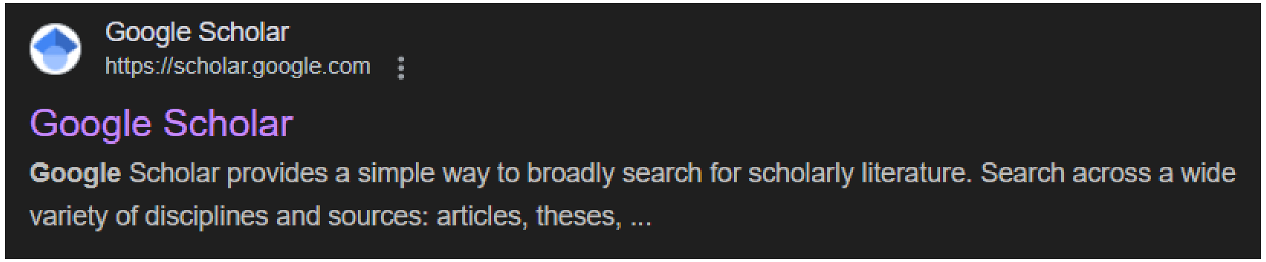
electronics
MDPI
Article
Privacy-Enhanced Sentiment Analysis in Mental Health: Federated Learning with Data Obfuscation and Bidirectional Encoder Representations from Transformers

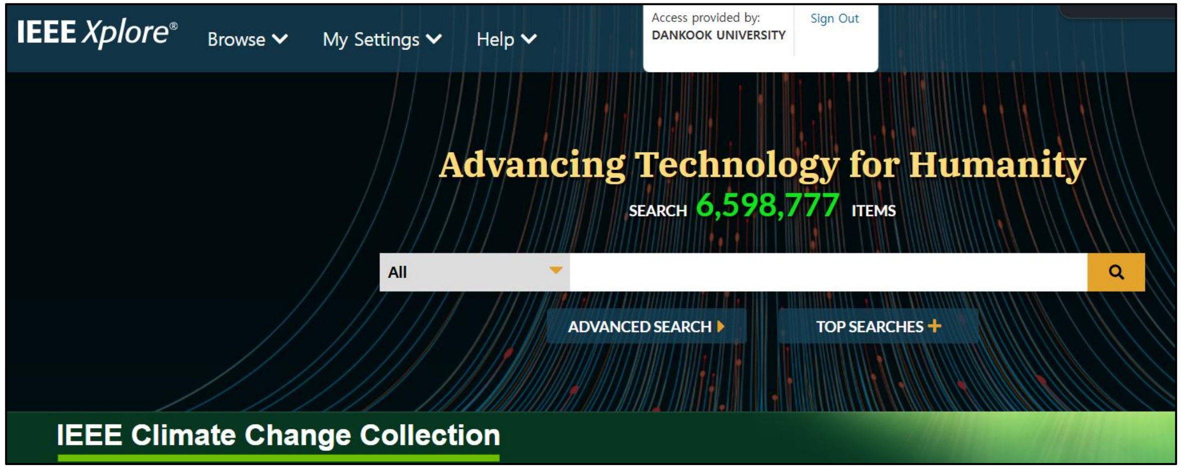
Shakil Ibne Ahsan *, Djamel Djenouri and Rakibul Haider

Federated Learnig with Data Obfuscation and Bidirectional Encoder Representations from Transformers Paper review

모바일시스템공학과
이승재

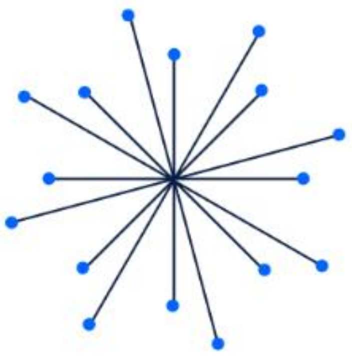
0. How to find a paper



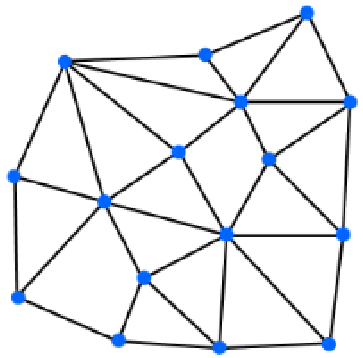


1. Introduction – Motivation

"As machine learning continues to gain popularity, sensitive user privacy has become an important issue, and FL has emerged, but it also has difficulties in ensuring complete privacy and accuracy."



Centralized



Distributed
(FL)

+

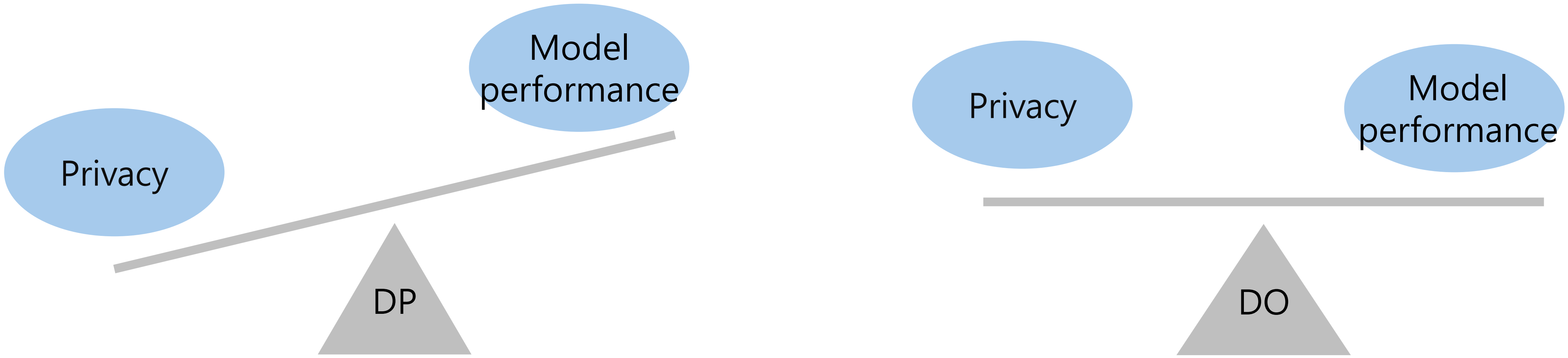
Differential privacy
(DP)

!

Data obfuscation
(DO)

1. Introduction – Motivation

"Existing methods such as DP can significantly degrade model performance by adding noise to data, making it difficult to balance privacy and accuracy, so new approaches for balanced models have been explored."



1. Introduction – Motivation

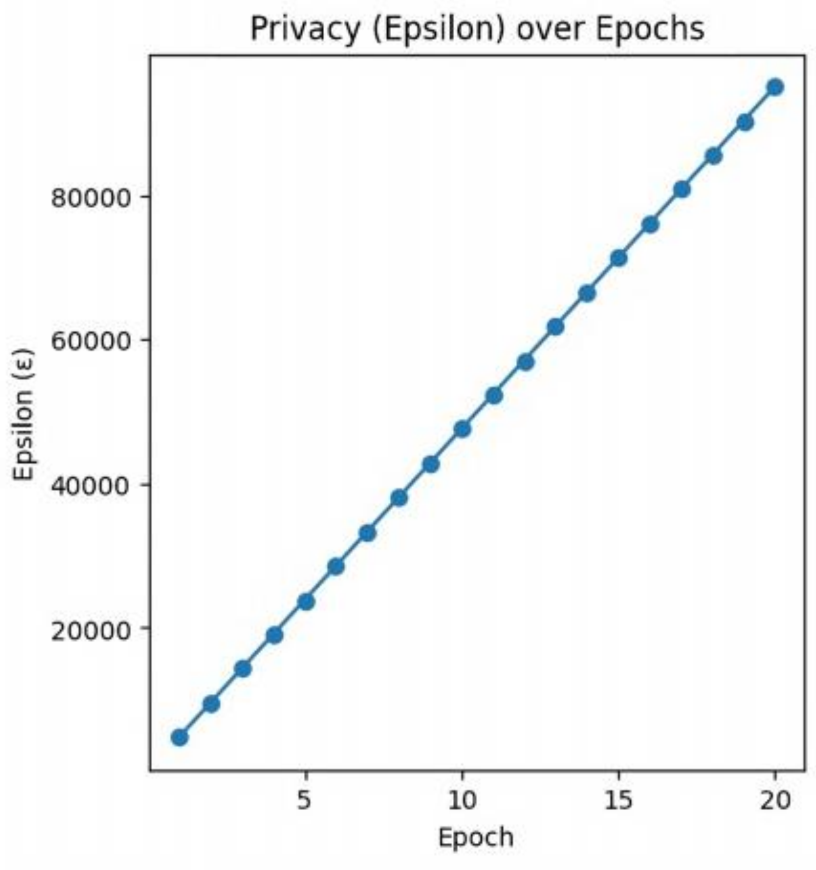


Figure 7. Comparative analysis of Epsilon (ϵ) vs. Epochs for FL-DP.

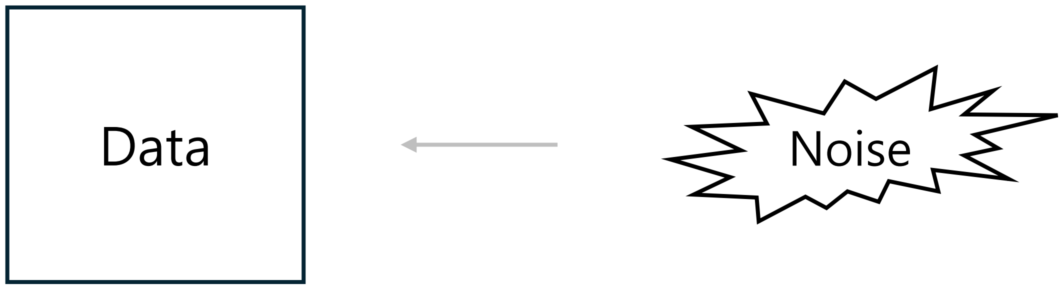
"DP not only has low model performance, but also reduces privacy protection strength as learning progresses."

Training time ↑	ε ↑	Privacy guarantees ↓

Data exposure risk when adding noise to data in DP

2. Background – DP & DO

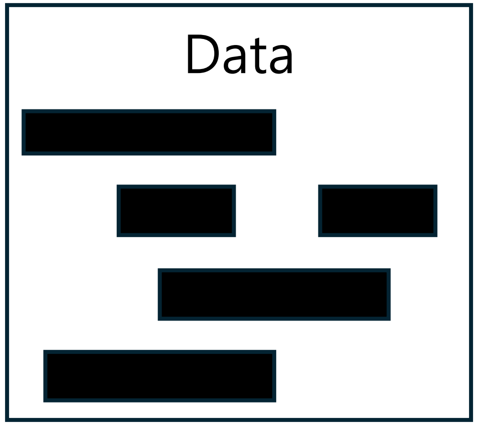
Differential Privacy



$$\mathbb{P}[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{A}(D_2) \in S].$$

https://seewoo5.tistory.com/23

Data Obfuscation



2. Background – More about DO

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.

2. Background – Federated Learning

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

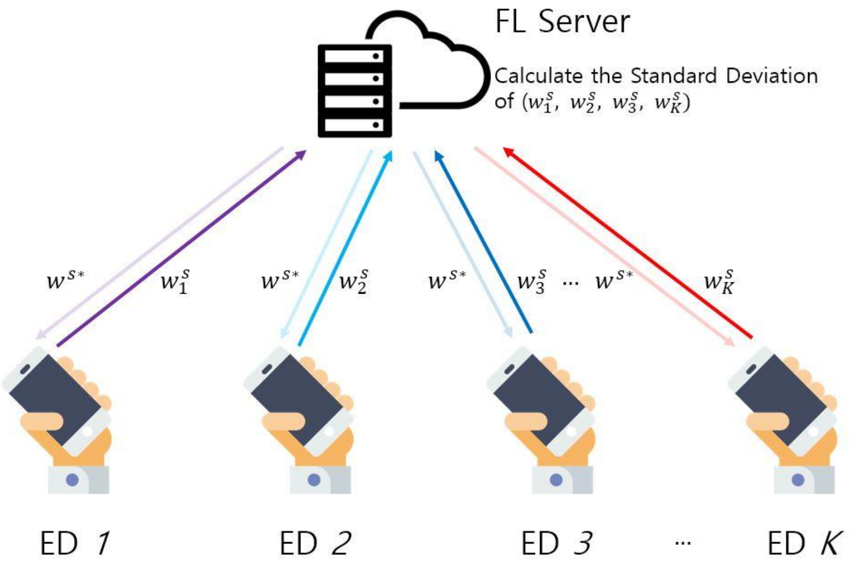
Server executes:
  initialize w_0
  for each round $t = 1, 2, \dots$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $m_t \leftarrow \sum_{k \in S_t} n_k$
    $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$ *// Erratum*[4]

ClientUpdate(k, w)**:** *// Run on client k*
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return w to server

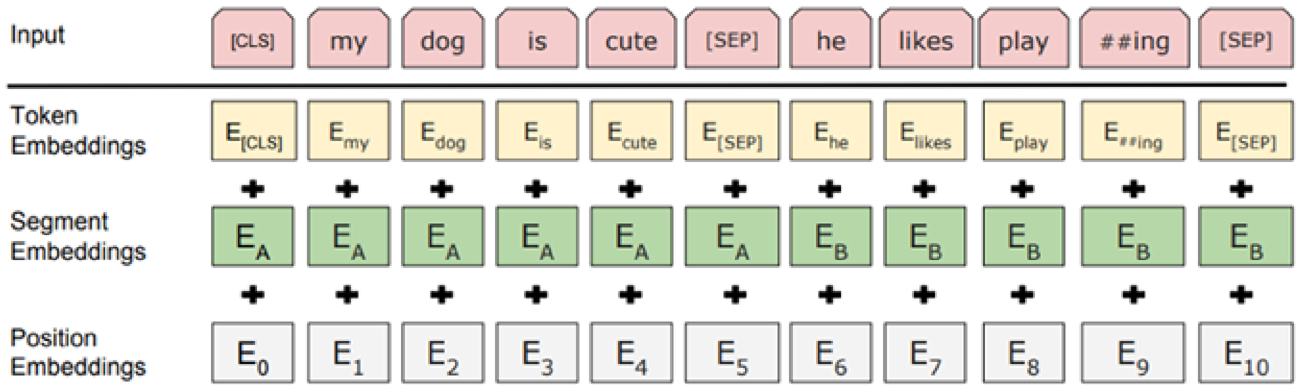
3. FederatedAveraging (FedAvg) 알고리즘

- **알고리즘 개요**: FedAvg는 각 클라이언트가 로컬 데이터로 여러 번의 스토캐스틱 경사 하강법(SGD)을 수행한 후, 중앙 서버가 이들을 평균하여 글로벌 모델을 업데이트하는 방식입니다.
- **단계별 과정**:
  1. **초기화**: 서버는 초기 글로벌 모델을 설정합니다.
  2. **클라이언트 선택**: 매 라운드마다 일부 클라이언트를 무작위로 선택합니다.
  3. **로컬 학습**: 선택된 클라이언트는 현재 글로벌 모델을 기반으로 로컬 데이터에서 여러 번의 SGD를 수행하여 모델을 업데이트합니다.
  4. **모델 업로드**: 업데이트된 로컬 모델 파라미터를 서버로 전송합니다.
  5. **모델 평균화**: 서버는 수신된 로컬 모델 파라미터를 평균하여 새로운 글로벌 모델을 생성합니다.
  6. **반복**: 원하는 성능에 도달할 때까지 2~5단계를 반복합니다.

"FL is a machine learning technique that enables multiple devices or systems to collaboratively train a model without sharing raw data. Instead of sending data to a central server, each device processes its data locally and shares only model updates (e.g., gradients) with the central server."



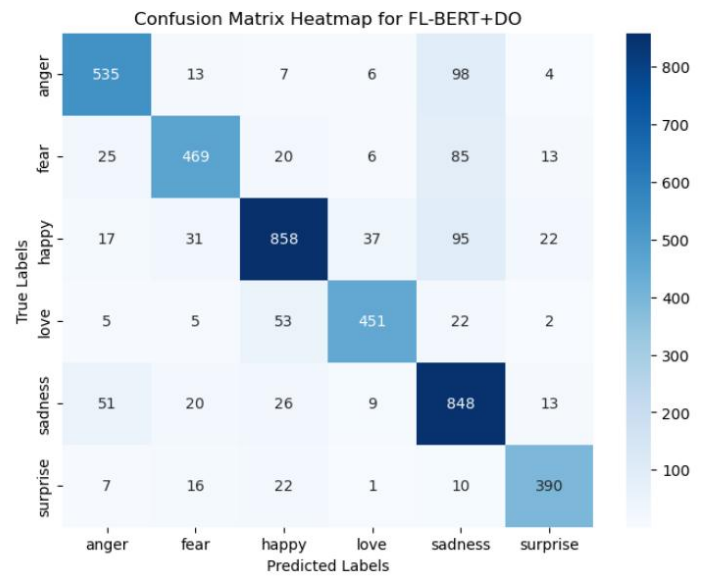
2. Background – BERT Configuration



"Optimization factors for training the BERT model in this paper"

- Cross-Entropy Loss
- Learning Rate : 1×10^{-5}
- AdamW
- Weight Decay
- Learning Rate Scheduler
- Mixed-Precision Training
- Batch Size = 16

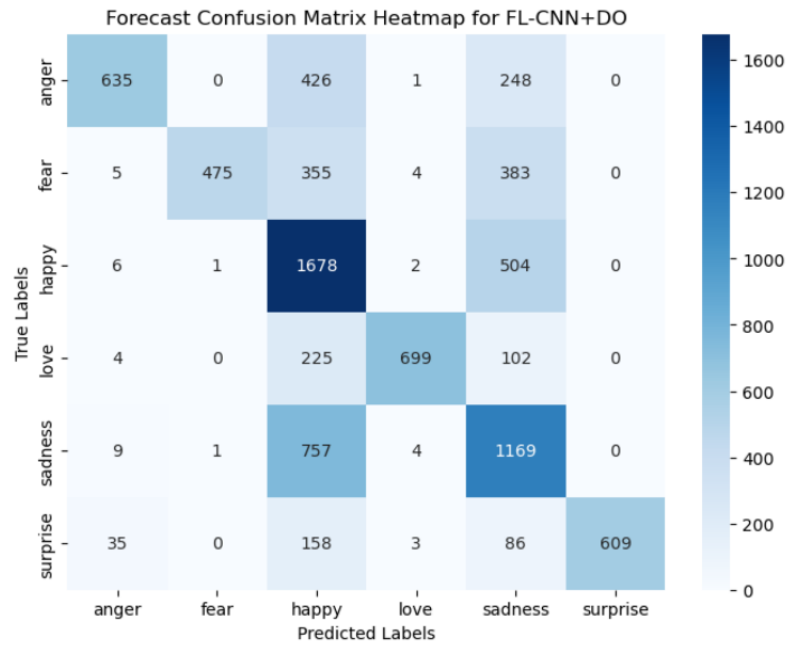
3. Experimentation – Emotional Perspective



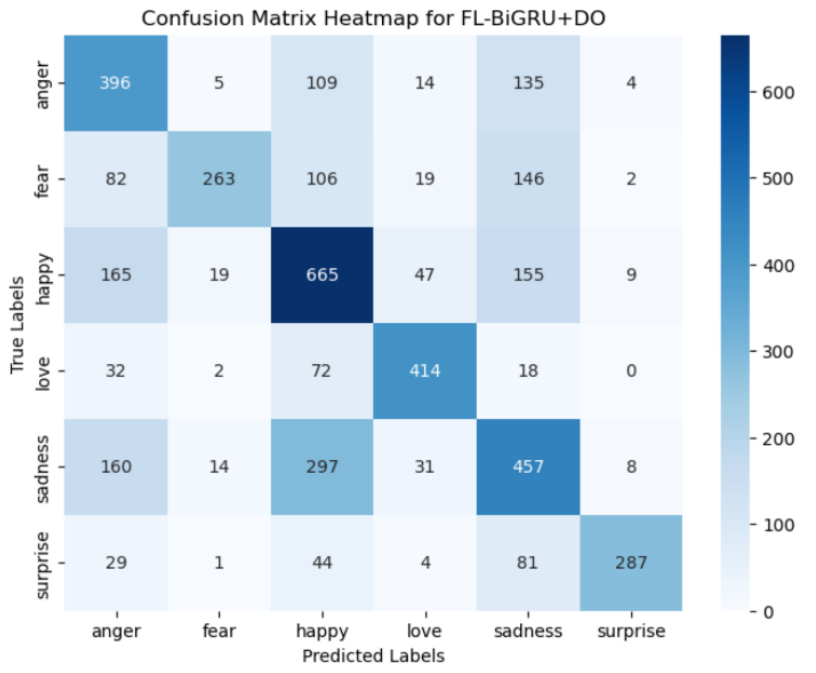
(a) Confusion matrix for BERT.

[FL-BERT+DO]

[FL-CNN+DO]



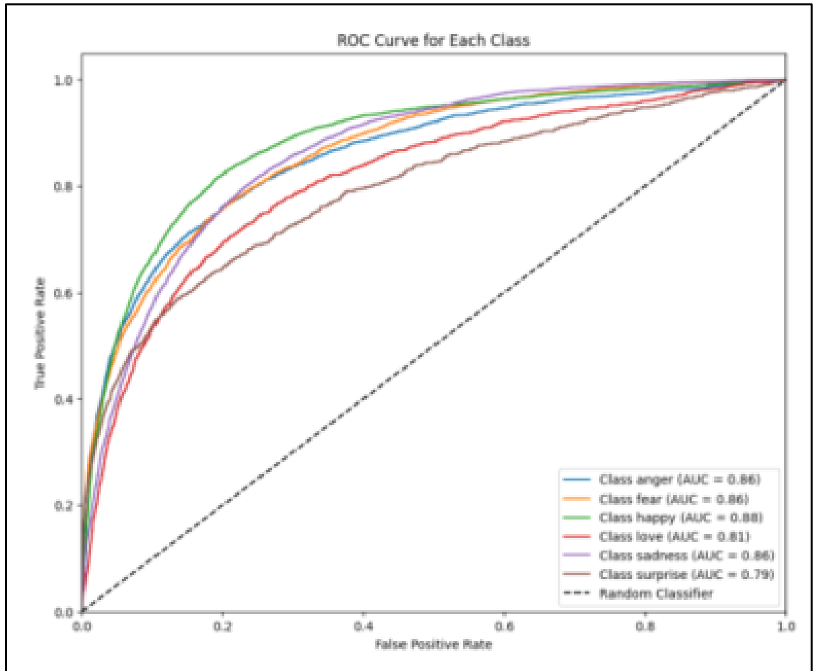
(b) Confusion matrix for CNN.



(c) Confusion matrix for BiGRU.

[FL-BiGRU+DO]

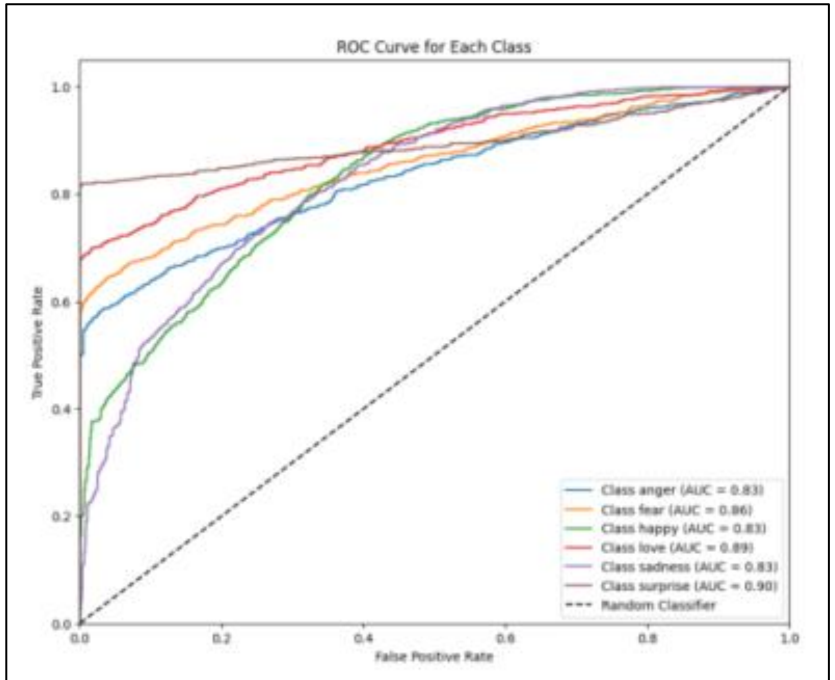
3. Experimentation – Emotional Perspective



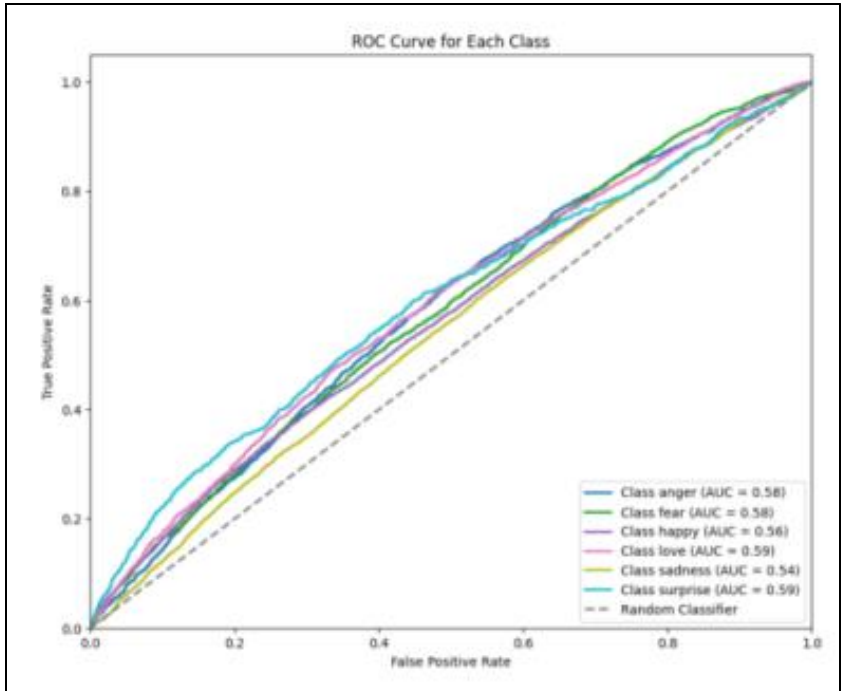
(a) ROC-AUC curve for BERT.

[FL-BERT+DO]

[FL-CNN+DO]



(b) ROC-AUC curve for CNN.



(c) ROC-AUC curve for BiGRU.

[FL-BiGRU+DO]

2. Experimentation – Emotional Perspective

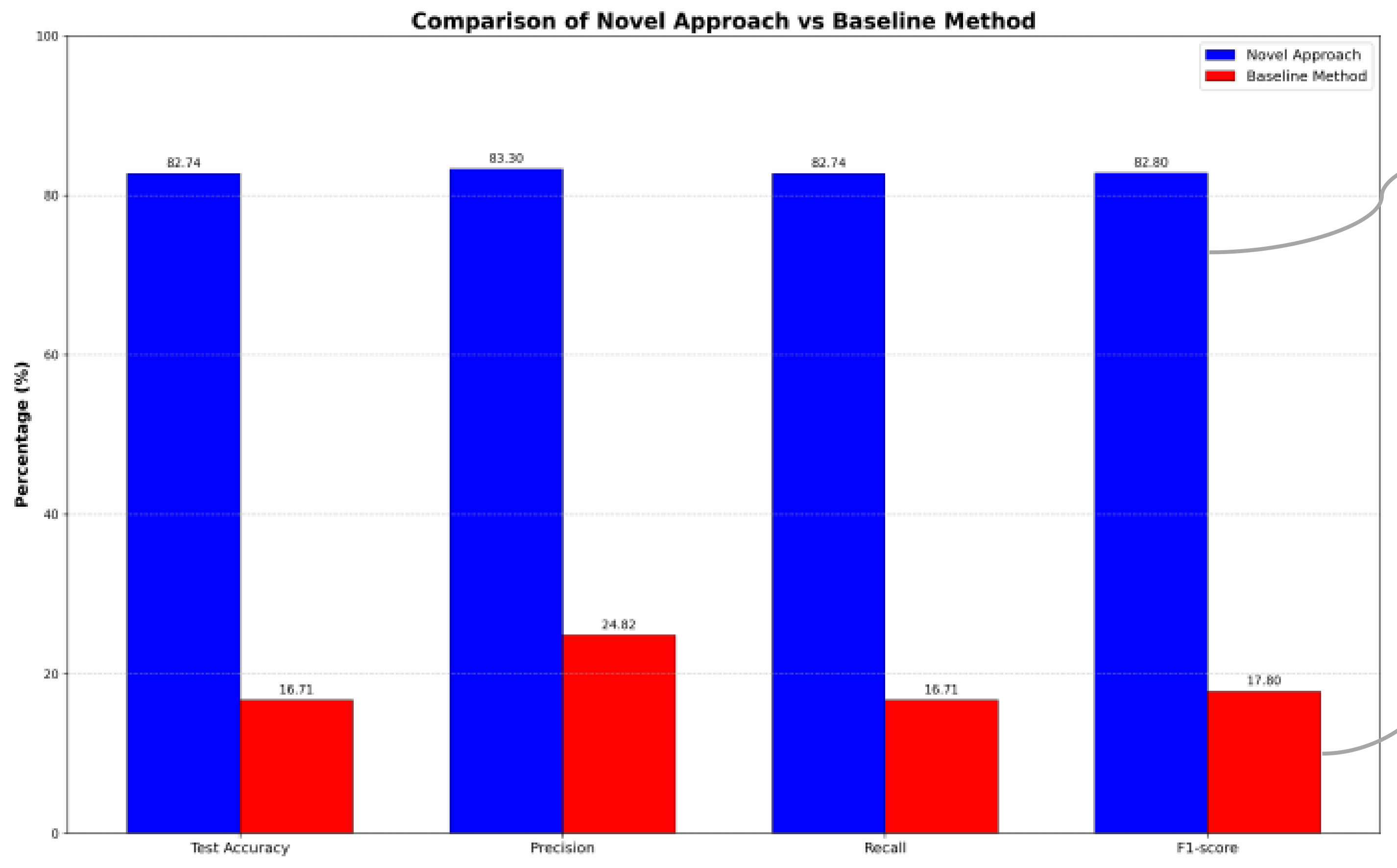


Figure 6. Comparative analysis of forecasting: our method vs. baseline methodology (DP).

FL-BERT + DO

FL-BERT + DP

2. Experimentation - Privacy Perspective

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

It has the lowest risk to privacy, but it is not balanced with the model performance as shown in the previous experiment.

3. Future Directions

"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."