning, differential privacy, and fairness and bias in detecting mental health conditions remains under-explored and would require developing datasets with demographic details, and thus presents a rich avenue for future work.

## References


Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on twitter. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 672–680.

Sulaiman Aftan and Habib Shah. 2023. A survey on bert and its applications. In *2023 20th Learning and Technology Conference (L&T)*, pages 161–166. IEEE.

Leah Ajmani, Logan Stapleton, Mo Houtti, and Stevie Chancellor. 2024. Data agency theory: A precise theory of justice for ai applications. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 631–641.

Maryam Mohammed Aldarwish and Hafiz Farooq Ahmad. 2017. Predicting depression levels using social media posts. In *2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)*, pages 277–280. IEEE.

Abdulrahman Aldkheel and Lina Zhou. 2024. Depression detection on social media: A classification

framework and research challenges and opportunities. *Journal of Healthcare Informatics Research*, 8(1):88–120.

Hatoon S AlSagri and Mourad Ykhlef. 2020. Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems*, 103(8):1825–1832.

Amin Aminifar, Fazle Rabbi, Violet Ka I Pun, and Yngve Lamo. Monitoring Motor Activity Data for Detecting Patients' Depression Using Data Augmentation and Privacy-Preserving Distributed Learning. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2163–2169. IEEE.

John W Ayers, Theodore L Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *NPJ digital medicine*, 1(1):30.

Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, Zumrut Muftuoglu, Sahib Singh, and Fatemehsadat Mireshghallah. 2021. Benchmarking differential privacy and federated learning for bert models. *arXiv preprint arXiv:2106.13973*.

Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical Research Protocols for Social Media Health Research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, Dirk Hovy, et al. 2017. Multitask learning for mental health conditions with limited social media data. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. Proceedings of Conference*. Association for Computational Linguistics.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Stevie Chancellor, Jessica L Feuston, and Jayhyun Chang. 2023. Contextual gaps in machine learning for mental illness prediction: The case of diagnostic disclosures. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–27.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117, San Diego, CA, USA. Association for Computational Linguistics.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366.

Lei Fu, Huanle Zhang, Ge Gao, Mi Zhang, and Xin Liu. 2023. Client selection in federated learning: Principles, challenges, and opportunities. *IEEE Internet of Things Journal*.

Jay Gala, Deep Gandhi, Jash Mehta, and Zeerak Talat. 2023. A Federated Approach for Hate Speech Detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3248–3259. Association for Computational Linguistics.

Deep Gandhi, Jash Mehta, Nirali Parekh, Karan Waghela, Lynette D'Mello, and Zeerak Talat. 2022. A Federated Approach to Predicting Emojis in Hindi Tweets. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 11951–11961, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*, pages 514–525.

Ala Gouissem, Zina Chkirbene, and Ridha Hamila. 2024. A comprehensive survey on client selections in federated learning. *Innovation and Technological Advances for Sustainability*, pages 417–428.

Kuhaneswaran AL Govindasamy and Naveen Palanichamy. 2021. Depression detection using machine learning techniques on twitter data. In *2021 5th international conference on intelligent computing and control systems (ICICCS)*, pages 960–966. IEEE.

Arti Gupta, Manish Kumar Maurya, Khyati Dhere, and Vijay Kumar Chaurasiya. 2024. Privacy-preserving hybrid federated learning framework for mental healthcare applications: Clustered and quantum approaches. *IEEE Access*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3774–3788.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. ArXiv:2106.09685 [cs].

Information Commissioner's Office ICO. 2024. What is special category data? Publisher: ICO.

Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6:1–12.

Abhik Jana and Chris Biemann. 2021. An investigation towards differentially private sequence tagging in a federated framework. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 30–35.

Malhar S Jere, Tyler Farnan, and Farinaz Koushanfar. 2020. A taxonomy of attacks on federated learning. *IEEE Security & Privacy*, 19(2):20–28.

Malhar S. Jere, Tyler Farnan, and Farinaz Koushanfar. 2021. A taxonomy of attacks on federated learning. *IEEE Security  Privacy*, 19(2):20–28.

Shaoxiong Ji, Guodong Long, Shirui Pan, Tianqing Zhu, Jing Jiang, and Sen Wang. 2019. Detecting suicidal ideation with data protection in online communities. In *Database Systems for Advanced Applications*, pages 225–229, Cham. Springer International Publishing.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. Domain-specific continued pretraining of language models for capturing long context in mental health. *arXiv preprint arXiv:2304.10447*.

Haihua Jiang, Bin Hu, Zhenyu Liu, Gang Wang, Lan Zhang, Xiaoyu Li, and Huanyu Kang. 2018. Detecting depression using an ensemble logistic regression model based on multiple speech features. *Computational and mathematical methods in medicine*, 2018(1):6508319.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Samar Samir Khalil, Noha S Tawfik, and Marco Spruit. 2024. Exploring the potential of federated learning in mental health research: a systematic literature review. *Applied Intelligence*, 54(2):1619–1636.

Yalan Kuang, Xiao Liao, Zekun Jiang, Yonghong Gu, Bo Liu, Chaowei Tan, Wei Zhang, and Kang Li. 2025. Federated learning-based prediction of depression among adolescents across multiple districts in china. *Journal of Affective Disorders*, 369:625–632.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. 2020a. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020b. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. Sensemood: Depression detection on social media. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 407–411, New York, NY, USA. Association for Computing Machinery.

Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua. 2017. Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):1820–1833.

Yang Liu. 2024. Depression clinical detection model based on social media: a federated deep learning approach. *The Journal of Supercomputing*, 80(6):7931–7954.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lingjuan Lyu, Han Yu, Jun Zhao, and Qiang Yang. 2020. Threats to federated learning. *Federated Learning: Privacy and Incentive*, pages 3–16.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

H Brendan McMahan, FX Yu, P Richtarik, AT Suresh, D Bacon, et al. 2016. Federated learning: Strategies for improving communication efficiency. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain*, pages 5–10.

Mohammad Moshawrab, Mehdi Adda, Abdenour Bouzouane, Hussein Ibrahim, and Ali Raad. 2023. Reviewing federated learning aggregation algorithms; strategies, contributions, limitations and future perspectives. *Electronics*, 12(10):2287.

Sachin R Pendse, Kate Niederhoffer, and Amit Sharma. 2019. Cross-cultural differences in the use of online mental health support forums. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–29.

Sachin R Pendse, Logan Stapleton, Neha Kumar, Munmun De Choudhury, and Stevie Chancellor. 2024. Advancing a consent-forward paradigm for digital mental health data. *Nature Mental Health*, pages 1–10.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Stephen R Pfohl, Andrew M Dai, and Katherine Heller. 2019. Federated and differentially private learning for electronic health records. *arXiv preprint arXiv:1911.05861*.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. 2020. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.

V Sanh. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Nobin Sarwar and Shubhashis Roy Dipta. 2025. Fedmentor: Domain-aware differential privacy for heterogeneous federated llms in mental health. *arXiv preprint arXiv:2509.14275*.

Fangfang Shan, Shiqi Mao, Yanlong Lu, and Shuaifeng Li. 2024. Differential privacy federated learning: A comprehensive review. *International Journal of Advanced Computer Science and Applications*, 15(7).

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018a. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018b. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Han-Chin Shing, Philip Resnik, and Douglas W Oard. 2020. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8124–8137.

Samantha J Teague, Adrian BR Shatte, Emmelyn Weller, Matthew Fuller-Tyszkiewicz, and Delyse M Hutchinson. 2022. Methods and applications of social media monitoring of mental health during disasters: scoping review. *JMIR mental health*, 9(2):e33058.

Florian Tramer and Dan Boneh. 2020. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*.

Arthur B Vasconcelos, Lúcia Maria de A Drummond, Rafaela C Brum, and Aline Paes. 2023. Exploring federated learning to trace depression in social media with language models. In *2023 International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW)*, pages 24–30. IEEE.

Ziyu Wang, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. 2024. Differential private federated transfer learning for mental health monitoring in everyday settings: A case study on stress detection. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469.

Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. 2023. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513–535.

Xiaohang Xu, Hao Peng, Lichao Sun, Md Zakirul Alam Bhuiyan, Lianzhong Liu, and Lifang He. 2021. Fedmood: Federated learning on mobile health data for mood detection. *arXiv preprint arXiv:2102.09342*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

# A Appendix

## A.1 Distribution of Tokens

Here, we look at the token distribution within the aggregated depression and suicide datasets. There's a high standard deviation across the treatment and control users in both behaviors (depression and suicide). In the depression dataset, the treatment users (STD=10686) have a slightly higher deviation from the control users (STD=10683). For the suicide dataset, the treatment users (STD=10835) have a higher standard deviation compared to the control (7994) counterparts. This significant variation in distribution shows the non-Independent and Identical Distribution (non-IID) of data among the clients. This shows our simulation of unique individual users as clients depicting the real-world use case of high risk behavior identification. (See Figures 2 and 3 for treatment and control depression users, respectively. Figures 4 and 5 for treatment and control suicide users, respectively)

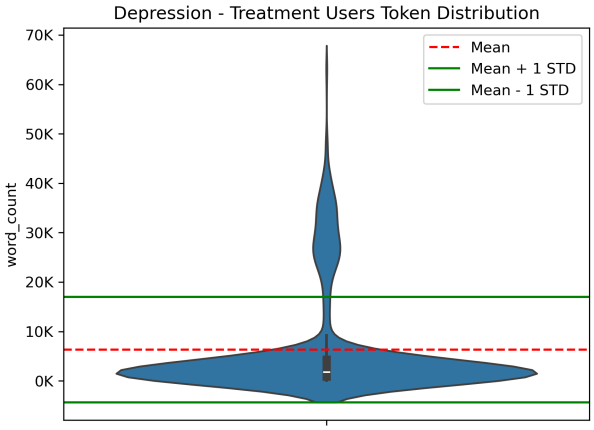


Figure 2: The violin plot illustrates the word count distributions of the treatment users in the depression detection dataset. Mean = 6324, Standard Deviation = 10686.

## A.2 Model Parameters and Experimental Setup

The learning rate is set to $4e-5$. Adam is used as an optimizer. We trained for 'epochs = 50' and applied an early stopping of 5. A batch size of 32 is used for all the models except for MentalLongformer, where we used 16 due to memory limits. We used the default Huggingface parameter values

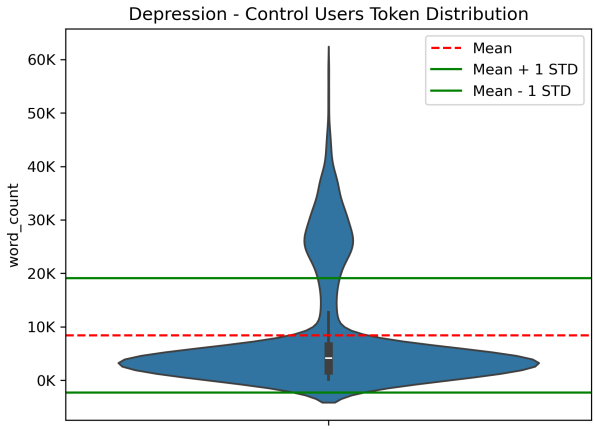


Figure 3: The violin plot illustrates the word count distributions of the control users in the depression detection dataset. Mean = 8402, Standard Deviation = 10683.

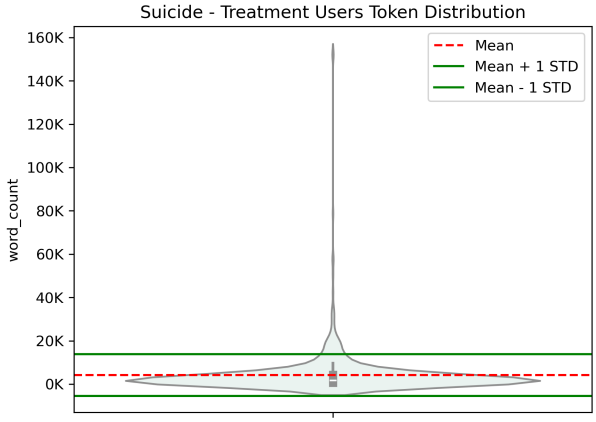


Figure 4: The violin plot illustrates the word count distributions of the treatment users in the suicidal thoughts behaviors identification dataset. Mean = 4282, Standard Deviation = 10835.

for the remaining hyperparameters for both centralized and federated settings.

For our **Differentially Private Federated Learning (DP-FL)** setup, we used finetuned BERT on the depression identification task. We set the clipping norm ($clip\_norm = 1$), the privacy budget is experimented ($\epsilon = 1, 5, 10, 50, 100$), we used the fedprox aggregation algorithm, client fraction (client_fraction = 0.1, 0.7, 10% and 70% of the data, respectively). Learning rate ($client\_lr = 4e-5$, $server\_lr = 1e-3$) provided the best performance after searching over higher learning rate space ($client\_lr = 1e-2$, $server\_lr = 1e-1$). Training for '$epochs = 50$' and '$rounds = 100$' yielded the best result (search space '$epochs = 50, 100$' and '$rounds = 50, 100, 200$'), and applied an early stopping of 5, with batch_size of 32. We used delta ($\delta = 1e-5$). The sigma is calculated with an adaptive Gaussian mechanism.

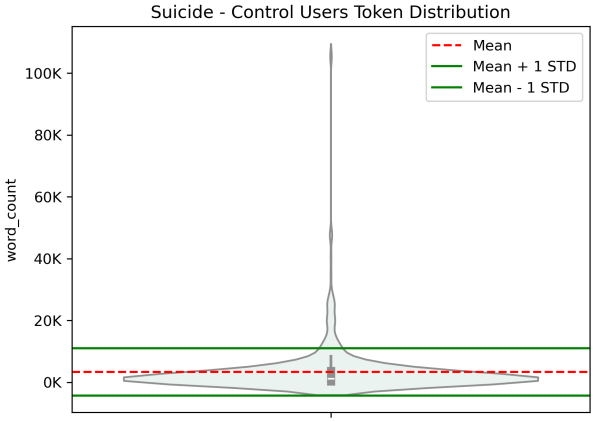


Figure 5: The violin plot illustrates the word count distributions of the control users in the suicidal thoughts behaviors identification dataset. Mean = 3378, Standard Deviation = 7994.

## B Depression and Suicidal Thoughts and Behaviors Detection Results

| Client Fraction | Logistic Regression | | MentalBERT | | MentalLongformer | | BERT | | RoBERTa | | DistilBERT | | DistilRoBERTa | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 |
| **c = 10%** | 59.39 | 59.39 | 76.89 | 76.85 | **82.99** | **83.16** | 73.96 | 73.93 | 75.06 | 75.00 | 74.25 | 74.25 | 72.89 | 72.98 |
| **c = 30%** | 55.47 | 55.46 | 77.02 | 77.01 | 82.29 | 82.40 | 74.06 | 73.48 | 75.30 | 75.21 | 74.15 | 74.20 | 73.31 | 73.37 |
| **c = 50%** | 56.05 | 56.05 | 76.62 | 76.59 | 81.98 | 82.07 | 76.64 | 76.40 | **76.85** | **76.83** | 75.12 | 75.13 | 73.81 | 73.85 |
| **c = 70%** | 55.95 | 55.94 | 75.57 | 75.34 | 82.48 | 82.63 | 77.03 | 76.83 | 75.92 | 75.88 | 75.29 | 75.37 | 74.18 | 74.21 |

Table 7: Federated approach on the **`depression dataset`** using **`fedAvg`** algorithm in different client fractions.

| Client Fraction | Logistic Regression | | MentalBERT | | MentalLongformer | | BERT | | RoBERTa | | DistilBERT | | DistilRoBERTa | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 |
| **c = 10%** | 55.91 | 55.90 | **77.25** | **77.08** | 78.19 | 78.35 | **77.62** | **77.57** | 76.69 | 76.71 | 74.78 | 74.82 | 73.34 | 73.06 |
| **c = 30%** | 55.05 | 54.90 | 75.29 | 75.00 | 80.02 | 80.11 | 76.75 | 76.41 | 76.14 | 76.08 | 74.91 | 74.78 | 73.81 | 73.76 |
| **c = 50%** | 55.77 | 55.78 | 75.81 | 75.60 | 81.72 | 81.80 | 76.16 | 75.79 | 75.41 | 75.41 | 75.33 | 75.37 | 73.67 | 73.75 |
| **c = 70%** | 55.76 | 55.75 | 75.71 | 75.51 | 79.93 | 80.06 | 76.37 | 76.13 | 75.73 | 75.65 | 74.57 | 74.55 | 73.24 | 73.31 |

Table 8: Federated approach on the **`depression dataset`** using **`fedProx`** algorithm in different client fractions.

| Client Fraction | Logistic Regression | | MentalBERT | | MentalLongformer | | BERT | | RoBERTa | | DistilBERT | | DistilRoBERTa | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 |
| **c = 10%** | 66.90 | 66.83 | 70.86 | 69.96 | 74.10 | 74.04 | 71.38 | 71.28 | 71.96 | 71.80 | **77.33** | **77.42** | 68.88 | 67.74 |
| **c = 30%** | 68.82 | 68.84 | 74.61 | 74.44 | 78.25 | 77.62 | 76.02 | 75.98 | 66.89 | 63.38 | 76.78 | 76.70 | 76.78 | 76.94 |
| **c = 50%** | **69.03** | **69.06** | 69.84 | 68.87 | 64.87 | 59.84 | 75.16 | 75.15 | 68.40 | 66.88 | 74.58 | 74.12 | 78.00 | 78.11 |
| **c = 70%** | 68.50 | 68.56 | 70.61 | 70.41 | 82.64 | 82.75 | 64.63 | 59.85 | 75.21 | 75.29 | 76.36 | 76.41 | **79.63** | **79.73** |

Table 9: Federated approach on the **`depression dataset`** using **`fedOpt`** algorithm in different client fractions.

| Client Fraction | Logistic Regression | | MentalBERT | | MentalLongformer | | BERT | | RoBERTa | | DistilBERT | | DistilRoBERTa | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 |
| **c = 10%** | 59.15 | 58.95 | 71.84 | 70.96 | 77.14 | 77.05 | 75.68 | 75.59 | 80.51 | 80.65 | 74.32 | 74.31 | 73.89 | 73.79 |
| **c = 30%** | 58.56 | 58.33 | 82.87 | 83.05 | 79.41 | 79.54 | 78.52 | 78.56 | **81.51** | **81.62** | 73.27 | 73.15 | 68.90 | 67.38 |
| **c = 50%** | 57.91 | 57.71 | **83.95** | **84.09** | **81.56** | **81.73** | 80.15 | 80.27 | 79.89 | 79.88 | 74.77 | 74.66 | 79.14 | 79.20 |
| **c = 70%** | 58.89 | 58.64 | 81.96 | 82.12 | 81.46 | 81.63 | **82.65** | **82.80** | 77.14 | 77.05 | 75.49 | 75.48 | 72.17 | 71.34 |

Table 10: Federated approach on the **`suicide dataset`** using **`fedAvg`** algorithm in different client fractions.

| Client Fraction | Logistic Regression | | MentalBERT | | MentalLongformer | | BERT | | RoBERTa | | DistilBERT | | DistilRoBERTa | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 |
| **c = 10%** | 59.15 | 58.95 | 75.97 | 75.88 | 69.20 | 67.66 | 78.38 | 78.48 | 78.88 | 78.92 | 74.29 | 74.26 | **81.63** | **81.79** |
| **c = 30%** | 58.57 | 58.33 | 82.35 | 82.50 | 70.93 | 70.01 | 78.00 | 78.07 | 81.44 | 81.58 | 73.20 | 73.02 | 70.54 | 69.39 |
| **c = 50%** | 57.91 | 57.72 | 81.53 | 81.70 | 74.66 | 74.46 | 76.88 | 76.82 | 80.67 | 80.78 | 73.09 | 72.80 | 77.89 | 77.92 |
| **c = 70%** | 58.89 | 58.64 | 82.87 | 83.05 | 79.10 | 79.15 | 79.82 | 79.93 | 79.91 | 80.03 | 73.20 | 73.16 | 77.15 | 77.05 |

Table 11: Federated approach on the **`suicide dataset`** using **`fedProx`** algorithm in different client fractions.

| Client Fraction | Logistic Regression | | MentalBERT | | MentalLongformer | | BERT | | RoBERTa | | DistilBERT | | DistilRoBERTa | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 | Recall | F1 |
| **c = 10%** | **65.24** | **65.12** | 72.46 | 72.41 | 54.58 | 43.93 | 70.14 | 69.88 | 67.73 | 66.30 | **77.81** | **77.75** | 73.05 | 72.46 |
| **c = 30%** | 64.86 | 64.48 | 76.44 | 76.45 | 68.51 | 68.54 | 65.94 | 64.55 | 70.74 | 70.73 | 69.26 | 68.78 | 72.36 | 72.23 |
| **c = 50%** | 62.68 | 62.32 | 76.20 | 74.96 | 62.52 | 60.80 | 71.24 | 70.99 | 70.85 | 70.87 | 69.26 | 68.78 | 69.85 | 69.60 |
| **c = 70%** | 62.50 | 62.34 | 77.77 | 77.74 | 57.36 | 50.45 | 63.16 | 59.74 | 74.10 | 74.04 | 74.94 | 74.93 | 69.85 | 69.60 |

Table 12: Federated approach on the **`suicide dataset`** using **`fedOpt`** algorithm in different client fractions.