electronics

MDPI

Article

Privacy-Enhanced Sentiment Analysis in Mental Health: Federated Learning with Data Obfuscation and Bidirectional Encoder Representations from Transformers

Shakil Ibne Ahsan *, Djamel Djenouri and Rakibul Haider

Department of Computer Science and Creative Technologies, University of the West of England, Bristol BS16 1QY, UK; djamel.djenouri@uwe.ac.uk (D.D.); rh.zico@gmail.com (R.H.)
* Correspondence: ahsan026@gmail.com; Tel.: +44-7576980870

Abstract: This research aims to find an optimal balance between privacy and performance in forecasting mental health sentiment. This paper investigates federated learning (FL) augmented with a novel data obfuscation (DO) technique, where synthetic data is used to "mask" real data points. Bidirectional Encoder Representations from Transformer (BERT) is used for sentiment analysis, forming a new framework, FL-BERT+DO, that addresses the privacy-performance trade-off. With FL, data remains decentralized, ensuring that user-sensitive information is retained on local devices rather than being shared with the FL server. The integration of BERT gives our system an enhanced feature of context sense-making from text conduct, and our model is extremely proficient in emotion categorization tasks. The experiments were performed on combined (real and replica synthetic) datasets containing emotions and showed significant enhancements compared to baseline methods. The proposed FL-BERT+DO framework shows the following metrics: prediction accuracy, 82.74%; precision, 83.30%; recall, 82.74%; F1-score, 82.80%. Further, we assessed its performance in the adversarial setup using membership inference and linkage attacks to ensure the privacy-preserved performance did not suffer deeply. It demonstrates that, even for large datasets, providing privacy-preserving prediction is possible and can significantly improve existing methods of addressing personal issues, like mental health support. Based on the results of our work, we can propose the development of secure decentralized learning systems that are capable of providing high accuracy of sentiment analysis and meeting strict privacy constraints.

Keywords: FL; data obfuscation; data privacy; predictive analytics; mental health support

check for updates

Citation: Ahsan, S.I.; Djenouri, D.; Haider, R. Privacy-Enhanced Sentiment Analysis in Mental Health: Federated Learning with Data Obfuscation and Bidirectional Encoder Representations from Transformers. *Electronics* **2024**, *13*, 4650. https://doi.org/10.3390/electronics13234650

Academic Editor: Andreas Mauthe

Received: 16 October 2024
Revised: 17 November 2024
Accepted: 22 November 2024
Published: 25 November 2024

1. Introduction

1.1. Background and Motivation

Digital platforms are increasingly taking a role in supporting healthcare, leveraging computer software, Internet of Things (IoT) devices, sensors, social media platforms, and emerging technologies to analyze individuals' online emotional expressions. Performing accurate sentiment analysis becomes an essential requirement for the success of these platforms. Sentiment analysis algorithms evaluate people's emotions in their social media posts and messages, offering assistance at 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 jeopardizes 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 (ML) 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 such as DP address this by adding noise to the data, which can lead to a noticeable drop in model performance, creating a challenging trade-off between privacy and accuracy. The authors of [1] defined some weaknesses in FL models, including data heterogeneity, communication problems, and privacy issues. They suggested a method to improve FL systems by increasing the efficiency of the aggregation technique and the security of the communication protocol.The work in [2] was primarily centred on decentralized FL (DFL), which offers a mechanism that minimizes the dependency on the central server and offers privacy improvements. The authors drew attention to features such as optimization, security, and scalability, advancing the network topologies and proposing adaptive algorithms to upgrade the efficiency of DFL. These papers offer very valuable information concerning further improvement of the FL frameworks for use in practice.

This paper explores a new approach to balancing privacy and performance in FL by incorporating DO. In contrast to FL with DP, where noise is added to protect data, FL with DO obscures specific details in the data themselves 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 BERT model.

Our study evaluates the effectiveness of FL-BERT+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 [3] and linkage attacks [4]. The results demonstrate that FL-BERT+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 the data used in interactions.

By characterizing mental health crises, the model's predictive capabilities may help anticipate such crises before they occur. 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 privacy-shielding advancements in the field of mental health treatment in the future.

1.2. Contributions

Overall, this research 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 presents a novel approach (FL-BERT+DO) that integrates DO [5] techniques into the local clients' dataset of FL [6] to enhance privacy protection while maintaining model performance and BERT for sentiment analysis.
- 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: A comprehensive evaluation of FL-BERT+DO's effectiveness demonstrates how it achieves a better balance between privacy and model

accuracy compared to the traditional FL approaches, particularly the baseline FL-DP (LDP+CDP) [7], achieving robust privacy protections against membership inference and linkage attacks, validated by ROC-AUC scores.

- Empirical evaluation, proof of concept, and POC evaluation: This 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

There still exist challenges when it comes to building and using sentiment analysis, especially in the context of social media text. A number of recent papers also underscore certain important methodological shortcomings that can hamper the sharpness and transportability of measures. These emerge from issues of their unstructured nature, language complexity, and computation, which make sentiment analysis a demanding task.

One of the main problems revealed is the problem of how to work with informal language, which often includes slang, misspelled words, and improper grammar, which are inherent in the texts of social networks. Challenges such as handling big and unstructured data were reviewed by [8] with reference to the differences in style, tone, and polarity that influence sentiments and emotions in written communication. In response to this, several approaches to sentiment analysis have been developed as adaptive models that will adjust models according to new data arriving and the new trends in the language being used. These techniques are applied where the target networks must be fine-tuned for applications of transfer learning, domain adaptation, and continual learning to the subtle differences within informal language conversations [9]. Moreover, the work with highly unstructured and noisy data remains an issue due to the lack of clear inter-model compatibility; i.e., the models might not perform well in a different setting even with a minimal level of retraining. The need to sustain high computational overhead and the continuous requirement for large labelled datasets make these systems challenging to maintain in the long term. Thus, the ability to achieve effective, highly flexible performance in the context of the informal and dynamically changing sample text remains only a subject for further experimentations and optimizations. This is made worse by the fact that there are few or no annotated datasets for these aspects of language, making sentiment analysis challenging at the global level [10]. Furthermore, different from traditional data in textual format, information on social media usually contains symbols, idioms, sarcasm, and other features that are not easy for the model to grasp [11].

The second main difficulty is inherent in the nature of short texts or those a few sentences long; this often causes the omission of word frequency and basic word associations. Widespread word occurrences and insufficiency are discouraging feature vector representations and frustrating for emotional recognition, as noted by [12]. Ref. [13] reported the limitations in deep learning (DL) frameworks to refine and capture language and enhance precision. Progress has been made in the recent work on ensemble models to overcome these issues, but the model still struggles with the contextual relationships of the words, especially for words that are outside of the vocabulary (or) often encountered in evolving contexts, including the COVID-19 pandemic. The choice of appropriate classification algorithms for SM data is still challenging, and the encoding should provide fine-granularity representation that covers the contextual meaning contained in texts [14]. Refs. [15,16] discussed multimodal sentiment and emotion analysis, with the use of physiological responses and stress data being a challenge. These models need complex techniques to estimate emotion in a step-by-step manner. Therefore, these datasets entail even more difficulties; in addition, it is also common to use data whose origin is diverse, and this also complicates the situation. Refs. [12,13] pointed out that recent developments in hybrid and ensemble methods can bring solutions to some of these issues. These methodologies are still far from perfect and are plagued by the variability in social media language and the computational cost of analyzing large amounts of data.

Emotional analysis, a vital component in natural language processing (NLP), has been investigated using DL approaches for different languages [17]. DL (DL) and ML (ML) are particularly effective at processing complex textual data related to mental health. FL fundamentally changes the training process by decentralizing 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 removing raw data from the central server, this decentralized approach improves privacy and lowers the danger of data breaches [18]. Additionally, FL allows for the addition of noise to model updates using methods such as DP. Because of this, it is more difficult to extract certain data points from the combined data [19]. 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 [18], where decentralized training protects user privacy by keeping data on local devices. Similarly, ref. [20] introduced the FedHome system, which integrates FL with cloud and edge computing for health monitoring, aiming to enhance privacy while reducing the communication burden.

As explained in [19], the impact of increased privacy protections often reduces NLP models' effectiveness. To tackle these limitations, ref. [21] proposed a privacy-preserving FL framework using bitwise quantization and local DP. Their framework supports NLP tasks, achieving a balance between privacy and accuracy. However, they did not discuss other performance metrics for evaluations. Other work, such as [17,22], has demonstrated notable improvements in sentiment analysis accuracy through the use of DL techniques like convolutional neural networks (CNNs) and long short-term memory (LSTMs), though often without addressing privacy. Furthermore, ref. [23] conducted a comprehensive review of ML algorithms, concluding that while models like support vector machines (SVMs) can achieve high accuracy, they may introduce significant privacy risks when relying on centralized data processing.

The interpretability issue is the biggest problem of using the Tensor Fusion Network for sentiment analysis because of its size and structure [24]. Dependence on accurate Multimodal Opinion Sentiment Intensity (CMUMOSI) data limits the practical use of these components, especially when dealing with noisy or sparse inputs. Also, exactly the speaking disturbances such as fuzziness or conflicting signals were shown to be detrimental to performance. This also limits its use, especially in environments with heavy computational constraints.

The unpredictability of FL can cause biases or inconsistencies in the final model. For that reason, ref. [25] 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 [20]. Additionally, IoT devices, which are frequently used in federated environments, pose distinct security issues. As noted by [26], IoT devices are often targets for cyberattacks. This situation makes it harder to protect FL models. This calls for robust security protocols and risk management strategies to support FL's application in privacy-sensitive domains like mental health. IoT sentiment analysis employs data from connected devices regarding the moods and opinions of users with the ultimate goal of enhancing service delivery. However, it has the following difficulties: One of the big concerns is related to data privacy because some data might be leaked during the analysis. Another challenge is the Non-IID (non-Independent and Identically Distributed) nature of data because they are collected from different users, which may lead to bias in sentiment interpretations. Real-time analysis is hampered by communication barriers, which include low bandwidth and latency. In addition, the information on which ML algorithms are based can be unbalanced or insufficient, which leads to rather high inaccuracy in determining positive or negative tonality in IoT applications and less effectiveness of sentiment analysis in this regard [20].

A practical way to categorize the approaches to sentiment analysis is to use categorical sentiment analysis and dimensional sentiment analysis, as they refer to different angles of the classification of sentiment. Categorical sentiment analysis involves categorizing text into arrangements, including positive, negative, or neutral sentiments, or sharpening up categorizations like anger, happy, sadness, fear, and so on. Earlier approaches used rule-based systems and conventional ML algorithms like naïve Bayes and the support vector machine (SVM), though basic sentiment categorizations were fairly accurate [17]. Further development has been incorporated for sentiment analysis using CNNs and LSTM for DL architectures since the ConvNet models and LSTM network capture text contexts and sequential presence to maximize sentiment forecasting.

In mental health applications, categorical sentiment analysis has been employed in several cases to identify emotional states and possible disorders from text content, blogs, forums, or chat, for instance, online therapeutic sessions. However, categorical data present their main disadvantage through the distinct separation of emotional reactions, which can oversimplify complex mental states.

Dimensional sentiment analysis on the other hand depicts feelings along the quantitative axes, including valence (positive–negative), arousal (calm–excited), and dominance (control–submission). This means there is another possibility to present the intensity and mixture of the values of emotions [27]. Emotional pattern gives a more detailed picture of how one feels, especially for the evaluation of mental health disorders, because emotions are often not binary but rather exist within a spectrum [28].

Dimensional sentiment analysis has the potential for use in mental health contexts since it can show changes in feeling within the period while spotting the basic intensity level. This freedom helps address the limitations of categorical models in covering certain ideas, including comorbidity and overlapping of symptoms in disorders like depression and anxiety [29]. Nevertheless, applying dimensional models to real-world problems is sometimes an intricate numerical process that may raise concerns in terms of understandability and data consumption.

The studies carried out in the recent past have made great advances in emotion classification and especially in mental health diagnosis. Previous conventional models, including the BoW model and LSA, failed in handling complex emotional contexts. Ref. [30] proposes a novel bi-directional LSTM and convolutional neural network (BiLSTM-CNN) structure in order to classify the emotion in psychiatric social texts where the authors reveal how the use of both LSTMs and CNNs can improve feature extraction by capturing temporal and spatial features at the same time. The addition of mechanisms such as GloVe and Word2Vec also enhanced the model performance since they offer an improved concept of word relations. However, to enhance the accuracy of emotion detection, recent works have incorporated multi-task learning, as well as attention mechanisms. However, there are problems that have not been solved yet; for example, one still can only use rather small and not very diverse datasets, and the set fits the model in a specific domain excessively often. The findings of this study suggest that there is still much to be accomplished to improve sentiment analysis research, including a focus on the language of the future, the need for better and more diverse corpora, and more refined frameworks, as well as the practical problems of interpreting and extrapolating the results of these techniques. These are significant developments for enhancing emotion detection systems, especially where practical use, such as in mental health, is contemplated [30].

A composite DP model incorporates DP with other privacy mechanisms, including cryptographic or statistical ones in order to improve the security level of the data. In FL, this framework guarantees the privacy of raw data in training the model across decentralized devices or IoT systems, and it addresses data leakage and inference attacks. The novel algorithm that is presented within this research work offers an optimal solution to the problems of data security, model accuracy, and computational complexity and is well suited to cross-IoT platform knowledge sharing. However, the primary limitation of this approach is the privacy vs. performance scenario trade-off. Strict privacy preservation sometimes entails the minimization of data attributes and thus produces noisy data that affect the best model performance. However, increasing the amount of exposure for performance optimization increases leakage and the model is susceptible to being attacked through an inference attack. This trade-off offers an important problem of privacy-preserving approaches for IoT applications because choosing the correct amount of privacy and performance is critical for a safe training model. On that account, although safeguard-enhancing frameworks of hybrid DP to offer protection to sensitive data in FL exist, there is the need to balance privacy and performance so that the FL model does not sacrifice efficiency for security and vice versa [31].

In response to these limitations, our research proposes FL-BERT+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.

While FL 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.

Table 1. Model performance metrics and privacy features across datasets.

Paper	Model	Accuracy (%)	Precision (%)	F1-Score (%)	Dataset	Privacy
[17]	Ensemble (CNN+LSTM)	65.05%	64.46%	64.46%	Arabic Tweets	No
[32]	Naïve Bayes	89%	30%	31%	AMASS	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
[23]	SVM	91.13%	-	-	Tweets	No
	Logistic Regression	89.78%	90%	90%	IMDB	No
	Naïve Bayes	89.28%	89%	89%		No
	Random Forest	85.08%	85%	85%		No
[22]	Single CNN Network	54%	41%	40%	Twitter	No
	Single LSTM Network	55%	58%	48%		No
	Individual CNN+LSTM	58%	60%	55%		No
	Multiple CNN+LSTM	58%	60%	55%		No
[24]	Random	50.2%	48.7%	1.88%	CMU-MOSI	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
Experimented	FL-CNN+DO	61.34%	72.79%	62.40%	Emotions in text	Yes
Experimented	FL-BiGRU+DO	57.83%	62.05%	58.32%	Emotions in text	Yes
Proposed	FL-BERT+DO	82.74%	83.30%	82.80%	Emotions in text	Yes
[24]	TFN	77.1%	77.9%	0.87%	CMU-MOSI	No
	Human	85.7%	87.5%	0.71%		No
[18]	FL	25.6%	-	-	-	Yes
[20]	SVM	77.25%	-	-	MobiAct	No
	KNN	80.85%	-	-		No
	RF	84.27%	-	-		No
	MLP	92.31%	-	-		No
	CNN	91.77%	-	-		No
	GCAE (FedHome)	92.02%	-	-	FL	Yes

Table 1. *Cont.*

Paper	Model	Accuracy (%)	Precision (%)	F1-Score (%)	Dataset	Privacy
	FL-MLP	89.28%	-	-	FL	Yes
	FL-CNN	85.07%	-	-	FL	Yes
	FL-CNN-Large	87.24%	-	-	FL	Yes
	FedHome-p	89.13%	-	-	FL	Yes
	FedHome	95.87%	-	-	FL	Yes
[26]	BT-b (single)	76.65%	73.4%	-	LAP14	No
	BT-b (union)	80.72%	76.87%	-	REST14	No
	FL (BT-b)	79.31%	75.11%	75.11%	TWITTER	Yes
	TM-FL (BT-b)	80.56%	76.78%	76.78%		Yes
	BT-l (single)	78.84%	74.73%	-		No
	BT-l (union)	82.6%	79.87%	-		No
	FL (BT-l)	81.35%	78.21%	-		Yes
	TM-FL (BT-l)	82.29%	79.25%	-	FL	Yes
[21]	FL RR-LDP (IMDB)	88.10%	-	-	IMDB	Yes
	FL RR-LDP (MovieLens)	68.10%	-	-	MovieLens	Yes

3. Methodology

3.1. Dataset Description

Two datasets were used in the study: a synthetic dataset created to resemble true emotional statements [33] and the original Emotions in Text dataset, as seen in Table 2. Text data labelled with different emotions, such as sadness, anger, 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. These synthetic data underwent similar cleaning processes and were 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. FL Framework

Traditional ML has issues concerning high computational costs, high communication overhead, and latency, which motivated FL. Recent advancements aim to eliminate these problems through the use of techniques such as gradient quantization, sparsification, as well as adaptive compression, to solve communication problems and latency. For example, the authors of [34] present the benefits of gradient quantization and sparsification in FL, and the authors of [35] provide privacy-preserving FL via Hybrid DP and Adaptive Compression, decreasing communication costs. The authors of [36] present a detailed survey on optimization techniques for improving FL performance in practical applications of IoT and the healthcare sector.

The FL framework was designed to allow multiple clients to train local models on their respective datasets without sharing raw data. The use of three clients for this experiment is beneficial because it divides a small dataset reasonably well so that each client has data for training and assessment. Even more clients mean fewer data for each and therefore fewer resources for training and measuring the model. Three 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, 1×10^{-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 were shared and protecting data privacy.

3.3. BERT Model Configuration

In this experiment's implementation, BERT was utilized for sentiment analysis sequence classification tasks. This model used the 'bert-base-uncased' version to process lowercase English text effectively. The AdamW optimizer 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 were tokenized. BERT's incorporation into the FL framework was made possible by the model design, which made it easier to understand complex emotional patterns in text.

BERT's base model consists of 12 blocks of transformer layer with 768 dimensions in each block and 12 attention heads. Because of this multi-head attention, BERT is able to attend to different parts of an input sequence in parallel and grasp the complexity of the contextual interactions of tokens. In each transformer block, information from the previous step is passed through a feed-forward neural network where the size of the hidden layer is 30,720 and layer normalization is used to maintain stability and speed up the training process.

The '[CLS]' token is also a token added at the beginning of an input sequence for the purpose of summarizing the whole input. When the BERT model is at work, the last sentence vector which represents this specific token is used to represent the overall input text, which then goes into a classification layer for work such as sentiment analysis. This transformation enables making BERT's bidirectional representations of constant length for further use in downstream applications.

To improve the prediction of the model, the cross-entropy loss is used for adjusting its functionality. To overcome overfitting, a learning rate of 1×10^{-5} is used, and AdamW optimization is used, together with weight decay for parameter tuning. Furthermore, a learning rate scheduler further adapts the learning rate with reference to the training parameters. To improve the speed, mixed-precision training is used for training DL models, which creates model accuracy with faster training on GPUs. In the model calibration, the batch size of 16 is used so that many samples can be processed at once during the learning step.

3.4. Data Obfuscation Techniques

DO 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 involves exchanging values between individual records. Additionally, nulling or deleting sensitive data replaces them 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

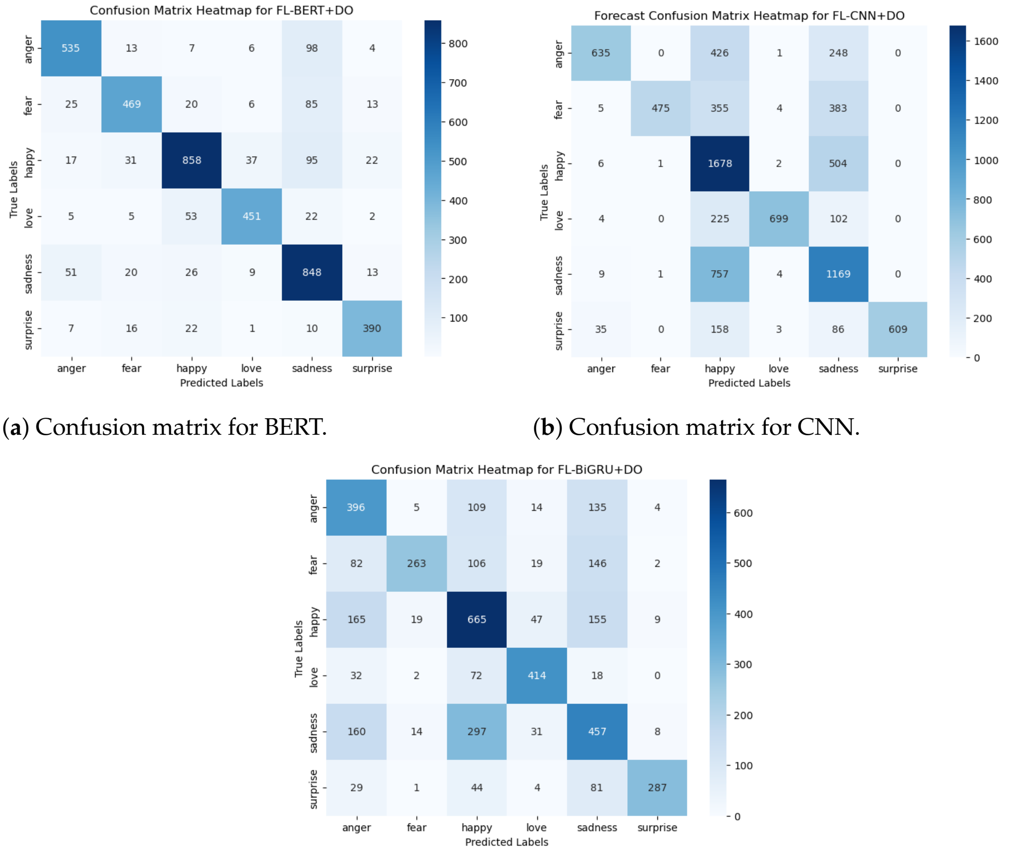
In order to generate a synthetic dataset, we used a rule-based technique to generate templates and keywords for six distinct moods. These templates and keywords were chosen at random to produce each phrase, resulting in 21,459 samples in total. Anonymity was ensured by using this dataset in an FL configuration, which enables clients to train local models on their data without disclosing them. 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 FL scenarios using fake data.

4.2. FL-BERT with Data Obfuscation

In this work, we classified emotions using a BERT-based model which capitalizes 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 five simulated clients using an FL 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 FL systems.

4.3. Performance Analysis

The robust performance of the FL-BERT+DO model in emotion classification for six of the six emotional categories is shown. Strong classification accuracy can be observed particularly for happy (858 correct predictions) and sadness (848 correct predictions), implying that the model does excel at classifying these emotionally differentiated states. Despite that, there are some notable misclassifications between semantically related emotions (e.g., fear and anger), where 25 fear instances were misclassified as anger and vice versa (13 anger and fear), Figure 1a. ROC curves provide further support for the models' effectiveness, with AUC scores from 0.79 to 0.88 in all emotion classes, Figure 2a. AUCs were highest for the 'happiness' class (highest at 0.88), which discriminated best between positive emotions, and lowest for the 'surprise' class (only 0.79); we have reason to think that 'surprise' might not be as easily separable as other emotions due to its contextual ambiguity and the possibility that it overlaps with other emotional states. It is shown that the model performs substantially better than the random classifier baseline (represented by the diagonal dashed line) for all the emotion classes, with AUCs in excess of 0.79. Unexpectedly, the model does well across most categories with AUCs within 0.1 of each other, including anger, fear, and sadness with AUCs of 0.86. The consistency of these results suggests that FL-BERT+DO's architecture is in fact able to capture the intricate features that differentiate one class of emotional expression from another, while there exists the opportunity for improvement in more subtle emotional distinctions, particularly between 'surprise' and 'love'.



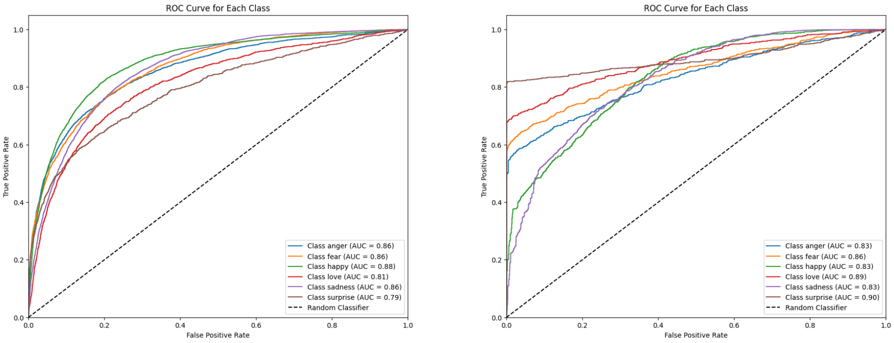
(**a**) Confusion matrix for BERT. (**b**) Confusion matrix for CNN.

(**c**) Confusion matrix for BiGRU.

Figure 1. Confusion Matrices for BERT, CNN, and BiGRU.

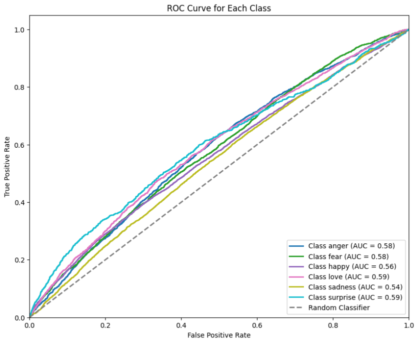
From a privacy perspective, the membership inference attack on the global model yields an AUC of 22.40%, indicating a lower risk of privacy leakage. However, the local model's AUC of 50.38% suggests a closer alignment to random guessing, signalling potential vulnerability. The AUC scores across individual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

The FL-BERT+DO model achieved a strong performance, achieving high overall test accuracy 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% (Table 3). The confusion matrix indicated that the federated model maintained high accuracy and generalization capabilities across different emotion categories.



(**a**) ROC-AUC curve for BERT. (**b**) ROC-AUC curve for CNN.

Figure 2. *Cont.*



(**c**) ROC-AUC curve for BiGRU.

Figure 2. ROC-AUC curves for BERT, CNN, and BiGRU.

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%

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. The training loss vs. accuracy for each client is demonstrated in Figures 3–5, showing the patterns in performance over the various training stages.

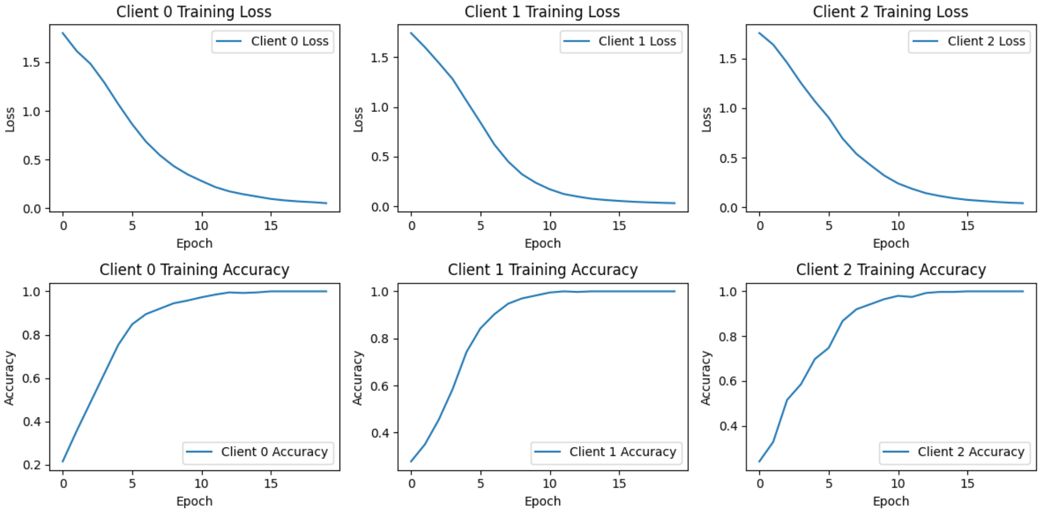


Figure 3. Training loss vs. Accuracy for BERT.

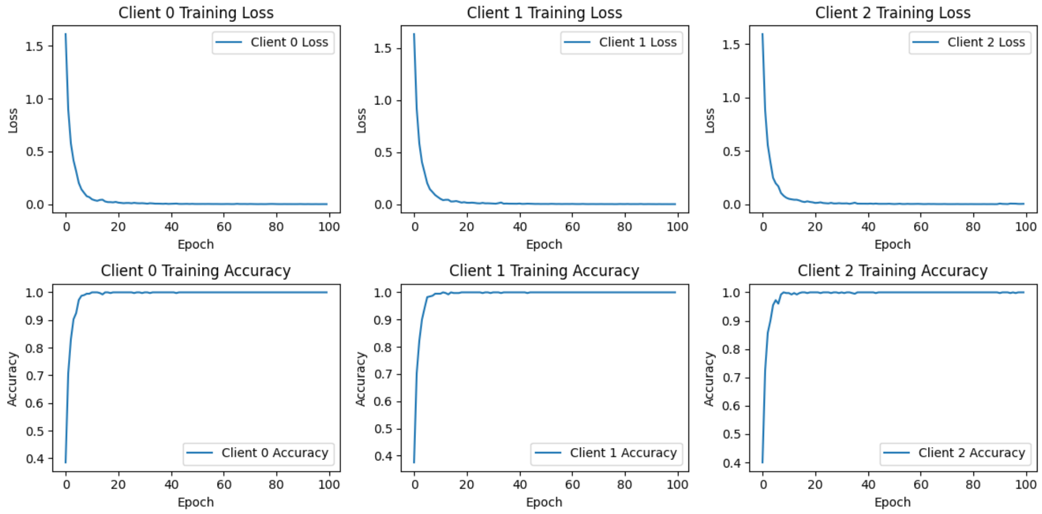


Figure 4. Training loss vs. accuracy for CNN.

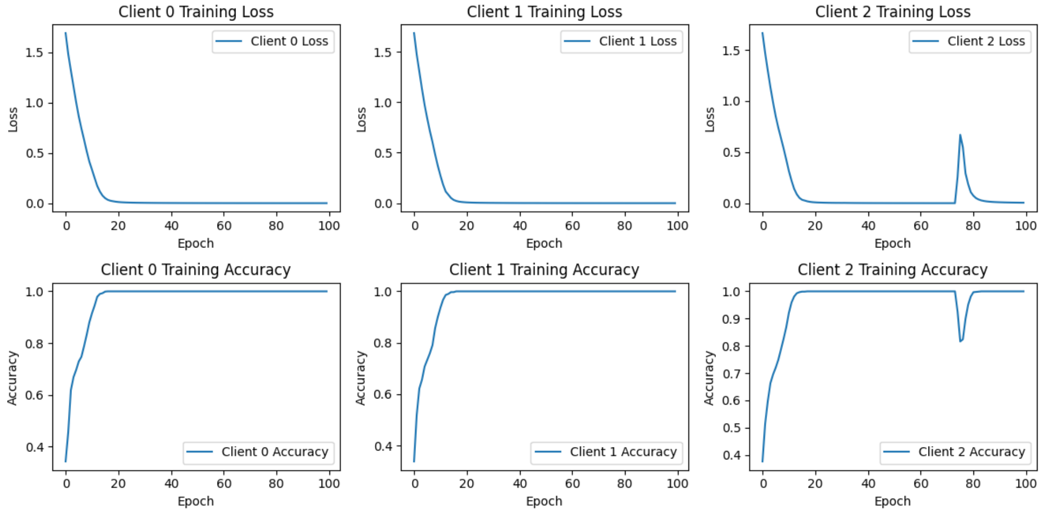


Figure 5. Training loss vs. accuracy for BiGRU.

The confusion matrix, Figure 1b, however, gives us some critical insights on prediction behaviours and some possible improvements. For example, the ‘happy’ class has a very good classification performance due to having a preponderance of true positives and a small number of classifications into other categories. Thus, the discriminative features represented by the FL and CNN model with DO (FL-CNN+DO) model for ‘happy’ have been effectively learned. Yet, they do appear to be confused, at least to some extent, as ’happy’ is frequently predicted as ‘anger’ (perhaps the same features or overlapping data patterns lead to both emotions being predicted that way). Just like this, there appears to be a notable misclassification trend between ‘fear’ and ‘happy’, possibly indicating feature similarity that could either benefit from some further separation of features or representation of data. Referring to the FL-CNN+DO presented in the confusion matrix, the model achieves high correct prediction for the happy and sadness classes with 1678 and 1169 respectively. This particular set of associations has been used to produce a misclassification of 426 instances of anger and 757 instances of sadness as happy.

The emphasis in the off-diagonal elements of the confusion matrix, i.e., misclassifications, is that there are opportunities for optimization, e.g., reducing false positives for ‘sadness’ and ‘love’. However, these challenges notwithstanding, the matrix shows a dominant diagonal overall, indicating strong overall predictive performance. Together, these visual analyses corroborate the performance of the FL-CNN+DO model by suggesting areas of potential future tuning and model improvement to reduce misclassification and improve class separation over complex emotional states (Figure 1b). Evaluation of the FL-CNN+DO for the classification of different emotions is provided by the ROC-AUC curve and the confusion matrix. The ROC-AUC curve shows that the model can reach robust discrimination ability with AUC values of 0.83 for 'anger', 'happiness' and 'sadness', but with a peak of 0.90 for 'surprise'. The curve shows that the model is very good at detecting a true positive rate compared to a false positive rate, so it is very good at performing most classes. The results of this emotion show such an exceptional sensitivity in identifying features that are unique to this emotion that the 'surprise' category stands out with an AUC of 0.90, Figure 2b.

The FL Bidirectional Gated Recurrent Unit model with DO (FL-BiGRU+DO) classification performance using several emotional categories such as anger, fear, happy, love, sadness, and surprise is shown in an ROC-AUC curve and confusion matrix (Figures 1c and 2c). This ROC-AUC plot demonstrates the discrimination capability of the model per emotion, and AUC values vary from 0.54 to 0.59 among classes, which indicates that the model can discriminate between the true positive and false positive rates relatively well. While AUC values do improve slightly for some emotions like surprise and love, approaching around 0.59, there are other emotions such as sadness that simply fail to reach 0.2, implying harder to identify emotions anyhow in federated domains experiencing obfuscation.

These performance nuances are shown even more clearly in the confusion matrix. For example, the model shows a relatively higher percentage of the 'happy' class being correctly predicted, amounting to 665 correctly predicted instances implying strong detection characteristic of this emotion. Notably, such misclassifications have yet to be fully resolved, including mislabelling of 'sadness'/'anger' manifestations as 'happy' or other labels, possibly because of shared features or imprecisely distinguishing between emotions under subverted data. It also holds information about the potential sources of improvements, such as confusion between 'anger' and 'fear', or the fact of mislabeled 'surprise' cases being labelled into other classes such as 'sadness'. The confusion matrix of FL-BiGRU+DO shows 165 instances of happy and 160 instances of sadness, which are misclassified as anger and 297 instances of sadness misclassified as happy, Figure 1c.

Overall, the sentiment analysis using obfuscated data in an FL context shows the promise of the FL-BiGRU+DO model but is challenged in terms of high discrimination for some emotions. Emotion detection in such privacy-preserving environments remains a complex problem; these results clearly indicate the difficulty of such a problem and underline the necessity for more fine-tuning and architectural improvements in order to reduce misclassification without compromising user data privacy.

4.3.1. Model Privacy Validation

This is a challenge when no common metrics have been identified for privacy performance comparisons between FL-BERT+DO and traditional FL-DP, Figure 6. 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. This 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 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 are 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 ML.

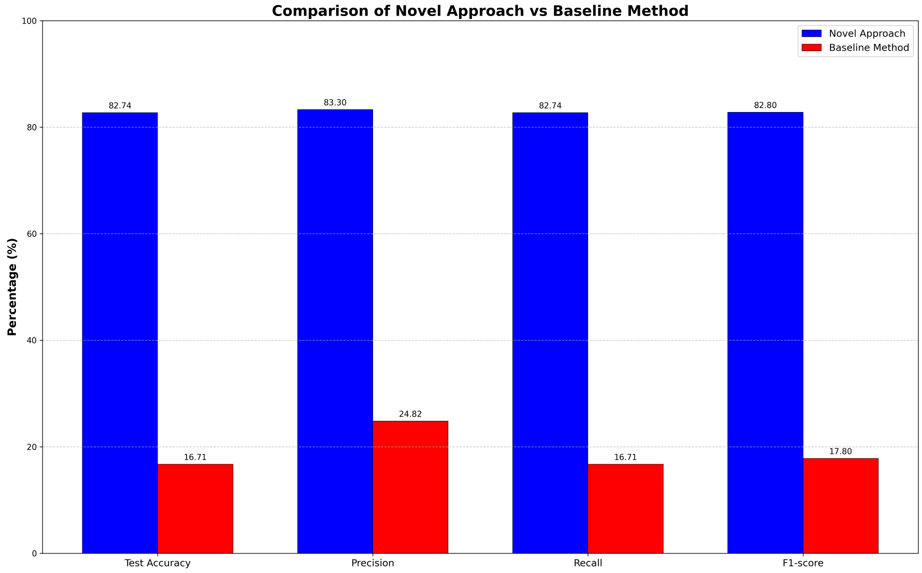


Figure 6. Comparative analysis of forecasting: our method vs. baseline methodology (DP).

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
	FL-BERT+DO		
Membership Inference	Global BERT	22.40%	Low
Membership Inference	Local BERT	50.38%	Moderate
Linkage Attack	Individual Clients (Macro-Avg.)	51.29%	Moderate
	FL-CNN+DO		
Membership Inference	Global CNN	37.36%	Low
Membership Inference	Local CNN	50.95%	Moderate
Linkage Attack	Individual Clients (Macro-Avg.)	50.72%	Moderate
	FL-BiGRU+DO		
Membership Inference	Global BiGRU	12.97%	Very Low
Membership Inference	Local BiGRU	31.48%	Low
Linkage Attack	Individual Clients (Macro-Avg.)	44.72%	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-BERT+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 [17,22], focused on enhancing precision using sophisticated DL methods such as CNNs and combined models that merge CNNs and LSTM. However, these studies frequently neglected confidentiality aspects. On the other hand, research such as [18,20] aimed to improve privacy through FL but often at the expense of consistent precision levels. For example, ref. [19] pointed out the inherent trade-offs associated with DP, 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. 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-BERT+DO approach effectively navigates these challenges, as shown by our experimental results. When compared to a baseline LSTM-based FL method that utilizes DP [7], our approach demonstrates considerable improvements in both accuracy and privacy assurances. As detailed in Table 1, our model reaches a forecast test 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, Figure 7. It shows a linearly growing epsilon, which means that privacy guarantees get steadily worse during training time.

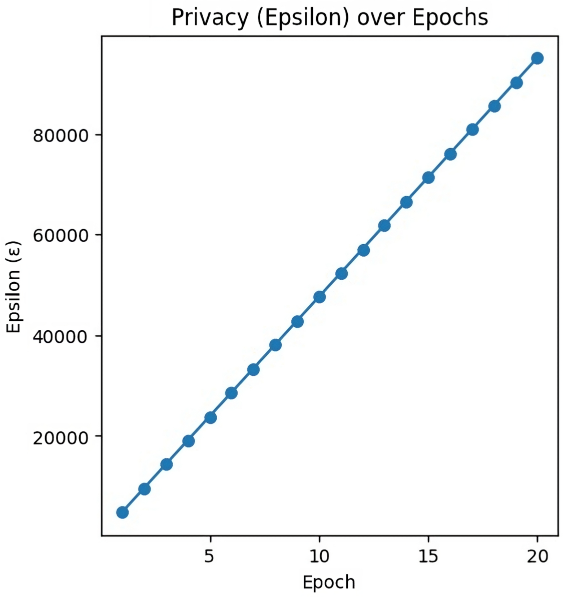


Figure 7. Comparative analysis of Epsilon (ϵ) vs. Epochs for FL-DP.

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 (FL-DP) is illustrated in Figure 6.

Table 1 reveals that FL-CNN+DO provides an effective measure and an accuracy level of 61.34%, a precision of 72.79%, and an F1-Score of 62.40%. Such findings suggest that current solutions have a fair accuracy in both detecting circumstances which lead to a high likelihood of reoffending and detecting cases where such predictions would be wrong. In terms of privacy protection, it has 37.36% global membership inference attacks, 50.95% local attacks, and 50.72% success in linkage attacks for each client (Table 4).

The accuracy of the given test set is 57.83%, showing that the FL-BiGRU+DO model has well-defined features for emotion classification; the precision of the given model is 62.05%; and the given model has an F1-score of 58.32%, Table 1. The model showed that it had a very low probability of global membership inference (12.97%), a low probability for local membership inference (31.48%), and, however, a moderate probability for linkage attack on individual clients (44.72%). From these outcomes, we can discern that the model exhibits high classification discriminant capabilities but moderate privacy resilience compared to leakage risks, especially at the client level (Table 4).

Through analyses of the test set metrics, we observe that the FL-BERT+DO model performs quite favourably for all considered metrics: accuracy = 82.74%, precision = 83.3%, recall = 82.74%, F1 score = 82.8%. On the other hand, FL-CNN+DO shows moderate performance and a slightly less score that is; a precision score of 72.79% and an F1 score of 62.4%. The use of the FL-BiGRU+DO model presents the lowest results in terms of accuracy and recall: 57.83%, and F-1 of 58.32%. The FL-BERT+DO approach demonstrates better general performance and predictive recovery, as seen in these results, for forecasting tasks. Comparative analyses of forecasting for different models are shown in Figure 8.

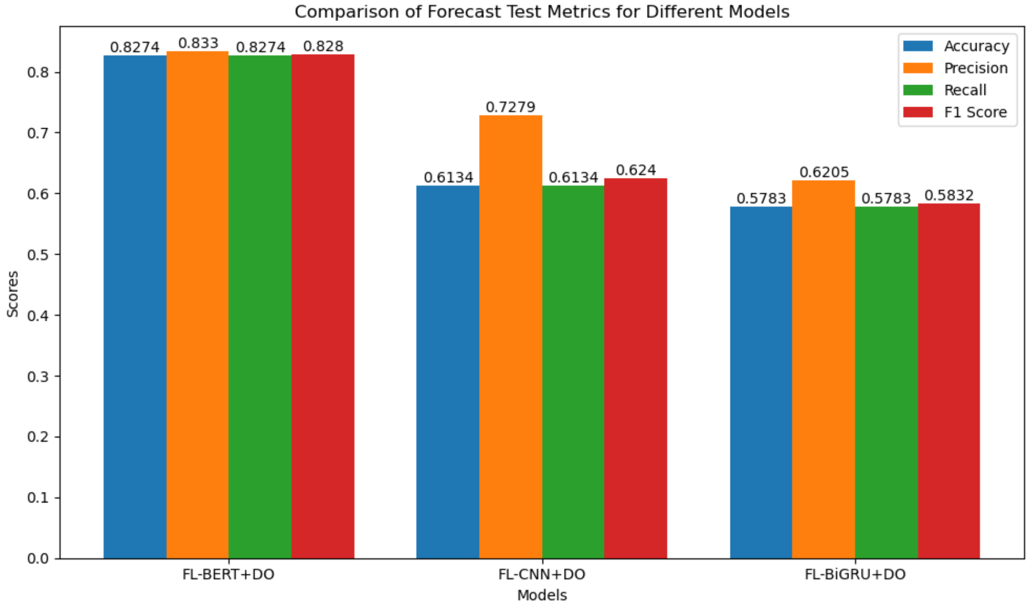


Figure 8. Comparative analysis of forecasting for different models.

5.2. Interpretation of Findings

From this understanding, the study's results assert the viability of the FL method alongside BERT for the field of sentiment analysis in mental health. Our model points out that FL 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 in 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 healthcare practitioners can be derived. The incorporation of FL means that patient mental health can be constantly, safely, and non-intrusively monitored by medical staff. This could mean that from 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 data scientists a road map for using FL in tandem with DP 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-BERT+DO vs. FL-BERT+DP

FL techniques employing DP have shown considerable potential for protecting the privacy of data; nevertheless, FL-BERT+DO, which we propose, offers a novel approach that eliminates the burdensome limitations of DP-based methods. The advantage of using FL-BERT+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, such as Active Personalized FL (ActPerFL) [37] and Topic Memory FL with BERT-Large TM-FL (BT-1) [26], our FL-BERT+DO approach is more focused on masking sensitive data right before model updates, which enhances data utility and reliability. FL-BERT+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-BERT+DO works better than ordinary FL-DP models in terms of predicted accuracy as described in Section 5.1.

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-BERT+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-BERT+DO against privacy breaches because it inherently protects the database from linkage and membership inference attacks. As it was revealed in testing, FL-BERT+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 DP (FL-DP) approaches.

5.4.3. Greater Robustness with Data-Level Privacy Protection

FL-BERT+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-BERT+DO's robustness against real-world FL scenarios, which are typical for the mental health domain and involve clients with potentially different data distributions or data quality. Consequently, FL-BERT+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 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-BERT+DO can provide fine-grained privacy control that can hide specific phrases or entities while preserving the semantics. FL-BERT+DO is more suitable for NLP tasks, where information semantics retention is critical to achieving accuracy and distinctiveness since it provides a thorough privacy mechanism that DP cannot.

6. Practical Applications and Limitations

The FL-BERT+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-BERT+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-BERT+DO will be able to function optimally even within limited resource conditions like consumer-based end devices, including those based on smartphones or the poorly energy-efficient IoT contents frequently involved in mental health practices, is difficult. Due to the need to meet the specifications of FL and the added challenge of data obfuscation, it can be awkward to implement FL-BERT+DO in such scenarios; as such, it is suggested that future research focuses on finding ways of improving the efficiency of FL-BERT+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, privacy is improved while permissive sentiment research is conducted.

Also, AI can play a highly important role in DO as well. The ML case comes when algorithms are developed to logically transform values and thus hide information while retaining 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.

Some prospects for future work with regard to data acquired for data obfuscation techniques include further exploration of the duality between privacy preservation and the usefulness of data. Future work might look into higher-level analogues of masking that would retain as much information as possible but would not compromise privacy. Moreover, further research has to be directed to cognitive obscuring techniques by considering the characteristics of data and the context in which it will be used. Moreover, procedures for normalizing and subsequently verifying that masked data are still appropriate for the learning algorithms would be useful.

The scalability of FL-BERT+DO strongly depends on the ability to work with a large number of devices with different computational and network resources. In order to scale up the systems, efficient means of data compression, privacy preservation, and realistic communication protocols with low latencies and bandwidth overheads are required. Moreover, modifying the framework to work with many kinds of mobile systems increases its usage and allows real-time data gathering to monitor mental health and help adopt AI models for individualized treatment.

8. Conclusions

In this work, the feasibility and effectiveness of employing BERT models and FL 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 dataset with good performance indicators. The requirement for strict data protection while developing mental health applications was well addressed by FL 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.

To illustrate how decentralized training can retain privacy and enhance accuracy, it begins with a case of using BERT on sentiment analysis in FL. 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 mixture 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 this work, specifically DP. Finally, this 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 emphasize the need for future studies and the application of privacy-preserving methods in healthcare 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 the practicability of 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.

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.; writing—editing, R.H.; 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: The original data presented in this study are openly available in FigShare at https://doi.org/10.6084/m9.figshare.27247368.v1.

Acknowledgments: Special thanks are given to Paul Yoo, Chair of the Threat Intelligence Lab at the School of Computing and Mathematical Sciences, Birkbeck, University of London, for invaluable support on research in the field of data privacy, including data obfuscation techniques.

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

References

1. Alsharif, M.H.; Kannadasan, R.; Wei, W.; Nisar, K.S.; Abdel-Aty, A.H. A contemporary survey of recent advances in federated learning: Taxonomies, applications, and challenges. *Internet Things* **2024**, *27*, 101251. [CrossRef]
2. Yuan, L.; Wang, Z.; Sun, L.; Yu, P.S.; Brinton, C.G. Decentralized federated learning: A survey and perspective. *IEEE Internet Things J.* **2024**, *11*, 34617–34638. [CrossRef]
3. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017.
4. Powar, J.; Beresford, A.R. SoK: Managing risks of linkage attacks on data privacy. *Proc. Priv. Enhancing Technol.* **2023**, *2023*, 97–116. [CrossRef]
5. Muralidhar, K.; Sarathy, R. Data shuffling—A new masking approach for numerical data. *Manag. Sci.* **2006**, *52*, 658–670. [CrossRef]
6. Banabilah, S.; Aloqaily, M.; Alsayed, E.; Malik, N.; Jararweh, Y. Federated learning review: Fundamentals, enabling technologies, and future applications. *Inf. Process. Manag.* **2022**, *59*, 103061. [CrossRef]
7. Naseri, M.; Hayes, J.; De Cristofaro, E. Local and Central Differential Privacy for robustness and privacy in federated learning. *arXiv* **2020**, arXiv:2009.03561.
8. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]

9. Huang, N.E.; Daubechies, I.; Hou, T.Y. Adaptive data analysis: Theory and applications. *Philos. Trans. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150207. [CrossRef]
10. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **2021**, *11*, 81. [CrossRef]
11. Singh, N.K.; Tomar, D.S.; Sangaiah, A.K. Sentiment analysis: A review and comparative analysis over social media. *J. Ambient Intell. Humaniz. Comput.* **2020**, *11*, 97–117. [CrossRef]
12. Khoshnam, F.; Baraani-Dastjerdi, A. A dual framework for implicit and explicit emotion recognition: An ensemble of language models and computational linguistics. *Expert Syst. Appl.* **2022**, *198*, 116686. [CrossRef]
13. Islam, M.S.; Kabir, M.N.; Ghani, N.A.; Zamli, K.Z.; Zulkifli, N.S.A.; Rahman, M.M.; Moni, M.A. Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach. *Artif. Intell. Rev.* **2024**, *57*, 62. [CrossRef]
14. Alsayat, A. Improving sentiment analysis for social media applications using an ensemble deep learning language model. *Arab. J. Sci. Eng.* **2022**, *47*, 2499–2511. [CrossRef] [PubMed]
15. Stappen, L.; Baird, A.; Christ, L.; Schumann, L.; Sertolli, B.; Messner, E.M.; Cambria, E.; Zhao, G.; Schuller, B.W. The MuSe 2021 multimodal Sentiment Analysis challenge: Sentiment, emotion, physiological-emotion, and stress. *arXiv* **2021**, arXiv:2104.07123.
16. Liu, F.; Hou, K.; Dong, Y. Deep parallel contextual analysis framework based emotion prediction in community wellness communications on social media. *Heliyon* **2024**, *10*, e31626. [CrossRef]
17. Heikal, M.; Torki, M.; El-Makky, N. Sentiment Analysis of Arabic Tweets using Deep Learning. *Procedia Comput. Sci.* **2018**, *142*, 114–122. [CrossRef]
18. Ramaswamy, S.; Mathews, R.; Rao, K.; Beaufays, F. Federated learning for emoji prediction in a mobile keyboard. *arXiv* **2019**, arXiv:1906.04329.
19. Basu, P.; Roy, T.S.; Naidu, R.; Muftuoglu, Z.; Singh, S.; Mireshghallah, F. Benchmarking Differential Privacy and Federated Learning for BERT Models. *arXiv* **2021**, arXiv:2106.13973.
20. 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. [CrossRef]
21. 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. *Knowl.-Based Syst.* **2023**, *268*, 110475. [CrossRef]
22. Goularas, D.; Kamis, S. Evaluation of deep learning techniques in sentiment analysis from twitter data. In Proceedings of the 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), Istanbul, Turkey, 26–28 August 2019.
23. Sinha, A.; Chakma, K. A comparative analysis of machine learning based sentiment analysis. In *Communications in Computer and Information Science*; Springer Nature: Cham, Switzerland, 2022; pp. 123–132.
24. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multi-modal Sentiment Analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 1731–1740.
25. Zhang, W.; Zhou, Y.; Chen, M.; Chen, J. Benefits and Challenges of Federated Learning in IoT Systems. *J. Netw. Comput. Appl.* **2022**, *204*, 103–114.
26. Qin, H.; Chen, G.; Tian, Y.; Song, Y. Improving Federated Learning for Aspect-based Sentiment Analysis via Topic Memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3942–3954.
27. Russell, J.A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
28. Rawat, T.; Jain, S. A dimensional representation of depressive text. In *Data Analytics and Management*; Springer: Singapore, 2021; pp. 175–187.
29. Hollander-Gijsman, M.E.; De Beurs, E. The Use of Dimensional Models for Monitoring Changes in Depression and Anxiety. *J. Affect. Disord* **2013**, *147*, 33–42.
30. Wu, J.L.; He, Y.; Yu, L.C.; Lai, K.R. Identifying emotion labels from psychiatric social texts using a bi-directional LSTM-CNN model. *IEEE Access* **2020**, *8*, 66638–66646. [CrossRef]
31. Ibrahim Khalaf, O.; Ashokkumar, S.R.; Algburi, S.; Anupallavi, S.; Selvaraj, D.; Sharif, M.S.; Elmedany, W. Federated learning with hybrid differential privacy for secure and reliable cross-IoT platform knowledge sharing. *Secur. Priv.* **2024**, *7*, e374. [CrossRef]
32. 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. [CrossRef]
33. Juyal, I. Emotions in Text. 2023. Available online: https://www.kaggle.com/datasets/ishantjuyal/emotions-in-text (accessed on 24 September 2024).
34. Zhang, C.; Zhang, W.; Wu, Q.; Fan, P.; Fan, Q.; Wang, J.; Letaief, K.B. Distributed deep reinforcement learning based gradient quantization for federated learning enabled vehicle edge computing. *IEEE Internet Things J.* 2024, *early access*. [CrossRef]
35. Jiang, B.; Li, J.; Wang, H.; Song, H. Privacy-preserving federated learning for industrial edge computing via hybrid differential privacy and adaptive compression. *IEEE Trans. Industr. Inform.* **2023**, *19*, 1136–1144. [CrossRef]

36. Taha, Z.K.; Yaw, C.T.; Koh, S.P.; Tiong, S.K.; Kadirgama, K.; Benedict, F.; Tan, J.D.; Balasubramaniam, Y.A.L. A survey of federated learning from data perspective in the healthcare domain: Challenges, methods, and future directions. *IEEE Access* **2023**, *11*, 45711–45735. [CrossRef]
37. Chen, H.; Ding, J.; Tramel, E.; Wu, S.; Sahu, A.K.; Avestimehr, S.; Zhang, T. ActPerFL: Active personalized federated learning. In Proceedings of the First Workshop on Federated Learning for Natural Language Processing (FL4NLP 2022), Dublin, Ireland, 27 May 2022; Lin, B.Y., He, C., Xie, C., Mireshghallah, F., Mehrabi, N., Li, T., Soltanolkotabi, M., Ren, X., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 1–5.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.