EVALUATION OF FEDERATED LEARNING APPROACHES FOR CLINICAL CONDITION NOTES CLASSIFICATION WITH LOGISTIC REGRESSION

Anna Aksenova

School of Science
Department of Computer Science
Aalto University, Espoo, Finland

ABSTRACT

Federated learning (FL) presents a promising approach for leveraging decentralized data in healthcare, enabling collaborative model training while preserving patient privacy. This paper explores the application of FL to text classification of patient conditions based on medical abstracts. By distributing the training process across multiple institutions, each maintaining local datasets, FL helps to avoid the need for centralized data aggregation, mitigating privacy concerns associated with sensitive medical information. We employed a federated text classification model using BERT (Bidirectional Encoder Representations from Transformers) and more traditional TF-IDF vectorization to classify patient conditions from a diverse set of medical abstracts. Our experimental results demonstrate that the federated model achieves comparable performance to traditional centralized approaches, with minimal performance degradation. The study also highlights the robustness of the FL framework against data heterogeneity and its potential for scalability in real-world healthcare settings. The findings suggest that FL is a viable and effective strategy for enhancing clinical decision support systems while ensuring data privacy and security. Future work will focus on optimizing model efficiency and exploring the integration of other advanced natural language processing techniques within the federated learning paradigm.

Index Terms— Federated Learning, Text Classification, Medical Abstracts, Patient Condition, BERT, Data Privacy, Clinical Decision Support Systems.

1. INTRODUCTION

In the healthcare domain, timely and accurate classification of patient conditions based on textual data such as medical abstracts is crucial for clinical decision-making and patient care management. Medical abstracts contain condensed information about patient diagnoses, treatments, and outcomes, and effectively analyzing this text can provide valuable insights into patient conditions. However, sharing and aggregating patient data across institutions pose significant privacy and security challenges. Federated learning (FL) offers a solution by enabling collaborative model training across multiple healthcare institutions without requiring data centralization. This approach ensures that sensitive patient information remains local, complying with privacy regulations such as HIPAA [1] and GDPR [2], while still benefiting from the collective knowledge of participating institutions. Federated learning is widely adopted for image and tabular data problems in healthcare [3, 4, 5], but not many text data problems are explored.

The traditional way of tackling text data vectorisation is TF-IDF (Term Frequency-Inverse Document Frequency) [6]. It is widely used in text mining and information retrieval to evaluate the importance of a word in a document relative to a corpus. It combines the frequency of a word in a specific document with its inverse frequency across all documents, highlighting terms that are significant in a particular context while downplaying common words. Despite its simplicity and effectiveness, TF-IDF has limitations in capturing semantic meanings and contextual relationships between words, which more advanced models

Devlin et al.'s paper on BERT [7] introduced a new way of using Transformer architecture to capture the context of words in a sentence from both directions. BERT creates powerful sentence representations by pre-training on a large text corpus. The model is commonly used as a backbone for fine-tuning on a task-specific dataset, however it can also be used as feature extraction method to obtain dense text and token representations.

Recent advancements in natural language processing (NLP) have significantly impacted the healthcare domain, particularly through the development of domain-specific language models. BioBERT [8], a first Transformer-based pre-trained biomedical language representation model, significantly enhances the performance of biomedical text mining tasks by being specifically trained on large-scale biomedical corpora, overcoming the limitations posed by word distribution shifts from general domain corpora. Compared to BERT and previous state-of-the-art models, BioBERT achieves notable improvements in biomedical named entity recognition, relation extraction, and question answering. This advancement demonstrates the importance of domain-specific pretraining in effectively understanding and extracting information from complex biomedical texts.

Complementing these advancements, Peng et al. [9] conducted an in-depth evaluation of federated learning (FL) for biomedical NLP, demonstrating that FL models not only outperform those trained on isolated datasets but also often rival those trained on aggregated data, while maintaining data privacy. Their findings indicate that FL, particularly with pre-trained transformer models, is robust against the challenges posed by decentralized data sources and regulatory constraints, making it a promising approach for collaborative medical research and applications.

Nevertheless, there is a lack of research on applications of federated learning approaches to medical condition classification problem, although the topic seems to be of high importance for clinical use-cases.

In this paper we are going to formulate the problem, i.e. discuss the dataset source and statistics in Section 2.1 and federated learning approaches in the scope in Section 2.2. The methodology and experimental setup is presented in Section 3. Section 4 presents the results obtained. The last part of this paper (Section 5) is dedicated to the discussion of findings and future improvements of the conducted work.

2. PROBLEM FORMULATION

As mentioned previously, the goal of this paper is to explore the application of federated learning methods towards medical text classification. In particular, the work will be focusing on clinical notes for 2 medical condition types applying 3 types of feature extraction techniques as a basis for FedSGD and FedAvg [10] algorithms.

2.1. Data

The dataset for this study consists of medical abstracts describing patient conditions across two categories: digestive system diseases and neoplasms. These abstracts are typically scanned by doctors during hospital rounds to quickly identify the patient's condition. The dataset was obtained from the paper [11].

It includes more than 10000 records, however for the sake of lowering computational cost of the project we focused on a training set with 2000 records and a test set with 100 records. The data distribution is presented in Table 1 and posits an imbalanced classification problem, therefore F1 score was chosen as the target metric for the classification task.

The F-score, also known as the F1-score, is a measure of a model's accuracy. It considers both the precision P and the recall R of the model predictions to compute the score. The formula for the F1-score is given by:

	Neoplasm	Digestive	Total
train	1354	646	2000
test	69	31	100

Table 1: Data distribution of the full dataset

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{1}$$

where precision P and recall R are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

Here, TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

2.2. Federated Learning Setup

We create an empirical graph, depicted as an undirected and unweighted graph $G = (V, E)$. Here, nodes $V := 1, ..., N$ correspond to local datasets $D^{(i)}$, where $i \in V$. In this work the full train dataset is evenly split into 5 local datasets with 400 train data points each. All nodes are connected as we have no information on geographic location of the hospitals in this setting.

The features of the dataset is vectors representing text data (more details on feature extraction techniques in Section 3). The labels for each data point are either 0 or 1 corresponding to neoplasm and digestive system issues respectively.

For the solution of the presented learning problem we decided to focus on the following federated learning approaches [12]:

- FedSGD for Logistic Regression

$$\hat{w} = \arg\min_{w \in W} \left(\sum_{i \in V} L_i(w^{(i)}) + \alpha \sum_{i,i' \in E} A_{i,i'}(w^{(i)}, w^{(i')}) \right) \tag{4}$$

- FedAvg for Logistic Regression

$$\hat{w} = \arg\min_{w \in \mathcal{C}} \sum_{i \in V} L_i(w^{(i)})$$

C = { w : w$^{(i)} = w^{(i')}$ for any edge $(i, i') \in E$}(5)

3. METHODS

3.1. Feature selection

Text data posits non-trivial feature selection problem as text should be converted into numerical features that will somehow represent the semantics. In this study, we utilized three feature extraction methods to represent medical abstracts: BERT, BioBERT, and TF-IDF. Each method has its own advantages and limitations.

1. TF-IDF

   TF-IDF is a simple and efficient method that highlights the importance of terms in documents, facilitating quick text representation.

   The TF-IDF for a term t in a document d is given by:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) \tag{6}$$

   where

$$\text{TF}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{7}$$

   and

$$\text{IDF}(t) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \tag{8}$$

   Here, $f_{t,d}$ is the frequency of term t in document d, N is the total number of documents, and $|\{d \in D : t \in d\}|$ is the number of documents containing the term t.

   For computing TF-IDF vectors we leveraged scikit-learn library[1]. We used word-based representation and removed the stop words (e.g. a, and, or) as they do not seem to be relevant for medical condition classification. As a result of this feature extraction approach we obtained an interpretable sparse feature vector of 100000 most important features where each feature position corresponds to vocabulary item importance for the particular document. The drawbacks of this method are that it does not capture semantic relationships between words and can result in very high-dimensional feature spaces, which may lead to sparsity and increased computational costs for training classification models.

2. BERT

   BERT is a Transformer-based feature extraction model pre-trained on large common-domain corpus. We use the tiny version of it to eliminate the issue of computation cost [2]. The pipeline is visualized in Figure 1. Such feature extraction involves tokenising text with BERT tokenizer and passing the input tokens through a pre-trained BERT model to obtain contextualized embeddings for each token in the text. Mean pooling then aggregates these token embeddings by taking the average across all tokens, producing a fixed-size vector representation that captures the overall semantic content of the input text. Therefore after applying the model we obtain a dense vector 312 non-interpretable features for each of the text snippets. Such vectorisation method should be able to capture word dependencies and grasp text semantics. In addition as the feature vector is dense and of smaller dimension, less parameters of the local models are required to be optimized.

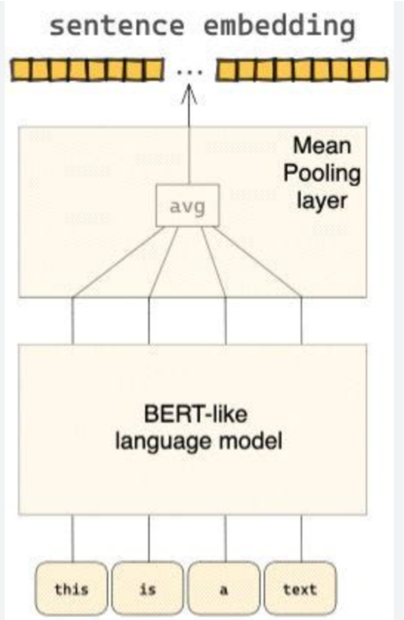

Fig. 1: BERT-like model feature extraction pipeline.

3. BioBERT

   BioBERT [3] is a Transformer-based feature extraction model specifically pre-trained on biomedical literature, allowing it to capture domain-specific nuances and contextual information effectively. The extraction pipeline for this model and number of features is the same, however as the model is claimed to perform better on clinical notes, we decided to include it into comparison.

3.2. Local models

We chose logistic regression as our local model due to its simplicity, efficiency, and fast training times. As it is a linear model it allows for for straightforward interpretation of results, which is particularly valuable in a medical context

[1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[2] https://huggingface.co/prajjwal1/bert-tiny

[3] https://huggingface.co/nlpie/tiny-biobert

where model transparency is crucial. Additionally, logistic regression performs well with high-dimensional data, making it a robust choice for our feature-rich representations from BioBERT, BERT, and TF-IDF.

Local models variation was estimated using the norm of the difference between local model parameters.

3.3. Loss function

We selected log-loss as our loss function because it is well-suited for classification tasks, providing a clear measure of the difference between predicted and actual class probabilities.

3.4. Model validation

In our validation process, each local training set was partitioned into 70% training and 30% validation subsets to assess model performance during training. The method with the lowest average validation loss across local datasets was selected for evaluation on the held-out test set (Table 1), ensuring that the final model's generalization ability was validated independently from the training data. This approach helps mitigate overfitting and provides a reliable estimate of the model's performance on unseen data.

One of the optimization parameters explored with the help of validation dataset was the learning rate parameter for both FedSGD and FedAvg algorithms. We evaluated values of 0.01, 0.1, 0.5, and 0.7 to determine their impact on the federated learning process. This allowed us to assess how different levels of gradient step size influenced model convergence, performance, and robustness across various values.

4. RESULTS

For all the experiments we used fully-connected graph with uniform edge weights equal to one. As the dataset split in our case is synthetic we do not want to prioritize any of the local datasets (hospitals) artificially.

We compare various parameters in the experimental setup, i.e. we would like to first select α values for FedSGD. Then we would like to compare three feature extraction methods that were suggested and lastly we compare FedSGD and FedAvg strategies. The models are evaluated on the test set using f1 score.

4.1. Hyperparameter selection

For the selection of hyperparameters in FedSGD and FedAvg, we compared learning rate values of 0.01, 0.1, 0.5, and 0.7. The best learning rate varies across all feature extraction methods, however most common best value was found to be 0.7, indicating that a larger gradient descent step facilitated better convergence and model performance. This optimal learning rate value was determined by achieving the lowest average validation score, ensuring robust and consistent model improvement during training.

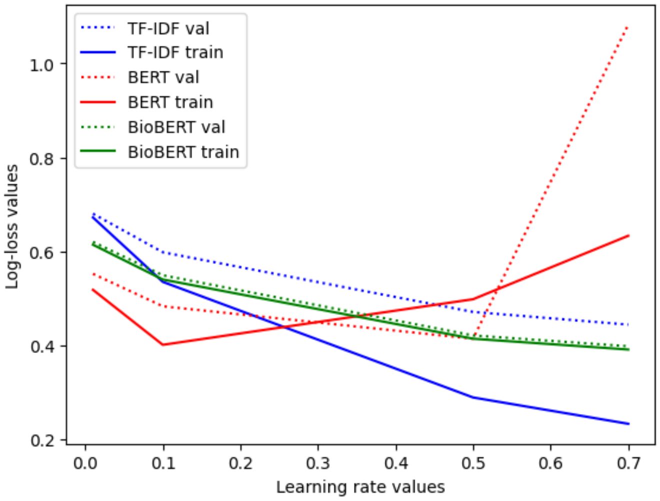

Fig. 2: Average train and validation loss depending on learning rate for FedSGD.

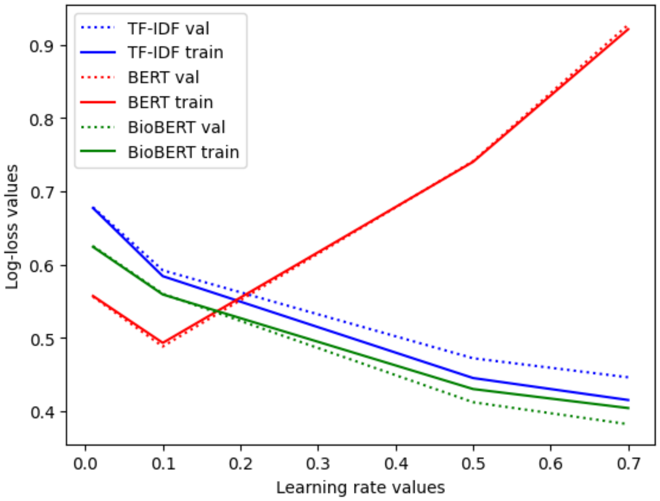

Fig. 3: Average train and validation loss depending on learning rate for FedAvg.

Figures 2 and 3 show the average train and validation losses for all feature extraction methods on FedSGD and FedAvg training respectively. It can be observed that only for BERT feature extraction large learning rate does not improve the training method performance. For the learning rate 0.7 the model overfitted. Colours on the plot correspond to different feature extraction methods while line types stand for validation and train errors. If comparing dotted and solid lines (train and validation errors) it could be observed that although on average FedAVG shows worse performance, the training of this method is more stable. For all learning rates and all algorithms validation error does not deviate much from the training error, therefore training could be potentially continued. As for FedSGD method, such pattern could be observed only for BioBERT embeddings.

4.2. Feature extraction

Although log loss is a good estimate for model performance in general, to evaluate the model performance in clinical practice one should use classification accuracy metrics. It is important to note the difference between log-loss and f1 score in terms of classification accuracy evaluation. Let's consider two different models f_1 and f_2. Both are trained to minimise logistic loss for classification problem. After training is finished $Log\_loss(f_1(x\_test), y\_test) > Log\_loss(f_2(x\_test), y\_test)$. Such comparison indicates that $f1$ predicts probabilities closer to the true class labels, leading to better overall likelihood estimates. However, it does not guarantee that f_1 predicts more accurate answers on test set. It could be true that $F\_score(f_1(x\_test), y\_test) < F\_score(f_2(x\_test), y\_test)$, suggesting that f_2 performs better in terms of the balance between precision and recall, which is crucial for imbalanced datasets and our specific application needs. This discrepancy can occur because log loss is sensitive to the predicted probabilities, while the F-score is more focused on the accuracy of the classification decisions, particularly in the context of true positives. This nuance is crucial in medical setting as we want the final model to be accurate, therefore test set f1 evaluation makes more sense then test set error evaluation.

Table 2 presents test set f1 score for the models chosen by the smallest validation error for each feature extraction and federated learning method.

Moreover, as the ultimate goal of federated learning is to leverage data from several local datasets to enhance trained model performance, we want our methods to perform at least as good as local models trained independently on the local train datasets. We conducted such evaluation and reported it under ”No FL” name.

	FedAvg	FedSGD	No FL
TF-IDF	0.884	*0.895*	0.894
BERT	0.822	*0.824*	0.819
BioBERT	0.833	0.834	*0.836*

Table 2: Comparison of feature extraction methods and Federated Learning strategies based on average test f1-score. *Italics* stands for the best score in a row.

For each of the nodes the same test set was used to ensure fair model comparison. Test set statistics can be found in Table 1. The final f1 score reported at the Table 2 is computed as average over local models f1 scores on the test set.

For TF-IDF and BioBERT feature extraction methods we report metrics for models trained with 0.7 learning rate, for BERT feature extraction, we report metrics for the model trained with 0.1 learning rate.

As for the feature extraction methods TF-IDF significantly outperforms BERT-based approaches. It could be explained by the fact that BERT embeddings are rarely used as pure feature extracion method, the models are quite often fine-tuned (i.e. weights of the whole model are updated) together with the last classification layer. Interestingly, BioBERT feature vectors show better score compared to common-domain model, showcasing the importance of in-domain pre-training.

4.3. Federated learning strategy

In our experimental setup, FedSGD consistently outperformed FedAvg across all feature extraction methods, though the performance difference was marginal. This suggests that both federated optimization methods are effective. For TF-IDF feature extraction method the difference is larger which could be explained by the fact that the features are explicit and correspond to relative word frequences, while for BERT-based methods the features are non-interpretable.

Both strategies perform on par with models trained on local datasets independently (Table 2, ”No FL” column). Comparison of federated learning approaches performance with the performance of independent models shows that only BioBERT-based feature extraction strategy is consistently worse when employed in Federated Learning training setup.

	node_1	node_2	node_3	node_4	node_5
train_loss	0.227	0.238	0.246	0.238	0.218
val_loss	0.452	0.421	0.464	0.457	0.426
test_f1	0.878	0.890	0.901	0.918	0.884

Table 3: Local datasets (empirical graph nodes) metrics of the best performing method

Table 3 provides train and validation loss values for each node in the empirical graph of the logistic regression trained on TF-IDF features with FedSGD method. In addition we provide test set f1 for each of the local models.

5. CONCLUSION

Our study examined the effectiveness of various hyperparameters, feature extraction methods, and federated learning strategies for classifying patient conditions based on medical abstracts. We found that across different feature extraction and federated learning methods large learning rate showed better results for our task, suggesting that a larger gradient descent step generally facilitated better convergence. Among the feature extraction methods, TF-IDF consistently outperformed BERT-based approaches, highlighting its efficiency in this context, while BioBERT showed better performance than standard BERT, showing the importance of domain-specific pre-training.

The results indicate that the problem is addressed satisfactorily with both FedSGD and FedAvg performing comparably to models trained on local datasets independently. However, there is definitely room for improvement as one could continue training FedAvg approach for larger number of steps. In addition, there should be more experiments done to select other hyperparameters such as batch size, number of training iterations, number of features in TF-IDF and smarter text preprocessing. To enhance BERT-based method performance one could focus on fine-tuning those models rather than using static embeddings. Additionally, using weighted loss to tackle class imbalance could improve the final model performance. Lastly, one could experiment more on empirical graph construction, using clustering methods to create graph edges rather than training the methods on fully connected graph.

Touching upon limitations of suggested methods one could mention computational complexity of BERT-based approaches and need for privacy protection in medical domain. None of the methods discussed gradient privacy, therefore additional methods should be leveraged for this purpose.

6. REFERENCES

[1] Accountability Act, “Health insurance portability and accountability act of 1996,” *Public law*, vol. 104, pp. 191, 1996.

[2] Paul Voigt and Axel Von dem Bussche, “The eu general data protection regulation (gdpr),” *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[3] Pallavi Dhade and Prajakta Shirke, “Federated learning for healthcare: A comprehensive review,” *Engineering Proceedings*, vol. 59, no. 1, pp. 230, 2024.

[4] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al., “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data,” *Scientific reports*, vol. 10, no. 1, pp. 12598, 2020.

[5] Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi, “Federated learning of predictive models from federated electronic health records,” *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.

[6] Karen Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[9] Le Peng, Gaoxiang Luo, sicheng zhou, jiandong chen, Rui Zhang, Ziyue Xu, and Ju Sun, “An in-depth evaluation of federated learning on biomedical natural language processing,” 2023.

[10] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[11] Tim Schopf, Daniel Braun, and Florian Matthes, “Evaluating unsupervised text classification: Zero-shot and similarity-based approaches,” in *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, New York, NY, USA, 2023, NLPIR ’22, p. 6–15, Association for Computing Machinery.

[12] A. Jung, “Federated learning,” *Lectures of the course CS-E4740*, 2024.