Lightweight Cross-Lingual Federated Prompt Tuning for Low-Resource Languages

Ubaid Azam[1], **Imran Razzak**[2], **Shoaib Jameel**[1]
[1]University of Southampton, Southampton, United Kingdom
[2]Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
u.azam@soton.ac.uk, imran.razzak@mbzuai.ac.ae, M.S.Jameel@southampton.ac.uk

Abstract

Multilingual NLP faces challenges of data heterogeneity, privacy, and limited computational resources, especially for low-resource languages. Centralised methods risk privacy breaches, while federated learning struggles with communication overhead and poor cross-lingual generalisation. We propose FLiP (Federated Lightweight Prompt-tuning), a privacy-preserving, resource-efficient, generalisable framework integrating prompt-based learning with federated optimisation. FLiP eliminates communication overhead, reduces trainable parameters to 16%, and cuts GPU memory use by 90%. Experiments show superior generalisation and efficiency under both IID and Non-IID settings, establishing FLiP as a scalable, privacy-aware solution for multilingual NLP, particularly in low-resource and indigenous language contexts.

Keywords: Federated Learning, Prompt learning, Heterogeneity, Generalisation

1. Introduction

Contemporary artificial intelligence has created unprecedented opportunities for natural language processing across diverse linguistic communities (Pakray et al., 2025), with multilingual language technologies and cross-lingual understanding systems representing impactful areas of advancement (Judijanto and Vandika, 2025; Piperno et al., 2025). Sophisticated neural architectures now handle intricate tasks including multilingual sentiment analysis (Azam et al., 2025), automatic language identification (Peng et al., 2024), linguistic evolution modeling (Ponti et al., 2019), cross-lingual document understanding (Wang, 2024), and anomaly detection in multilingual datasets (Foufa, 2020).

Despite progress, major challenges limit the effective integration of machine learning for multilingual and low-resource languages. The primary constraint is scarcity of large, high-quality datasets due to limited digitisation and restricted data sharing (Magueresse et al., 2020). Achieving robust generalisation across diverse languages, dialects, and scripts is difficult given variations in orthographies, grammatical systems, and cultural contexts (Maaz et al., 2024). These challenges intensify for minority and indigenous languages, where resources are fragmented and culturally sensitive (Zhao et al., 2024). Cultural protocols, ethical considerations, and community ownership further limit access to linguistic materials (Lim et al., 2020), while variations in documentation practices and dialectal differences create additional barriers (Yang et al., 2024).

Federated Learning (FL) offers a promising solution by enabling communities to train models locally with native language resources (Nagy et al., 2023). Instead of sharing raw texts, only model parameters are transmitted to a coordinating server for integration into a global model. This preserves cultural privacy, sovereignty, and community control (Singh and Thakur, 2024). However, FL faces challenges including high communication costs, computational demands, heterogeneous data distributions, and privacy risks from parameter inference attacks.

Communication and computational overhead represent critical limitations. Continuous parameter exchange creates substantial latency, particularly problematic for time-critical applications like hate speech detection (Rauniyar et al., 2023). Each participant requires independent local training capabilities, often without high-performance computing resources (Wen et al., 2023). This challenge is pronounced in resource-constrained environments, remote linguistic communities or under-funded institutions with insufficient technological infrastructure and connectivity (Imteaj et al., 2021).

Managing linguistic heterogeneity while preserving privacy presents another significant obstacle (Chen et al., 2024; Li et al., 2025). Multilingual datasets are inherently non-IID, varying in linguistic structure, orthographic norms, data collection methods, and cultural contexts. Disparities arise from language contact, code-switching, sociolinguistic influences, language vitality differences, orthographic variations, dialectal diversity, and cultural taboos (Yang et al., 2024). Such imbalances hinder generalisation and reduce accuracy for minority languages (Zhang et al., 2022). Additionally, parameters may leak sensitive cultural or linguistic information (Li et al., 2023). While differential privacy and secure aggregation mitigate risks (El Ouadrhiri and Abdelhadi, 2022; Fereidooni et al., 2021), they often sacrifice efficiency and accuracy.

To address these challenges, we propose Feder-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3304–3316
11–16 May 2026.

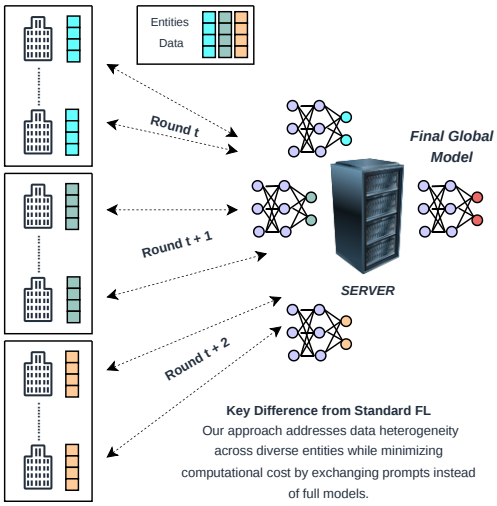

Figure 1: Federated Lightweight Prompt-tuning (FLiP) Architecture.

ated Lightweight Prompt-tuning (FLiP), a framework built on prompt-driven federated learning (Zhao et al., 2023) to enhance computational efficiency, privacy, and performance. By fine-tuning only limited prompt parameters while keeping the core model frozen (Guo et al., 2023), FLiP minimises computation and communication costs while strengthening privacy through compact prompt updates. It further integrates a linguistically-aware design to handle multilingual data heterogeneity, reduce resource consumption, and support deployment in constrained environments. This adaptive architecture enables communities with limited computational capacity to collaborate effectively with better-resourced participants.

FLiP's main contributions are fourfold: (1) it addresses linguistic data heterogeneity, including cultural variation, label imbalance, code-switching, and typological differences, by federating heterogeneous data sources to generalise across languages and scripts without centralization; (2) it cuts communication overhead by transmitting only 16% of trainable parameters, achieving nearly 100% reduction compared to traditional methods; (3) it enhances privacy by sharing lightweight prompts instead of raw data, protecting sensitive linguistic information; and (4) it lowers local computational demands by 90% in GPU memory through prompt-based adaptation, enabling efficient, low-resource training without full model retraining.

2. Related Work

This section reviews prior research in three key areas that inform our proposed FLiP framework.

2.1. Federated Learning for Multilingual Languages

Federated Learning (FL), introduced by McMahan et al. (McMahan et al., 2017), enables collaborative model training without centralising data, allowing entities to retain local datasets while contributing to a shared global model. In each communication round, participating entities perform local gradient-based optimisation on their private data and transmit only the resulting parameter updates to a central coordinator, which aggregates these updates, typically via weighted averaging, to refine the global model. This decentralised approach ensures privacy and supports ethical data practices, particularly important for minority and indigenous languages where cultural ownership and community consent are essential (Voigt and Von dem Bussche, 2017).

FL has been extensively explored for multilingual tasks (Gamal et al., 2023; Wang et al., 2022). Khalil et al. (Khalil et al., 2024) applied FL for multilingual depression detection, while Riedel et al. (Riedel et al., 2023) evaluated FL algorithms for multilingual protest news detection, both achieving strong results under privacy-preserving settings. Singh et al. (Singh and Thakur, 2024) proposed a fair selection FL approach for multilingual hate speech detection in low-resource Indian languages, demonstrating superior performance over baselines.

However, FL faces persistent challenges, including high communication overhead (Asad et al., 2023), security vulnerabilities (Mothukuri et al., 2021), and statistical heterogeneity among entities (Jafarigol et al., 2024). These issues, coupled with fairness and scalability constraints, pose barriers to real-world deployment, particularly in resource-

limited multilingual contexts where efficient, secure, and equitable learning remains critical.

2.2. Prompt Learning for Multilingual Languages

Prompt learning has emerged as an efficient approach for adapting large pre-trained language models (LLMs) to specific tasks with minimal parameter updates (Sahoo et al., 2024). Rather than updating all model parameters, prompt learning prepends a small set of learnable tokens, known as soft prompts, to the input sequence, optimising only these prompt embeddings while keeping the pre-trained backbone frozen, thereby considerably reducing computational and memory requirements. Crafting structured or learned prompts enables knowledge transfer across diverse linguistic and contextual settings without extensive fine-tuning. This paradigm has proven valuable for multilingual applications where data scarcity and linguistic diversity pose major challenges (Wang et al., 2024; Ullah et al., 2025; Vatsal et al., 2025).

Feng et al. (Feng et al., 2024) proposed a zero-shot cross-lingual classification framework using language-agnostic prompts to enable semantic transfer. Qiu et al. (Qiu et al., 2024) demonstrated how multilingual prompts facilitate cross-lingual transfer by disentangling task knowledge from language-specific features. Zhao et al. (Zhao et al., 2024) extended this concept through Federated Prompt Tuning, introducing parameter-efficient fine-tuning for multilingual environments.

Despite these advancements, prompt learning faces notable challenges. Studies by Geroimenko et al. (Geroimenko, 2025) and Gupta et al. (Gupta et al., 2025) highlight persistent issues, including bias, limited generalisability, contextual drift, and domain-specific knowledge gaps. Addressing these constraints is essential for ensuring fair, reliable, and ethical use of prompt learning, particularly in low-resource multilingual contexts where maintaining cultural and linguistic integrity is critical.

2.3. Cross-Lingual Generalisation

Achieving robust generalisation in multilingual settings remains challenging due to inherent linguistic diversity and variability in data distribution, emphasising the need for domain-agnostic, language-flexible frameworks capable of adapting across multiple languages and cultural contexts. Manias et al. (Manias et al., 2023) conducted a comparative study exploring multilingual model effectiveness for classification tasks, while Cui et al. (Cui et al., 2022) examined compositional generalisation across languages. Kim et al. (Kim et al., 2025) proposed probabilistic content masking and language-aware batching to improve generalisation across languages and domains. The XTREME benchmark by Hu et al. (Hu et al., 2020) revealed substantial performance gaps in cross-lingual transfer, underscoring persistent limitations of existing multilingual systems. These findings highlight the need for more robust, adaptive, and fair approaches to enhance generalisability in multilingual AI.

In this context, our proposed FLiP framework integrates the efficiency of prompt learning with the privacy-preserving and collaborative strengths of federated learning, specifically addressing the non-IID and heterogeneous nature of multilingual data. FLiP delivers a scalable, low-latency, and computationally efficient solution tailored to underrepresented and low-resource linguistic environments. To our knowledge, this is among the first approaches to jointly tackle data heterogeneity, privacy preservation, and efficiency in multilingual contexts, offering a step toward truly inclusive and generalisable multilingual AI.

3. Our Novel FLiP Framework

We develop a novel computational model Federated Lightweight Prompt-tuning (FLiP), designed to achieve robust generalisation while managing linguistic heterogeneity in multilingual environments. Considering the sensitivity and cultural ownership of indigenous and minority language data, we leverage FL to maintain data decentralisation across linguistic communities, eliminating raw data exchange. Unlike traditional FL approaches, FLiP introduces a novel prompt-based learning within a federated architecture, substantially reducing training overhead while improving model adaptability across diverse linguistic datasets. As depicted in Figure 1, this approach minimises local computational demands, making it suitable for deployment in low-resource language communities and edge devices. By reducing trainable parameters by over 84%, FLiP dramatically lowers computational costs while preserving performance, enhancing both scalability and practical feasibility for real-world multilingual applications.

3.1. Technical Framework

The FLiP framework (Algorithm 1) operates through two sequential stages: linguistic community organisation and collaborative prompt optimisation, achieving parameter-efficient learning in distributed settings while addressing cross-lingual heterogeneity and bandwidth limitations.

3.1.1. Linguistic Community Organisation

The initial stage organises participating entities based on linguistic characteristics of their data. Each entity e_j provides a representative subset

r_j from which we derive an embedding v_j. Linguistic affinity between entities is measured using a commonly used cosine similarity measure: affinity$(e_j, e_k) = \frac{v_j \cdot v_k}{\|v_j\| \|v_k\|}$. Entities with affinity exceeding threshold τ form communities $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_h\}$. The threshold τ was determined empirically, balancing within-community consistency and community count. High thresholds led to excessive fragmentation with poor training efficiency, while low thresholds created heterogeneous communities unsuitable for collaborative learning.

This organisation strategy mitigates linguistic diversity challenges by facilitating cooperation among linguistically similar entities. To prevent imbalanced community formation, we enforce constraints requiring 5–15 entities per community, ensuring adequate sample representation, mitigating overfitting risks, constraining computational requirements, and balancing community-level specialisation with system-wide scalability. Proper community sizing and affinity thresholds minimise negative knowledge transfer while improving global model resilience to linguistic variance.

3.1.2. Soft Prompt Mechanism

We introduce a soft prompt learning mechanism that enhances adaptation capabilities while maintaining efficiency. Soft prompts are parameterised as learnable embeddings $\Theta \in \mathbb{R}^{\ell \times d}$, where ℓ indicates the sequence length of prompt tokens and d denotes the embedding space dimensionality. Our design uses vocabulary-informed initialisation that leverages semantic knowledge from pre-trained multilingual encoders, transforming prompt optimisation into a meta-learning framework supporting adaptive prefix conditioning with minimal parameter overhead. By parameterising prompts as trainable embeddings, we enable flexible task adaptation while keeping pre-trained model weights frozen.

3.1.3. FLiP Training Protocol

FLiP executes over R communication cycles, maintaining a universal prompt Θ_{univ} and community-specific prompts $\Theta_{\mathcal{C}}$ for each community $\mathcal{C}$. During initialization, Θ_{univ} is randomly initialized, with each $\Theta_{\mathcal{C}}$ initialized to match it. Entity contribution weights ω_j are computed based on dataset characteristics, ensuring equitable aggregation.

During cycle r, community $\mathcal{C}_r$ is activated and K entities are sampled, forming subset $\mathcal{S}_r$. Each entity e_j initialises its local prompt from the community prompt: $\Theta_{j,r} = \Theta_{\mathcal{C}_r}$ and executes I optimisation iterations. The objective function incorporates proximity regularization weight, ensuring coherence between local and community prompts:

$$\mathcal{L}_j = \mathcal{L}(\Theta_{j,r}, \mathcal{B}_j) + \frac{\lambda}{2}\|\Theta_{j,r} - \Theta_{\mathcal{C}_r}\|^2, \qquad (1)$$

Algorithm 1 Our FLiP Model

Require:
Entities $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$
Communities $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_h\}$
Local corpus $\mathcal{D}_j$ for entity e_j
Step size γ, Total cycles R, Sample size K
Affinity threshold τ, Regularization weight λ
Community prompt $\Theta_{\mathcal{C}}$, Universal prompt Θ_{univ}
procedure STAGE 1: COMMUNITY ORGANIZATION
  Input: Representative subsets r_j from entity e_j
  Output: Communities $\mathcal{C}$
  for each entity $e_j \in \mathcal{E}$ **do**
    Derive embedding v_j from subset r_j
  for each pair (e_j, e_k) **do**
    Calculate: affinity$(e_j, e_k) = \frac{v_j \cdot v_k}{\|v_j\| \|v_k\|}$
  Construct $\mathcal{C}$ where $\forall e_j, e_k \in \mathcal{C}_i$: affinity$(e_j, e_k) \geq \tau$
  return $\mathcal{C}$
procedure STAGE 2: COLLABORATIVE OPTIMIZATION
  Input: $\mathcal{C}, \{\mathcal{D}_j\}, R, \lambda$
  Output: Optimized universal prompt
  Initialize Θ_{univ} randomly
  for each $\mathcal{C}_i \in \mathcal{C}$ **do**
    $\Theta_{\mathcal{C}_i} \leftarrow \Theta_{\text{univ}}$
  Compute contribution weights ω_j for entities
  for cycle $r = 1$ to R **do**
    $\mathcal{C}_r \leftarrow$ SelectCommunity$(\mathcal{C}, r)$
    $\mathcal{S}_r \leftarrow$ SampleEntities$(\mathcal{C}_r, K)$
  Distributed optimization (parallel):
    for each entity $e_j \in \mathcal{S}_r$ **do**
      Initialize: $\Theta_{j,r} \leftarrow \Theta_{\mathcal{C}_r}$
      for iteration = 1 to I **do**
        Draw minibatch $\mathcal{B}_j$ from $\mathcal{D}_j$
        Compute: $\mathcal{L}_j = \mathcal{L}(\Theta_{j,r}, \mathcal{B}_j) + \frac{\lambda}{2}\|\Theta_{j,r} - \Theta_{\mathcal{C}_r}\|^2$
        Gradient step: $\Theta_{j,r} \leftarrow \Theta_{j,r} - \gamma \nabla \mathcal{L}_j$
      Transmit $\Delta\Theta_{j,r} = \Theta_{j,r} - \Theta_{\mathcal{C}_r}$ to coordinator
  Coordinator aggregation:
    $\Delta\Theta_r = \sum_{e_j \in \mathcal{S}_r} \omega_j \Delta\Theta_{j,r}$, where $\sum \omega_j = 1$
    Refine community prompt: $\Theta_{\mathcal{C}_r} \leftarrow \Theta_{\mathcal{C}_r} + \Delta\Theta_r$
    Refine universal prompt: $\Theta_{\text{univ}} \leftarrow \rho\Theta_{\text{univ}} + (1-\rho)\Theta_{\mathcal{C}_r}$
  return Θ_{univ}

where $\mathcal{L}(\Theta_{j,r}, \mathcal{B}_j)$ is the task-specific loss that optimizes the prompt for entity e_j's local data, $\mathcal{B}_j$ denotes a minibatch from entity e_j's corpus $\mathcal{D}_j$, and the regularization term $\frac{\lambda}{2}\|\Theta_{j,r} - \Theta_{\mathcal{C}_r}\|^2$ constrains the local prompt to remain proximate to the community prompt, preventing catastrophic forgetting and ensuring stable convergence. The hyperparameter λ is a regularisation weight that controls the trade-off between local adaptation and community-level consistency. After optimisation, each entity computes prompt adjustments $\Delta\Theta_{j,r}$ relative to the community baseline.

The coordinator aggregates entity adjustments through a weighted combination:

$$\Delta\Theta_r = \sum_{e_j \in \mathcal{S}_r} \omega_j \, \Delta\Theta_{j,r}, \quad \text{where} \quad \sum \omega_j = 1. \qquad (2)$$

The community prompt is refined using these aggregated adjustments: $\Theta_{\mathcal{C}_r} \leftarrow \Theta_{\mathcal{C}_r} + \Delta\Theta_r$, while the universal prompt undergoes exponential moving average update for stable convergence: $\Theta_{\text{univ}} \leftarrow \rho\,\Theta_{\text{univ}} + (1-\rho)\Theta_{\mathcal{C}_r}$, where ρ is a momentum parameter that balances historical and current information.

The protocol incorporates several architectural features promoting efficiency and robustness: multi-tier prompt organisation (universal, community, and entity-level), facilitating adaptation across linguistic distributions; proximity regularization weight stabilising distributed optimisation; weighted aggregation accommodating heterogeneous entity contributions; and exponential moving average updates ensuring convergence stability. This design achieves equilibrium between universal generalisation and community-specific adaptation while preserving communication efficiency through hierarchical organisation. The community organisation ensures systematic rotation through all communities, providing equal representation to each linguistic distribution and preventing bias.

Our protocol achieves effectiveness through systematic training organisation. Each cycle involves a randomly sampled entity subset from a designated community performing distributed optimisation, with successive cycles activating different communities with fresh entity samples (Figure 1). This embodies domain adaptation principles, where each community represents a distinct linguistic domain. After each cycle, distributed prompt adjustments aggregate to refine the universal prompt, which subsequently propagates to entities in different linguistic domains, paralleling transfer learning mechanisms where linguistic knowledge acquired from one domain transfers and specialises across varied datasets. Through systematic cycling, the model progressively strengthens generalisability while maintaining equilibrium across all entity distributions. Crucially, only lightweight prompt parameters are exchanged, substantially reducing computational overhead and enabling deployment on resource-constrained devices.

4. Experiments and Results

Our experiments were designed to (1) compare the proposed framework with established benchmarks, including state-of-the-art language models, (2) quantitatively assess improvements in computational efficiency and resource utilisation, and (3) evaluate the effectiveness of entity grouping through performance on text classification.

4.1. Experimental Setup

We addressed hate speech detection in the multilingual domain owing to its heterogeneous and culturally sensitive nature. This heterogeneity arises from cross-lingual variations, cultural interpretations of offensiveness, and demographic diversity across online communities, complicating label consistency and semantic alignment while introducing distributional shifts that limit generalisation (Al-Badani et al., 2025). Moreover, hate speech detection is complex, constrained by limited language resources, uneven data, and privacy considerations, especially in low-resource or underrepresented linguistic contexts (Kumar et al., 2025).

Datasets: We utilised three publicly available multilingual hate speech datasets, each representing a distinct low-resource language: the Urdu Offensive Dataset (UOD) (Akhter et al., 2020), the Pashto Offensive Language Dataset (POLD) (Haq et al., 2023), and the Bengali Hate Speech Dataset (BHD) (Romim et al., 2021). The UOD comprises 2,170 instances annotated for offensive and non-offensive content in Urdu. The POLD contains 34,400 tweets in Pashto categorised as offensive or non-offensive, representing one of the most extensive resources for this low-resource language. The BHD consists of 30,000 instances in Bengali labelled for hate and non-hate speech, covering diverse linguistic and cultural contexts. These datasets capture the cross-lingual, cultural, and contextual diversity in multilingual hate speech, providing a robust benchmark for evaluating model generalisation across underrepresented languages.

Comparative Models: To assess our framework's effectiveness, we conducted experiments across three setups for multilingual hate speech detection. To benchmark FLiP against state-of-the-art federated learning strategies, we compared it with FedAvg (baseline) (McMahan et al., 2017), FedProx (which incorporates heterogeneity-aware proximal regularization) (Li et al., 2020), FedOpt (which features server-side adaptive optimisation) (Reddi et al., 2020), and FedPrompt (which applies prompt tuning in federated learning) (Zhao et al., 2023). **1. Traditional Centralised Training (No FL):** All multilingual datasets were combined into a single centralised dataset, and the model was trained using standard, non-federated methods. This baseline serves to evaluate the effectiveness of our FL-based approach by comparing its performance against fully centralised training. Experiments were carried out under both balanced and imbalanced conditions, with class distributions set at a 1:3 ratio. **2. Federated Learning with IID and Non-IID Data:** The combined dataset was distributed among multiple entities, each representing a distinct linguistic source. We evaluated two conditions: (i) IID, where data was uniformly sampled across entities to maintain class balance, and (ii) Non-IID, where entities received skewed class distributions reflecting realistic linguistic disparities (1:3 ratio). **3. FLiP with IID and Non-IID Data:** We implemented our proposed Federated Lightweight Prompt-tuning (FLiP) framework (see Algorithm 1) under the same IID and Non-IID conditions as baseline FL setups.

To more accurately reflect real-world conditions in both IID and non-IID federated learning settings,

Model	FedAvg		FedProx		FedOpt		FedPrompt		FLiP	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
IID Results										
BERT-M	0.865	0.857	0.870	0.861	0.864	0.855	0.849	0.848	**0.881**	0.872
XLM-R	0.863	0.856	0.866	0.860	0.863	0.851	0.841	0.845	0.876	**0.875**
DistilBERT	0.849	0.841	0.850	0.841	0.841	0.840	0.825	0.823	0.852	0.844
Mini-LM	0.851	0.851	0.856	0.853	0.850	0.849	0.833	0.832	0.858	0.855
Non-IID Results										
BERT-M	0.851	0.853	0.867	0.863	0.852	0.851	0.834	0.831	**0.876**	**0.868**
XLM-R	0.850	0.851	0.855	0.854	0.851	0.850	0.831	0.830	0.863	0.861
DistilBERT	0.833	0.831	0.839	0.837	0.835	0.835	0.817	0.810	0.841	0.839
Mini-LM	0.838	0.838	0.842	0.840	0.839	0.836	0.827	0.822	0.846	0.843

Table 1: Comparison of FLiP with strong baselines.

we assumed that each entity possessed data from only one language domain, capturing the decentralised and diverse characteristics of multilingual social media content. For consistency and to enable fair comparisons, the validation and test sets were kept the same across the centralised, FL, and FLiP configurations.

Language Models: This study utilized four multilingual transformer-based models: BERT-base-multilingual-cased (Devlin et al., 2018), DistilBERT-base-multilingual-cased (Sanh et al., 2019), XLM-RoBERTa-base (Conneau et al., 2019), and Multilingual-MiniLM (Wang et al., 2020). These models were chosen for their extensive use in multilingual NLP research, open accessibility, and suitability for our computational setup.

Parameter Settings: To ensure efficiency across multilingual entities with limited resources, we adopted a lightweight fine-tuning strategy by updating only the final four layers of each transformer model while integrating prompt tuning. The remaining layers were frozen to retain cross-lingual knowledge in the pre-trained representations. This approach capitalises on lower layers' ability to encode general linguistic features (Talebpour et al., 2023). By fine-tuning only upper layers alongside prompt parameters, the method significantly reduces computational and memory costs, making it suitable for federated training in low-resource multilingual environments while minimising catastrophic forgetting.

For optimisation, we used the Adam optimiser with an epsilon of 1e-8 and a learning rate of 2e-5. The Affinity threshold τ for entity grouping was set to 0.7, while the prompt length was fixed at 20 tokens. Both the regularization weight coefficient (λ) and momentum parameter (ρ) were assigned values of 0.0009. These hyperparameters were tuned via a grid search on a validation subset, with search ranges: prompt length $\in 10, 20, 30, 50$, $\tau \in 0.5, 0.7, 0.9$, and $\lambda, \rho \in 0.001, 0.0001, 0.009, 0.0009$.

Within each federated group, 90% of entities participated per round, and the data heterogeneity parameter (α) was set to 0.35 to simulate realistic multilingual imbalances. Implementation was in PyTorch 1.13 and Transformers 4.28, leveraging open-source multilingual models from Hugging Face[1]. All experiments were executed on an NVIDIA A100 GPU (40GB) with 128GB RAM. Datasets were divided into 70% training, 10% validation, and 20% test splits, stratified by language source. Due to federated memory constraints, batch size was set to 8, and each entity group consisted of 5 to 15 entities, as described in Section 3.1.1. Comprehensive sensitivity analysis of key hyperparameters is presented in Section 4.4.

4.2. Evaluation Metrics

To thoroughly assess the performance and computational efficiency of the proposed FLiP framework against federated and centralised baselines, we used diverse performance and efficiency metrics.

Performance Metrics: The effectiveness of multilingual classification was evaluated using Accuracy and F1-score, providing a balanced assessment of prediction quality across majority and minority classes. These metrics are defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Accuracy indicates the overall proportion of correctly predicted instances, while the F1-score harmonises Precision and Recall, making it particularly suitable for imbalanced multilingual datasets where the distribution of samples varies across languages and regions.

Efficiency Metrics: To evaluate the computational efficiency of the proposed FLiP framework, we employed three complementary metrics. First, **Communication Cost**, which quantifies the total volume of data transmitted between entities and the central server across all communication rounds, where S_t

[1] https://huggingface.co/

Dataset	FedAvg		FedProx		FedOpt		FedPrompt		FLiP	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
IID Results										
UOD	0.925	0.923	0.931	0.930	0.918	0.917	0.895	0.891	0.939	0.939
POLD	0.877	0.873	0.887	0.875	0.872	0.869	0.865	0.861	0.899	0.889
BHD	0.839	0.822	0.845	0.837	0.831	0.838	0.829	0.817	0.857	0.847
Non-IID Results										
UOD	0.917	0.910	0.926	0.925	0.919	0.917	0.885	0.881	0.932	0.932
POLD	0.879	0.871	0.889	0.878	0.871	0.867	0.860	0.857	0.900	0.892
BHD	0.822	0.819	0.837	0.824	0.827	0.826	0.817	0.810	0.845	0.836

Table 2: FLiP compared with baselines on individual datasets.

	# Trainable Params	Communication Cost	GPU memory Usage
Centralized	177,854,978	-	3438.65 MB
Federated	177,854,978	49557.94 MB	15844.02 MB
FLiP	28,368,386	5.62 MB	1493.68 MB

Table 3: Resource efficiency comparison.

Model	IID		Non-IID	
	Acc	F1	Acc	F1
Bert-M	0.885	0.876	0.850	0.849
XLM-R	0.902	0.896	0.889	0.883
DistilBERT	0.900	0.892	0.885	0.884
Mini-LM	0.878	0.875	0.866	0.861

Table 4: Centralised baseline settings.

Dataset	IID		Non-IID	
	Acc	F1	Acc	F1
UOD	0.939	0.939	0.915	0.915
POLD	0.917	0.911	0.909	0.900
BHD	0.883	0.874	0.878	0.859

Table 5: Centralised baseline results per dataset.

represents the participating entity set at round t and $\text{Size}(\Delta P_{i,t})$ is the model update size from entity i:

$$C_{\text{comm}} = \sum_{t=1}^{T} \sum_{i \in S_t} \text{Size}(\Delta P_{i,t}) \tag{5}$$

Second, **Trainable Parameter Ratio**, indicating the percentage of parameters updated during training, with lower values reflecting reduced computational and memory overhead:

$$R_{\text{params}} = \frac{P_{\text{train}}}{P_{\text{total}}} \times 100 \tag{6}$$

Third, **Memory Utilization**, M_{total}, representing the total GPU memory consumed during training, including both model parameters and intermediate tensors. Together, these efficiency metrics provide a holistic view of FLiP's scalability, resource optimisation, and communication efficiency, highlighting its suitability for low-resource multilingual federated environments with limited bandwidth and computational capacity.

4.3. Discussions

We evaluated FLiP across multiple models under both IID and Non-IID data distributions to assess its generalisation and personalisation capabilities. Generalisation refers to performance on the combined multilingual dataset, reflecting how well the model captures shared cross-lingual patterns (Table 1). In contrast, personalisation measures how effectively the model adapts to the characteristics of each individual language dataset, thereby capturing language-specific nuances. To evaluate this, we tested the best-performing model separately on each dataset (Table 2). Centralised training results (Tables 4 and 5) provide additional context for FLiP's advantages.

As shown in Table 1, overall performance declines under non-IID conditions due to class imbalance and distribution variability, yet FLiP consistently surpasses all federated baselines in Accuracy and F1-score. Multilingual BERT achieves the strongest results in both settings. Importantly, while overall performance decreases under non-IID conditions compared to IID ones, the reduction in accuracy for our proposed FLiP model is considerably smaller than that of the baseline FL methods. This enhanced robustness stems from the inclusion of the regularization weight coefficient λ, which helps preserve consistency with the community prompt and prevents substantial divergence during training.

Per-dataset evaluation (Table 2) shows FLiP consistently outperforms baseline FL models, confirming federated prompt learning benefits on domain-specific partitions. While global aggregation enhances generalisation, performance varies across datasets due to textual diversity, linguistic varia-

Dataset	Example Text (English Translation)	Context
UOD	*"Kerry brother, today the [offensive word] fans of that [offensive word] will come on your page, be careful."*	Toxic / insult; interpersonal hostility on social media.
POLD	*"The Taliban are a group of Punjabi [offensive word] trained in [offensive word]; anyone who supports them is not Afghan, he is a [offensive word]."*	Political hate speech; ethnic tension.
BHD	*"I see the presenter herself is a [offensive word]; an ideal Muslim is enough to [offensive word] these four atheists."*	Gendered hate and religious intolerance.

Table 6: Example posts illustrating heterogeneity across datasets.

Model	FLiP (zero shot)				FLiP (unfreeze layers = 2)				FLiP (unfreeze layers = 4)				FLiP (unfreeze layers = 12)			
	IID		Non-IID		IID		Non-IID		IID		Non-IID		IID		Non-IID	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Bert-M	0.785	0.771	0.768	0.752	0.857	0.849	0.844	0.841	0.881	0.872	0.876	0.868	0.885	0.876	0.879	0.871
XLM-R	0.772	0.768	0.754	0.751	0.847	0.837	0.837	0.835	0.876	0.875	0.863	0.861	0.881	0.879	0.870	0.865
DistilBERT	0.742	0.738	0.728	0.711	0.813	0.825	0.801	0.800	0.852	0.844	0.841	0.839	0.866	0.857	0.843	0.842
Mini-LM	0.750	0.745	0.733	0.727	0.826	0.817	0.819	0.804	0.858	0.855	0.846	0.843	0.864	0.859	0.851	0.844

Table 7: Layer-wise performance

tions, cultural context, and platform-specific language use (as depicted in Table 6). Our proposed approach tackles this challenge by improving the model's ability to generalise across different individual datasets. Through an entity grouping strategy, we cluster entities with similar characteristics, facilitating more efficient knowledge sharing within each group. This enables the model to utilise domain-specific information and refine it across varied datasets, ultimately enhancing its adaptability.

We further assessed our model's efficiency with respect to both communication and computation, as summarised in Table 3. It reduces trainable parameters by 84% and local GPU memory usage by up to 90% through prompt-based fine-tuning, enabling lightweight updates without full retraining. This design accelerates training, lowers resource use, and mitigates overfitting, making FLiP ideal for resource-limited environments. In terms of communication, traditional FL required roughly 50 GB of data exchange, while FLiP nearly eliminates this overhead, achieving almost a 100% reduction and enabling faster, bandwidth-efficient training for low-connectivity settings like rural areas and edge devices.

To provide a broader comparison, we also conducted experiments under a centralised learning setup, with results summarised in Tables 4 and 5. As expected, centralised models achieved slightly higher accuracy since they benefit from full access to all data, enabling more comprehensive learning. In contrast, federated entities operate on domain-specific, heterogeneous datasets, which inherently limit performance. When FL entities train on mixed-domain data, FLiP matches or slightly surpasses the centralised baseline. For realistic evaluation, however, experiments assume each entity accesses only one domain's data under both IID and non-IID settings. Importantly, under realistic conditions where each FL entity accessed only a single domain, FLiP still outperformed traditional

FLiP	α=0.25		α=0.30		α=0.35	
Model	Acc	F1	Acc	F1	Acc	F1
Bert-M	0.870	0.862	0.878	0.865	0.876	0.868
XLM-R	0.859	0.857	0.861	0.860	0.863	0.861
DistilBERT	0.843	0.835	0.847	0.842	0.841	0.839
Mini-LM	0.821	0.811	0.835	0.829	0.846	0.843

Table 8: FLiP compared on different data heterogeneity.

FL baselines across both IID and non-IID setups, confirming its robustness, scalability, and efficiency in real-world, heterogeneous environments.

We also performed an ablation study to examine the effect of unfreezing different numbers of layers during training, as shown in Table 7. While performance generally improved with more trainable layers, we found that unfreezing only the last four layers combined with prompt learning already outperformed standard FL baselines. This strategy, grounded in transfer learning principles, preserves the general linguistic knowledge in frozen lower layers while allowing the upper layers and prompts to adapt to task-specific and domain-specific variations. The approach minimises trainable parameters and computational cost while preventing catastrophic forgetting, maintaining stability across heterogeneous entity data. Consequently, this lightweight fine-tuning setup achieves strong performance and high efficiency, making it ideal for real-world, resource-limited federated environments.

Our framework outperforms baselines through the synergy of its core components. Soft prompts enable efficient task adaptation with minimal training overhead, while entity grouping reduces negative transfer and enhances generalisation. Proximal regularization weight stabilises updates, and momentum-based aggregation ensures smooth convergence. Selectively unfreezing top encoder layers allows precise task adaptation with low computation. Together, these mechanisms create a cohesive, resource-efficient system that balances accuracy, stability, and scalability for robust multi-

FLiP	Non IID (α=0.35)					
	E = 1		E = 3		E = 5	
Model	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Bert-M	0.871	0.865	0.876	0.868	0.875	0.864
XLM-R	0.855	0.854	0.863	0.861	0.862	0.862
DistilBERT	0.829	0.826	0.841	0.839	0.845	0.844
Mini-LM	0.837	0.834	0.846	0.843	0.847	0.847

Table 9: FLiP with different local epochs

FLiP	Non IID (α=0.35)					
	$\mathcal{E}$ = 50%		$\mathcal{E}$ = 70%		$\mathcal{E}$ = 90%	
Model	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Bert-M	0.871	0.870	0.877	0.867	0.876	0.868
XLM-R	0.841	0.836	0.855	0.853	0.863	0.861
DistilBERT	0.849	0.837	0.835	0.836	0.841	0.839
Mini-LM	0.819	0.815	0.827	0.822	0.846	0.843

Table 10: FLiP with different participation ratios.

lingual federated learning.

4.4. Sensitivity Analysis

This section examines the sensitivity of the proposed FLiP framework to key hyperparameter variations. Table 8 reports results across different levels of data heterogeneity, controlled by the parameter α, with an entity participation rate of 90% and three local epochs. A smaller α value indicates greater class imbalance and non-IID distribution. As shown, model performance exhibits a mild decline with increasing heterogeneity, reflecting the expected challenges of imbalanced data distribution across entities.

Table 9 shows the model's performance across varying local training epochs with a 90% entity participation rate. While increasing epochs generally enhances performance, occasional drops suggest overfitting to local entity data. Similarly, Table 10 illustrates the impact of different entity participation ratios ($\mathcal{E}$) with fixed local epochs ($E = 3$). Higher participation consistently improves results, as involving more entities enhances knowledge aggregation and model generalisation.

FLiP maintains consistent performance across varying hyperparameters, data heterogeneity, and entity participation, demonstrating strong robustness. Its stability stems from the regularization weight coefficient aligning community prompts and momentum-based averaging ensuring smooth convergence, confirming FLiP's effectiveness in heterogeneous, resource-limited multilingual settings.

5. Conclusions

We have proposed a novel framework that addresses data heterogeneity, privacy preservation, and computational efficiency in multilingual tasks through federated prompt tuning. Unlike conventional FL methods facing communication bottlenecks and weak generalisation, FLiP incorporates lightweight prompt sharing and cross-lingual adaptation for effective knowledge transfer across underrepresented languages, drastically reducing communication cost and memory usage. This establishes a scalable, privacy-conscious, resource-efficient paradigm for multilingual learning, providing a foundation for future research on cross-lingual fairness, robustness, and multimodal extensions in decentralised NLP.

6. Ethical Considerations

Our experiments employed publicly available datasets with appropriate citations. Although FLiP maintains privacy through decentralised data storage, practical implementation necessitates transparent communication with entities regarding data usage and model update procedures. The framework deliberately incorporates low-resource entities to address ethical concerns surrounding equitable AI access and bridge the digital divide. Nevertheless, careful consideration must be given to prevent inadvertent exclusion of participants facing connectivity or infrastructure constraints. Given that our framework targets indigenous and minority language communities, we emphasise the importance of obtaining informed community consent before deployment. Language data carries cultural significance, and any real-world application of FLiP should adhere to indigenous data sovereignty principles, ensuring communities retain ownership and control over their linguistic resources.

7. Limitations

FLiP demonstrates strong performance and efficiency in multilingual hate speech detection, which serves as a representative and challenging task for evaluating the framework in low-resource and cross-lingual settings. While this focused scope enables a rigorous assessment, future work can extend FLiP to additional multilingual NLP tasks such as named entity recognition, machine translation, and question answering. Furthermore, the evaluation includes three low-resource languages, namely Urdu, Pashto, and Bengali, selected to reflect linguistic diversity and realistic deployment contexts. Future studies may consider incorporating a wider range of typologically diverse languages to further examine and enrich understanding of the framework's generalisability and robustness across varied linguistic settings.

8. Acknowledgements

This work was supported by the Turing's Defence and Security programme through a partnership with

the UK government in accordance with the framework agreement between GCHQ & The Alan Turing Institute. This study was also funded by the University of Southampton (grant number 522886110).

9. Bibliographical References

Ghadeer Al-Badani, Muneer Hazaa Alsurori, and Akram Alsubari. 2025. Transfer learning and cross-linguistic generalization in multilingual hate speech detection: Approaches and challenges. In *2025 5th International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, pages 1–8. IEEE.

Muhammad Asad, Saima Shaukat, Dou Hu, Zekun Wang, Ehsan Javanmardi, Jin Nakazato, and Manabu Tsukada. 2023. Limitations and future aspects of communication costs in federated learning: A survey. *Sensors*, 23(17):7358.

Ubaid Azam, Imran Razzak, Shelly Vishwakarma, and Shoaib Jameel. 2025. Uncertainty modelling in under-represented languages with bayesian deep gaussian processes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1438–1450.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. 2024. Feddat: An approach for foundation model finetuning in multimodal heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11285–11293.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2022. Compositional generalization in multilingual semantic parsing over wikidata. *Transactions of the Association for Computational Linguistics*, 10:937–955.

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

Ahmed El Ouadrhiri and Ahmed Abdelhadi. 2022. Differential privacy for deep and federated learning: A survey. *IEEE access*, 10:22359–22380.

Kai Feng, Lan Huang, Kangping Wang, Wei Wei, and Rui Zhang. 2024. Prompt-based learning framework for zero-shot cross-lingual text classification. *Engineering Applications of Artificial Intelligence*, 133:108481.

Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Helen Möllering, Thien Duc Nguyen, Phillip Rieger, Ahmad-Reza Sadeghi, Thomas Schneider, Hossein Yalame, et al. 2021. Safelearn: Secure aggregation for private federated learning. In *2021 IEEE security and privacy workshops (SPW)*, pages 56–62. IEEE.

Mastafa Foufa. 2020. Anomaly detection across multiple languages.

Karim Gamal, Ahmed Gaber, and Hossam Amer. 2023. Federated learning based multilingual emoji prediction in clean and attack scenarios. *arXiv preprint arXiv:2304.01005*.

Vladimir Geroimenko. 2025. Key challenges in prompt engineering. In *The Essential Guide to Prompt Engineering: Key Principles, Techniques, Challenges, and Security Risks*, pages 85–102. Springer.

Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. 2023. Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 23(5):5179–5194.

Rajan Gupta, Sanju Tiwari, and Poonam Chaudhary. 2025. Prompt engineering. In *Generative AI: Techniques, Models and Applications*, pages 163–186. Springer.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Ahmed Imteaj, Urmish Thakker, Shiqiang Wang, Jian Li, and M Hadi Amini. 2021. A survey on federated learning for resource-constrained iot devices. *IEEE Internet of Things Journal*, 9(1):1–24.

Elaheh Jafarigol, Theodore B Trafalis, Talayeh Razzaghi, and Mona Zamankhani. 2024. Exploring machine learning models for federated learning: A review of approaches, performance, and limitations. *Dynamics of Disasters: From Natural Phenomena to Human Activity*, pages 87–121.

Loso Judijanto and Arnes Yuli Vandika. 2025. Emerging research trends in natural language processing for multilingual ai. *The Eastasouth*

Journal of Information System and Computer Science, 2(03):187–199.

Samar Samir Khalil, Noha S Tawfik, and Marco Spruit. 2024. Federated learning for privacy-preserving depression detection with multilingual language models in social media posts. *Patterns*, 5(7).

Junghwan Kim, Haotian Zhang, and David Jurgens. 2025. Leveraging multilingual training for authorship representation: Enhancing generalization across languages and domains. *arXiv preprint arXiv:2509.16531*.

Mohit Kumar et al. 2025. Exploring hate speech detection: challenges, resources, current research and future directions. *Multimedia Tools and Applications*, pages 1–37.

Hao Li, Chengcheng Li, Jian Wang, Aimin Yang, Zezhong Ma, Zunqian Zhang, and Dianbo Hua. 2023. Review on security of federated learning and its application in healthcare. *Future Generation Computer Systems*, 144:271–290.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

Ying Li, Xingwei Wang, Rongfei Zeng, Praveen Kumar Donta, Ilir Murturi, Min Huang, and Schahram Dustdar. 2025. Federated domain generalization: A survey. *Proceedings of the IEEE*.

Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. 2020. Federated learning in mobile edge networks: A comprehensive survey. *IEEE communications surveys & tutorials*, 22(3):2031–2063.

Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. 2024. Palo: A polyglot large multimodal model for 5b people. *arXiv preprint arXiv:2402.14818*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

George Manias, Argyro Mavrogiorgou, Athanasios Kiourtis, Chrysostomos Symvoulidis, and Dimosthenis Kyriazis. 2023. Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. *Neural Computing and Applications*, 35(29):21415–21431.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. 2021. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640.

Balázs Nagy, István Hegedűs, Noémi Sándor, Balázs Egedi, Haaris Mehmood, Karthikeyan Saravanan, Gábor Lóki, and Ákos Kiss. 2023. Privacy-preserving federated learning and its application to natural language processing. *Knowledge-Based Systems*, 268:110475.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Yifan Peng, Yui Sudo, Muhammad Shakeel, and Shinji Watanabe. 2024. Owsm-ctc: An open encoder-only speech foundation model for speech recognition, translation, and language identification. *arXiv preprint arXiv:2402.12654*.

Ruben Piperno, Luca Bacco, Felice Dell'Orletta, Mario Merone, and Leandro Pecchia. 2025. Cross-lingual distillation for domain knowledge transfer with sentence transformers. *Knowledge-Based Systems*, 311:113079.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Xiaoyu Qiu, Yuechen Wang, Jiaxin Shi, Wengang Zhou, and Houqiang Li. 2024. Cross-lingual transfer for natural language inference via multilingual prompt translator. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Ashish Rauniyar, Desta Haileselassie Hagos, Debesh Jha, Jan Erik Håkegård, Ulas Bagci, Danda B Rawat, and Vladimir Vlassov. 2023. Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal*.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. 2020. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.

Pascal Riedel, Manfred Reichert, Reinhold Von Schwerin, Alexander Hafner, Daniel Schaudt, and Gaurav Singh. 2023. Performance analysis of federated learning algorithms for multilingual protest news detection using pre-trained distilbert and bert. *IEEe Access*, 11:134009–134022.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Akshay Singh and Rahul Thakur. 2024. Generalizable multilingual hate speech detection on low resource indian languages using fair selection in federated learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7204–7214.

Mozhgan Talebpour, Alba García Seco de Herrera, and Shoaib Jameel. 2023. Topics in contextualised attention embeddings. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*, page 221–238, Berlin, Heidelberg. Springer-Verlag.

Faizad Ullah, Safiullah Faizullah, Imdad Ullah Khan, Turki Alghamdi, Toqeer Ali Syed, Ahmad B Alkhodre, Muhammad Sohaib Ayub, and Asim Karim. 2025. Prompt-based fine-tuning with multilingual transformers for language-independent sentiment analysis. *Scientific Reports*, 15(1):20834.

Shubham Vatsal, Harsh Dubey, and Aditi Singh. 2025. Multilingual prompt engineering in large language models: A survey across nlp tasks. *arXiv preprint arXiv:2505.11665*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Haoyu Wang, Handong Zhao, Yaqing Wang, Tong Yu, Jiuxiang Gu, and Jing Gao. 2022. Fedkc: Federated knowledge composition for multilingual natural language understanding. In *Proceedings of the ACM Web Conference 2022*, pages 1839–1850.

Lihua Wang. 2024. Cross-lingual nlp: Bridging language barriers with multilingual model. In *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, pages 1005–1012. IEEE.

Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. 2024. Large language models are good multi-lingual learners: When llms meet cross-lingual prompts. *arXiv preprint arXiv:2409.11056*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. 2023. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513–535.

Jenny Yang, Nguyen Thanh Dung, Pham Ngoc Thach, Nguyen Thanh Phong, Vu Dinh Phu, Khiem Dong Phu, Lam Minh Yen, Doan Bui Xuan Thy, Andrew AS Soltan, Louise Thwaites, et al. 2024. Generalizability assessment of ai models across hospitals in a low-middle and high income country. *Nature Communications*, 15(1):8270.

Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. 2022. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pages 26311–26329. PMLR.

Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. 2023. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wanru Zhao, Yihong Chen, Royson Lee, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. 2024. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages. In *The Twelfth International Conference on Learning Representations*.

10. Language Resource References

Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed Abdelmajeed, and Muhammad Tariq Sadiq. 2020. Automatic detection of offensive language for urdu and roman urdu. *IEEE Access*, 8:91213–91226.

Ijazul Haq, Weidong Qiu, Jie Guo, and Peng Tang. 2023. Pashto offensive language detection: a benchmark dataset and monolingual pashto bert. *PeerJ Computer Science*, 9:e1617.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 457–468. Springer.