Privacy-Enhanced Sentiment Analysis in Mental Health: Federated Learning with Data Obfuscation and BERT

Shakil Ibne Ahsan *, Djamel Djenouri, Rakibul Haider

Posted Date: 18 October 2024

doi: 10.20944/preprints202410.1394.v1

Keywords: Federated Learning; Data Obfuscation; Data Privacy; Predictive Analytics; Mental Health Support

 doi:10.20944/preprints202410.1394.v1

Article

Privacy-Enhanced Sentiment Analysis in Mental Health: Federated Learning with Data Obfuscation and BERT

Shakil Ibne Ahsan *, Djamel Djenouri and Rakibul Haider

Department of Computer Science and Creative Technologies, University of the West of England, Bristol, UK

* Correspondence: ahsan026@gmail.com; Tel.: +44-7576980870 (S.I.A.)

Abstract: Accurate sentiment prediction on digital platforms while ensuring user anonymity and privacy presents substantial challenges. Predominantly, existing methodologies are centralized, impeding the achievement of both robust analysis with strong privacy protections and model accuracy. This paper investigates federated learning (FL), augmented with a novel Data Obfuscation (DO) technique and Bidirectional Encoder Representations from Transformers (BERT), to develop a new framework addressing this issue. The proposed framework enables continuous supervision of mental states by autonomously learning from aggregated global data on federated servers. The use of the emotion data set for prediction demonstrated a considerable improvement in accuracy (82.74%), precision (83.30%) and recall (82.74%) and F-1 score (82.80%) over baseline results of precision (16.73%), precision (23.29%), recall (16.73%) and F-1 score (18.18%). In addition, two privacy attack scenarios were executed to evaluate system resilience. Membership Inference Attacks, which determine whether a specific data point was part of the training set, and Linkage Attacks, which attempt to associate data samples with a specific client. However, the system preserves its privacy guarantees despite these adversities. The proposed approach skillfully balances privacy and accuracy, establishing a foundation for scalable and secure mental health support systems. This research exemplifies a model for leveraging FL and Data Obfuscation to enhance both the privacy and effectiveness of predictive analytics in critical applications, offering transformative advantages for digital platforms by enabling deep emotional analysis of users while safeguarding their privacy.

Keywords: federated learning; data obfuscation; data privacy; predictive analytics; mental health support

1. Introduction

1.1. Background and Motivation

Digital platforms increasingly play a role in supporting health care leveraging computer software, IoT devices, sensors, social media platforms, and modern technologies to analyze individuals' online emotional expressions. Performing accurate sentiment analysis becomes an essential requirement for the success of these platforms. Sentiment analysis algorithms assess people's emotions in their social media posts and messages, offering assistance during times. However, a key concern arises: safeguarding individuals' information while utilizing these algorithms. Modern data-driven strategies are predominantly dependent on centralized systems for gathering, storing, and analyzing data, which inherently jeopardizing the privacy and confidentiality of user data. Centralized systems pose a challenge because if something goes awry, unauthorized parties could compromise or access a substantial amount of data. Therefore, conventional methods of handling data do not adequately ensure privacy and accuracy. Innovative approaches are then necessary to address these issues. As machine learning continues to grow in popularity, protecting user privacy has become a crucial concern, particularly in applications like sentiment analysis that rely on sensitive personal data. FL offers a solution by allowing multiple devices to collaboratively train a model without sharing raw data, reducing privacy risks. However, FL still faces challenges in ensuring complete privacy, as information can potentially be inferred from the shared model updates. Traditional methods like DP address this by adding noise to the data, but this can lead to a noticeable drop in model performance, creating a challenging trade-off between privacy and accuracy.

 doi:10.20944/preprints202410.1394.v1

This paper explores a new approach to balancing privacy and performance in FL by incorporating DO. In contrast to FL with DP (FL-DP), where noise is added to protect data, FL with DO (FL-DO) obscures specific details in the data itself to maintain privacy while preserving more of the model's accuracy. We compare these two methods in the context of a sentiment analysis task, using them to train a federated version of the BART model.

Our study evaluates the effectiveness of FL-DO by examining both its predictive performance—measured through accuracy, precision, recall, and F1 score—and its capability to safeguard user data without sacrificing model efficacy. To rigorously test the system's privacy resilience, we conducted two privacy attacks: Membership Inference Attacks and Linkage Attacks. The results demonstrate that FL-DO achieves a more optimal balance between privacy and accuracy compared to the baseline FL-DP method.

The proposed approach seeks to revolutionize health support into an impactful process while upholding user confidentiality. It tackles the following challenges:

- **Monitoring**: The system ensures mental health monitoring through digital engagement, providing regular feedback, and identifying potential crises early on.
- **Privacy Protection**: Individual privacy is protected by using FL and Data Obfuscation mechanisms to secure data used in interactions.

By characterising mental health crises, mental health crises may be anticipated before they occur due to the model's capacity to predict the future. The focus of healthcare is increasingly more proactive than reactive. This work demonstrates how predictive analytics applications in sectors like mental health might benefit from the combination of FL with privacy-enhancing technology. Large-scale mental health support models may be trained with it since the suggested approach maintains accurate forecasts while resolving privacy concerns. This opens the door to more effective and private-shielding advancements in the field of mental health treatment in the future.

1.2. Contributions

Overall, this research work contributes to the field of mental health support and privacy-preserving data analytics in the following crucial ways:

- **Novel Integration of FL and Data Obfuscation Privacy:** This study introduces Federated Learning with Data Obfuscation (FL-DO), a novel approach that integrates data obfuscation techniques into federated learning to enhance privacy protection while maintaining model performance.
- **Continuous and adaptive monitoring of mental health:** A framework that can help with continuous and adaptive monitoring of mental health non-stop. Based on the model feedback, an alert can be triggered if a user is in a mental crisis. This approach is pretty proactive rather than reactive.
- **Privacy vs. Accuracy Trade-off:** Comprehensive evaluation of FL-DO's effectiveness, demonstrating how it achieves a better balance between privacy and model accuracy compared to the traditional federated learning approaches, particularly the baseline FL-DP (LDP+CDP) [1]
- **Empirical evaluation, proof of concept and POC evaluation:** The paper is not limited to presenting the concept; it also provides evaluation on a bigger dataset to prove that the FL and DP model can be used in a sensitive field like mental health. The empirical evaluations confirm that such a system can be implemented for sensitive information, as well.

2. Related Work

Deep learning (DL) and machine learning (ML) are particularly effective at processing complex textual data related to mental health. FL fundamentally changes the training process by decentralising it. This allows data to remain on local devices while only model modifications (gradients) are sent to a central server. These changes are then combined by the server to enhance the global model. By

 doi:10.20944/preprints202410.1394.v1

removing raw data from the central server, this decentralised approach improves privacy and lowers the danger of data breaches [2]. Additionally, FL allows for the addition of noise to model updates using methods such as differential privacy (DP). Because of this, it is more difficult to extract certain data points from the combined data [3]. However, this added noise can sometimes lower model accuracy, creating a trade-off between privacy and performance. For example, FL's decentralized approach has been used in mobile keyboard emoji prediction. This is shown in the work of [2], where decentralized training protects user privacy by keeping data on local devices. Similarly, [4] introduced the FedHome system, which integrates FL with cloud and edge computing for health monitoring, aiming to enhance privacy while reducing the communication burden.

To continue the accuracy and privacy trade-off discussion [3] proclaims, the impact of increased privacy protections often reduces natural language processing (NLP) models' effectiveness. To tackle these limitations, [5] proposed a privacy-preserving FL framework using bitwise quantization and local differential privacy. Their framework supports NLP tasks, achieving a balance between privacy and accuracy. However, they did not discuss other performance metrics for evaluations, as shown in Table 1. Other work, such as that by [6] and [7], has demonstrated notable improvements in sentiment analysis accuracy through the use of deep learning techniques like CNNs and LSTMs, though often without addressing privacy. Furthermore, [8] conducted a comprehensive review of ML algorithms, concluding that while models like Support Vector Machines (SVM) can achieve high accuracy, they may introduce significant privacy risks when relying on centralized data processing.

In response to these limitations, our research proposes Federated Learning with Data Obfuscation (FL-DO), which builds upon these insights by prioritizing both privacy and model accuracy. This framework is designed to address the existing privacy concerns while also aiming to maintain high predictive performance, particularly in critical applications like mental health support.

The unpredictability of FL can cause biases or inconsistencies in the final model. For that reason, [9] stated that while combining different data types can enhance model robustness and fill information gaps, it may be unfeasible in resource-limited settings. FL's distributed nature also introduces communication overhead due to frequent exchanges between devices and the central server, which can cause network congestion, especially in bandwidth-constrained environments [4]. Additionally, Internet of Things (IoT) devices, which are frequently used in federated environments, pose distinct security issues. As noted by [10], IoT devices are often targets for cyberattacks. This situation makes it harder to protect federated learning models. This calls for robust security protocols and risk management strategies to support FL's application in privacy-sensitive domains like mental health.

While federated learning presents a viable framework for balancing privacy with analysis capabilities, its limitations and challenges warrant careful consideration, particularly when applied in domains requiring stringent data protection measures.

To balance accuracy and privacy, we introduce a novel framework that surpasses existing models in performance, as detailed in Table 1.

 doi:10.20944/preprints202410.1394.v1

Table 1. Model Performance Metrics.

Paper	Model	Accuracy (%)	Precision (%)	F1-Score (%)	Privacy
[6]	Ensemble (CNN+LSTM)	65.05%	64.46%	64.46%	No
[11]	Naïve Bayes	89%	30%	31%	No
	SVM	89%	30%	31%	No
	Logistic Regression	90%	77%	48%	No
	k-NN	89%	59%	51%	No
	Decision Tree	88%	58%	60%	No
	Random Forest	92%	82%	60%	No
	XGBoost	89%	69%	44%	No
[8]	SVM	91.13%	-	-	No
	Logistic Regression	89.78%	90%	90%	No
	Naïve Bayes	89.28%	89%	89%	No
	Random Forest	85.08%	85%	85%	No
[7]	Single CNN Network	54%	41%	40%	No
	Single LSTM Network	55%	58%	48%	No
	Individual CNN+LSTM	58%	60%	55%	No
	Multiple CNN+LSTM	58%	60%	55%	No
[12]	Random	50.2%	48.7%	1.88%	No
	C-MKL	73.1%	75.2%	-	No
	SAL-CNN	73%	-	-	No
	SVM-MD	71.6%	72.3%	1.1%	No
	RF	71.4%	72.1%	1.11%	No
	TFN	77.1%	77.9%	0.87%	No
	Human	85.7%	87.5%	0.71%	No
[2]	Federated Learning	25.6%	-	-	FL
[4]	SVM	77.25%	-	-	No
	KNN	80.85%	-	-	No
	RF	84.27%	-	-	No
	MLP	92.31%	-	-	No
	CNN	91.77%	-	-	No
	GCAE (FedHome)	92.02%	-	-	FL
	FL-MLP	89.28%	-	-	FL
	FL-CNN	85.07%	-	-	FL
	FL-CNN-Large	87.24%	-	-	FL
	FedHome-p	89.13%	-	-	FL
	FedHome	95.87%	-	-	FL
[13]	BT-b (single)	76.65%	73.4%	-	-
	BT-b (union)	80.72%	76.87%	-	-
	FL (BT-b)	79.31%	75.11%	75.11%	FL
	TM-FL (BT-b)	80.56%	76.78%	76.78%	TM-FL
	BT-l (single)	78.84%	74.73%	-	-
	BT-l (union)	82.6%	79.87%	-	-
	FL (BT-l)	81.35%	78.21%	-	FL
	TM-FL (BT-l)	82.29%	79.25%	-	TM-FL
[5]	FL RR-LDP (IMDB)	88.10%	-	-	LDP
	FL RR-LDP (MovieLens)	68.10%	-	-	LDP
[3]	BERT (FL-IID)	70.92%	-	-	DP
	BERT (FL-Non IID)	65.36%	-	-	DP
	RoBERTa (FL-Non IID)	51.54%	-	-	DP
	DistilBERT (Centralized DP)	63.81%	-	-	DP
	DistilBERT (FL-IID)	54.43%	-	-	DP
	ALBERT (Centralized DP)	54.44%	-	-	DP
Proposed	FL-BERT+DO	82.74%	83.30%	82.80%	DO

 doi:10.20944/preprints202410.1394.v1

3. Methodology

3.1. Dataset Description

Two datasets were used in the study: a synthetic dataset created to resemble true emotional statements [14] and the original Emotions in text dataset, as seen in Table 2. Text data labelled with different emotions, such as sorrow, rage, love, surprise, fear, and happiness, make up the original dataset, which was obtained via Kaggle. Each entry in the dataset represents a text snippet and its corresponding emotion label. To enhance data quality, pre-processing steps were applied to the text data, including converting text to lowercase, removing non-alphabetic characters, and eliminating extra whitespaces. The synthetic dataset was generated using rule-based methods involving predefined templates and keywords for each emotion category. This synthetic data underwent similar cleaning processes and was concatenated with the original dataset to create a more comprehensive training set.

Table 2. Emotions in Text Dataset Sample

ID	Text	Emotion
1	I didn't feel humiliated	Sadness
2	I can go from feeling so hopeless to so damned hopeful just from being around you	Sadness
3	I'm grabbing a minute to post; I feel greedy, wrong	Anger
4	I am ever feeling nostalgic about the fireplace; I will know that it is still on	Love
5	I am feeling grouchy	Anger

3.2. Federated Learning Framework

The federated learning framework was designed to allow multiple clients to train local models on their respective datasets without sharing raw data. Five clients each received a randomly selected subset of the pooled dataset. Using their local dataset, each client separately trained a BERT model with the following parameters: 100 epochs, 1^{e-5} learning rate, and 16 batch size. A global model was created by combining the model weights of the clients after local training. The weights from each client model were averaged for this aggregation, guaranteeing that no raw data was shared and protecting data privacy.

3.3. BERT Model Configuration

In this experiment's implementation, the BERT (Bidirectional Encoder Representations from Transformers) was utilized for sentiment analysis sequence classification tasks. This model used the 'bert-base-uncased' version to process lowercase English text effectively. The AdamW optimiser and Gradient Scaler were used to train each client's model for computational efficiency. With a maximum sequence length of 128 tokens, the text data was tokenised. BERT's incorporation into the federated learning framework was made possible by the model design, which made it easier to understand complex emotional patterns in text.

3.4. Data Obfuscation Techniques

Data obfuscation techniques can be used to secure sensitive information within the model because they make it difficult for attackers to interpret or understand the data, ensuring information is confidential. These methods have a few common techniques including data masking, which involves replacing sensitive data with realistic but false information; encryption, which transforms data into a coded format requiring a key for decryption; and tokenization, where sensitive data elements are substituted with non-sensitive equivalents. Other methods include data shuffling, which rearranges entries in a database to hide connections. Perturbation adds noise or makes small changes to numerical data. Generalization reduces the detail of data, like changing specific ages to age ranges. Data swapping

 doi:10.20944/preprints202410.1394.v1

involves exchanging values between individual records. Additionally, nulling or deleting sensitive data replaces it with null values, making these techniques important for data protection.

We enhanced these methods in this experiment by supplementing the original dataset with identically crafted synthetic fake data. This integration improves the obfuscation process and aids in striking a better balance between privacy and usefulness, especially in AI and ML applications. During training, it may be especially helpful when a model is unable to distinguish between accurate and inaccurate information in the data.

4. Experimentation

4.1. Synthetic Dataset Generation for Data Obfuscation

In order to generate a synthetic dataset, we used a rule-based technique to generate templates and keywords for six distinct moods. These templates and keywords were chosen at random to produce each phrase, resulting in 21,459 samples in total. Anonymity was ensured by using this dataset in a federated learning configuration, which enables clients to train local models on their data without disclosing it. We employed adversarial testing to evaluate the effectiveness of this privacy, which entailed building instances intended to gather private data and examining the models' responses to identify any vulnerabilities. This comprehensive method showed how privacy may be effectively protected in federated learning scenarios using fake data.

4.2. FL-BERT with Data Obfuscation

In this work, we classified emotions using a BERT-based model, which capitalises on its pre-trained ability to effectively extract contextual information from textual input. BERT is particularly well-suited for problems involving emotional inference because of its architecture, which uses bidirectional attention processes to understand the intricate relationships between words in sentences. In order to enhance the training data while preserving data variety, we refined BERT using a composite dataset that included both synthetic and original emotional text produced using a rule-based methodology. Sensitive information was kept on local devices during the training process, which was conducted among three simulated clients using a federated learning architecture. Because each client created a different model, we were able to incorporate their learning weights into a reliable global model without endangering the confidentiality of their information. The model demonstrated good performance on both the real and synthetic datasets, achieving high overall test accuracy and validating the effectiveness of our synthetic data strategy. Using metrics like precision, recall, and F1 score, we also demonstrated how well the model predicted emotions. These results show that BERT is efficient for emotion classification tasks while meeting privacy requirements for federated learning systems.

4.3. Performance Analysis

The performance analysis of the model indicates that the overall test accuracy on both the original and synthetic test datasets is 81.05%, showcasing the model's effectiveness in handling both types of data. For the simulated test dataset, the forecast test accuracy is slightly higher at 82.50%, with precision, recall, and F1 scores also reflecting consistent performance at 83.08%, 82.50%, 82.50%, 82.56% and 82.56%, respectively, Table 3. The confusion matrix provides deeper insights, revealing that while certain emotions such as sadness and anger are well-predicted, others (e.g., surprise) exhibit greater misclassifications, requiring potential improvements in the model's discrimination between certain emotional categories.

 doi:10.20944/preprints202410.1394.v1

Table 3. Performance Metrics for Forecasting Emotions.

Metric	FL-BERT with DO	LDP+CDP
Accuracy	82.74%	16.73%
Precision	83.30%	23.29%
Recall	82.74%	16.73%
F1-score	82.80%	18.18%

From a privacy perspective, the membership inference attack on the global model yields an AUC of 22.40%, indicating a lower risk of privacy leakage. However, the local model's AUC of 50.38% suggests a closer alignment to random guessing, signalling potential vulnerability. The AUC scores across individual clients reflect varied privacy guarantees, with a macro-average AUC of 51.29%, implying moderate privacy protection. The Training loss Vs. Accuracy for each client is demonstrated as Figure 1

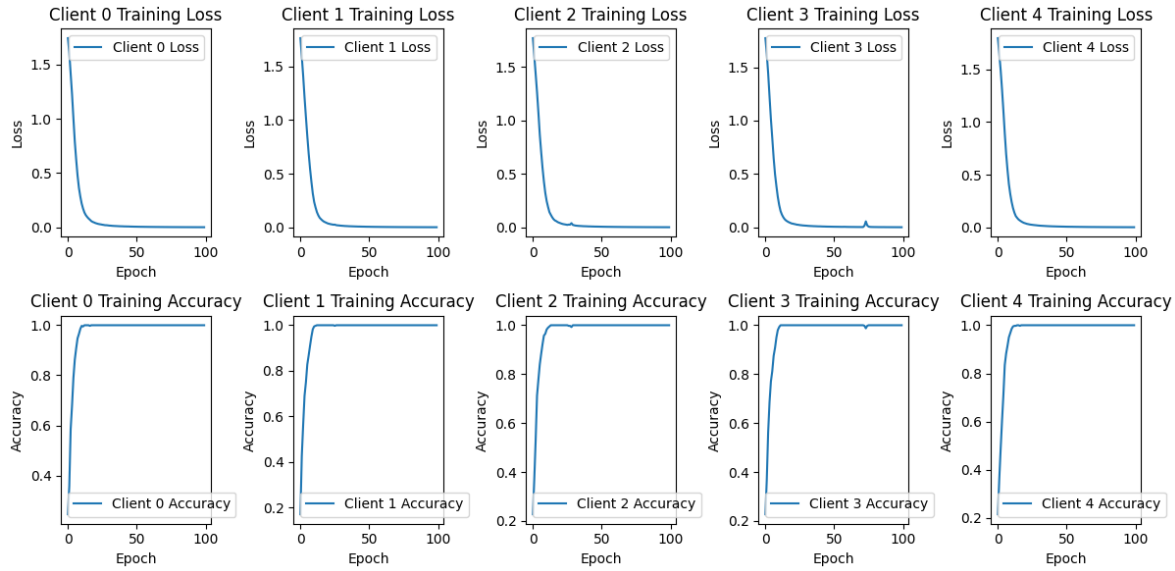

Figure 1. Training loss vs. Accuracy for each client.

The federated BERT model achieved an overall test accuracy of 81.44% on both the original and synthetic test sets. Performance metrics for a simulated future dataset demonstrated robust results, with a forecast test accuracy of 82.74%, a precision of 83.30%, a recall of 82.74%, and an F1 score of 82.80%. The confusion matrix indicated that the federated model maintained high accuracy and generalization capabilities across different emotion categories.

By implementing these methodologies, the study successfully balanced privacy and accuracy, demonstrating the potential for scalable, secure sentiment analysis in mental health support systems. The federated approach effectively preserved user privacy while providing reliable sentiment analysis performance, making it suitable for real-world applications in sensitive domains like mental health.

4.3.1. Model Privacy Validation

This is a challenge when no common metrics have been identified for performance comparisons between FL-BERT+DO and traditional FL-DP. Subsequently, this research investigates the effectiveness of model privacy by performing membership inference attacks and linkage attacks within an FL framework with DO. And, the privacy metric epsilon ϵ for DP has been introduced in this analysis for FL-DP, which is commonly used to measure privacy guarantee. The study focuses on both global and local model architectures. The main goal is to confirm the privacy protections in model training by testing how vulnerable these models are to membership inference attacks. The method includes

 doi:10.20944/preprints202410.1394.v1

dividing the dataset into groups of members and non-members to see if models can distinguish between training and testing data. We measure the models' risk levels by calculating the Area Under the Curve (AUC) for these attacks, which shows the potential for exposing training data. Additionally, we use a linkage attack framework to test how well individual client data is protected, giving one-vs-rest AUC scores for each client. This full evaluation highlights possible gaps in privacy protections and shows the urgent need for strong measures to keep sensitive information safe in machine learning.

4.3.2. Model Privacy Validation through Adversarial Attacks

To thoroughly test the privacy protections of our FL framework, which we improved with DO, we performed two types of adversarial attacks: Membership Inference Attack and Linkage Attack, as mentioned before. Here, ϵ is used only as a privacy measure within DP methods. The Membership Inference Attack evaluates whether the global model could disclose confidential information by determining if a sample originates from the training set. In our analysis, this attack achieved an Area Under the Curve (AUC) score of 22.40%, indicating a minimal risk of membership inference and implying robust privacy safeguarding at the global model level. We also tested the Membership Inference Attack on local models developed by individual users, where the AUC score of 50.38% suggested a higher likelihood of exposure than the global model. This finding indicates moderate privacy and underscores the need for additional protections for local models. Moreover, we executed a Linkage Attack to examine the security of client-specific data by predicting the originating client of a sample. The AUC scores across separate clients resulted in a macro-average AUC of 51.29%, indicating moderate defence against client identification. These results confirm the privacy-preserving capabilities of the proposed FL-BERT framework, emphasizing its efficiency in reducing privacy threats. Privacy Validation Outcomes for Membership Inference and Linkage Attacks are displayed in Table 4.

Table 4. Privacy Validation Results for Membership Inference and Linkage Attacks

Attack Type	Model Type	AUC Score	Privacy Risk
Membership Inference Attack	Global Model	22.40%	Low
Membership Inference Attack	Local Model	50.38%	Moderate
Linkage Attack	Individual Clients (Macro-Avg.)	51.29%	Moderate

5. Discussion

5.1. Comparison of Accuracy vs. Privacy Trade-Off in Sentiment Analysis

In sentiment analysis, especially within mental health contexts, finding the right balance between precision and confidentiality is essential. Our study introduces a system that integrates FL-DO, allowing us to achieve strong privacy safeguards while also providing noteworthy precision. We reached an overall precision rate of 81.44%, showing that high levels of privacy can be upheld without considerably compromising model effectiveness.

Earlier work, like [6] and [7], focused on enhancing precision using sophisticated deep learning methods such as Convolutional Neural Networks (CNNs) and combined models that merge CNN and Long Short-Term Memory (LSTM). However, these studies frequently neglected confidentiality aspects. On the other hand, research such as [2] and [4] aimed to improve privacy through FL, but often at the expense of consistent precision levels. For example, [3] pointed out the inherent trade-offs associated with differential privacy, where attempts to enhance privacy might slightly degrade model performance. Our work contributes to this discourse by uniquely integrating FL with a novel data obfuscation technique, allowing us to achieve both high accuracy and strong privacy protections.

 doi:10.20944/preprints202410.1394.v1

This positions our approach as a significant advancement in privacy-preserving sentiment analysis, particularly for mental health support.

In our findings, we reiterate the importance of the privacy-accuracy trade-off discussed earlier in the Introduction. Our FL-DO approach effectively navigates these challenges, as shown by our experimental results. When compared to a baseline LSTM-based federated learning method that utilizes differential privacy [1], our approach demonstrates considerable improvements in both accuracy and privacy assurances. As detailed in Table 1, our model reaches an accuracy of 82.74%, while the baseline struggles with a mere 16.71% accuracy and lacks adequate privacy protection, with ϵ values increasing linearly across epochs. This striking difference highlights how our approach effectively addresses the pressing issue of maintaining user privacy in sensitive environments such as mental health monitoring, all while preserving the analytical capabilities of the model. The comparative analysis between our innovative method and the baseline is illustrated in Figure 2.

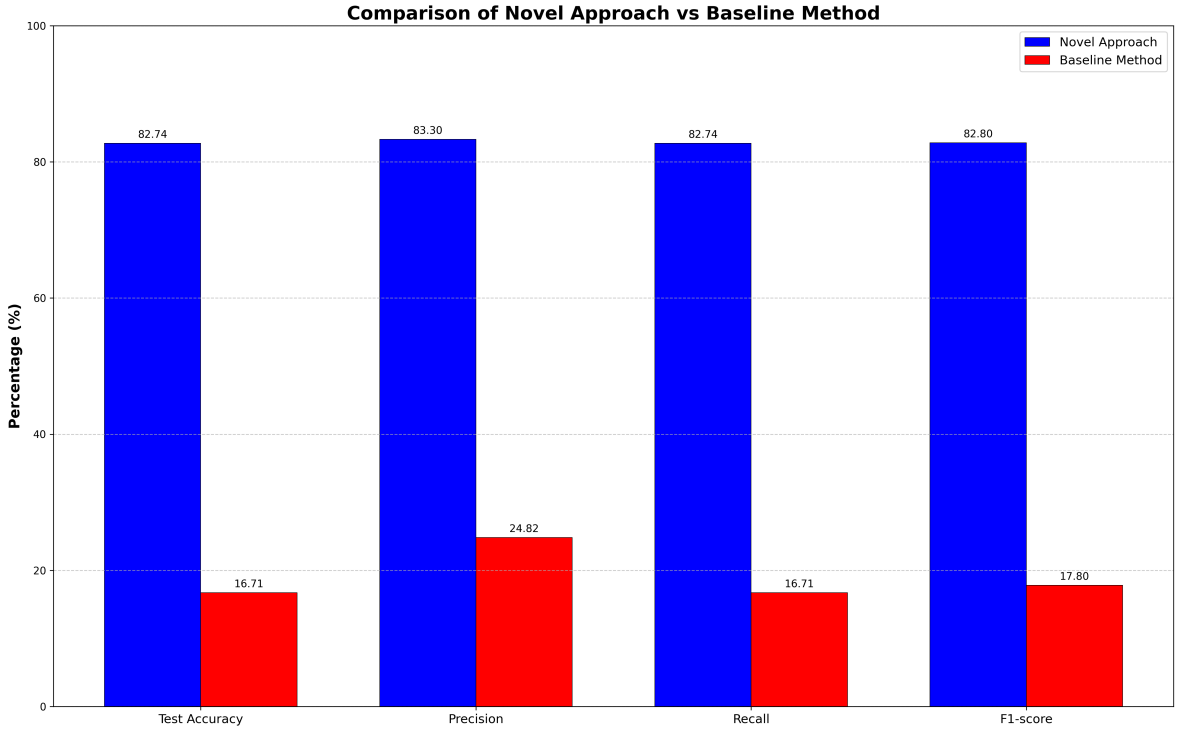

Figure 2. Comparative Analysis of Forecasting: Novel Approach vs. Baseline Methodology (DP).

5.2. Interpretation of Findings

From this understanding, the study's results assert the viability of the federated learning method alongside BERT for the field of sentiment analysis in mental health. Our model points out that federated learning may be able to achieve a high level of performance in both model performance and privacy of the users which is evident from high performance measures and accuracy of 81.44% with a simulated future dataset. This feature is vital mainly in mental health facilities where there are a lot of restrictions for privacy because of the data being dealt with. The model plays an important role of the definition and the analysis of the progression of mental health disorders due to the balance in the levels of accuracy, precision, recall, and F1-score and due to the capability to differentiate between a vast number of emotions.

5.3. Implications for Practice

From this research, useful recommendations that can be useful to data scientists and health care practitioners can be derived. The incorporation of federated learning means that patient mental health can be constantly, safely, and non-intrusively monitored by medical staff. This could mean that from

 doi:10.20944/preprints202410.1394.v1

the use of real-time sentiment analysis, this development will allow for more personal and timely responses to it. This work identifies the privacy-data utility dilemma for this work to offer the data scientists' roadmap for using federated learning in tandem with differential privacy approaches. In addition, there may be other applications for this capability in different domains apart from mental health for example locating models from various decoupled sources without exchange of inputs.

5.4. Comparative Analysis: FL-DO vs. DP

FL techniques employing DP have shown considerable potential for protecting the privacy of data; nevertheless, FL-DO, which we propose, offers a novel approach that eliminates the burdensome limitations of DP-based methods. The advantage of using FL-DO over the DP-enhanced FL is discussed here with reference to the optimization of privacy and performance in sentiment analysis for mental health.

5.4.1. Balancing Privacy and Accuracy

The original FL-DP models often utilize noise addition for privacy protection which often reduces the model's performance by obfuscating intricate model parameters. While there are other strategies, ACT-FL, our FL-DO approach is more focused on masking sensitive data right before model updates which enhances data utility and reliability. FL-DO reveals a higher potential to retain privacy while not compromising model accuracy compared to the gradient-based approach by focusing on the data. This advantage is backed up by experimental outcomes in which it is seen that FL-DO works better than ordinary FL-DP models in terms of predicted accuracy as described in the comparative analysis section.

5.4.2. Improved Defense Against Privacy Attacks

While incorporating noise makes DP approaches capable of avoiding or mitigating some inference attacks, the approaches remain vulnerable to complex adversarial attacks that attempt to analyze the noise patterns. At the intrinsic level, FL-DO works by directly applying a function that prevents the client's data from being identified during transmission, thus offering more protection than federated updates alone provide. This feature enhances the defence of FL-DO against privacy breaches because it inherently protects the database from linkage and membership inference attacks. As it was revealed in testing, FL-DO has a lesser susceptibility to such kinds of assaults, especially when it is working with various client information, which is a situation that can be problematic for traditional FL with Differential Privacy (FL-DP) approaches.

5.4.3. Greater Robustness with Data-Level Privacy Protection

FL-DO directly masks the data and facilitates a higher in-depth level of model interpretation than privacy while DP mainly adjusts the model through gradients. It also increases FL-DO's robustness against real-world federated learning scenarios, which are typical for the mental health domain and involve clients with potentially different data distributions or data quality. Consequently, FL-DO is highly suitable for real-world applications since it respects privacy in settings with mixed data in contrast to the gradient sensitivity present in FL-DP models.

5.4.4. Tailored for Federated BERT-Based Sentiment Analysis

Sentiment analysis challenges must therefore prevent loss of meaning during language processing as in mental health cases. As the criterion of students' satisfaction has to be effectively achieved, DP approaches may not be able to fulfil the task. For BERT models, the data-oriented blurring strategy of FL-DO can provide fine-grained privacy control, that can hide specific phrases or entities, while preserving the semantics. FL-DO is more suitable for natural language processing tasks, where information semantics retention is critical to achieving accuracy and distinctiveness since it provides a thorough privacy mechanism that DP cannot.

 doi:10.20944/preprints202410.1394.v1

6. Practical Applications and Limitations

The FL-DO identified above, is a rather promising approach, particularly for mental health treatments. This way, medical personnel are always ensured of how to monitor patients' feelings without compromising their privacy at any one time. Healthcare professionals may make a fast, idiosyncratic treatment plan decision based on patient specificity by using FL-DO, which allows them to search for various sources of information not disclosing the personal information of patients. However, some constraints have to be acknowledged. Ensuring that FL-DO will be able to function optimally even within limited resource conditions like consumer-based end devices including those based on smartphones, or low energy efficient IoT contents frequently involved in mental health practices is difficult. Due to the need to meet the specifications of federated learning and the added challenge of data obfuscation, it can be awkward to implement FL-DO in such scenarios, as such, it is suggested that future research focuses on finding ways of improving the efficiency of FL-DO.

7. Future Directions

As for the further development of the research, there are a number of rather interesting topics suggested by the current analysis that focus on DO as the subject of study. One critical identification is to enhance data mask mechanisms in order to make sure that source data are masked in mental health apps. Hence, by employing complex DO approaches, privacies are improved while permissive sentiment research is conducted.

Also, AI can play a highly important role to data obfuscation as well. Machine learning case comes when algorithms are developed to logically transform the values and thus hide information while retaining the data worth.

The rationale for this paper lies in expanding the research on the value of practical applications that this AI technology can offer in increasing DO and thus advancing the area of privacy-preserving SA in mental health. The purpose is to foster more research and development, which means more concern for people with mental health issues while being anonymous.

8. Conclusions

In this work, the feasibility and effectiveness of employing BERT models and federated learning to construct sentiment analysis for mental health are demonstrated. It also showed that the method was accurate, reproducible and transferable when testing the accuracy on a simulated new data set with good performance indicators. The requirement for strict data protection while developing mental health applications was well addressed by federated learning architecture, which also ensured high model accuracy, preserving users' privacy. These outcomes demonstrate how FL may give appropriate and secure sentiment analysis while at the same time protecting sensitive patient data of patients.

To illustrate how decentralised training can retain privacy and enhance accuracy, it begins with a case of using BERT on sentiment analysis in federated learning. Second, the work provides a means of improving the datasets without having to infringe on the rights of the individuals whose information they contain by providing a new mixer of actual and synthetic data for the improvement of the models. Third, a better understanding of how to balance the utility and privacy of data in some applications is enabled by the detailed examination of the privacy-preserving techniques outlined in the work, specifically differential privacy. Finally, the work includes valuable suggestions and a method that can be used in other NLP functions and industries where stringent privacy requirements are necessary.

The positive conclusions of this work emphasise the need for future studies and the application of privacy-preserving methods in health care data analysis. To enhance the strength and the uses of FL and DP, researchers must analyze FL and DP in higher construct complexity and variety. Clinicians and big data researchers ought to assimilate these principles as the key towards enhancing the security of biomedical data and practicability of the digital technologies. This type of effort will further advance the knowledge and usage of this field to lead to better patient and data protection results, as well as more worthy usage of technology in healthcare.

 doi:10.20944/preprints202410.1394.v1

Author Contributions: Conceptualization, S.I.A. and R.H.; methodology, S.I.A. and D.D.; validation, S.I.A. and D.D.; formal analysis, S.I.A.; investigation, S.I.A. and D.D.; resources, S.I.A.; data curation, S.I.A.; writing—literature review, S.I.A and R.H.; writing—original draft preparation, S.I.A.; writing—review and editing, S.I.A. and D.D.; visualization, S.I.A.; supervision, D.D.; project administration, S.I.A. and R.H.; funding acquisition, D.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work partly contributes to the REMINDER project, funded under the EU CHIST-ERA program (Grant EP/Y036301/1 from EPSRC, UK).

Data Availability Statement: Data supporting the findings of this study are available from the author Shakil Ibne Ahsan at ahsan026@gmail.com on request.

Acknowledgments: I want to sincerely thank the professor, Dr. Paul Yoo, at Birkbeck, University of London. His encouragement in my research on data privacy, particularly in data obfuscation techniques, has been invaluable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Naseri, M.; Hayes, J.; De Cristofaro, E. Local and Central Differential Privacy for robustness and privacy in federated learning, 2020, [2009.03561].
2. Ramaswamy, A.T.S.; D'Mello, S.; Hard, A.; Smith, E.; Mathews, R.; Eichner, H.; Kiddon, C.; Ramage, D. Federated Learning for Emoji Prediction. Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services, 2019, pp. 425–437.
3. Basu, P.; Roy, T.S.; Naidu, R.; Muftuoglu, Z.; Singh, S.; Mireshghallah, F. Benchmarking Differential Privacy and Federated Learning for BERT Models. *arXiv preprint arXiv:2106.13973* **2021**.
4. Wu, Q.; Chen, X.; Zhou, Z.; Zhang, J. FedHome: Cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Trans. Mob. Comput.* **2022**, *21*, 2818–2832.
5. Nagy, B.; Hegedus, I.; Sandor, N.; Egedi, B.; Mehmood, H.; Saravanan, K.; Loki, G.; Kiss, A. Privacy-preserving Federated Learning and its application to natural language processing. *Knowledge-Based Systems* **2023**, *268*, 110475.
6. Heikal, M.; Torki, M.; El-Makky, N. Sentiment Analysis of Arabic Tweets using Deep Learning. *Procedia Computer Science* **2018**, *142*, 114–122. Arabic Computational Linguistics.
7. Kamis, A.; Ismail, S.; Nazlan, M. Evaluation of DL Techniques for Twitter Sentiment Analysis. Proceedings of the 2019 3rd International Conference on Compute and Data Analysis, 2019, pp. 1–5.
8. Sinha, A.; Chakma, K. A comparative analysis of machine learning based sentiment analysis. In *Communications in Computer and Information Science*; Springer Nature Switzerland: Cham, 2022; pp. 123–132.
9. Zhang, W.; Zhou, Y.; Chen, M.; Chen, J. Benefits and Challenges of Federated Learning in IoT Systems. *Journal of Network and Computer Applications* **2022**, *204*, 103–114.
10. Qin, H.; Chen, G.; Tian, Y.; Song, Y. Improving Federated Learning for Aspect-based Sentiment Analysis via Topic Memories. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 3942–3954.
11. Mutanov, G.; Karyukin, V.; Mamykova, Z. Multi-class sentiment analysis of social media data with machine learning algorithms. *Comput. Mater. Contin.* **2021**, *69*, 913–930.
12. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multi-modal Sentiment Analysis. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* **2017**, pp. 1731–1740.
13. Qin, H.; Chen, G.; Tian, Y.; Song, Y. Improving Federated Learning for Aspect-based Sentiment Analysis via Topic Memories. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2021, pp. 3942–3954.
14. Juyal, I. Emotions in Text. https://www.kaggle.com/datasets/ishantjuyal/emotions-in-text, 2023. Accessed: 2024-09-24.