Exploring Federated Learning to Trace Depression in Social Media with Language Models

Arthur B. Vasconcelos
Universidade Federal Fluminense
Institute of Computing
Niterói, Brazil
athurbittencourt@id.uff.br

Lúcia Maria de A. Drummond
Universidade Federal Fluminense
Institute of Computing
Niterói, Brazil
lucia@ic.uff.br

Rafaela C. Brum
Universidade Federal Fluminense
Institute of Computing
Niterói, Brazil
Sorbonne Université, LIP6
Paris, France
rafaelabrum@id.uff.br

Aline Paes
Universidade Federal Fluminense
Institute of Computing
Niterói, Brazil
alinepaes@ic.uff.br

***Abstract*—Due to the rising numbers of depression cases in recent years, many initiatives have investigated the use of Machine Learning models to detect depressive symptoms from the individual's presence on social media. To train these models, a dataset is needed. An adequate way of collecting reliable data is to elicit volunteers to agree to share their posts for research. Usually, the volunteer is also requested to answer a depressive inventory to provide the required depression label. However, this data is often sensitive and cannot be shared between research groups, harming reproducibility and collaboration. To address that problem, in this manuscript, we investigate Federated Learning techniques to train a classifier depression method while still preserving the data privacy of individuals. Since social media posts are primarily text-based, we fine-tune language models induced by the Transformer architecture to our task. In our experiments, we simulate the common heterogeneity across clients. Our experiments show that Federated Learning achieves competitive models compared to the centralized version.**

***Index Terms*—Federated Learning, Machine Learning, BERT, Transformers, Depression Classifier, Social Media**

I. Introduction

Depression comprises mood disorders observed from the presence of sad, empty, or irritable mood, accompanied by somatic and cognitive changes that significantly affect the individual's capacity to function [1]. According to the World Health Organization's (WHO) global health estimates, cases of depression have been on the rise worldwide [2]. To aggravate those already alarming rates, the organization has also recently pointed out that COVID-19 pandemic has triggered a rise in the prevalence of mental health issues, including anxiety and depression[1]. The reasons for such an increase remain under investigation, as well as how to identify individuals at risk and adequately treat them. Unfortunately, depression is still not fully understood by society as a health issue, even being stigmatized and mixed with other mental health symptoms and states of mind. Even worse, acquiring a correct diagnosis and proper treatment is not available to all individuals due to a lack of knowledge, strength, or financial resources. Consequently, many initiatives have investigated using Machine Learning models for depression detection in social media.

To train such models, a dataset is needed. For that, there are two main ways to gather this data. The first is using openly available posts from a social media platform annotated by experts. The second is to ask volunteers to share their posts for research and answer depression questionnaires for labeling the data. The second method is more adequate, given that the annotation for the examples is more sound and based on the psychology literature. Usually, the volunteers get a score for depression symptoms based on a series of questions, employing tools widely used in psychology, such as the Beck's Depression Inventory (BDI) [3]. However, this data is often sensitive and cannot be shared between research groups.

To address that problem, in this paper, we investigate Federated Learning techniques to train depression screening models. Furthermore, we investigate how to screen for depression using data posted by individuals on social media. Given the textual nature of our task, we explore Transformer-based language models [4] to encode the posted publications as numerical vectors. Specifically, we chose BERT [5] as our language model, given that it has obtained successful results on the same problem in previous work [6]. However, while Transformer-based models constitute the state-of-the-art for Natural Language Processing-based tasks, they also pose additional challenges for Federated Learning [7]. Besides the usual concern of Federated Learning research with learning in heterogeneous and non-iid environments, Transformer-based models have huge sizes and are heavily based on the attention mechanism. Sharing large models and aggregating attention weights is not trivial.

This paper presents preliminary results concerning the following research question: *"How does transformer-based*

[1] https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

979-8-3503-8160-3/23/$31.00 ©2023 IEEE
DOI 10.1109/SBAC-PADW60351.2023.00014

federated learning compare to its centralized version to create privacy-preserving depression classifiers?”. To answer the question, we design a set of experiments simulating class preserving homogeneous data distribution for automatic classification of depression level according to BDI. Our preliminary results point out that we can leverage Federated Learning formulation to preserve the sensitivity of the data while still achieving competitive results compared to the centralized version.

The rest of the paper is organized as follows: Section II presents the related works, Section III brings the basic concepts we built our contribution. Section IV brings the methodology and formulation proposed in this paper. Section V presents the experimental results and discusses them, and Section VI concludes the paper.

II. Related work

The significant increase in mental disorders worldwide in recent years has made it challenging to clinically analyze all the possible cases to identify those requiring immediate care. Thus, several works have proposed automatic methods for detecting depression based on user-generated data from social media [8], [9]. Most of the works focused on post-based classification, while others propose formulating the problem by looking at the individual and its set of posts [6], [10].

Moreover, different initiatives have investigated using Federated Learning as a solution for privacy preservation in automatic depression detection. Some have explored its use on speech-based depression detection models [11]–[13]. The technique presented in [11] uses the DAIC-WOZ dataset to train different models in a controlled environment and deploys them in a smartphone to assess performance overhead. In [12], a framework to extend an existing machine learning model to work in a federated setting was proposed. In addition, that work also investigates the effect of a more extensive dataset facilitated by Federated Learning compared to local training. The work presented in [13] uses the DAIC-WOZ dataset but focuses on improving the privacy protection of a Federated Learning setting, using techniques such as differential privacy and norm bounding.

Others, like our work, have also investigated individuals’ written texts but focusing on asynchronous Federated Learning for the task of depression detection [14], [15]. Additionally, in [16], different BERT models with IID and Non-IID distributions using differential privacy were investigated. They focus on benchmarking differential privacy between models and dataset distributions. In [17], a model that makes detection daily using a recurrent neural network (RNN) was proposed. This work focuses on state-of-the-art Federated Transformers to deal with written texts.

III. Key Concepts

A. Federated Learning

Federated Learning is a decentralized form of Machine Learning where we consider a federation of clients that respond to a central server [18]. Federated learning introduces a new stage into the learning procedure called a round. A round is described as follows: each client trains locally on their data and communicates their learned parameters to the central server; the central server aggregates the received training, sends back these updates to the clients for testing, and starts the next round.

The process of receiving the client’s training is called sampling while updating a client’s parameters with the central server’s parameters is called updating. A Federated Learning process may be configured to sample and update from a fraction of the clients. Whether a fraction or all clients participate in the sampling or updating depends on the Federated Learning type, which can be Cross-Device or Cross-Silo.

Cross-Device Federated Learning is mainly used when the federation’s clients are numerous, have low power, and can store limited data (such as a cellphone or a similar edge device). In this arrangement, the communication must be failure-resistant, as clients can shut down at any moment, and do not need to consider all clients when sampling and updating. Thus, the central server samples and updates a fraction of these clients between rounds.

Cross-Silo Federated Learning is an arrangement mainly used when the federation’s clients are limited but hold large amounts of data; normally the clients are data centers or other similar nodes with a high amount of computing power. In this case, sampling and updating are made on all clients due to the limited number of clients and volume of data each client holds.

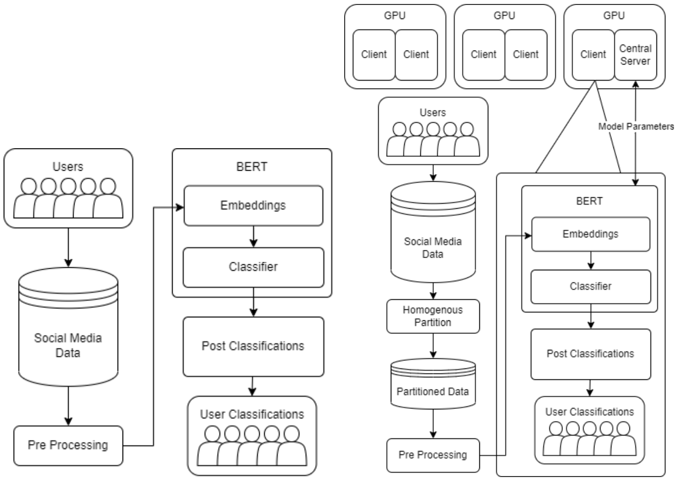
Flower: Flower [19] is a model and library-agnostic Federated Learning framework. It supports any Machine Learning (ML) library underneath it (e.g., TensorFlow, PyTorch, HuggingFace, custom one). It allows users to adapt any ML model to learn in a federated way.

Two modules compose Flower: the server and the client. As explained earlier, the server module acts as the central server of the federation. Flower allows users to set up different aggregation strategies, add model persistence, and change server configurations, such as the number of expected clients, the percentage of clients participating in each round, and the number of communication rounds. The client module represents each client in the federation, which executes the training in their local dataset and sends the weights to the server. In Flower, users can set the ML model clients train on, the hyperparameters, and how clients send the weights to the server, with or without cryptography.

B. Encoder-based Transformers

A Transformer [4] is a neural network architecture composed of two main stacks: the encoder and the decoder. Both are composed of sublayers encompassing embeddings and positional layers, multiple self-attention heads, normalization, and feed-forward networks. The encoder and decoder components are connected with an attention mechanism.

Several architectures have been proposed from the original transformers. Some offer different training regimes and components while still relying on both encoder and decoder



(a) Centralized Scheme (b) Federated Scheme

Fig. 1: Centralized and Federated Training Schemes experimented in this work.

components [20]. However, it is quite common to propose architectures focused only on the encoder or the decoder components. Encoding models such as BERT (Bidirectional Encoder Representation Transformers)-family [5] rely on the encoder component, while decoder models such as GPT (Generative Pre-Trained Transformer) [21] are essentially a Transformer-decoder. In this work, we formulate our task as a classification problem that relies on textual social media data. Based on that, we selected BERT as it is the most used and state-of-the-art encoder of texts into numerical representations.

IV. Experimental methodology

To test the capabilities of training a depression-level classifier based on a language model and social media data in a Federated learning regime, we have trained BERT in three different settings:

1) Centralized. First, we trained BERT with the default setting in a centralized server to compare it to Federated Learning.
2) Federated with five clients and 50 rounds. Each client executes one epoch.
3) Federated with five clients and ten rounds. Each client executes five epochs.

Settings 2 and 3 were tested multiple times with different versions of the dataset, as further detailed in Subsection IV-A1. As shown in Figure 1, user posts form the dataset. Then, the dataset is pre-processed to exclude posts without textual content. Next, they are fed into BERT's tokenizer. As our task is classification, we insert a final neural network classification layer on top of BERT. The tokenized input is used to finetune both the classification layer and the pre-trained BERT parameters. The same regime is employed in the federated version, however, with proper divisions of the dataset to simulate a privacy-preserving collaborative environment.

We formulated the classification task in two ways as follows.

- Post-based: The input is a single post collected from social media, and the output is a binary value indicating whether the post has been written by an individual presenting depression symptoms.
- User-based: The classification model executes as before, but the final classification is true, indicating an individual with depression symptoms only if more than half of their posts are also classified as true.

Subsection IV-B presents further hyperparameters used for each training fashion and server parameters in the case of Federated Learning.

All experiments were done in an Nvidia DGX-1 server. It has 8 GPUs Tesla V100 with 16GB of memory, 512GB of RAM, and two 20-Cores Intel Xeon E5-2698 v4.

A. Dataset

The dataset used for this task is eRisk2021 [22]. It was created for various mental health-related tasks, including the detection and severity classification of depressed individuals. This dataset holds Reddit [2] posts that may contain images, videos, and textual content individuals publish. The dataset annotates each post with the user's BDI score. However, this score is acquired by a peer responding to the questionnaire rather than the volunteer. The dataset was split into training, testing, and a validation split using the SciKit Learn library [23] and manually rearranged so that users do not appear between splits (i.e., user A's posts will only be on the validation split and User B's posts will only be in the training split). Table I shows the number of labels and the number of each class in each split.

TABLE I: eRisk2021 Statistics by split

Split	Labels	Class 0	Class 1	Class 0 %	Class 1 %
Train	19757	7912	11845	40.05%	59.95%
Test	6065	1490	4575	24.57%	75.43%
Validation	5304	1336	3968	25.19%	74.81%

For the task addressed in this paper, we consider only the textual content of these posts. This way, posts that contain no text have been filtered out. As said before, we formulate the classification task as binary: a post receiving the class false has been written by an individual that has been scored with $BDI < 20$; conversely, a post is labeled as true when the individual has been scored with $BDI \geq 20$, this cutoff value is based on the original questionnaire.

To better understand this dataset, we have analyzed the distribution of the number of tokens generated by the posts when tokenized[3] by BERT-tokenizer. All three splits have similar token distribution, where most posts in the dataset are

[2] https://www.reddit.com/. Reddit is a social news and forum website where content is socially curated and promoted by site members through voting.

[3] The tokenization process breaks the input texts into small pieces, such as words, sub-words, or even characters. BERT relies on a tokenization process called word piece that returns sub-words to account for out-of-vocabulary words during inference.

from the year 2020, and most posts have under 100 tokens, as shown by the histograms in Figure 2.

1) Federated Partitioning: For the Federated Learning experiments, we have partitioned the original eRisk2021 dataset's training split, keeping the restriction that all user posts belong to a single dataset partition. In total, we have created three versions of the partitioned dataset, varying class ratio imbalance as follows:

- Homogeneous: This partition was kept as close as possible to the original dataset class ratios, simulating the best-case scenario for a Federated Learning setting.
- Heterogeneous: This partition was kept as far as possible to the original dataset class ratios and with as much variation in class imbalance between clients as possible. Simulating the worst-case scenario for a Federated Learning setting.
- Midway: This partition was made by introducing a moderate class-ratio imbalance between the original dataset and between clients.

Figure 3 illustrates the class ratio imbalance between clients for each version of the dataset partition.

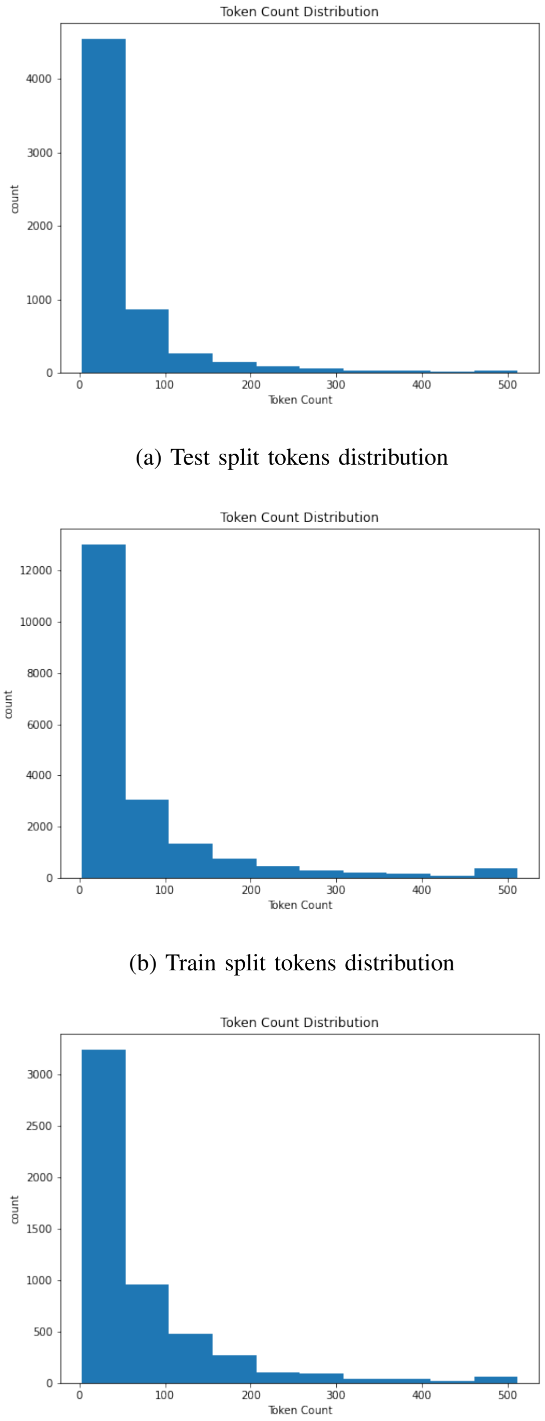
B. Training Methods

The three training configurations explored in this paper can be separated into two groups: The centralized method was used as a baseline for the following two Federated Learning training methods.

1) Centralized and Client Hyper-Parameters: We use the same neural network hyper-parameters to train the centralized and federated models for fairness, except for the number of epochs, as stated previously. The optimizer is AdamW, with a learning rate of 0.00001 and batches of size six. This relatively small batch size is due to the Federated Learning arrangement, which is further detailed in the following subsection. The training method was set up to update all layers of the language model. The model was the pre-trained bert-base-uncased, provided by the huggingface transformers library [24]. We chose this model instead of a larger one, such as bert-large-uncased, to account for hardware constraints and due to the token distributions showing that most posts generate under 100 tokens, as detailed in Section IV-A.

We have also used the same training methods with mental/mental-bert-uncased [25] to investigate how a domain-specific model differs from a generalist model. This pre-trained BERT model was trained using mental health-related posts from Reddit.

2) Federated Learning Hyper Parameters: Both Federated Learning training methods followed the same arrangement of five clients. Each client shared a GPU except for the last one, which shared a GPU with the central server, as illustrated in Figure 1b. We followed this regime to save hardware resources. We have used FedOpt as the aggregation strategy implemented by the Flower framework, given that it reports less communication overhead and better privacy strategies than other methods [26]. In essence, FedOpt has each client send their compressed and encrypted gradient to the central server.

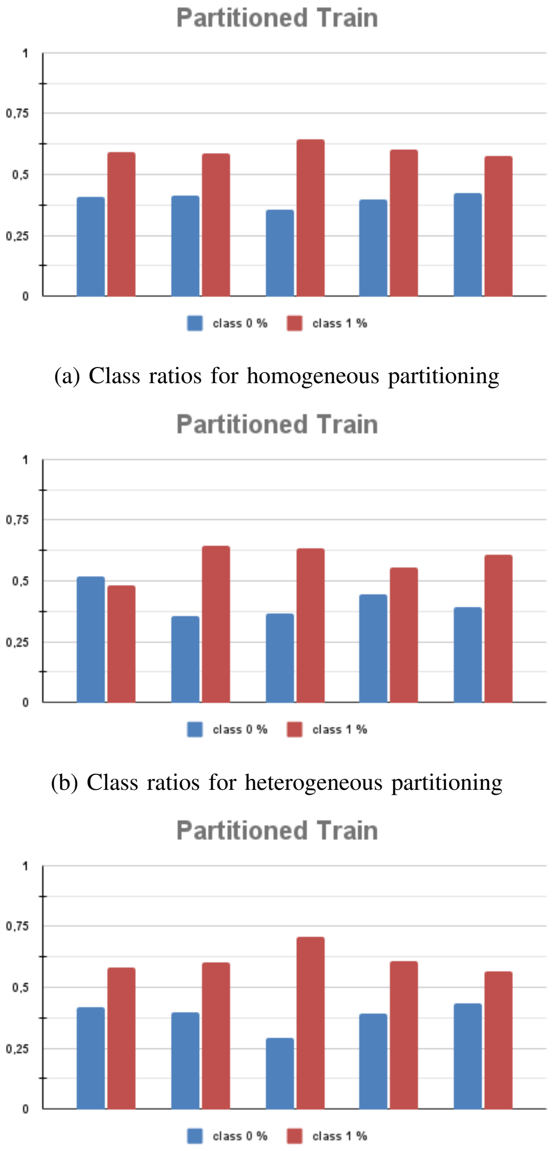


(a) Test split tokens distribution

(b) Train split tokens distribution

(c) Validation split tokens distribution

Fig. 2: Token Count Distribution between eRisk2021 splits



(a) Class ratios for homogeneous partitioning

(b) Class ratios for heterogeneous partitioning

(c) Class ratios for midway partitioning

Fig. 3: Class ratios between partitioned versions

The encryption used is additively homomorphic, meaning that the central server may aggregate the encrypted gradients without decrypting them first. As previously mentioned, we have partitioned the original eRisk2021 dataset as each client's local data. Each client received both the training and test splits.

V. Experimental Results

We have observed that in the centralized training setting, the BERT model presents worse performance every epoch after the first with both pre-trainings experimented. Figure 4b illustrates the evaluation results of every epoch on the centralized learning setting with the bert-base-uncased pre-training. The same behavior was observed on a per-round evaluation of both Federated settings using the homogeneous partitioning, though this evaluation was done using the Test split, as can be seen by the loss graphs in Figures 4c and 4d. The model shows steady performance degradation in the training settings done in this study. This behavior is lessened by using a domain-specific pre-training, as shown in Figure 4a. However, our experiments show that the domain-specific pre-training resulted in worse-performing models.

The Federated Learning settings have resulted in models that perform closely to the baseline, with a loss of 6.946% in Accuracy and 13.106% in Recall compared to the centralized scores, in the worst-case scenario (Homogeneous with ten rounds and five epochs). However, given the simple strategy used for aggregating post-classifications for user-based classification, the results were inadequate as it is sensitive to post-based classifier inaccuracies. Thus, with our current training dataset having a majority of depressed posts, the model classifies most users as showing depressive symptoms.

We observe that results have been insensitive to dataset heterogeneity for the most part, given that all versions have achieved similar scores in at least one method.

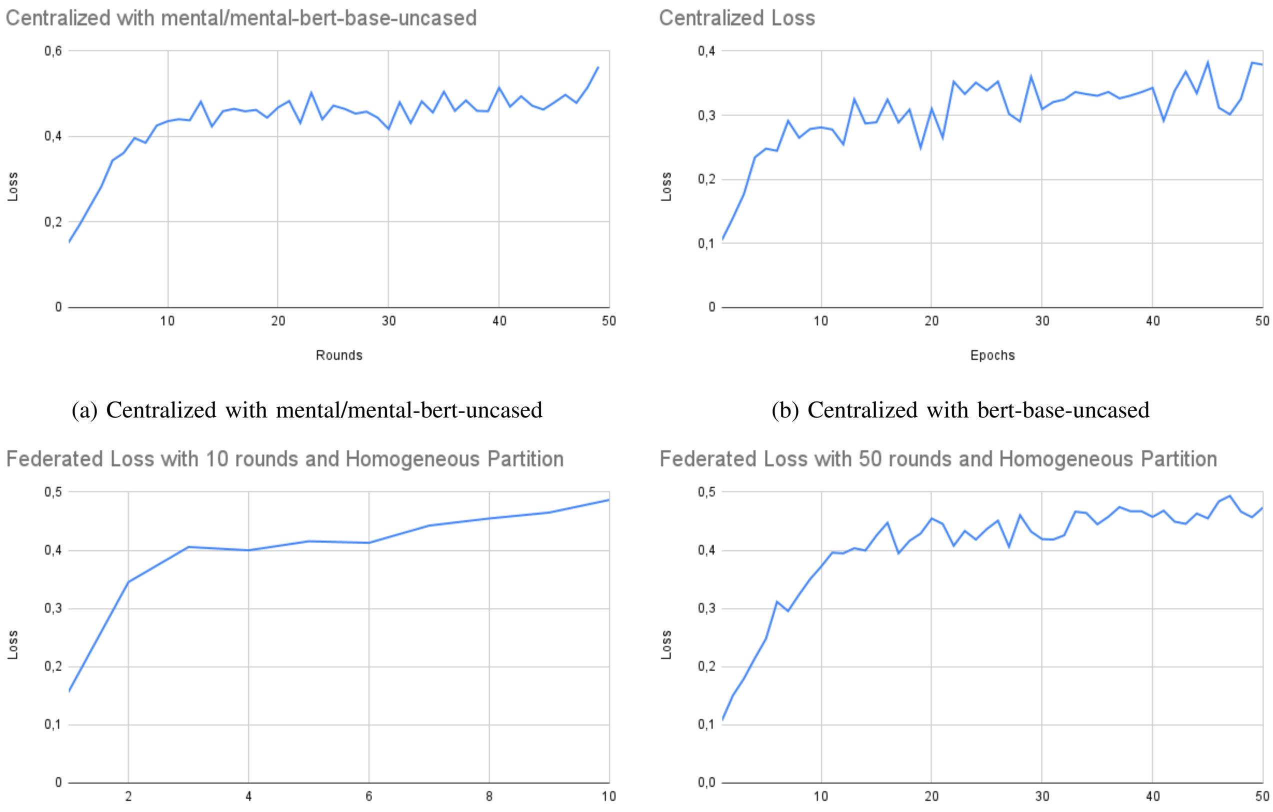
Tables II and III show the three training methods evaluating predictive accuracy, precision, recall, F1-Score and loss in the Post-based strategy using the bert-base-uncased and mental/mental-bert-base-uncased pre-trainings, respectively. Columns “User Precision”, “User Recall”, and “User F1” evaluate the method's Precision, Recall, and F1 score in the User-based strategy. All models were evaluated on the same Test split from the original eRisk2021 dataset. The experiments done using mental/mental-bert-uncased were not made with all three partitioned versions as this pre-training has shown significantly poorer performance concerning the bert-base-uncased pre-training.

VI. Concluding Remarks

This work has explored a federated learning technique for screening for depression in social media using Transformers. We used a pre-trained language model and fine-tuned the model with an additional classifier layer using a broadly employed depression dataset collected from Reddit in English.

The experiments simulated different forms of distribution among clients to observe the impact of adopting other numbers of rounds and epochs within various levels of client heterogeneity. The results showed that Federated Learning techniques achieve competitive models compared to the centralized version. We also observe that aggregating the classifier's output to execute a user-based strategy was inadequate, given the complexity of splitting users' data only among clients.

The following steps include testing other language models and elaborating post-aggregation strategies to realize the user-based classification. In addition, a promising investigation is verifying each transformer component's role during the federation aggregation procedure. One might suspect that feed-forward components and attention weights should be aggregated with distinct strategies.



(a) Centralized with mental/mental-bert-uncased

(b) Centralized with bert-base-uncased

(c) Federated with 10 rounds with homogeneous partition

(d) 50 rounds with homogeneous partition

Fig. 4: Loss Graphs

TABLE II: Methods Comparison with bert-base-uncased

	Method	Accuracy	Precision	Recall	F1 Score	Loss	User Precision	User Recall	User F1
Centralized	50 epochs	0.619	0.747	0.747	0.747	0.480	0.367	0.458	0.407
Homogeneous	Federated 50 Rounds, 1 Epoch Federated 10 Rounds, 5 Epochs	0.629 0.576	0.749 0.747	0.763 0.663	0.756 0.702	0.474 0.487	0.375 0.346	0.500 0.375	0.429 0.360
Heterogeneous	Federated 50 Rounds, 1 Epoch Federated 10 Rounds, 5 Epochs	0.611 0.623	0.747 0.753	0.732 0.745	0.740 0.749	0.479 0.434	0.375 0.375	0.500 0.500	0.429 0.429
Midway	Federated 50 Rounds, 1 Epoch Federated 10 Rounds, 5 Epochs	0.624 0.611	0.748 0.754	0.756 0.719	0.752 0.736	0.459 0.419	0.375 0.375	0.500 0.500	0.429 0.429

TABLE III: Methods Comparison with mental/mental-bert-base-uncased

	Method	Accuracy	Precision	Recall	F1 Score	Loss	User Precision	User Recall	User F1
Centalized	50 epochs	0.574	0.745	0.663	0.701	0.421	0.367	0.458	0.407
Homogeneous	Federated 50 Rounds, 1 Epoch Federated 10 Rounds, 5 Epochs	0.585 0.597	0.745 0.744	0.682 0.710	0.712 0.726	0.563 0.474	0.357 0.357	0.417 0.417	0.385 0.385

References

[1] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders: DSM-5*, 5th ed. Washington, DC: Autor, 2013.

[2] World Health Organization, Depression and Other Common Mental Disorders: Global Health Estimates. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO , Tech. Rep., 2017.

[3] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An Inventory for Measuring Depression," *Archives of General Psychiatry*, vol. 4, no. 6, pp. 561–571, 06 1961. [Online]. Available: https://doi.org/10.1001/archpsyc.1961.01710120031004

[4] V. Ashish, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[5] D. Jacob, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[6] P. Mann, E. H. Matsushima, and A. Paes, "Detecting depression from social media data as a multiple-instance learning task," in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2022, pp. 1–8.

[7] Y. Tian, Y. Wan, L. Lyu, D. Yao, H. Jin, and L. Sun, "Fedbert: When federated learning meets pre-training," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 4, aug 2022. [Online]. Available: https://doi.org/10.1145/3510033

[8] R. Skaik and D. Inkpen, "Using social media for mental health surveillance: A review," *ACM Comput. Surv.*, vol. 53, 6 2020. [Online]. Available: https://doi.org/10.1145/3422824

[9] S. Dhelim, L. Chen, S. K. Das, H. Ning, C. Nugent, G. Leavey, D. Pesch, E. Bantry-White, and D. Burns, "Detecting mental distresses using social behavior analysis in the context of covid-19: A survey," *ACM Comput. Surv.*, 6 2023. [Online]. Available: https://doi.org/10.1145/3589784

[10] A.-M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, "It's just a matter of time: Detecting depression with time-enriched multimodal transformers," in *Advances in Information Retrieval*, J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, and A. Caputo, Eds. Cham: Springer Nature Switzerland, 2023, pp. 200–215.

[11] S. Bn and S. Abdullah, "Privacy sensitive speech analysis using federated learning to assess depression," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6272–6276.

[12] X. Xu, H. Peng, M. Z. A. Bhuiyan, Z. Hao, L. Liu, L. Sun, and L. He, "Privacy-preserving federated depression detection from multi-source mobile health data," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 7, pp. 4788–4797, 2021.

[13] Y. Cui, Z. Li, L. Liu, J. Zhang, and J. Liu, "Privacy-preserving speech-based depression diagnosis via federated learning," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 1371–1374.

[14] J. Li, R. Zhang, M. Cen, X. Wang, and M. Jiang, "Depression detection using asynchronous federated optimization," in *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2021, pp. 758–765.

[15] J. Li, M. Jiang, Y. Qin, R. Zhang, and S. H. Ling, "Intelligent depression detection with asynchronous federated optimization," *Complex & Intelligent Systems*, pp. 1–17, 2022.

[16] P. Basu, T. S. Roy, R. Naidu, Z. Muftuoglu, S. Singh, and F. Mireshghallah, "Benchmarking differential privacy and federated learning for bert models," *arXiv preprint arXiv:2106.13973*, 2021.

[17] M. A. M. Pranto and N. Al Asad, "A comprehensive model to monitor mental health based on federated learning and deep learning," in *2021 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*. IEEE, 2021, pp. 18–21.

[18] M. Brendan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[19] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, H. L. Kwing, T. Parcollet, P. P. d. Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[20] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 7871–7880. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.703

[21] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[22] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "Overview of erisk 2021: Early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2021, p. 324–344. [Online]. Available: https://doi.org/10.1007/978-3-030-85251-1_22

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. null, p. 2825–2830, nov 2011.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[25] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare," in *Proceedings of LREC*, 2022.

[26] M. Asad, A. Moustafa, and T. Ito, "Fedopt: Towards communication efficiency and privacy preservation in federated learning," *Applied Sciences*, vol. 10, no. 8, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/8/2864