FAIRNESS-AWARE FEDERATED LEARNING FOR TEXT CLASSIFICATION

B. Tech. Project Report

Submitted by

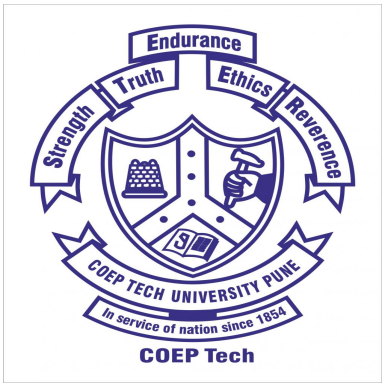
Shubham Gupta **112103046**

Manas Jorvekar **112103058**

Under the guidance of

Dr.Y.V.Haribhakta

COEP Technological University, Pune



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

COEP TECHNOLOGICAL UNIVERSITY

(COEP TECH), PUNE - 5

May 2025

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,

COEP TECHNOLOGICAL UNIVERSITY, PUNE - 5

CERTIFICATE

Certified that this project titled, "FAIRNESS-AWARE FEDERATED LEARNING FOR TEXT CLASSIFICATION" has been successfully completed by

Shubham Gupta 112103046

Manas Jorvekar 112103058

and is approved for the partial fulfillment of the requirements for the degree of "B. Tech. Computer Engineering".

Dr.Y.V.Haribhakta
Project Guide
Department of CSE
COEP Tech Pune,
Shivajinagar, Pune - 5.

Dr. P.K. Deshmukh
Head
Department of CSE
COEP Tech Pune,
Shivajinagar, Pune - 5.

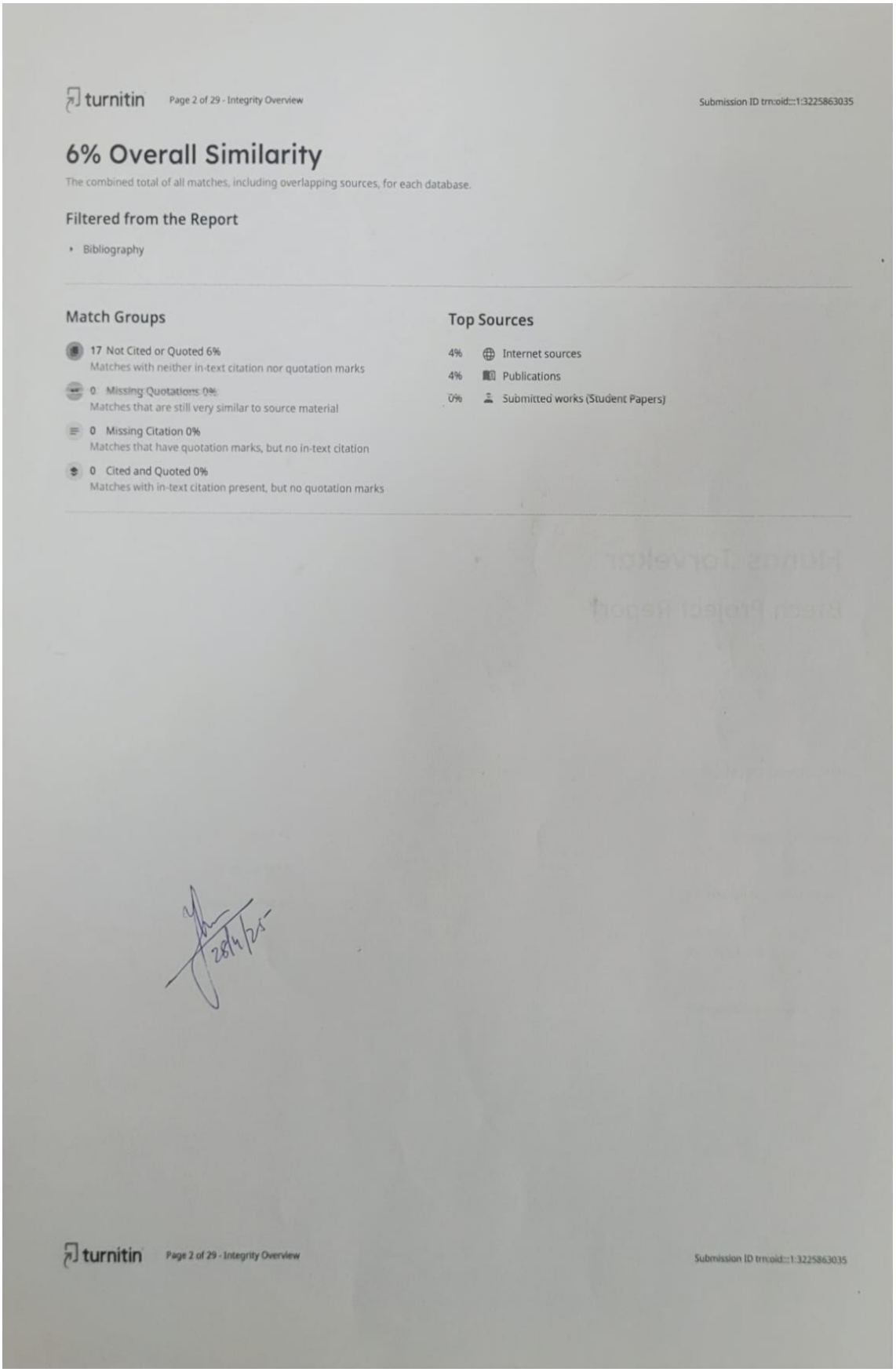


Figure 1: Similarity Report (Turnitin).

Abstract

Federated Learning (FL) is a privacy-preserving mechanism of machine learning across decentralized clients finding its applications in various NLP tasks. However, FL algorithms struggle in performance and fairness when operating in Non-independent and identically distributed (non-IID) data of real-world scenarios.

This work introduces **DualMetric-Adaptive FL** as a new algorithm enhancing global performance and client-level fairness in multi-class text classification under different data distributions. This method ensures better generalization by using a dual-metric monitoring system and preserves fairness among the clients on the basis of intra-client and inter-client performance analysis, thus offering a robust privacy preserving system for NLP tasks in federated settings with diverse distribution.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Federated learning [1] is introduced as a communication efficient and privacy preserving approach for deep learning. A shared global model is trained from data that is distributed across many clients. [14] highlights that the learned model does not reveal whether a certain data point was used during training. This ensures that local data points are not easily reverse-engineered to obtain raw data.

Federated learning operates as a distributed machine learning system which maintains data privacy because it keeps raw information solely present on end-user devices. Independent models are trained on each device locally and insights learned from the local data are shared with a central server [19]. This completely eliminates data centralization to achieve user privacy. The basic FL problem can be cast as one of empirical minimization of a global loss objective, which is decomposable as a sum of device-level empirical loss objectives [11].

Reference [2] presents various contributions and advancements in FL to enable readers to understand recent developments. The study mentions applications in healthcare, IoT networks, Retail recommendation systems, and Intrusion detection systems. The study demonstrates how FL creates cross-

disciplinary effects between distributed optimization and cryptography and security and differential privacy fields.

Most existing studies on text classification focus on the effectiveness of Convolutional Neural Networks (CNNs) in learning hierarchical features [6]. The research acts as foundation for our project which involves developing CNN-based text classification algorithms for non-IID federated settings that handle data heterogeneity and categorical imbalance across multiple clients. The effects of FL principles on NLP tasks show a tradeoff between model performance and the local data privacy [4]. Solving this issue demands personalization along with optimization methods which serve crucially important for real-world problem solutions.

Chapter 2

Literature Review

FedAvg [1], a basic aggregation algorithm that helps the global server combine knowledge gained from clients into one common shared global model. It performs weighted average based on the number of data points at each client for all parameters. The algorithm demonstrates effective performance with balanced IID data but its functionality declines when processing heterogeneous environments. The algorithm, however, needs an improved version to process data distributions occurring in real-world situations.

Reference [3] introduces an aggregation algorithm FedProx that is capable of handling heterogeneity across clients under realistic scenarios. FedProx achieves higher and more stable accuracy compared to the traditional FedAvg approach even under extreme heterogeneity. Proximal term in the aggregation determines how far the local updates can deviate from the global model. The aim is to achieve a convergent performance across clients for better overall performance. However, it sometimes leads to overfitting because of server-side proximation especially when distributions are very skewed.

Reference [11] recognizes the bottleneck faced in federated scenarios due to heterogeneity across clients. The authors propose introducing a dynamic regularization term to increase client level personalization allowing each client

to gain more insights from the global knowledge. This study lays the foundation for FedNova [10], which is another successful aggregation algorithm that tackles the problem posed by Non–IID data. In many non-IID cases, FedNova outperforms FedProx in personalized metrics.

Reference [5] systematically evaluates federated learning (FL) algorithms on non-IID data settings and provides a benchmark framework to analyze their performance under different non-IID scenarios. The study helps us identify weaknesses in existing aggregation methods FedAvg, FedProx and FedNova when handling Non-IID situations.

The advantages of FedProx and FedNova do not include adaptive local training capabilities for handling dynamic hyperparameter variations. The proposed solution involves dynamic regularization applied at the client side rather than the server to allow clients to personalize their learning rate along with training epochs and regularization parameters.

Reference [2] also highlights that Adaptive algorithms for client-specific model updates are required to handle heterogeneity in data. This motivated us to create a modified aggregation approach that can handle robust environments. In this study, we choose to work with text classification on e-commerce data.

Chapter 3

Proposed Methodology

Our initial evaluation on the FedAvg [1] algorithm in non-IID data settings exhibited major problems centered around consistency and fairness among clients.In FedAvg the model was biased towards certain clients and also the global model had a decent accuracy. FedProx [16] addresses this issue by introducing a fixed regularization parameter that helped to increase the global accuracy but the client fairness was still a major problem in non-IID settings. The accuracy results from different clients indicated wide variations showing that using static regularization is not enough to handle the differences between clients in a non-IID setup.

The research demonstrates two main problems with present approaches.

i. **Static Regularization**: A fixed μ prevents the model from adapting to changes in client data during training [11].

ii. **Fairness Tradeoff**: Improving global accuracy often comes at the expense of fairness between clients wherein the model often tends to be biased towards some clients.

We need a method which upholds high global accuracy while maintaining fairness across different client data distributions, therefore, we introduce

DualMetric-Adaptive FL that combines performance monitoring with dynamic proximal regularization to solve existing method limitations [3].

3.1 Dual-Metric Monitoring

We use dual-metric monitoring that measures accuracy in two ways to understand individual client behavior:

1. Global Accuracy: Evaluates how well the global model performs on completely unseen data, representing overall generalization capability of the model.

   Global Accuracy (GA):

   $$\text{GA} = \frac{1}{N_g} \sum_{i=1}^{N_g} \mathbb{I}(f_g(x_i) = y_i)$$

   where N_g is the number of examples in the global training set, $f_g(x_i)$ is the global model's prediction for i$^{\text{th}}$ example and y_i is the actual correct label.

2. Local Accuracy: Tracks how well the global model performs on individual client data after local updates, revealing fairness and client-specific performance.

   Local Accuracy (LA):

   For client k:

   $$\text{LA}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{I}(f_k(x_i^k) = y_i^k)$$

   where n_k is the number of examples in client k's local training set,$f_k(x_i)$ is the locally updated model's prediction for i$^{\text{th}}$ example and y_i is the actual correct label.

These evaluation metrics provide us with the capability to achieve optimization of generalization and fairness together.

3.2 Adaptive Proximal Regularization

Fedprox [3] makes use of a proximal term to restrict clients from sending updates that are drastically different from the global model.

Proximal Regularization:

$$\min_{w_k} \mathcal{L}_k(w_k) + \frac{\mu}{2} \left\| w_k - w_g^{(t)} \right\|^2$$

where $\mathcal{L}_k(w_k)$ is the local loss for client k, μ is the proximal term for regularization, $(w_k - w_g^t)$ is the penalty for deviation of client model from global model during its local updates at round t.

Our work improves and modifies Fedprox through a dynamic system that modulates the regularization parameter (μ) as the traditional static regularization does not adapt to changing client behaviour patterns during training particularly in non-IID settings.

For each client we make two types of comparisons:

1. Intra-Client Comparison: Measures the change in client performance on local training data between round t-1 and round t, and compares this with the change in global accuracy over the same rounds for overall model improvement.

Intra-Client Analysis:

$$\mu \leftarrow \begin{cases} 1.005\mu & \text{if } \Delta_{\text{local}} < -0.03 \wedge \Delta_{\text{global}} \leq 0 \\ 0.995\mu & \text{if } \Delta_{\text{local}} > 0.05 \wedge \Delta_{\text{global}} \geq 0 \\ 1.05\mu & \text{if } \Delta_{\text{local}} > 0.05 \wedge \Delta_{\text{global}} < -0.03 \\ 0.95\mu & \text{if } \Delta_{\text{local}} < -0.03 \wedge \Delta_{\text{global}} > 0.05 \end{cases}$$

where:

- $\Delta_{\text{local}} = acc_t^{\text{local}} - acc_{t-1}^{\text{local}}$
- $\Delta_{\text{global}} = acc_t^{\text{global}} - acc_{t-1}^{\text{global}}$

2. Inter-Client Comparison: Compare client performance on local training data with the average local accuracy of all k clients at round t to identify disparities among clients.

   Inter-Client Analysis:

$$\mu \leftarrow \begin{cases} 0.85\mu & \text{if } \delta_{\text{disparity}} < -0.05 \\ 0.95\mu & \text{if } -0.05 < \delta_{\text{disparity}} < -0.02 \\ 1.15\mu & \text{if } \delta_{\text{disparity}} > 0.05 \\ 1.05\mu & \text{if } 0.02 < \delta_{\text{disparity}} < 0.05 \end{cases}$$

   where:

   - $\delta_{\text{disparity}} = acc_t^{\text{local}} - \frac{1}{k}\sum_{i=1}^{k} acc_t^{\text{local}}$

Proposed **DualMetric-Adaptive FL** mechanism enables automatic balance between performance accuracy and distributed fairness measures through an automated process by adjusting the regularization parameter μ on the basis

of two scaling factors f_{intra} and f_{inter} which helps us to control client contributions to the global model [11]. These scaling factors increase or decrease μ based on changes observed in local and global accuracies across clients.The system provides lower constraint (lower μ) to underperforming clients while enforcing higher constraints (higher μ) to clients who tend to overfit in order to protect the integrity of the global model and maintain fairness.The final update combines both perspectives to give a new regularization parameter:

DualMetric-Adaptive FL Regularization term:

$$\mu_k^{(t+1)} = \mu_k^{(t)} * f_{\text{intra}}(\Delta_{local}, \Delta_{global}) * f_{\text{inter}}(\delta)$$

where $\mu_k^{(t+1)}$ is the updated proximal term for client k, μ_k^t is previous proximal term, f_{intra} adjusts μ based on intra client comparison and f_{inter} adjusts μ based on inter client comparison.

Chapter 4

Experimental Setup

We use the Bitext retail ecommerce LLM chatbot training dataset to perform and showcase text classification in our federated learning setup. The dataset consists of 45k rows of text across 46 categories for customer responses and queries on an e-commerce website. The multi label classification task will help us to visualize a real world use-case of Federated learning in NLP [12]. We split the overall dataset into training and testing in a 9:1 ratio. The training set is further split among 11 clients in an uneven way to simulate non-IID scenario for studying how federated learning principles work with the split. The testing set is used for global testing which stays the same for all clients.

The non-iid settings also help us to understand the influence of label distribution and skewness [15] on performance of different aggregation algorithms. We further observe how different federated learning algorithms perform on our data distribution under the same conditions[16].

We conduct our Federated learning simulations and experimentations using the Flower framework [7]. Flower offers the infrastructure to support easy communication between client and server in federated settings and also provides robustness to changes on client side or data distributions.

Chapter 5

Results and Discussion

For evaluation and demonstration of our DualMetric-Adaptive FL framework, we conducted extensive experimentations for over 500 communication rounds. Our framework achieves high global accuracy and also reduces the gap between client's local accuracies, achieving a lower variance. The same experimental setup was utilized throughout all our experiments. The research findings are presented according to performance benchmarks, bias reduction analysis and tradeoff management strategies.

5.1 Performance Comparison

The DualMetric-Adaptive FL framework was evaluated on the parameters of global accuracy and fairness variance using FedProx as the baseline across 500 communication rounds. The obtained results have been showcased in Table 5.1 and Table 5.2.

The framework demonstrates better performance across all checkpoints compared to the baseline FedProx model. It achieves higher global accuracy and lower variance which ensures proper generalization of the overall model. This reveals both, performance improvements and equitable learning con-

Rounds	FedProx	DualMetric-Adaptive FL	Difference (in %)
100	73.68	79.49	+5.81
200	81.23	85.43	+4.20
300	83.55	87.86	+4.31
400	85.45	89.79	+4.34
500	87.14	92.39	+5.25

Table 5.1: Comparison of Global Accuracy for FedProx and Our model(%) Across Rounds

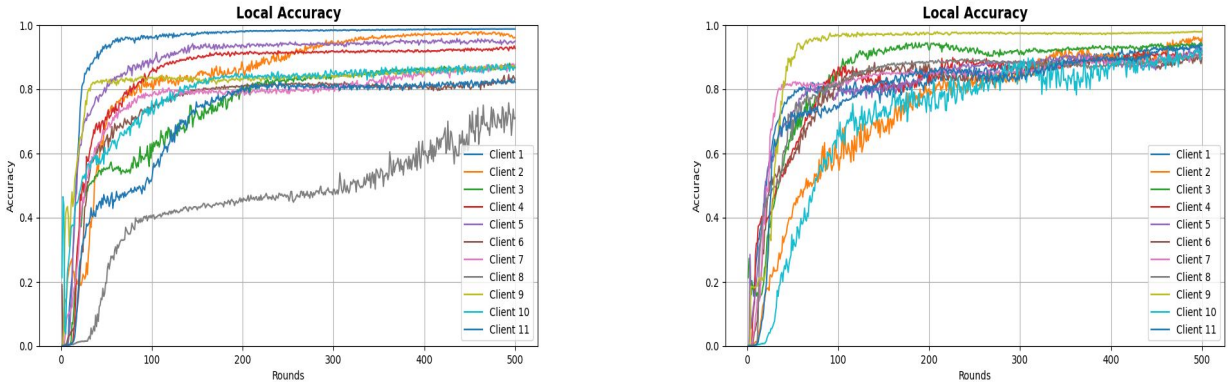
Rounds	FedProx	DualMetric-Adaptive FL	Difference (in %)
100	0.0253	0.0099	-60.87%
200	0.0173	0.0032	-81.50%
300	0.0159	0.0012	-92.45%
400	0.0099	0.0007	-92.92%
500	0.0055	0.0006	-89.09%

Table 5.2: Comparison of Variance for FedProx and Our model(%) Across Rounds

sideration for clients allowing elimination of bias and prevention of any one client from controlling the global model.

5.2 Bias Reduction Analysis

The reduction in bias among clients, a major improvement of our framework, can be observed in Figure 5.1 that shows a clear difference in the local accuracies of clients over time. Our framework uses intra-client and inter-client adaptation strategies that achieve better client accuracy alignment and clustering when compared to FedProx. The framework demonstrated these different fairness and performance scenarios during training in our subsequent observations:



(a) FedProx

(b) DualMetric-Adaptive FL

Figure 5.1: Comparison of local accuracy across 500 training rounds.

1. Lagging Client Recovery: DualMetric-Adaptive FL delivers substantial benefits to clients who maintain low local accuracy. For example, Client 2 in Figure 5.1b had a local accuracy of just 60% at round 100 which triggered both intra-client and inter-client mechanisms — a drop in local accuracy($\Delta_{local} = -0.04$) along with minimal global improvement($\Delta_{global} = -0.01$) combined with a large accuracy gap compared to other clients($\delta = -0.12$) resulted in the reduction of regularization parameter μ giving the client more flexibility to adapt and recover. Client 2 achieved an 84% local accuracy in round 300 through our effective framework thereby boosting underperforming client accuracy.

2. Early Fairness Enforcement: One of the key differences between FedProx and our method appears early in training. In FedProx, client accuracies quickly diverged, creating a wide spread by round 100 and continued this trend even till round 500 whereas, our framework's inter-client mechanism intervened proactively. For example the adjustment made in round 100 where the μ was lowered for Clients 2 and 10 to contribute more and increasing μ for overfitting client like Client 3 and 9 ($\Delta_{local} = +0.05 and \Delta_{global} = -0.02$) were made so that they didn't dom-

inate the global model.These timely adjustments kept 7 out of 11 clients within a 70–85% accuracy range by round 100 (Figure 5.1b), compared to FedProx's much wider spread of 45–85% (Figure 5.1a).

5.3 Tradeoff Management

FedProx produces a traditional accuracy versus fairness tradeoff because of its static regularization parameter that cannot adapt in real time to adjust according to the needs of the individual client.

Our DualMetric-Adaptive FL framework uses adaptive regularization to effectively balance the global accuracy against local fairness of clients during training. In early rounds, the intra-client mechanism becomes more aggressive so that the global model learns faster to establish an initial foundation. Eventually, when there is not much change in global accuracy in the middle rounds, the inter-client mechanisms become more aggressive so that the underperforming client updates are taken more into consideration in the global model. By the final rounds,we achieve a stable federated learning model with all clients achieving high local accuracy without forcing the high-performing clients to compromise. This dynamic strategy helps our framework achieve the balance between the accuracy and fairness among clients.

Chapter 6

Conclusion

Using dynamically adjusted regularization parameters for each client according to intra-client and inter-client trends makes DualMetric-Adaptive FL successful in minimizing bias while avoiding overfitting among clients. The proposed approach delivers improved global client performance alongside tight local accuracy clusters which results in better collaborative text classification than standard FedProx. This work demonstrates the need for solutions that implement NLP based federated learning systems in realistic deployments.

Bibliography

[1] McMahan, Brendan, 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, pp. 1273-1282. PMLR, 2017.

[2] Kairouz, Peter, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz et al. "Advances and open problems in federated learning." Foundations and trends® in machine learning 14, no. 1–2 (2021): 1-210.

[3] Li, Tian, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. "Federated optimization in heterogeneous networks." Proceedings of Machine learning and systems 2 (2020): 429-450.

[4] Zhu, Xinghua, Jianzong Wang, Zhenhou Hong, and Jing Xiao. "Empirical studies of institutional federated learning for natural language processing." In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 625-634. 2020.

[5] Li, Qinbin, Yiqun Diao, Quan Chen, and Bingsheng He. "Federated learning on non-iid data silos: An experimental study." In 2022 IEEE 38th international conference on data engineering (ICDE), pp. 965-978. IEEE, 2022.

[6] Chen, Yahui. "Convolutional neural network for sentence classification." Master's thesis, University of Waterloo, 2015.

[7] Beutel, Daniel J., Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani et al. "Flower: A friendly federated learning research framework." arXiv preprint arXiv:2007.14390 (2020).

[8] Elkordy, Ahmed Roushdy, and A. Salman Avestimehr. "Secure aggregation with heterogeneous quantization in federated learning." arXiv preprint arXiv:2009.14388 (2020).

[9] Ramaswamy, Swaroop, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. "Federated learning for emoji prediction in a mobile keyboard." arXiv preprint arXiv:1906.04329 (2019).

[10] Wang, Jianyu, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. "Tackling the objective inconsistency problem in heterogeneous federated optimization." Advances in neural information processing systems 33 (2020): 7611-7623.

[11] Acar, Durmus Alp Emre, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. "Federated learning based on dynamic regularization." arXiv preprint arXiv:2111.04263 (2021).

[12] Lin, Bill Yuchen, Chaoyang He, Zihang Zeng, Hulin Wang, Yufen Huang, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. "Fednlp: Benchmarking federated learning methods for natural language processing tasks." arXiv preprint arXiv:2104.08815 (2021).

[13] Zhang, Zhuo, Xiangjing Hu, Lizhen Qu, Qifan Wang, and Zenglin Xu. "Federated model decomposition with private vocabulary for text classification." In Empirical Methods in Natural Language Processing 2022, pp. 6413-6425. Association for Computational Linguistics (ACL), 2022.

[14] Geyer, Robin C., Tassilo Klein, and Moin Nabi. "Differentially private federated learning: A client level perspective." arXiv preprint arXiv:1712.07557 (2017).

[15] Francis, Sumam, Kanimozhi Uma, and Marie-Francine Moens. "Understanding the impact of label skewness and optimization on federated learning for text classification." In Companion Proceedings of the ACM Web Conference 2023, pp. 1161-1166. 2023.

[16] Delehouzée, Mathis, Xavier Lessage, Théo Reginster, and Saïd Mahmoudi. "Performance Analysis of Aggregation Algorithms in Cross-Silo Federated Learning for Non-IID Data." In 2024 4th International Conference on Embedded & Distributed Systems (EDiS), pp. 74-79. IEEE, 2024.

[17] Liu, Lifeng, Fengda Zhang, Jun Xiao, and Chao Wu. "Evaluation framework for large-scale federated learning." arXiv preprint arXiv:2003.01575 (2020).

[18] Leroy, David, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau. "Federated learning for keyword spotting." In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6341-6345. IEEE, 2019.

[19] Learning, Federated. "Collaborative machine learning without centralized training data." Publication date: Thursday 6 (2017).