Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Closed-Loop Fairness and Privacy Optimization for Federated Text Classification under Non-IID Data

Lihong Guo
Harbin Finance University
Mingkai Dai
Harbin University of Science and Technology
Luogang Zhang
Harbin University of Science and Technology
Chao Ma
machao@hrbust.edu.cn

Harbin University of Science and Technology

Research Article

Keywords: Federated learning, Differential privacy, Fairness-aware learning, Low-rank representation learning, Natural language processing, Text classification

Posted Date: July 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10011467/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Closed-Loop Fairness and Privacy Optimization for Federated Text Classification under Non-IID Data

Abstract

Federated text classification under non-IID data faces three coupled challenges: utility degradation, group fairness mismatch, and privacy leakage. To address these issues, this paper proposes FedMOJO, a closed-loop collaborative optimization framework for fair and privacy-preserving federated classification. On the client side, FedMOJO integrates low-rank representation learning, differentially private local training, and fairness–privacy collaborative optimization to reduce redundant sensitive information while preserving task-relevant semantics. On the server side, it performs low-rank structural alignment, global fairness-state fusion, state-driven control generation, and robust aggregation, thereby forming a cross-round closed loop between global coordination and local execution. Experiments on Twitter Sentiment, Bias in Bios, and Adult Income show that FedMOJO achieves a more favorable trade-off among utility, fairness, and privacy than representative federated baselines. Compared with DPFairFed, FedMOJO improves accuracy by up to 4.89 percentage points, while consistently reducing fairness disparity and privacy leakage and increasing MDL. These results indicate that closed-loop client–server coordination can effectively balance classification performance, group fairness, and privacy preservation without sharing raw data, providing a promising solution for privacy-sensitive federated learning scenarios.

Keywords: Federated learning; Differential privacy; Fairness-aware learning; Low-rank representation learning; Natural language processing; Text classification

1 Introduction

Text classification has become a fundamental task in natural language processing and has been extensively applied to sentiment analysis, public opinion monitoring, recruitment screening, clinical decision support, and risk assessment (Taha et al., 2024). With the increasing deployment of text-based models in real-world decision-making systems, the criteria used to assess such models have expanded well beyond predictive accuracy. In high-stakes scenarios, a reliable text classification model is expected not only to achieve strong task performance, but also to avoid systematic discrimination against sensitive groups and to reduce the risk of privacy leakage from training data or intermediate representations. Recent studies have shown that language technologies can inherit historical bias and distributional imbalance from training corpora,

thereby yielding disparate performance across groups defined by gender, race, region, or occupation (Gallegos et al., 2024; Leteno et al., 2025; Yang et al., 2024). Meanwhile, textual data often contain rich personal information, making trained models vulnerable to membership inference, attribute inference, and related privacy threats (Shahriar et al., 2025; Y. Wang et al., 2023).

To address these concerns, prior research has mainly evolved along two complementary directions. The first direction focuses on fairness-aware learning for text classification, where existing methods have gradually shifted from data balancing and post-processing correction toward representation-level debiasing, adversarial learning, and transferable fairness-aware modeling (Gallegos et al., 2024; Leteno et al., 2025). The second direction concerns privacy-preserving learning, among which differential privacy has become one of the most representative formal mechanisms because it can bound the influence of each individual training sample through perturbation and privacy accounting (Shahriar et al., 2025; Y. Wang et al., 2023). Despite the effectiveness of these two lines of research, fairness, privacy, and utility are intrinsically intertwined rather than independently optimizable. Differential privacy may weaken already scarce learning signals for minority groups, which can further exacerbate group disparities, while fairness-oriented constraints may alter the way a model exploits sensitive patterns and consequently reshape the privacy–utility trade-off (Kim et al., 2025; Rafi et al., 2024). Therefore, methods that optimize only one or two of these objectives in isolation often fail to achieve a robust overall balance.

Federated learning provides a promising paradigm for privacy-sensitive text analytics because it enables multiple parties to collaboratively train a shared model without directly centralizing raw data (Khan et al., 2024). This property is particularly attractive in domains such as healthcare, finance, and cross-organizational language intelligence, where textual records are both sensitive and institutionally siloed. However, federated learning does not automatically guarantee either fairness or privacy. Although raw data remain local, information leakage may still occur through gradients or model updates. More importantly, federated optimization is widely affected by statistical heterogeneity, label skew, resource asymmetry, and unstable client participation (Khan et al., 2024; B. Liu et al., 2024; Lu et al., 2024). These challenges are especially severe in text classification, where clients may differ substantially in topic distribution, writing style, label composition, and sensitive-group coverage. As a consequence, non-independent and identically distributed (non-IID) data drift becomes a major obstacle to stable global optimization and may further amplify fairness mismatch across clients (Lu et al., 2024; C. Wang et al., 2024).

In recent years, fair federated learning has emerged as an active research topic. Existing surveys indicate that fairness in federated environments is tightly coupled with privacy preservation, client selection, aggregation strategy, and heterogeneity mitigation (Kim et al., 2025; Mukhtiar et al., 2025; Rafi et al., 2024). Recent studies have explored fairness-aware dynamic client scheduling in heterogeneous systems (Mao et al., 2024) and differentially private low-rank adaptation under federated learning (X.-Y. Liu et al., 2025), suggesting that both fairness regulation and privacy-efficient parameter updating are critical for practical deployment. Nevertheless, current methods still exhibit several limitations when applied to federated text classification. First,

many approaches treat fairness and privacy merely as auxiliary regularizers imposed on otherwise standard federated optimization, rather than modeling them as jointly structured objectives over the learned text representation. Second, most methods rely on static trade-off weights, fixed perturbation policies, or one-shot local correction schemes, which are often inadequate when fairness risk, convergence behavior, and privacy-budget consumption evolve continuously across communication rounds. Third, local fairness improvement on individual clients does not necessarily lead to global fairness enhancement when the server lacks an explicit mechanism to aggregate fairness-relevant signals and regulate the training process at the system level under heterogeneous data distributions (Kim et al., 2025; Mukhtiar et al., 2025; Rafi et al., 2024).

Motivated by these observations, this paper investigates fair and privacy-preserving federated text classification from the perspective of closed-loop collaborative optimization under non-IID data. Rather than regarding fairness control, privacy preservation, and federated optimization as loosely connected components, we formulate them as a coupled dynamic system spanning client-side local learning and server-side global coordination. On the client side, low-rank representation learning is introduced to compress redundant semantics and suppress sensitive interference before locally differentially private optimization is carried out. In addition, each client constructs uploadable fairness statistics and local state summaries so that information relevant to global coordination can be preserved without exposing raw data. On the server side, conventional model aggregation is further enhanced by low-rank structural consistency maintenance, global fairness-statistics fusion, and state-driven control generation. Through this design, the overall training procedure forms a cross-round closed loop consisting of global state aggregation, control update, and local execution. Such a framework is intended to improve robustness against non-IID distributions, system heterogeneity, and cumulative privacy-budget consumption, while achieving a better overall trade-off among classification utility, group fairness, and privacy preservation.

The main contributions of this work are summarized as follows.

1. A closed-loop collaborative framework is proposed for federated text classification, which unifies low-rank representation learning, differentially private local training, multi-objective coordination, and server-side global regulation within a single optimization pipeline.
2. A client–server cooperative mechanism is developed for global fairness-state control through uploadable fairness statistics, low-rank structural alignment, and cross-round control-signal generation, thereby enabling the transition from locally isolated optimization to globally coordinated closed-loop learning.
3. Experiments on multiple public datasets demonstrate that the proposed method achieves a more robust balance among classification utility, group fairness, and privacy preservation in heterogeneous federated settings.

2 Related Work

2.1 Fair text classification

Text classification has been extensively studied in recent years, and recent surveys have shown that modern text classification systems have rapidly evolved from conventional statistical learning pipelines to deep neural and transformer-based architectures deployed across a wide range of application domains (Taha et al., 2024). However, as text classifiers are increasingly used in socially sensitive scenarios, fairness has become a central concern in addition to predictive performance. Recent literature on fairness in language technologies indicates that text models can encode and amplify harmful social bias, thereby producing systematic disparities across sensitive groups. In particular, Gallegos et al. (2024) show that fairness research in NLP has gradually evolved from coarse-grained dataset-level analysis toward a more structured understanding of harms, evaluation metrics, benchmark design, and mitigation strategies. This trend suggests that fairness in text classification is no longer viewed as a peripheral issue, but as an essential property of deployable language systems.

Methodologically, existing fair text classification approaches can be broadly grouped into data-level, objective-level, and representation-level interventions. Data-level methods typically attempt to mitigate imbalance or spurious correlations through augmentation, reweighting, or counterfactual construction, whereas objective-level methods incorporate fairness constraints directly into the training loss. More recent efforts increasingly emphasize representation-level control, since sensitive information often remains implicitly encoded even when explicit demographic variables are removed. A representative recent example is provided by Leteno et al. (2025), who address fair text classification through transferable representations and Wasserstein-based dependency control in the latent space. Their study highlights an important recent direction: fairness can be improved by explicitly regularizing learned representations rather than relying solely on output-level correction. Nevertheless, most existing fair text classification methods remain primarily focused on centralized settings and do not explicitly consider privacy preservation or decentralized optimization during training (Gallegos et al., 2024; Leteno et al., 2025).

2.2 Privacy-preserving text learning

Privacy-preserving learning for text data has also attracted increasing attention, largely because textual corpora frequently contain sensitive personal information that may be exposed through memorization, inference attacks, or representation leakage. Shahriar et al. (2025) provide a recent review of text data privacy and classify the literature into multiple privacy-risk categories, including unintentional memorability, membership inference, exposure and re-identification, privacy risks in language models and embeddings, authorship attribution, and collaborative processing. Their review highlights that privacy risks in text learning arise not only from raw data sharing, but also from intermediate representations and collaborative modeling pipelines, which is particularly relevant to modern NLP systems.

Among existing privacy-preserving paradigms, differential privacy remains one of the most widely adopted formal mechanisms. Y. Wang et al. (2023) show that recent work on differential privacy in deep learning has concentrated on integrating perturbation mechanisms with gradient-based training while also exploring broader roles of differential privacy in fairness, robustness, and regularization. Although differential privacy provides a mathematically grounded privacy guarantee, applying it to NLP tasks often introduces a significant utility cost, especially when textual representations are high-dimensional and semantically dense. This challenge is further complicated when privacy requirements coexist with fairness constraints, because noise injection may disproportionately weaken minority-group learning signals. As a result, privacy-preserving text learning increasingly requires more structured control over where and how perturbation is applied, rather than relying exclusively on uniform privacy mechanisms (Shahriar et al., 2025; Y. Wang et al., 2023).

2.3 Federated text learning and fair federated learning

Federated learning offers an appealing alternative to centralized training when text data cannot be directly pooled across institutions or platforms. In the NLP domain, Khan et al. (2024) provide a systematic review showing that federated learning has been applied to a growing range of language tasks, while also emphasizing persistent issues related to convergence, communication overhead, privacy attacks, and deployment complexity. More generally, recent federated learning surveys have consistently identified non-IID data, client heterogeneity, and unstable participation as fundamental obstacles to robust global optimization (B. Liu et al., 2024). These problems are particularly severe for text classification, where clients may differ markedly in topic distribution, label composition, linguistic style, and sensitive-group coverage.

The intersection of fairness and privacy in federated learning has consequently become an active research area. Rafi et al. (2024) present a dedicated survey on fairness and privacy preservation in federated learning and show that these two objectives are deeply intertwined rather than independently solvable. More recently, Kim et al. (2025) provide a comprehensive review of fair federated learning and further confirm that data partitioning, privacy mechanisms, heterogeneity mitigation, and system architecture all influence fairness outcomes in federated environments. On the methodological side, Zhang et al. (2025) propose SFFL, a self-aware fairness federated learning framework that improves fairness and performance under heterogeneous data distributions without requiring clients' sensitive attributes. Their study demonstrates that fairness-aware local training can partially alleviate the mismatch caused by inconsistent local objectives under non-IID settings. In parallel, X.-Y. Liu et al. (2025) investigate differentially private low-rank adaptation for federated fine-tuning of large language models, suggesting that low-rank parameterization may offer a more communication-efficient and privacy-aware alternative to full-parameter federated optimization.

Despite these advances, several limitations remain for federated text classification. First, many existing studies treat fairness, privacy, and federated optimization as loosely coupled add-on components rather than jointly structured objectives over text representations. Second, most approaches still rely on static regularization strengths, fixed perturbation policies, or local fairness adjustments, which may be insufficient

when fairness risk, convergence behavior, and privacy-budget consumption evolve across communication rounds. Third, although some methods improve local fairness or privacy efficiency, they generally provide limited mechanisms for server-side aggregation of fairness-relevant signals and closed-loop regulation at the global level under heterogeneous client distributions. Taken together, the existing literature suggests that a more integrated framework is needed to jointly model representation compression, privacy-aware local training, and global fairness-state coordination in non-IID federated text classification (Khan et al., 2024; Kim et al., 2025; B. Liu et al., 2024; X.-Y. Liu et al., 2025; Rafi et al., 2024; Zhang et al., 2025).

3 Methodology

This section presents the proposed closed-loop collaborative optimization framework for fair and privacy-preserving federated text classification under non-IID data. Different from conventional federated methods that mainly rely on local fairness regularization, local differential privacy, or direct combinations of the two, the proposed method formulates federated fair and privacy-preserving learning as a unified client–server coordination problem. The overall framework is first introduced, followed by the client-side local optimization mechanism and the server-side global coordination process.

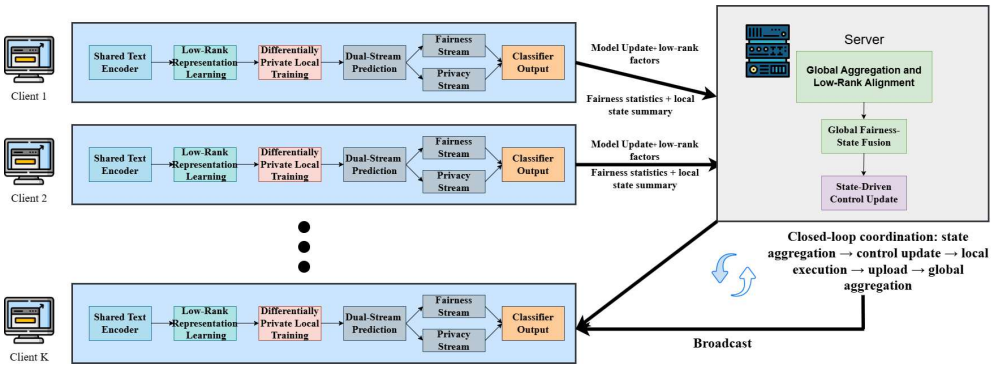


Fig. 1: Overall framework of the proposed closed-loop collaborative optimization method for fair and privacy-preserving federated text classification. Each client performs low-rank representation learning, differentially private local training, and dual-stream fairness–privacy prediction, while uploading model updates, aligned low-rank factors, fairness statistics, and local state summaries to the server. The server conducts global aggregation, low-rank alignment, fairness-state fusion, and state-driven control update, forming a cross-round closed loop of global coordination and local execution.

3.1 Overall Framework

Consider a federated system with K clients. The local dataset of client k is denoted by

$$\mathcal{D}_k = \{(x_i, y_i, a_i)\}_{i=1}^{N_k}, \tag{1}$$

where x_i, y_i, and a_i denote the input text, task label, and sensitive attribute, respectively, and N_k is the number of local samples. At communication round t, the server maintains the global model $\Theta^{(t)}$ and broadcasts it together with a round-specific control vector

$$\boldsymbol{\phi}^{(t)} = \left[\lambda_f^{(t)}, \lambda_r^{(t)}, \lambda_c^{(t)}, \mu^{(t)}, \sigma^{(t)}\right], \tag{2}$$

where $\lambda_f^{(t)}$, $\lambda_r^{(t)}$, and $\lambda_c^{(t)}$ denote the strengths of fairness regularization, low-rank structural regularization, and consistency stabilization, respectively, $\mu^{(t)}$ is the proximal coefficient for controlling local drift, and $\sigma^{(t)}$ is the noise scale for differentially private local training.

The global optimization objective is formulated as

$$\min_{\Theta} \sum_{k=1}^{K} \frac{N_k}{\sum_{j=1}^{K} N_j} \mathcal{L}_k(\Theta), \tag{3}$$

where $\mathcal{L}_k(\Theta)$ is the local multi-objective loss of client k. Unlike standard federated empirical risk minimization, the proposed framework explicitly treats utility, fairness, and privacy as jointly coordinated objectives.

As shown in Fig. 1, each client starts from a shared text encoder and performs three sequential steps: low-rank representation learning, differentially private local training, and dual-stream fairness–privacy prediction. In addition to model updates, each client constructs uploadable fairness statistics and local state summaries that characterize fairness risk, privacy-budget consumption, and local optimization stability. On the server side, the received local packets are used not only for model aggregation, but also for low-rank structural consistency maintenance, global fairness-state fusion, and state-driven control generation. Therefore, the overall training process forms a cross-round closed loop of global state aggregation, control update, and local execution.

To make the interaction explicit, let

$$\mathcal{P}_k^{(t)} = \left\{\widetilde{\Delta\Theta}_{k,\mathrm{den}}^{(t)}, \widetilde{A}_k^{(t)}, \widetilde{B}_k^{(t)}, \widetilde{\Xi}_k^{(t)}, u_k^{(t)}\right\} \tag{4}$$

denote the upload packet of client k at round t, where $\widetilde{\Delta\Theta}_{k,\mathrm{den}}^{(t)}$ is the compressed dense update, $\widetilde{A}_k^{(t)}$ and $\widetilde{B}_k^{(t)}$ are the aligned low-rank adapter factors, $\widetilde{\Xi}_k^{(t)}$ denotes privacy-aware fairness statistics, and $u_k^{(t)}$ is the local state summary. Based on the packets from active clients $\mathcal{S}^{(t)}$, the server estimates the global state as

$$s^{(t)} = \mathcal{F}_{\mathrm{agg}}\left(\left\{\mathcal{P}_k^{(t)}\right\}_{k\in\mathcal{S}^{(t)}}\right), \tag{5}$$

where $\mathcal{F}_{\text{agg}}(\cdot)$ denotes the aggregation-and-fusion operator. The control vector is then updated by

$$\phi^{(t+1)} = \mathcal{G}_{\phi}\left(s^{(t)}\right), \tag{6}$$

where $\mathcal{G}_{\phi}(\cdot)$ denotes the state-driven control generator. The updated control vector is broadcast to clients for the next round of local execution.

3.2 Client-Side Local Optimization

At communication round t, each active client $k \in \mathcal{S}^{(t)}$ receives the global model $\Theta^{(t)}$ and the control vector $\boldsymbol{\phi}^{(t)}$, and then performs local optimization on $\mathcal{D}_k$. The local objective is defined as

$$\mathcal{L}_k^{(t)} = \mathcal{L}_{\text{cls},k}^{(t)} + \lambda_f^{(t)}\mathcal{L}_{\text{fair},k}^{(t)} + \lambda_r^{(t)}\mathcal{L}_{\text{rank},k}^{(t)} + \lambda_c^{(t)}\mathcal{L}_{\text{cons},k}^{(t)} + \mu^{(t)}\mathcal{L}_{\text{prox},k}^{(t)}, \tag{7}$$

where $\mathcal{L}_{\text{cls},k}^{(t)}$ is the classification loss, $\mathcal{L}_{\text{fair},k}^{(t)}$ is the local fairness regularization, $\mathcal{L}_{\text{rank},k}^{(t)}$ is the low-rank structural loss, $\mathcal{L}_{\text{cons},k}^{(t)}$ is the consistency stabilization term, and $\mathcal{L}_{\text{prox},k}^{(t)}$ is the proximal regularizer that mitigates excessive local drift.

3.2.1 Low-rank local learning and fairness-aware prediction

Dual-stream representation and gated fusion.

To explicitly coordinate fairness-aware representation learning and privacy-preserving local prediction, each client adopts a dual-stream representation module before the final classifier. Given an input sample x_i, the shared encoder first maps it into a latent representation:

$$h_i = f_{\theta_e}(x_i), \qquad h_i \in \mathbb{R}^{d_h}, \tag{8}$$

where $f_{\theta_e}(\cdot)$ denotes the shared text encoder. Based on h_i, two parallel representation streams are constructed. The fairness-enhanced stream is defined as

$$h_i^f = F_f(h_i; \theta_f), \tag{9}$$

where $F_f(\cdot)$ aims to preserve task-relevant information while supporting the fairness regularization defined later. The privacy-preserving stream is defined as

$$h_i^p = F_p(h_i; \theta_p, A_k^{(t)}, B_k^{(t)}), \tag{10}$$

where $F_p(\cdot)$ is implemented with the low-rank adapted transformation parameterized by $A_k^{(t)}$ and $B_k^{(t)}$. In this stream, privacy preservation is achieved through the combination of low-rank structural constraint, differentially private local training, and perturbation-aware optimization, rather than through an additional adversarial privacy classifier.

The two streams are adaptively fused by a gated mechanism. Specifically, the sample-wise gate vector is computed as

$$g_i^{(t)} = \sigma\left(W_g^{(t)}[h_i^f; h_i^p] + b_g^{(t)}\right), \tag{11}$$

where $[h_i^f; h_i^p]$ denotes concatenation, $\sigma(\cdot)$ is the sigmoid function, and $g_i^{(t)} \in (0,1)^{d_h}$ controls the contribution of the two streams. The fused representation is then obtained by

$$h_i^{\text{fused}} = g_i^{(t)} \odot h_i^f + \left(1 - g_i^{(t)}\right) \odot h_i^p, \tag{12}$$

where $\odot$ denotes element-wise multiplication. Finally, the prediction probability is computed as

$$\hat{y}_i = \text{softmax}\left(W_c h_i^{\text{fused}} + b_c\right). \tag{13}$$

For binary classification, $\hat{p}_{i,1}^{(t)}$ denotes the predicted probability of the positive class. For multi-class classification, $\hat{y}_{i,c}$ denotes the predicted probability of class c.

To describe the round-level stability of the gated fusion module, the client-level average gate vector is further defined as

$$\bar{g}_k^{(t)} = \frac{1}{N_k} \sum_{i=1}^{N_k} g_i^{(t)}. \tag{14}$$

Based on the fused representation h_i^{fused}, the classification loss is defined as the cross-entropy loss

$$\mathcal{L}_{\text{cls},k}^{(t)} = -\frac{1}{N_k} \sum_{i=1}^{N_k} \sum_{c=1}^{C} \mathbf{1}(y_i = c) \log \hat{y}_{i,c}, \tag{15}$$

where $\hat{y}_{i,c}$ is the predicted probability of class c for sample x_i.

For binary classification tasks, the local fairness regularization adopts a differentiable proxy of TPR disparity. Let $\hat{p}_{i,1}^{(t)}$ denote the predicted probability that sample x_i belongs to the positive class at round t. The soft TPR of sensitive group g on client k is defined as

$$\widetilde{\text{TPR}}_{k,g}^{(t)} = \frac{\sum_{i=1}^{N_k} \mathbf{1}(a_i = g)\mathbf{1}(y_i = 1)\hat{p}_{i,1}^{(t)}}{\sum_{i=1}^{N_k} \mathbf{1}(a_i = g)\mathbf{1}(y_i = 1) + \epsilon_0}, \tag{16}$$

where ϵ_0 is a small constant used for numerical stability. The client-level mean soft TPR is computed as

$$\overline{\text{TPR}}_k^{(t)} = \frac{1}{G} \sum_{g=1}^{G} \widetilde{\text{TPR}}_{k,g}^{(t)}. \tag{17}$$

Accordingly, the binary fairness regularization term is defined as

$$\mathcal{L}_{\text{fair},k}^{(t)} = \frac{1}{G} \sum_{g=1}^{G} \left(\widetilde{\text{TPR}}_{k,g}^{(t)} - \overline{\text{TPR}}_k^{(t)}\right)^2. \tag{18}$$

This formulation uses predicted probabilities rather than hard true-positive counts, and therefore provides a differentiable approximation of group-level TPR disparity during local training.

For multi-class tasks, the local fairness term is written as

$$\mathcal{L}_{\text{fair},k}^{(t)} = \frac{1}{GC} \sum_{g=1}^{G} \sum_{c=1}^{C} \left(\bar{p}_{k,g,c}^{(t)} - \bar{p}_{k,c}^{(t)}\right)^2, \tag{19}$$

where $\bar{p}_{k,g,c}^{(t)}$ is the average predicted probability of class c for group g, and $\bar{p}_{k,c}^{(t)}$ is the client-level mean predicted probability of class c.

To stabilize representation learning under non-IID local data, we introduce a parameter-level low-rank adapter into the shared representation transformation. Let $W^{(t)} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ denote a trainable transformation matrix in the text classifier. On client k, the local low-rank adapter is defined as

$$\Delta W_{k,\text{lr}}^{(t)} = A_k^{(t)} \left(B_k^{(t)}\right)^\top, \quad A_k^{(t)} \in \mathbb{R}^{d_{\text{out}} \times r}, \quad B_k^{(t)} \in \mathbb{R}^{d_{\text{in}} \times r}, \tag{20}$$

where $r \ll \min(d_{\text{out}}, d_{\text{in}})$ is the low-rank dimension. The adapted transformation used by client k is then written as

$$W_k^{(t)} = W^{(t)} + \Delta W_{k,\text{lr}}^{(t)}. \tag{21}$$

The low-rank structural regularization is defined as

$$\begin{aligned} \mathcal{L}_{\text{rank},k}^{(t)} = &\left\| \Delta W_{k,\text{lr}}^{(t)} - \Delta W_{\text{lr}}^{(t)} \right\|_F^2 + \eta_A \left\| A_k^{(t)} \right\|_F^2 + \eta_B \left\| B_k^{(t)} \right\|_F^2 \\ &+ \eta_o \left(\left\| \left(A_k^{(t)}\right)^\top A_k^{(t)} - I_r \right\|_F^2 + \left\| \left(B_k^{(t)}\right)^\top B_k^{(t)} - I_r \right\|_F^2 \right), \end{aligned} \tag{22}$$

where $\Delta W_{\text{lr}}^{(t)} = A^{(t)} \left(B^{(t)}\right)^\top$ denotes the global low-rank adapter broadcast by the server, $A^{(t)}$ and $B^{(t)}$ are the corresponding global low-rank factors, and η_A, η_B, and η_o are regularization coefficients. Unlike sample-level representation factorization, this parameter-level formulation makes the low-rank factors independent of the number of local samples N_k, thereby allowing the aligned low-rank factors to be validly aggregated across clients.

To reduce dual-stream divergence and gate oscillation, the consistency loss is defined as

$$L_{\text{cons},k}^{(t)} = \frac{1}{N_k} \sum_{i=1}^{N_k} \left\| h_i^f - h_i^p \right\|_2^2 + \gamma \left\| \bar{g}_k^{(t)} - \bar{g}_k^{(t-1)} \right\|_2^2, \tag{23}$$

where h_i^f and h_i^p are the outputs of the fairness-enhanced stream and the privacy-preserving stream, respectively, $\bar{g}_k^{(t)}$ denotes the client-level average gate vector at round t, and γ is the smoothing coefficient. The first term prevents excessive divergence between the two representation streams, while the second term suppresses abrupt gate oscillation across communication rounds. The proximal regularizer is

$$\mathcal{L}_{\text{prox},k}^{(t)} = \frac{1}{2} \|\Theta_k - \Theta^{(t)}\|_2^2. \tag{24}$$

3.2.2 Differentially private local training

Local optimization is performed under differential privacy using DP-SGD. For the b-th mini-batch on client k at communication round t, let the per-sample gradient be

$$g_{k,i}^{(t,b)} = \nabla_\Theta \ell(x_i, y_i, a_i; \Theta). \tag{25}$$

The gradient is clipped by the threshold $C^{(t)}$:

$$\bar{g}_{k,i}^{(t,b)} = \frac{g_{k,i}^{(t,b)}}{\max\left(1, \left\| g_{k,i}^{(t,b)} \right\|_2 / C^{(t)}\right)}. \tag{26}$$

The noisy batch gradient is then computed as

$$\widetilde{g}_k^{(t,b)} = \frac{1}{B_k} \sum_{i=1}^{B_k} \bar{g}_{k,i}^{(t,b)} + \mathcal{N}\left(0, \left(\frac{\sigma^{(t)} C^{(t)}}{B_k}\right)^2 I\right), \tag{27}$$

and the local update is

$$\Theta_k^{(t,b+1)} = \Theta_k^{(t,b)} - \eta_k \widetilde{g}_k^{(t,b)}, \tag{28}$$

where η_k is the local learning rate.

For privacy accounting, let $q_k = B_k/N_k$ denote the effective mini-batch sampling rate on client k, where B_k is the mini-batch size and N_k is the number of local training samples. Let $T_k^{(t)}$ denote the number of local DP-SGD optimization steps performed by client k in round t. In this work, RDP accounting is conducted over the order set

$$\mathcal{A} = \{2, 3, \ldots, 64\}. \tag{29}$$

For an integer order $\alpha \in \mathcal{A}$, the per-step RDP cost of the subsampled Gaussian mechanism is computed as

$$\rho_\alpha^{\mathrm{SGM}}\left(q_k, \sigma^{(t)}\right) = \frac{1}{\alpha - 1} \log\left[\sum_{j=0}^{\alpha} \binom{\alpha}{j} (1 - q_k)^{\alpha - j} q_k^j \exp\left(\frac{j(j-1)}{2(\sigma^{(t)})^2}\right)\right]. \tag{30}$$

Therefore, the DP-SGD RDP increment of client k in round t is

$$\bar{\rho}_{k,\alpha,\mathrm{sgd}}^{(t)} = T_k^{(t)} \rho_\alpha^{\mathrm{SGM}}\left(q_k, \sigma^{(t)}\right). \tag{31}$$

In addition to DP-SGD, the uploaded fairness statistics are also perturbed before transmission. Let $\bar{\rho}_{k,\alpha,\mathrm{stat}}^{(t)}$ denote the RDP increment introduced by the Gaussian perturbation on the uploaded statistics. The cumulative RDP loss of client k up to

round t is then written as

$$\rho_{k,\alpha}^{(t)} = \sum_{\tau=1}^{t} \mathbf{1}\left(k \in \mathcal{S}^{(\tau)}\right) \left[\bar{\rho}_{k,\alpha,\mathrm{sgd}}^{(\tau)} + \bar{\rho}_{k,\alpha,\mathrm{stat}}^{(\tau)}\right]. \tag{32}$$

Finally, the cumulative RDP guarantee is converted to (ϵ, δ)-DP by

$$\epsilon_k^{(t)} = \min_{\alpha \in \mathcal{A}} \left(\rho_{k,\alpha}^{(t)} + \frac{\log(1/\delta)}{\alpha - 1}\right). \tag{33}$$

The adaptive noise scale $\sigma^{(t)}$ only affects subsequent communication rounds and does not reduce the privacy loss already accumulated in previous rounds.

3.2.3 Uploadable fairness statistics and local state summaries

Besides local parameter learning, each client constructs uploadable fairness statistics. For binary tasks, the uploaded group statistics are

$$\Xi_k^{(t)} = \left\{ \left(n_{k,g}^{(t)}, p_{k,g}^{(t)}, tp_{k,g}^{(t)}, \bar{s}_{k,g}^{(t)}\right) \,\middle|\, g = 1, \ldots, G \right\}, \tag{34}$$

where $n_{k,g}^{(t)}$, $p_{k,g}^{(t)}$, $tp_{k,g}^{(t)}$, and $\bar{s}_{k,g}^{(t)}$ denote the sample count, positive count, true-positive count, and mean prediction score of group g, respectively. For multi-class tasks, the uploaded statistics are

$$\Xi_k^{(t)} = \left\{ \left(n_{k,g,c}^{(t)}, \bar{p}_{k,g,c}^{(t)}\right) \,\middle|\, g = 1, \ldots, G,\ c = 1, \ldots, C \right\}. \tag{35}$$

The local state summary is defined as

$$\mathbf{u}_k^{(t)} = \left[\Delta\mathcal{L}_{\mathrm{cls},k}^{(t)}, \Delta\mathcal{L}_{\mathrm{fair},k}^{(t)}, \Delta v_k^{(t)}, \Delta\varepsilon_k^{(t)}\right], \tag{36}$$

where

$$\Delta\mathcal{L}_{\mathrm{cls},k}^{(t)} = \mathcal{L}_{\mathrm{cls},k}^{(t)} - \mathcal{L}_{\mathrm{cls},k}^{(t-1)}, \quad \Delta\mathcal{L}_{\mathrm{fair},k}^{(t)} = \mathcal{L}_{\mathrm{fair},k}^{(t)} - \mathcal{L}_{\mathrm{fair},k}^{(t-1)}, \tag{37}$$

and

$$v_k^{(t)} = \left\| \bar{g}_k^{(t)} - \bar{g}_k^{(t-1)} \right\|_2^2, \qquad \Delta\epsilon_k^{(t)} = \epsilon_k^{(t)} - \epsilon_k^{(t-1)}. \tag{38}$$

To reduce the risk of disclosing local group information, the uploaded fairness statistics are clipped and perturbed before transmission. Let $\mathrm{clip}_{\Xi}(\cdot)$ denote the clipping operator applied to the fairness statistics. Count-based statistics are clipped to a bounded non-negative range, and score-based statistics are clipped to $[0, 1]$. The clipped statistics are defined as

$$\widehat{\Xi}_k^{(t)} = \mathrm{clip}_{\Xi}\left(\Xi_k^{(t)}\right). \tag{39}$$

Let C_Ξ be the ℓ_2 sensitivity bound of the clipped fairness statistics. The transmitted fairness statistics are generated by

$$\widetilde{\Xi}_k^{(t)} = \Pi_{\mathcal{D}_\Xi}\left[\widehat{\Xi}_k^{(t)} + \mathcal{N}\left(0, (\sigma_{\text{stat}} C_\Xi)^2 I\right)\right], \tag{40}$$

where σ_{stat} is the statistics noise multiplier and $\Pi_{\mathcal{D}_\Xi}(\cdot)$ projects the perturbed statistics back to the valid domain. Specifically, count-based entries are projected to non-negative values, while score-based entries are projected to $[0, 1]$.

Since the uploaded statistics may reveal local sensitive-group composition, their privacy cost is included in the cumulative RDP accounting. Under the Gaussian mechanism with noise standard deviation $\sigma_{\text{stat}} C_\Xi$ and sensitivity C_Ξ, the RDP increment of the statistics perturbation is

$$\bar{\rho}_{k,\alpha,\text{stat}}^{(t)} = \frac{\alpha}{2\sigma_{\text{stat}}^2}. \tag{41}$$

3.2.4 Communication-efficient upload

To reduce communication overhead, compressed dense updates and aligned low-rank factors are uploaded separately. Let the local model after E local epochs be $\Theta_k^{(t,E)}$. The dense update is

$$\Delta\Theta_k^{(t)} = \Theta_k^{(t,E)} - \Theta^{(t)}. \tag{42}$$

Using a compression operator $\mathcal{C}(\cdot)$ with error feedback, the transmitted dense update is

$$\tilde{\Delta\Theta}_k^{(t)} = \mathcal{C}\left(\Delta\Theta_k^{(t)} + \mathbf{r}_k^{(t-1)}\right), \tag{43}$$

and the residual is updated by

$$\mathbf{r}_k^{(t)} = \mathbf{r}_k^{(t-1)} + \Delta\Theta_k^{(t)} - \tilde{\Delta\Theta}_k^{(t)}. \tag{44}$$

For the low-rank adapter block, the local low-rank update is represented as

$$\Delta W_{k,\text{lr}}^{(t)} = A_k^{(t)}\left(B_k^{(t)}\right)^\top. \tag{45}$$

Since low-rank factorization is invariant to orthogonal rotations, i.e., $\Delta W_{k,\text{lr}}^{(t)} = A_k^{(t)} R \left(B_k^{(t)} R\right)^\top$ for any orthogonal matrix R, the client aligns its local low-rank adapter factors to the global reference factors before uploading. Specifically, the alignment matrix is obtained by solving

$$R_k^{(t)} = \underset{R^\top R = I}{\arg\min} \left(\left\| A_k^{(t)} R - A^{(t)} \right\|_F^2 + \left\| B_k^{(t)} R - B^{(t)} \right\|_F^2 \right), \tag{46}$$

where $A^{(t)}$ and $B^{(t)}$ are the global low-rank adapter factors broadcast by the server. The aligned low-rank adapter factors are then computed as

$$\tilde{A}_k^{(t)} = A_k^{(t)} R_k^{(t)}, \qquad \tilde{B}_k^{(t)} = B_k^{(t)} R_k^{(t)}. \tag{47}$$

Finally, each client uploads the packet in Eq. (4), which contains the compressed dense update, the aligned low-rank adapter factors, fairness statistics, and local state summaries for global coordination.

3.3 Server-Side Global Coordination

The server acts as a global coordination center rather than a simple parameter aggregator. In fair and privacy-preserving federated text classification, utility, fairness, and privacy evolve jointly across rounds and are strongly affected by non-IID distributions, local optimization drift, and cumulative privacy-budget consumption. Therefore, the server integrates multiple types of information uploaded by clients and uses them to regulate the next round of local learning.

3.3.1 Global parameter aggregation with low-rank consistency maintenance

Let $\mathcal{S}^{(t)}$ denote the set of active clients at round t, and let

$$N^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} N_k \tag{48}$$

be the total number of samples involved in this round. The basic aggregation weight is

$$\omega_k^{(t)} = \frac{N_k}{N^{(t)}}. \tag{49}$$

The global model is decomposed into a dense block and a low-rank adapter block:

$$\Theta^{(t)} = \left\{ \Theta_{\text{den}}^{(t)}, A^{(t)}, B^{(t)} \right\}, \qquad \Delta W_{\text{lr}}^{(t)} = A^{(t)} \left(B^{(t)} \right)^{\top}. \tag{50}$$

The aggregated dense update is

$$\bar{\Delta\Theta}_{\text{den}}^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} \omega_k^{(t)} \Delta\tilde{\Theta}_{k,\text{den}}^{(t)}, \tag{51}$$

and the dense block is updated by

$$\Theta_{\text{den}}^{(t+1)} = \Theta_{\text{den}}^{(t)} + \bar{\Delta\Theta}_{\text{den}}^{(t)}. \tag{52}$$

For the low-rank adapter block, the aligned low-rank factors are aggregated separately:

$$\bar{A}^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} \omega_k^{(t)} \tilde{A}_k^{(t)}, \qquad \bar{B}^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} \omega_k^{(t)} \tilde{B}_k^{(t)}. \tag{53}$$

To prevent scale instability in the low-rank factors, a column-wise normalization operator is applied:

$$A^{(t+1)} = \mathcal{N}_A\left(\bar{A}^{(t)}\right), \qquad B^{(t+1)} = \mathcal{N}_B\left(\bar{B}^{(t)}\right), \tag{54}$$

where $\mathcal{N}_A(\cdot)$ and $\mathcal{N}_B(\cdot)$ denote normalization operators that stabilize the scale of the aggregated low-rank factors. The updated global low-rank adapter is then obtained as

$$\Delta W_{\mathrm{lr}}^{(t+1)} = A^{(t+1)}\left(B^{(t+1)}\right)^{\top}. \tag{55}$$

Thus, the updated global model becomes

$$\Theta^{(t+1)} = \left\{\Theta_{\mathrm{den}}^{(t+1)}, A^{(t+1)}, B^{(t+1)}\right\}. \tag{56}$$

3.3.2 Global fairness-state fusion

After model aggregation, the server reconstructs the global fairness state from uploaded statistics. For binary tasks, the group-level counts are fused as

$$n_g^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} n_{k,g}^{(t)}, \qquad p_g^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} p_{k,g}^{(t)}, \qquad tp_g^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} tp_{k,g}^{(t)}. \tag{57}$$

The global TPR of group g is

$$\mathrm{TPR}_g^{(t)} = \frac{tp_g^{(t)}}{p_g^{(t)} + \epsilon_0}, \tag{58}$$

and the global mean TPR is

$$\overline{\mathrm{TPR}}^{(t)} = \frac{1}{G}\sum_{g=1}^{G} \mathrm{TPR}_g^{(t)}. \tag{59}$$

Accordingly, the global fairness gap is

$$\Delta_{\mathrm{fair}}^{(t)} = \frac{1}{G}\sum_{g=1}^{G} \left|\mathrm{TPR}_g^{(t)} - \overline{\mathrm{TPR}}^{(t)}\right|. \tag{60}$$

For multi-class tasks, a GRMS-style global fairness proxy is constructed from the fused class-group statistics:

$$\Delta_{\text{fair}}^{(t)} = \sqrt{\frac{1}{GC}\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\bar{p}_{g,c}^{(t)} - \bar{p}_{c}^{(t)}\right)^2}. \tag{61}$$

This design enables the server to estimate a global fairness state even though each client can only observe its local group distribution.

3.3.3 Global state estimation and control generation

In addition to fairness statistics, the server aggregates local state summaries:

$$\bar{\mathbf{u}}^{(t)} = \sum_{k\in\mathcal{S}^{(t)}} \omega_k^{(t)}\mathbf{u}_k^{(t)} = \left[\Delta\bar{\mathcal{L}}_{\text{cls}}^{(t)}, \Delta\bar{\mathcal{L}}_{\text{fair}}^{(t)}, \bar{v}^{(t)}, \bar{\varepsilon}^{(t)}\right]. \tag{62}$$

The average privacy level and participation coverage are

$$\bar{\varepsilon}^{(t)} = \sum_{k\in\mathcal{S}^{(t)}} \omega_k^{(t)}\varepsilon_k^{(t)}, \qquad \rho_{\text{part}}^{(t)} = \frac{|\mathcal{S}^{(t)}|}{K}. \tag{63}$$

The global state vector is then defined as

$$\mathbf{s}^{(t)} = \left[\Delta_{\text{fair}}^{(t)}, \bar{v}^{(t)}, \Delta\bar{\mathcal{L}}_{\text{cls}}^{(t)}, \bar{\varepsilon}^{(t)}, \rho_{\text{part}}^{(t)}\right]. \tag{64}$$

Based on $s^{(t)}$, the server updates the control vector through

$$\phi^{(t+1)} = \Pi_{\Omega}\left(\phi^{(t)} + M_{\phi}s^{(t)}\right), \tag{65}$$

where M_ϕ is the regulation matrix for state-driven control and $\Pi_\Omega(\cdot)$ projects the updated vector onto the feasible set Ω.

In component form, the control variables are updated by

$$\lambda_f^{(t+1)} = \text{clip}\left(\lambda_f^{(t)} + \eta_f(\Delta_{\text{fair}}^{(t)} - \tau_f), \lambda_f^{\min}, \lambda_f^{\max}\right), \tag{66}$$

$$\lambda_r^{(t+1)} = \text{clip}\left(\lambda_r^{(t)} + \eta_r(d_{\text{lr}}^{(t)} - \tau_r), \lambda_r^{\min}, \lambda_r^{\max}\right), \tag{67}$$

$$\lambda_c^{(t+1)} = \text{clip}\left(\lambda_c^{(t)} + \eta_c(\bar{v}^{(t)} - \tau_c), \lambda_c^{\min}, \lambda_c^{\max}\right), \tag{68}$$

$$\mu^{(t+1)} = \text{clip}\left(\mu^{(t)} + \eta_\mu(d_{\text{drift}}^{(t)} - \tau_\mu), \mu^{\min}, \mu^{\max}\right), \tag{69}$$

and

$$\sigma^{(t+1)} = \text{clip}\left(\sigma^{(t)} + \eta_\sigma(\bar{\varepsilon}^{(t)} - \tau_\varepsilon), \sigma^{\min}, \sigma^{\max}\right). \tag{70}$$

3.3.4 Robust aggregation under non-IID data and system heterogeneity

To improve robustness, each transmitted dense update is first clipped:

$$\check{\Delta\Theta}_k^{(t)} = \frac{\tilde{\Delta\Theta}_k^{(t)}}{\max\left(1, \|\tilde{\Delta\Theta}_k^{(t)}\|_2 / C_{\text{agg}}\right)}, \tag{71}$$

where C_{agg} is the server-side aggregation threshold. A baseline average is then computed as

$$\bar{\Delta\Theta}_{\text{base}}^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} \omega_k^{(t)} \check{\Delta\Theta}_k^{(t)}. \tag{72}$$

The update drift of client k is

$$d_k^{(t)} = \left\| \check{\Delta\Theta}_k^{(t)} - \bar{\Delta\Theta}_{\text{base}}^{(t)} \right\|_2, \tag{73}$$

and the global average drift is

$$d_{\text{drift}}^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} \omega_k^{(t)} d_k^{(t)}. \tag{74}$$

For the low-rank adapter block, the server also measures the structural drift of the aligned low-rank updates:

$$d_{\text{lr}}^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} \omega_k^{(t)} \left\| \tilde{A}_k^{(t)} \left(\tilde{B}_k^{(t)}\right)^\top - A^{(t+1)} \left(B^{(t+1)}\right)^\top \right\|_F. \tag{75}$$

A stability-aware reliability score is assigned to each client:

$$r_k^{(t)} = \exp\left(-\alpha d_k^{(t)} - \beta \|\mathbf{u}_k^{(t)} - \bar{\mathbf{u}}^{(t)}\|_2^2\right), \tag{76}$$

where α and β control the contributions of update drift and local instability. If participation dominance is considered, let

$$\nu_k^{(t)} = \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}(k \in \mathcal{S}^{(\tau)}) \tag{77}$$

be the participation ratio of client k up to round t. The frequency-corrected reliability score is

$$\tilde{r}_k^{(t)} = \frac{r_k^{(t)}}{1 + \xi \nu_k^{(t)}}, \tag{78}$$

where ξ is the participation-frequency penalty. The corrected aggregation weight becomes

$$\hat{\omega}_k^{(t)} = \frac{N_k \tilde{r}_k^{(t)}}{\sum_{j \in \mathcal{S}^{(t)}} N_j \tilde{r}_j^{(t)}}. \tag{79}$$

Accordingly, the robust dense aggregation is

$$\Delta\Theta_{\text{rob}}^{(t)} = \sum_{k \in \mathcal{S}^{(t)}} \hat{\omega}_k^{(t)} \Delta\check{\Theta}_k^{(t)}. \tag{80}$$

The same corrected weights can also replace $\omega_k^{(t)}$ in Eq. (53), yielding robust low-rank fusion under non-IID data and system heterogeneity.

Overall, the server-side procedure integrates global parameter aggregation, low-rank structural consistency maintenance, fairness-state reconstruction, dynamic control generation, and robust weighting into a unified cross-round coordination process. As a result, the proposed method transforms federated text classification from static model averaging into globally regulated closed-loop collaborative optimization under fairness and privacy constraints.

4 Experiments

This section evaluates the proposed method in terms of classification utility, group fairness, and privacy preservation under heterogeneous federated settings. We first describe the experimental setup, and then report the comparison results with representative federated baselines, followed by an ablation study and discussion.

4.1 Experimental Setup

To improve reproducibility, we summarize the experimental settings from two aspects: dataset preprocessing and implementation configuration. The same data partitioning strategy, local training protocol, privacy accounting procedure, and evaluation metrics were used for all compared methods unless otherwise specified.

Experiments were conducted on three public datasets, namely Twitter Sentiment, Bias in Bios, and Adult Income. Twitter Sentiment and Bias in Bios were used as text classification benchmarks, while Adult Income was included as an additional fairness-sensitive classification benchmark to evaluate the generality of the proposed fairness–privacy coordination mechanism beyond text-only scenarios. The dataset and preprocessing settings are summarized in Table 1.

The model was built on top of the corresponding input representation of each dataset. Twitter Sentiment and Adult Income used a two-layer MLP encoder, while Bias in Bios used a deeper MLP encoder because of its larger label space and higher-dimensional text representation. The model architecture, local optimization, federated training protocol, and closed-loop control settings are summarized in Table 2. The differential privacy configuration and privacy evaluation settings are reported in Table 3, and the final privacy accounting results are summarized in Table 4.

Table 1: Dataset and preprocessing settings.

Setting	Twitter Sentiment	Bias in Bios	Adult Income
Task	Binary sentiment classification	28-class occupation prediction	Binary income prediction
Sensitive attribute	Race-related dialect attribute, AAE vs. SAE	Gender, male vs. female	Sex, male vs. female
Input type	Text representation	Text representation	Tabular features
Input representation	Precomputed DeepMoji representation	BERT-base `[CLS]` representation	One-hot categorical features and normalized continuous features
Input dimension	2304	768	Dataset-dependent after one-hot encoding
Tokenizer	Not applicable	BERT-base tokenizer	Not applicable
Maximum sequence length	Not applicable	128 tokens	Not applicable
Data split	70% / 10% / 20%	Official train / validation / test split	60% / 20% / 20%
Evaluation metrics	Accuracy, TPR-gap, Leakage, MDL	Accuracy, GRMS, Leakage, MDL	Accuracy, TPR-gap, Leakage, MDL

The final privacy accounting results are summarized in Table 4. The reported values are computed using the RDP accountant with the order set $\mathcal{A} = \{2, 3, \ldots, 64\}$. The accumulated local DP steps are estimated under 60 communication rounds and a client participation ratio of 0.25; therefore, each client participates in approximately 15 rounds on average.

Following the evaluation protocol adopted throughout this study, Accuracy was used to measure task utility, TPR-gap or GRMS was used to quantify group fairness disparity, and Leakage together with MDL was used to assess privacy risk. Higher Accuracy and MDL indicate better performance, whereas lower TPR-gap/GRMS and Leakage indicate better fairness and privacy preservation.

Table 2: Federated training and model settings.

Category	Setting	Value
Model architecture	Twitter Sentiment encoder	MLP: $2304 \to 512 \to 256$ with ReLU activation
Model architecture	Bias in Bios encoder	MLP: $768 \to 512 \to 256$ with ReLU activation
Model architecture	Adult Income encoder	MLP: $d_{\text{in}} \to 128 \to 64$ with ReLU activation
Model architecture	Classifier head	Linear classifier for binary tasks; MLP prediction head for Bias in Bios
Model architecture	Dropout rate	0.10
Local optimization	Optimizer	Adam
Local optimization	Learning rate	1×10^{-3}
Local optimization	Batch size	64
Local optimization	Weight decay	1×10^{-5}
Local optimization	Local epochs	1
Local optimization	Low-rank adapter dimension r	64
Local optimization	Gradient clipping threshold	1.0
Federated training	Number of clients K	20
Federated training	Non-IID partition strategy	Dirichlet label distribution
Federated training	Dirichlet concentration parameter	$\alpha = 0.10$
Federated training	Client participation ratio	0.25
Federated training	Active clients per round	5
Federated training	Communication rounds	60
Federated training	Aggregation weight	$\omega_k^{(t)} = N_k / \sum_{j \in \mathcal{S}^{(t)}} N_j$
Closed-loop control	Initial fairness coefficient	0.20
Closed-loop control	Initial privacy coefficient	0.15
Closed-loop control	Global-state smoothing coefficient	0.90
Closed-loop control	Update-drift suppression coefficient	0.50
Closed-loop control	State-anomaly suppression coefficient	0.30
Closed-loop control	Participation-frequency penalty	0.20
Closed-loop control	Server-side aggregation clipping threshold	1.0

Table 3: Differential privacy and privacy evaluation settings.

Category	Setting	Value
Differential privacy	DP mechanism	DP-SGD
Differential privacy	Privacy accountant	RDP accountant
Differential privacy	Noise multiplier	1.1
Differential privacy	Sampling scheme	Poisson mini-batch sampling
Differential privacy	Effective sampling rate	Reported in Table 4
Differential privacy	RDP orders	$\mathcal{A} = \{2, 3, \ldots, 64\}$
Differential privacy	Target privacy budget	$(\epsilon, \delta) = (4.0, 10^{-5})$
Differential privacy	Statistics perturbation	Gaussian perturbation on uploaded fairness statistics
Differential privacy	Statistics noise multiplier	$\sigma_{\text{stat}} = 20.0$
Differential privacy	Final privacy budget	Reported in Table 4
Privacy evaluation	Leakage attacker input	Learned representation before the final task classifier
Privacy evaluation	Leakage attacker architecture	Two-layer MLP attacker with hidden dimension 128
Privacy evaluation	Attacker optimizer	Adam with learning rate 1×10^{-3} and batch size 128
Privacy evaluation	Attacker training	20 epochs with early stopping
Privacy evaluation	Leakage metric	Sensitive-attribute prediction accuracy of the attacker
Privacy evaluation	MDL protocol	Online coding / prequential evaluation
Privacy evaluation	MDL training fractions	1%, 2%, 5%, 10%, 20%, 50%, and 100%

Table 4: Final privacy accounting results.

Dataset	Noise mult.	Sampling rate	DP steps	Stat. noise mult.	δ	Final ϵ
Twitter Sentiment	1.1	0.0225	667	20.0	10^{-5}	3.996
Bias in Bios	1.1	0.0200	750	20.0	10^{-5}	3.754
Adult Income	1.1	0.0215	698	20.0	10^{-5}	3.910

Note: Noise mult. denotes the DP-SGD noise multiplier, Sampling rate denotes the effective mini-batch sampling rate, DP steps denotes the accumulated local DP-SGD optimization steps, and Stat. noise mult. denotes the noise multiplier used for perturbing uploaded fairness statistics.

4.2 Comparison with Federated Baselines

We compared the proposed method with four representative federated baselines: FedAvg (McMahan et al., 2017), FairFed (Ezzeldin et al., 2023), DPFed, and DPFairFed. FedAvg serves as the standard federated optimization baseline without

Table 5: Comparison of different federated methods in terms of utility, fairness, and privacy on the three datasets.

Dataset	Method	Accuracy	TPR-gap / GRMS	Leakage	MDL
Twitter Sentiment	FedAvg	72.44 ± 0.55	25.83 ± 1.15	86.16 ± 0.80	16.03 ± 0.62
Twitter Sentiment	FairFed	71.65 ± 0.61	12.45 ± 0.92	84.95 ± 0.92	16.65 ± 0.71
Twitter Sentiment	DPFed	68.91 ± 0.72	27.37 ± 1.42	63.27 ± 2.05	22.84 ± 0.96
Twitter Sentiment	DPFairFed	68.43 ± 0.68	10.60 ± 0.88	62.75 ± 1.86	23.16 ± 1.04
Twitter Sentiment	FedMOJO	**73.32** ± 0.42	**4.95** ± 0.36	**59.26** ± 1.32	**25.86** ± 0.58
Bias in Bios	FedAvg	**78.64** ± 0.34	16.78 ± 0.78	77.85 ± 1.61	176.20 ± 4.85
Bias in Bios	FairFed	77.72 ± 0.39	12.25 ± 0.66	76.93 ± 1.72	183.40 ± 5.18
Bias in Bios	DPFed	74.63 ± 0.47	18.19 ± 0.82	62.47 ± 1.08	229.50 ± 3.46
Bias in Bios	DPFairFed	74.11 ± 0.45	10.95 ± 0.61	61.80 ± 1.02	232.50 ± 3.25
Bias in Bios	FedMOJO	76.92 ± 0.31	**9.24** ± 0.43	**56.13** ± 0.92	**247.60** ± 2.10
Adult Income	FedAvg	**84.74** ± 0.28	16.50 ± 1.72	83.75 ± 0.95	19.65 ± 0.48
Adult Income	FairFed	84.10 ± 0.35	7.12 ± 0.86	82.84 ± 1.05	20.42 ± 0.52
Adult Income	DPFed	81.95 ± 0.44	8.85 ± 0.74	70.65 ± 1.22	27.45 ± 0.68
Adult Income	DPFairFed	81.47 ± 0.42	6.43 ± 0.58	70.10 ± 1.17	27.96 ± 0.61
Adult Income	FedMOJO	84.28 ± 0.30	**4.13** ± 0.39	**64.26** ± 0.88	**30.37** ± 0.44

Note: Values are reported as mean ± standard deviation. Bold values denote the best mean results. Higher Accuracy and MDL indicate better performance, whereas lower TPR-gap/GRMS and Leakage indicate better performance.

fairness or privacy constraints. FairFed augments federated training with fairness-aware aggregation. DPFed is implemented as a privacy-preserving FedAvg variant by applying local DP-SGD during client-side training (Abadi et al., 2016). DPFairFed further combines this private local updating scheme with fairness regularization, thereby representing a direct combination of federated optimization, differential privacy, and fairness constraints. The proposed method, denoted as FedMOJO, integrates low-rank representation learning, differentially private local training, global fairness-statistics fusion, low-rank structural consistency maintenance, and cross-round state-driven regulation into a unified federated closed-loop framework. All reported results are presented as mean ± standard deviation over three independent runs.

As reported in Table 5, the compared methods exhibit a clear utility–fairness–privacy trade-off under non-IID federated conditions. FedAvg usually maintains relatively high classification accuracy, but suffers from the worst fairness disparity and privacy leakage, indicating that optimizing only the main task in heterogeneous federated environments tends to preserve both group-bias signals and sensitive information. FairFed significantly reduces fairness disparity, yet its improvement in privacy protection remains limited, suggesting that fairness regularization alone is insufficient to effectively suppress sensitive-attribute recoverability. In contrast, DPFed and DPFairFed achieve noticeably lower leakage and higher MDL, but this comes at the cost of clear accuracy degradation, implying that simply applying local DP-SGD or directly stacking fairness and privacy constraints is still inadequate for stable three-objective coordination in federated learning.

Compared with all baselines, FedMOJO achieves a more favorable overall trade-off across the three datasets. On Twitter Sentiment, FedMOJO obtains the highest accuracy (73.32%) while reducing TPR-gap to 4.95, lowering leakage to 59.26, and increasing MDL to 25.86. On Bias in Bios, although FedMOJO is not the best on

every single metric, it provides the most balanced combination of utility, fairness, and privacy, with 76.92% accuracy, 9.24 GRMS, 56.13 leakage, and 247.60 MDL. On Adult Income, FedMOJO maintains a high accuracy of 84.28% while reducing TPR-gap to 4.13 and leakage to 64.26, together with an MDL of 30.37. Relative to DPFairFed, FedMOJO improves accuracy by 4.89, 2.81, and 2.81 percentage points on the three datasets, reduces TPR-gap/GRMS by 5.65, 1.71, and 2.30, lowers leakage by 3.49, 5.67, and 5.84, and increases MDL by 2.70, 15.10, and 2.41, respectively.

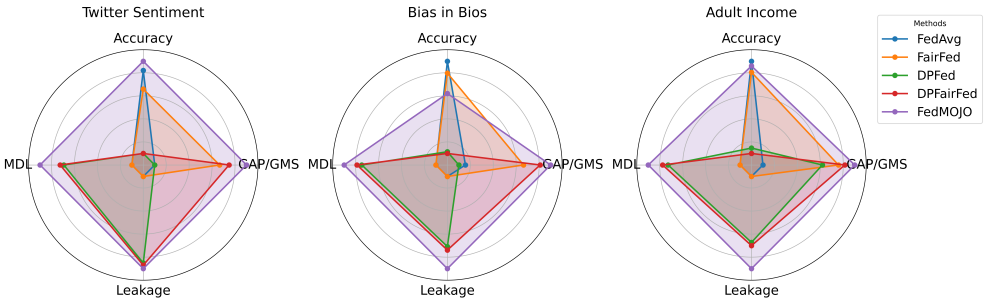


Fig. 2: Radar-chart comparison of different federated methods with respect to utility, fairness, and privacy on the three datasets.

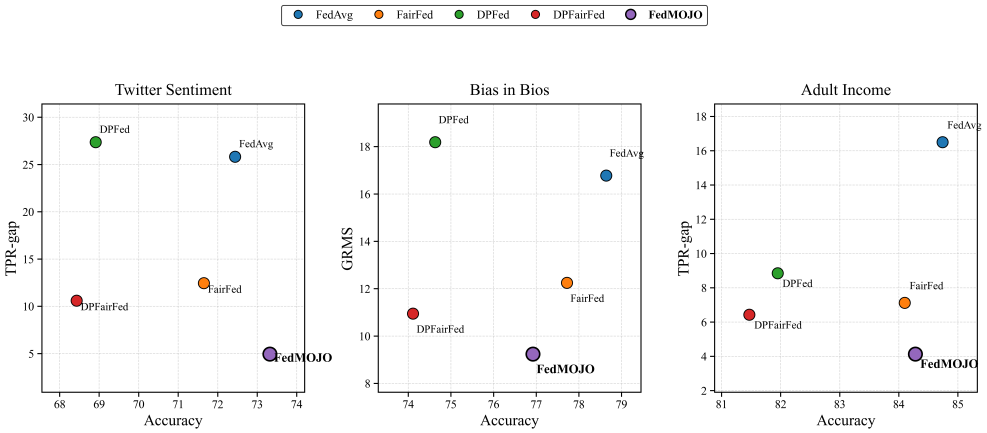


Fig. 3: Scatter plots of accuracy versus TPR-gap/GRMS for different federated methods on the three datasets.

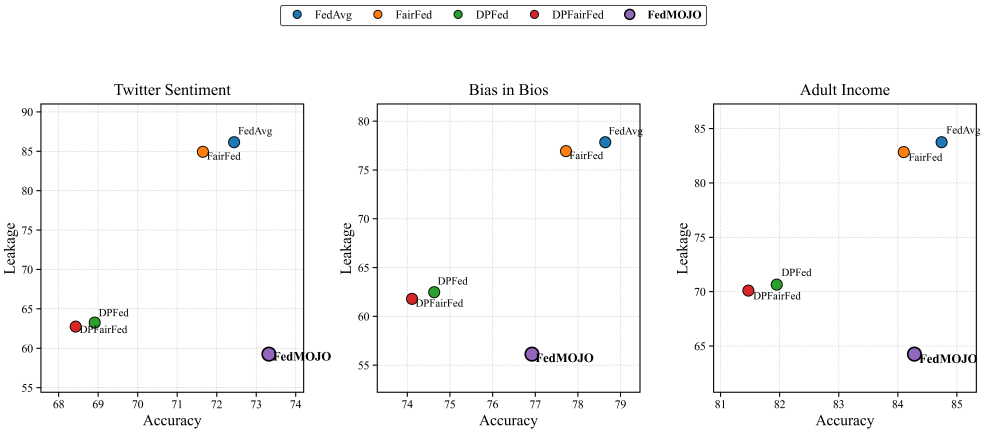


Fig. 4: Scatter plots of accuracy versus privacy leakage for different federated methods on the three datasets.

The visual comparisons in Fig. 2, Fig. 3, and Fig. 4 further support the quantitative findings. The radar chart shows that FedMOJO consistently yields a more balanced multi-metric profile than the baselines. The Accuracy–Fairness scatter plots indicate that FedMOJO is located closer to the favorable region with higher accuracy and lower fairness disparity, especially on the two binary tasks. Similarly, the Accuracy–Leakage scatter plots show that FedMOJO is also closer to the desirable high-utility and low-leakage region, suggesting that its superiority does not stem from merely increasing local perturbation strength, but from the joint effect of low-rank structured representation, global fairness-statistics fusion, and state-driven regulation across rounds.

4.3 Ablation Study

To further evaluate the contribution of each key component in the proposed federated framework, we performed an ablation study by removing the following modules one at a time while keeping the data partition, number of clients, communication rounds, local training steps, and privacy-accounting protocol unchanged: low-rank structured representation (w/o LR), global fairness-statistics fusion (w/o GFS), cross-round MOJO dynamic regulation (w/o MOJO), and alignment plus robust aggregation (w/o ARA). The complete model is denoted as FedMOJO.

Table 6: Ablation results of the proposed FedMOJO under different component removal settings on the three datasets.

Dataset	Method	Accuracy	TPR-gap / GRMS	Leakage	MDL
Twitter Sentiment	FedMOJO	73.32 ± 0.42	**4.95** ± 0.36	**59.26** ± 1.32	**25.86** ± 0.58
Twitter Sentiment	w/o LR	72.61 ± 0.47	6.93 ± 0.54	62.85 ± 1.48	24.16 ± 0.67
Twitter Sentiment	w/o GFS	**73.43** ± 0.39	9.83 ± 0.83	60.53 ± 1.24	25.23 ± 0.62
Twitter Sentiment	w/o MOJO	72.95 ± 0.48	6.22 ± 0.59	61.11 ± 1.35	24.95 ± 0.66
Twitter Sentiment	w/o ARA	72.71 ± 0.50	6.84 ± 0.62	60.32 ± 1.41	25.04 ± 0.70
Bias in Bios	FedMOJO	76.92 ± 0.31	**9.24** ± 0.43	**56.13** ± 0.92	**247.60** ± 2.10
Bias in Bios	w/o LR	76.22 ± 0.36	11.33 ± 0.58	59.41 ± 1.18	240.01 ± 2.64
Bias in Bios	w/o GFS	**77.15** ± 0.34	13.85 ± 0.76	56.91 ± 0.98	246.03 ± 2.22
Bias in Bios	w/o MOJO	76.63 ± 0.38	10.64 ± 0.55	58.22 ± 1.14	241.54 ± 2.78
Bias in Bios	w/o ARA	76.17 ± 0.40	10.94 ± 0.61	57.43 ± 1.21	244.22 ± 2.51
Adult Income	FedMOJO	84.28 ± 0.30	**4.13** ± 0.39	**64.26** ± 0.88	**30.37** ± 0.44
Adult Income	w/o LR	83.62 ± 0.35	5.16 ± 0.49	67.52 ± 1.03	28.90 ± 0.55
Adult Income	w/o GFS	**84.34** ± 0.31	6.94 ± 0.66	64.75 ± 0.91	30.03 ± 0.49
Adult Income	w/o MOJO	83.46 ± 0.37	5.87 ± 0.54	65.08 ± 0.98	29.86 ± 0.52
Adult Income	w/o ARA	83.51 ± 0.38	5.60 ± 0.52	65.21 ± 1.04	29.64 ± 0.56

Note: Values are reported as mean ± standard deviation. Bold values denote the best mean results. Higher Accuracy and MDL indicate better performance, whereas lower TPR-gap/GRMS and Leakage indicate better performance.

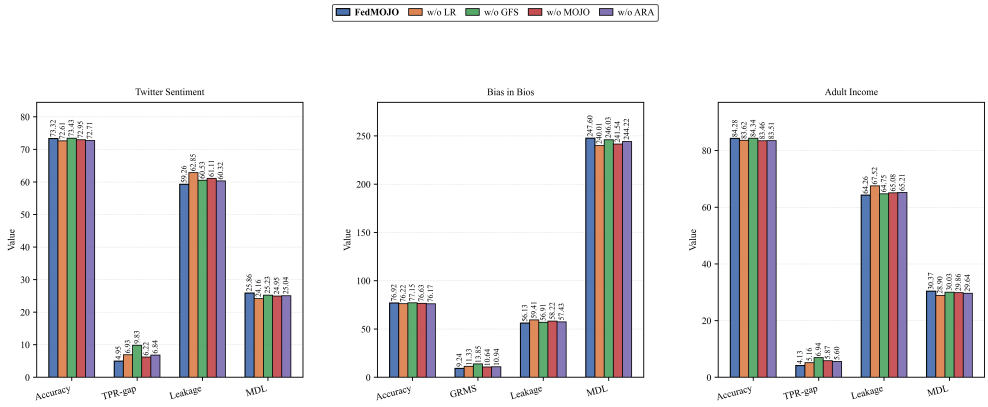


Fig. 5: Bar-chart visualization of the ablation results of the proposed FedMOJO on the three datasets.

The ablation results in Table 6 and Fig. 5 show that each module contributes to the final performance from a different perspective. Removing global fairness-statistics fusion causes the most obvious increase in fairness disparity, e.g., TPR-gap/GRMS rises from 4.95 to 9.83 on Twitter Sentiment, from 9.24 to 13.85 on Bias in Bios, and from 4.13 to 6.94 on Adult Income. This result indicates that, under heterogeneous federated distributions, local fairness constraints alone are insufficient to establish a globally consistent correction direction, and local fairness improvement does not necessarily translate into global fairness improvement.

Removing the low-rank structured representation module leads to simultaneous degradation in accuracy and privacy-related metrics, which suggests that, in the presence of DP noise, high-dimensional parameter updates are more likely to lose useful discriminative information and also become less stable. Removing the MOJO dynamic regulation module yields moderate but consistent degradation on all three datasets, showing that state-driven regulation is necessary to coordinate cumulative privacy consumption, varying client participation, and fluctuating fairness/privacy risk across rounds. Finally, removing alignment and robust aggregation harms both accuracy and fairness under non-IID conditions, confirming that low-rank consistency maintenance and robust aggregation are essential to suppress local drift and stabilize federated optimization. Overall, the gain of FedMOJO does not come from a single module, but from the collaboration of low-rank representation learning, global fairness-state modeling, cross-round dynamic regulation, and robust federated aggregation within a unified closed-loop framework.

4.4 Discussion

Overall, the experimental results indicate that the proposed FedMOJO achieves a more robust balance among classification utility, group fairness, and privacy preservation than representative federated baselines. Both the comparison experiments and the ablation study confirm that the advantage of FedMOJO in balancing fairness and privacy does not stem from a single isolated design, but from the joint effect of low-rank structured local learning, global fairness-state fusion, cross-round dynamic regulation, and robust federated aggregation.

5 Limitations and Future Work

The current study mainly focuses on federated classification tasks with fairness and privacy constraints. Although two datasets are based on textual representations, Adult Income is included as an additional tabular benchmark to examine whether the proposed coordination mechanism can generalize beyond text-only inputs. Nevertheless, the method has not yet been validated on multimodal data or large-scale generative language models. Future work will extend the framework to multimodal federated learning scenarios and investigate more complex sensitive-attribute structures, including multiple and intersectional sensitive attributes.

6 Conclusion

This paper presented FedMOJO, a closed-loop collaborative optimization framework for fair and privacy-preserving federated text classification under non-IID data. Different from conventional federated methods that mainly rely on local fairness regularization, local differential privacy, or direct combinations of the two, the proposed approach formulates federated fair and privacy-preserving learning as a unified client–server coordination problem. On the client side, low-rank structured representation learning, differentially private local training, and fairness–privacy collaborative local optimization are jointly performed. On the server side, global aggregation is enhanced

through low-rank structural consistency maintenance, global fairness-statistics fusion, state-driven control generation, and robust aggregation under heterogeneous federated conditions. Through this design, the proposed method transforms federated text classification from static round-wise model averaging into globally regulated closed-loop collaborative optimization.

Experimental results on Twitter Sentiment, Bias in Bios, and Adult Income demonstrate that the proposed method achieves a more robust overall trade-off among classification utility, group fairness, and privacy preservation than representative federated baselines. In particular, compared with DPFairFed, FedMOJO improves Accuracy by 4.89, 2.81, and 2.81 percentage points, reduces TPR-gap/GRMS by 5.65, 1.71, and 2.30, lowers Leakage by 3.49, 5.67, and 5.84, and increases MDL by 2.70, 15.10, and 2.41, respectively. These results indicate that local fairness constraints and local differential privacy alone are insufficient for stable three-objective coordination in heterogeneous federated environments, whereas the proposed closed-loop mechanism can more effectively coordinate utility, fairness, and privacy without sharing raw data.

The ablation study further confirms that the performance gain of FedMOJO does not come from a single isolated design choice. Instead, it arises from the collaboration of several essential components, including low-rank structured local learning, global fairness-state fusion, cross-round MOJO dynamic regulation, and alignment-aware robust aggregation. Removing any one of these modules leads to noticeable degradation in at least one of the three aspects, namely utility, fairness, and privacy, which validates the necessity of jointly modeling local optimization and global coordination in federated fair and privacy-preserving text classification.

Although the proposed method has shown promising results, several directions remain worthy of further investigation. First, the current framework can be extended to larger pre-trained language models and more complex cross-silo federated NLP scenarios. Second, more refined privacy-accounting and fairness-auditing mechanisms may further improve the coordination between cumulative privacy budgets and global fairness control. Third, future work may explore more communication-efficient client–server interaction schemes and adaptive participation strategies for highly heterogeneous real-world environments. Overall, this work provides a systematic step toward robust federated text classification that jointly considers predictive performance, group fairness, and privacy preservation.

Funding

This work was supported by the Heilongjiang Provincial Natural Science Foundation of China (Grant No. ZL2025F005) and the National Natural Science Foundation of China (Grant No. 62172123).

Ethics declaration

Not applicable.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308–318).

Ezzeldin, Y.H., Yan, S., He, C., Ferrara, E., Avestimehr, A.S. (2023). FairFed: Enabling group fairness in federated learning. *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, pp. 7494–7502).

Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., ... Ahmed, N.K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179, https://doi.org/10.1162/coli_a_00524

Khan, Y., Sánchez, D., Domingo-Ferrer, J. (2024). Federated learning-based natural language processing: A systematic literature review. *Artificial Intelligence Review*, *57*(12), 320, https://doi.org/10.1007/s10462-024-10970-5

Kim, D., Oh, K., Lee, Y., Woo, H. (2025). Overview of fair federated learning for fairness and privacy preservation. *Expert Systems with Applications*, *293*, 128568, https://doi.org/10.1016/j.eswa.2025.128568

Leteno, T., Perrot, M., Laclau, C., Gourru, A., Gravier, C. (2025). Fair text classification via transferable representations. *Journal of Machine Learning Research*, *26*(239), 1–47, Retrieved from https://www.jmlr.org/papers/v26/25-0485.html

Liu, B., Lv, N., Guo, Y., Li, Y. (2024). Recent advances on federated learning: A systematic survey. *Neurocomputing*, *597*, 128019, https://doi.org/10.1016/j.neucom.2024.128019

Liu, X.-Y., Zhu, R., Zha, D., Gao, J., Zhong, S., White, M., Qiu, M. (2025). Differentially private low-rank adaptation of large language model using federated learning. *ACM Transactions on Management Information Systems*, *16*(2), 1–24, https://doi.org/10.1145/3682068

Lu, Z., Pan, H., Dai, Y., Si, X., Zhang, Y. (2024). Federated learning with non-IID data: A survey. *IEEE Internet of Things Journal*, *11*(11), 19188–19209, https://doi.org/10.1109/JIOT.2024.3376548

Mao, Y.-C., Shen, L.-J., Wu, J., et al. (2024). Federated dynamic client selection for fairness guarantee in heterogeneous edge computing. *Journal of Computer Science and Technology*, *39*(1), 139–158, https://doi.org/10.1007/s11390-023-2972-9

McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th international conference on artificial intelligence and statistics* (Vol. 54, pp. 1273–1282).

Mukhtiar, N., Mahmood, A., Sheng, Q.Z. (2025). Fairness in federated learning: Trends, challenges, and opportunities. *Advanced Intelligent Systems*, *7*(6), 2400836, https://doi.org/10.1002/aisy.202400836

Rafi, T.H., Noor, F.A., Hussain, T., Chae, D.-K. (2024). Fairness and privacy preserving in federated learning: A survey. *Information Fusion*, *105*, 102198, https://doi.org/10.1016/j.inffus.2023.102198

Shahriar, S., Dara, R., Akalu, R. (2025). A comprehensive review of current trends, challenges, and opportunities in text data privacy. *Computers & Security*, *151*, 104358, https://doi.org/10.1016/j.cose.2025.104358

Taha, K., Yoo, P.D., Yeun, C., Homouz, D., Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review*, *54*, 100664, https://doi.org/10.1016/j.cosrev.2024.100664

Wang, C., Xia, H., Xu, S., Chi, H., Zhang, R., Hu, C. (2024). FedBnR: Mitigating federated learning non-IID problem by breaking the skewed task and reconstructing representation. *Future Generation Computer Systems*, *153*, 1–11, https://doi.org/10.1016/j.future.2023.11.020

Wang, Y., Wang, Q., Zhao, L., Wang, C. (2023). Differential privacy in deep learning: Privacy and beyond. *Future Generation Computer Systems*, *148*, 408–424, https://doi.org/10.1016/j.future.2023.06.010

Yang, Y., Lin, M., Zhao, H., Peng, Y., Huang, F., Lu, Z. (2024). A survey of recent methods for addressing AI fairness and bias in biomedicine. *Journal of Biomedical Informatics*, *154*, 104646, https://doi.org/10.1016/j.jbi.2024.104646

Zhang, J., Li, Y., Wu, D., Zhao, Y., Palaiahnakote, S. (2025). SFFL: Self-aware fairness federated learning framework for heterogeneous data distributions. *Expert Systems with Applications*, *269*, 126418, https://doi.org/10.1016/j.eswa.2025.126418