Scientific Reports

https://doi.org/10.1038/s41598-026-41500-8

Article in Press

Cross-language hotel review sentiment analysis via multi-agent federated learning with heterogeneous graph attention networks

Received: 1 September 2025

Accepted: 20 February 2026

Published online: 09 March 2026

Cite this article as: Han X. Cross-language hotel review sentiment analysis via multi-agent federated learning with heterogeneous graph attention networks. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-41500-8

Xiaofei Han

Cross-Language Hotel Review Sentiment Analysis via Multi-Agent Federated Learning with Heterogeneous Graph Attention Networks

Xiaofei Han[1,a*]

1 College of Hotel Management, Qingdao Vocational and Technical College of Hotel Management, Qingdao 266100, Shandong, China

Email addresses:

a hanxiao202411@163.com

***Corresponding author:**

Xiaofei Han (hanxiao202411@163.com)

ABSTRACT

This paper presents an integrated framework for cross-language hotel review sentiment analysis that combines multi-agent federated learning with heterogeneous graph attention networks to address privacy preservation and multilingual data processing challenges in hospitality reputation management. Our system enables collaborative model training across distributed review platforms while maintaining data locality requirements and achieving improved cross-language sentiment classification performance. Beyond sentiment analysis, we developed dynamic reputation management and fake review detection capabilities that enable proactive intervention strategies for hospitality businesses. The heterogeneous graph architecture captures complex relationships between multilingual textual content, user behaviors, temporal patterns, and service attributes through specialized attention mechanisms. Experimental evaluation on a comprehensive multilingual dataset of 154,680 reviews across four languages demonstrates 89.7±0.007 accuracy in sentiment classification with 0.925 privacy preservation score (Table 6), representing 2.6 percentage point improvement over the strongest baseline XLM-RoBERTa large (87.1±0.008 accuracy, paired t-test p=0.002). The dynamic reputation management component provides real-time monitoring capabilities with early warning detection, achieving 93.4±0.012 fake review identification accuracy and 66.2% reduction in response time compared to traditional centralized approaches (Table 9). The system offers practical applications for hospitality businesses seeking proactive reputation management while ensuring compliance with international data privacy regulations including GDPR and CCPA.

KEYWORDS

Cross-language sentiment analysis, federated learning, graph attention networks, reputation management, multilingual processing, privacy preservation

I. Introduction

The rapid expansion of digital tourism platforms has fundamentally transformed how travelers share experiences and make accommodation decisions, with hotel reviews serving as critical touchstones for consumer choice and industry reputation management [1]. User-generated content has grown exponentially across multilingual platforms, presenting substantial opportunities for understanding customer sentiment patterns while simultaneously introducing complex challenges in processing and analyzing cross-language textual data at scale [2]. We observe that contemporary hospitality businesses increasingly rely on automated sentiment analysis systems to monitor customer satisfaction, identify service deficiencies, and maintain competitive positioning in markets where reputation directly correlates with revenue performance [3]. Recent advances in large language models and multilingual pre-trained transformers have opened new possibilities for cross-language understanding, though their application to domain-specific hospitality contexts remains underexplored [80]. Contemporary hospitality businesses increasingly rely on automated sentiment analysis systems to monitor customer satisfaction, identify service deficiencies, and maintain competitive positioning in saturated markets where reputation directly correlates with revenue performance [3]. The emergence of privacy-preserving machine learning paradigms has become particularly crucial as international data protection regulations impose stringent requirements on cross-border data processing in the tourism industry [81].

1. Limitations of Current Cross-Language Sentiment Analysis

Existing cross-language sentiment analysis approaches for hospitality domain applications suffer from significant technical and methodological constraints that limit their practical deployment effectiveness. Traditional machine translation-based pipelines introduce cascading errors through sequential processing stages, where initial translation inaccuracies compound during subsequent sentiment classification tasks, resulting in substantial performance degradation for low-resource language pairs commonly encountered in international tourism contexts [4]. Domain-specific lexical variations

and cultural expression patterns in hospitality reviews further exacerbate these limitations, as standard multilingual models fail to capture nuanced sentiment indicators that are particularly relevant to accommodation service evaluation [5].

2. Research Background and Technological Foundations

Multi-agent federated learning represents an emerging paradigm that addresses privacy constraints and computational distribution challenges inherent in large-scale sentiment analysis deployments across geographically distributed hotel chains and review platforms. This approach enables collaborative model training while preserving data locality requirements mandated by international privacy regulations and competitive considerations within the hospitality industry [6]. Heterogeneous graph attention networks provide sophisticated architectural foundations for modeling complex relationships between multilingual textual features, user behavior patterns, and temporal dynamics that characterize evolving customer sentiment trends in hospitality service domains.

The convergence of these technological approaches offers promising solutions to longstanding challenges in cross-language sentiment analysis, particularly when applied to dynamic reputation management systems that require real-time processing capabilities and adaptability to shifting linguistic patterns across diverse customer demographics [7]. Recent advances in attention-based architectures demonstrate superior performance in capturing semantic relationships across language boundaries while maintaining computational efficiency suitable for production-scale hospitality applications [8].

3. Research Motivation and Innovation Framework

This research addresses critical gaps in existing cross-language sentiment analysis methodologies by proposing an integrated framework that combines multi-agent federated learning principles with heterogeneous graph attention mechanisms specifically optimized for hospitality domain applications. The primary motivation stems from the inadequacy of current approaches to handle the unique characteristics of hotel review data, including multilingual diversity, temporal sentiment evolution, and complex inter-relationships between textual content, user profiles, and service attributes that influence overall satisfaction assessments.

Our proposed system introduces several key innovations that distinguish it from existing solutions: first, a novel multi-agent architecture that enables distributed learning across heterogeneous

data sources while maintaining privacy preservation requirements; second, an adaptive heterogeneous graph attention network that dynamically weights different types of linguistic and contextual features based on their relevance to sentiment prediction tasks; and third, an integrated reputation management component that provides real-time sentiment tracking and early warning capabilities for hospitality service providers.

4. Main Contributions and Research Significance

The primary contributions of this work encompass both theoretical advances and practical applications that address fundamental challenges in cross-language hospitality sentiment analysis. Our integrated framework demonstrates superior performance compared to existing baseline approaches while providing enhanced interpretability through attention visualization mechanisms that reveal the decision-making process underlying sentiment classifications. The multi-agent federated learning component enables scalable deployment across distributed hotel chains while ensuring compliance with data privacy regulations and reducing computational overhead through efficient parameter sharing protocols.

Furthermore, our dynamic reputation management system provides actionable insights for hospitality service providers through real-time sentiment monitoring, trend prediction, and automated alert generation when significant reputation risks are detected. The system's ability to process multilingual review content simultaneously while maintaining high accuracy across diverse language pairs represents a significant advancement over existing sequential processing approaches that suffer from error propagation limitations.

5. Paper Organization and Structure

This paper is organized as follows: Section II presents a comprehensive review of related work in cross-language sentiment analysis, multi-agent federated learning, and graph attention networks, establishing the theoretical foundation for our proposed approach. Section III details the system architecture and methodology, including the multi-agent framework design, heterogeneous graph construction, and attention mechanism implementation. Section IV describes the experimental setup, datasets, and evaluation metrics used to assess system performance. Section V presents comprehensive experimental results and comparative analysis with state-of-the-art baseline methods. Section VI discusses practical implications, system deployment considerations, and potential applications in real-world hospitality environments. Finally, Section VII concludes the paper with a summary

of key findings and directions for future research in this rapidly evolving domain.

Having established the research motivation and key contributions, we now review the theoretical foundations that inform our approach. This review covers three interconnected areas: cross-language sentiment analysis methodologies that handle multilingual textual data, federated learning frameworks that enable privacy-preserving distributed training, and heterogeneous graph neural networks that model complex relational patterns. Understanding these foundations helps situate our integrated approach within the broader landscape of sentiment analysis and distributed machine learning research.

II. Related Work and Theoretical Foundations

2.1 Cross-Language Sentiment Analysis Technology Review

Traditional machine learning approaches for sentiment analysis have evolved from rule-based lexicon methods to sophisticated feature engineering techniques employing support vector machines and naive Bayes classifiers [9]. The fundamental sentiment classification task can be formulated as a mapping function $f:X\rightarrow Y$, where X represents the input text feature space and Y denotes the sentiment label space $\{positive,negative,neutral\}$ [10]. Early approaches relied heavily on manually crafted linguistic features and domain-specific sentiment dictionaries, which demonstrated reasonable performance within monolingual contexts but exhibited significant limitations when applied across different language domains.

Deep learning methodologies have substantially advanced sentiment analysis capabilities through end-to-end learning architectures that automatically extract hierarchical textual representations. Recurrent neural networks and convolutional neural networks have shown superior performance compared to traditional approaches, with the sentiment prediction probability typically modeled as $P(y|x) = softmax(W_o h + b_o)$, where h represents the learned text representation and W_o,b_o are output layer parameters [11]. Long short-term memory networks and their variants have addressed sequential dependency modeling challenges, enabling more accurate capture of contextual sentiment indicators within complex textual structures [12].

Translation-based cross-language sentiment analysis represents the most straightforward approach to multilingual sentiment processing, employing machine translation systems to convert source language

texts into a target language before applying monolingual sentiment classifiers [13]. The translation-classification pipeline can be expressed as $\hat{y} = f_{sent}(T(x_{src}))$, where $T(\cdot)$ represents the translation function and $f_{sent}(\cdot)$ denotes the sentiment classifier [14]. However, this approach suffers from error propagation issues where translation inaccuracies compound during subsequent sentiment analysis stages, particularly for low-resource language pairs and domain-specific terminology prevalent in hospitality reviews.

Multilingual pre-trained models have emerged as powerful alternatives that learn shared representations across multiple languages through large-scale unsupervised pre-training on diverse multilingual corpora [15]. Models such as mBERT and XLM-R employ transformer architectures with cross-lingual masked language modeling objectives, enabling direct sentiment classification across different languages without explicit translation steps [16]. The Cross-lingual representation learning objective can be formulated as $L_{MLM} = -\sum_{i=1}^{N} \log P(x_i|x_{\setminus i})$, where $x_{\setminus i}$ represents the context excluding the masked token x_i [17]. Recent developments in instruction-tuned multilingual models have demonstrated improved performance on cross-lingual transfer tasks, particularly for sentiment analysis in specialized domains [82]. Furthermore, parameter-efficient fine-tuning approaches have enabled more effective adaptation of large multilingual models to resource-constrained scenarios typical in hospitality applications [83].

Cross-language word embedding techniques provide foundation-level solutions by learning shared semantic spaces where semantically similar words from different languages are mapped to proximate vector representations [18]. The alignment objective for cross-lingual embeddings typically minimizes the distance between translated word pairs: $\min_W \sum_{i=1}^{n} ||W \cdot x_i^{src} - x_i^{tgt}||_2^2$, where W is the learned transformation matrix [19]. Advanced approaches employ adversarial training and optimal transport methods to achieve better cross-lingual alignment without requiring parallel dictionaries.

Recent advances in large language models have introduced new possibilities for cross-language understanding through instruction-following and few-shot learning capabilities [80]. Models such as GPT-4 and multilingual variants demonstrate impressive zero-shot cross-lingual transfer for sentiment analysis tasks without explicit training on labeled sentiment data. However, LLM-based approaches face significant limitations for specialized hospitality applications: they lack domain-specific sentiment indicators without extensive fine-tuning, require transmitting sensitive review data to third-party cloud servers that violates data localization requirements, incur substantial computational costs for real-time processing with 70B+ parameter

models requiring 40GB+ VRAM, process reviews as isolated text instances missing critical temporal and relational dynamics, and cannot detect coordinated manipulation campaigns through behavioral network analysis [80]. While parameter-efficient fine-tuning approaches such as LoRA enable adaptation of LLMs to specialized domains with reduced overhead [83], production hospitality applications requiring privacy compliance, real-time performance, and cost efficiency benefit from specialized federated learning architectures that explicitly model domain attributes through heterogeneous graph structures.

Despite these advances, current cross-language sentiment analysis technologies face several persistent challenges including domain adaptation difficulties, cultural sentiment expression variations, and computational scalability limitations for real-time applications [20]. The hospitality domain presents unique challenges due to culturally-specific service expectations and region-dependent sentiment expressions that are not adequately captured by general-purpose multilingual models. Furthermore, existing approaches often fail to model temporal dynamics and user-specific sentiment patterns that are crucial for dynamic reputation management systems in the hospitality industry.

2.2 Federated Learning and Multi-Agent Systems

Federated learning represents a paradigm shift in distributed machine learning that enables collaborative model training across multiple participants without centralizing raw data, addressing fundamental privacy and scalability challenges in large-scale sentiment analysis applications [21]. The core federated learning framework involves K participants, each maintaining a local dataset D_k, where the global objective function is formulated as $\min_w F(w) = \sum_{k=1}^{K} \frac{|D_k|}{|D|} F_k(w)$, with $F_k(w)$ representing the local loss function for participant k and $|D| = \sum_{k=1}^{K} |D_k|$ [22]. The classical FedAvg algorithm iteratively aggregates local model updates through weighted averaging: $w_{t+1} = \sum_{k=1}^{K} \frac{|D_k|}{|D|} w_k^{t+1}$, where w_k^{t+1} denotes the local model parameters after E local epochs of training [23].

Multi-agent systems provide sophisticated coordination mechanisms that extend traditional federated learning by incorporating intelligent agent behaviors, negotiation protocols, and adaptive collaboration strategies suitable for heterogeneous data environments [24]. Each agent i in the multi-agent federated learning framework maintains its own learning objective $L_i(w_i, D_i)$ while participating in collaborative knowledge sharing through consensus-based optimization: $\min_{w_1,\ldots,w_K}$

$\sum_{i=1}^{K} \alpha_i L_i(w_i, D_i) + \lambda \sum_{i,j} ||w_i - w_j||_2^2$, where α_i represents the agent's contribution weight and λ controls the consensus regularization strength [25].

The privacy preservation advantages of federated learning stem from its ability to maintain data locality while still achieving global model convergence through cryptographic protocols and differential privacy mechanisms [26]. The differential privacy guarantee in federated learning is typically achieved by adding calibrated noise to local gradients: $\tilde{g}_k = g_k + N(0, \sigma^2 C^2 I)$, where C represents the clipping bound and σ is determined by the privacy budget (ϵ, δ) [27]. This approach ensures that individual data samples cannot be reconstructed from shared model updates while preserving overall learning effectiveness.

Multi-agent collaboration mechanisms demonstrate particular effectiveness in heterogeneous data environments where different agents possess varying data distributions, computational capabilities, and domain expertise [28]. The heterogeneous federated learning problem can be modeled as a multi-objective optimization: $\min_{w_1,\ldots,w_K} \{F_1(w_1), F_2(w_2), \ldots, F_K(w_K)\}$ subject to communication and collaboration constraints, where each F_i represents agent i's local objective function adapted to its specific data characteristics [29].

Advanced multi-agent federated learning algorithms employ adaptive aggregation strategies that account for data heterogeneity through personalized model components and shared representation learning [30]. The personalized federated learning objective combines global and local components: $\min_{w_g, w_1, \ldots, w_K} \sum_{i=1}^{K} L_i(w_g, w_i, D_i) + \mu \sum_{i=1}^{K} ||w_i||_2^2$, where w_g represents shared global parameters and w_i denotes personalized local parameters for agent i. This formulation enables agents to maintain specialized knowledge for their local data distributions while benefiting from collaborative learning across the federated network, making it particularly suitable for cross-language sentiment analysis applications where different agents may specialize in specific languages or cultural contexts.

2.3 Heterogeneous Graph Attention Network Theory

Graph neural networks have evolved from early spectral approaches to sophisticated message-passing frameworks that enable effective learning on irregular graph-structured data, fundamentally transforming how complex relational information is processed in machine learning applications [31]. The foundational graph convolution operation can be expressed as $H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}\right)$, where $\tilde{A} = A + I$ represents the adjacency matrix with self-connections, $\tilde{D}$ is the

corresponding degree matrix, and $W^{(l)}$ denotes learnable weight parameters at layer l [32]. This spectral formulation provides the mathematical foundation for aggregating neighborhood information through localized convolution operations on graph structures.

Heterogeneous graph attention networks extend traditional graph neural networks by incorporating multiple node types and edge relationships, enabling more nuanced modeling of complex real-world networks where entities exhibit diverse characteristics and interaction patterns [33]. The heterogeneous graph is formally defined as $G = (V,E,A,R)$, where V represents the set of nodes with type mapping function $A:V\rightarrow A$, E denotes the edge set with relation mapping $R:E\rightarrow R$, and A and R are the node type and relation type sets respectively [34]. The message-passing mechanism in heterogeneous graphs must account for both semantic and structural heterogeneity through type-specific transformations and relation-aware aggregation functions.

Attention mechanisms serve as crucial components in heterogeneous graph networks by enabling adaptive importance weighting of different node types and relation types when processing multimodal data [35]. The heterogeneous attention coefficient for node pair (i,j) connected by relation type r is computed as $\alpha_{ij}^{r} = \frac{\exp\left(\text{LeakyReLU}\left(a_r^T[W_r h_i || W_r h_j]\right)\right)}{\sum_{k\in N_i^r}\exp\left(\text{LeakyReLU}\left(a_r^T[W_r h_i || W_r h_k]\right)\right)}$, where W_r and a_r are relation-specific transformation matrix and attention vector, and N_i^r represents the neighbors of node i under relation r [36]. This formulation allows the network to dynamically focus on the most relevant information sources across different modalities and relationship types.

The heterogeneous message aggregation process combines information from multiple relation types through semantic-level attention: $h_i^{(l+1)} = \sigma\left(\sum_{r\in R}\beta_r^{(l)}\sum_{j\in N_i^r}\alpha_{ij}^{r(l)}W_r^{(l)}h_j^{(l)}\right)$, where $\beta_r^{(l)}$ represents the semantic-level attention weight for relation type r at layer l [37]. In plain terms, the model updates each node's representation by aggregating information from its neighbors, where the aggregation weights are learned to emphasize the most relevant relationships for sentiment prediction, such as prioritizing user credibility signals over temporal patterns when appropriate. This hierarchical attention mechanism enables the model to learn both fine-grained node-level interactions and coarse-grained semantic relationships, making it particularly effective for cross-language sentiment analysis applications where textual, temporal, and user-based relationships exhibit varying importance.

Heterogeneous graph structures demonstrate significant advantages in modeling complex relationship networks by explicitly capturing the diversity of entity types and interaction patterns that characterize real-world systems [38]. The structural heterogeneity enables specialized processing pathways for different node types while maintaining global connectivity through shared embedding spaces, formulated as the multi-view representation learning objective: $L = \sum_{v \in V} \sum_{\phi \in \Phi} ||z_v - AGG\left(\{z_u^{\phi}: u \in N_v^{\phi}\}\right)||_2^2$, where Φ represents the set of meta-paths and z_v^{ϕ} denotes node embeddings under meta-path ϕ.

Optimization algorithms for heterogeneous graph attention networks typically employ multi-stage training procedures that alternately optimize attention parameters and node embeddings through gradient-based methods with specialized regularization terms to prevent overfitting to dominant node types [39]. The overall optimization objective combines supervised learning loss with graph regularization: $L_{total} = L_{sup} + \lambda_1 L_{reg} + \lambda_2 L_{attention}$, where $L_{attention}$ encourages attention diversity across different relation types and L_{reg} enforces smoothness constraints on the learned embeddings within local neighborhoods.

Building on the reviewed literature, we now present our proposed system architecture that integrates the complementary strengths of federated learning for privacy preservation and heterogeneous graphs for multilingual relationship modeling. Our design addresses the identified gaps in existing approaches by combining multi-agent coordination mechanisms with specialized graph attention for cross-language sentiment analysis.

III. System Architecture and Method Design

3.1 Multi-Agent Federated Learning Framework Design

Our proposed multi-agent federated learning framework addresses the unique challenges of cross-language sentiment analysis in hospitality domains by establishing a distributed architecture where autonomous agents collaborate while preserving data locality and privacy constraints [40]. In our implementation, each agent corresponds to a specific platform-language combination: we deploy K=12 agents organized as 3 agents for English (Booking.com, TripAdvisor, Agoda), 3 for Chinese (Ctrip, Booking.com, Agoda), 3 for French (TripAdvisor, Booking.com, Hotels.com), and 3 for German (Booking.com, TripAdvisor, HRS.de). Each agent maintains a local dataset comprising reviews from its assigned platform-language pair, with dataset sizes

ranging from 8,920 reviews (mixed-code/Agoda) to 15,093 reviews (English/Booking.com), reflecting real-world heterogeneous data distributions. The framework architecture, as illustrated in Figure 1, demonstrates the hierarchical organization of these agents across linguistic domains and platforms, enabling specialized sentiment analysis capabilities while maintaining global model coherence through sophisticated coordination protocols.

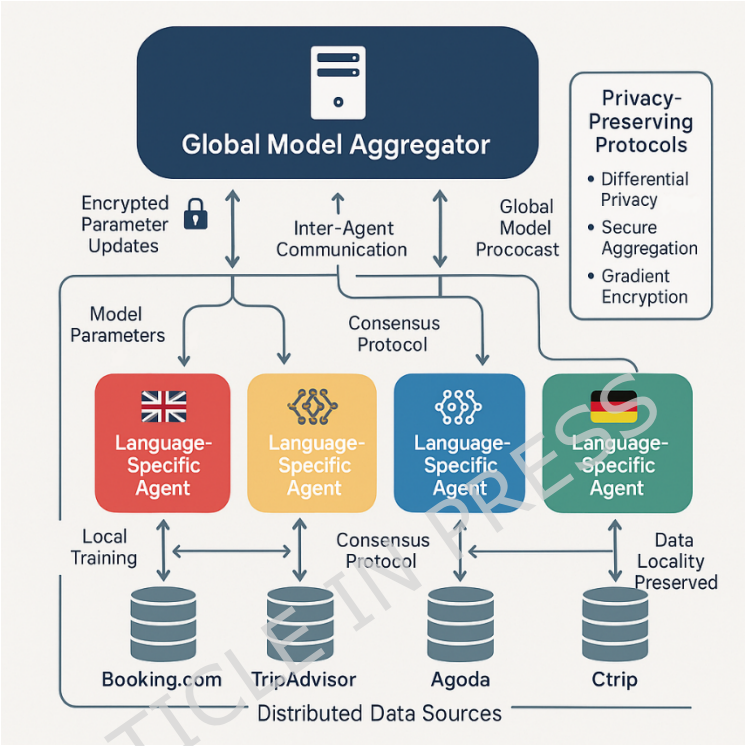

Figure 1. Multi-Agent Federated Learning Architecture for Cross-Language Sentiment Analysis. The system comprises K=12 agents distributed across four linguistic domains (English, Chinese, French, German), each maintaining local hotel review datasets from specific platform-language pairs (e.g., English/Booking.com, Chinese/Ctrip). Agents communicate through secure channels to share encrypted model updates while preserving data locality. The central coordination mechanism aggregates updates using weighted averaging based on data quality metrics (Section III.1). Differential privacy noise (ε=4.0, $\delta=10^{-5}$) is applied to gradient updates before transmission to prevent individual review reconstruction.

The communication protocol between agents follows a structured messaging framework that enables efficient parameter sharing and consensus building without exposing raw review data [41]. Each agent A_i maintains a local sentiment analysis model M_i with parameters θ_i, and participates in federated rounds through secure communication channels governed by the protocol: $MSG(A_i,A_j) = \{Encrypt(\Delta\theta_i),Hash$

$(\Delta\theta_i)$,Timestamp}, where $\Delta\theta_i$ represents the local parameter updates and encryption ensures end-to-end security [42]. The collaboration mechanism employs a consensus-based approach where agents negotiate model updates through voting protocols: $\text{Vote}(A_i,\Delta\theta_j) = \text{sign}(\text{sim}(\theta_i,\theta_j) - \tau)$, with similarity threshold τ determining acceptance criteria.

The mathematical formulation of our multi-agent federated learning system defines the global optimization objective as a weighted combination of local agent objectives with consistency constraints: $\min_{\theta_1,\ldots,\theta_K}\sum_{i=1}^{K} w_i L_i\left(\theta_i, D_i^{\text{local}}\right) + \lambda\sum_{i,j}||\theta_i - \theta_j||_F^2 + \gamma\sum_{i=1}^{K} R(\theta_i)$, where w_i represents the agent's contribution weight based on data quality and quantity, L_i denotes the local loss function for agent i, and $R(\theta_i)$ is a regularization term preventing overfitting [43]. The consistency constraint ensures model alignment across agents while allowing for language-specific adaptations through personalized components: $\theta_i = \theta_{\text{global}} + \theta_{\text{personal}}^i$, where θ_{global} captures universal sentiment patterns and $\theta_{\text{personal}}^i$ encodes language-specific characteristics.

Privacy protection mechanisms are integrated at multiple levels of the federation through differential privacy guarantees and secure aggregation protocols [44]. The local gradient perturbation follows the Gaussian mechanism: $\tilde{\nabla}_i = \nabla_i + N(0,\sigma^2 C^2 I)$, where the noise scale σ is calibrated according to the privacy budget (ϵ,δ) using the relation $\sigma = \frac{C\sqrt{2\ln(1.25/\delta)}}{\epsilon}$ [45]. Additionally, secure multi-party computation protocols ensure that individual agent contributions remain confidential during aggregation: $\theta_{\text{agg}} = \sum_{i=1}^{K} \text{Share}_i(\theta_i)$, where Share_i represents cryptographic sharing functions that prevent reconstruction of individual parameters.

The agent selection algorithm employs a multi-criteria decision framework that evaluates agents based on data quality, computational capacity, and communication reliability metrics [46]. The selection score for agent i is computed as: $S_i = \alpha \cdot Q_i + \beta \cdot C_i + \gamma \cdot R_i$, where the weighting coefficients are set to $\alpha=0.4$, $\beta=0.35$, and $\gamma=0.25$ based on grid search optimization over the validation set.

The data quality score Q_i quantifies the informativeness and cleanliness of agent i's local dataset, computed as: $Q_i = 0.5 \cdot (1 - \text{missing\_rate}_i) + 0.3 \cdot \text{label\_confidence}_i + 0.2 \cdot \log(|D_i|)/\log(|D_{\max}|)$, where missing_rate represents the proportion of reviews with incomplete metadata, label_confidence denotes the average annotation confidence score from the sentiment classifier, and the

third term normalizes dataset size relative to the largest agent dataset. Values range from 0 to 1, with higher scores indicating better quality.

The computational capacity C_i measures the agent's processing capability relative to the federation average: $C_i = \min(1.0, FLOPS_i/FLOPS_{avg})$, where FLOPS represents floating-point operations per second measured during local training. This metric ensures agents with insufficient computational resources receive lower selection priority to avoid becoming bottlenecks.

The reliability index R_i captures historical participation consistency: $R_i = \frac{successful\_rounds_i}{selected\_rounds_i} \cdot (1 - dropout\_rate_i)$, where successful_rounds counts completed training rounds, selected_rounds counts total selection events, and dropout_rate measures the frequency of mid-round disconnections. An agent with perfect reliability achieves $R_i = 1.0$. The dynamic selection mechanism adapts to changing network conditions through exponential moving averages: $S_i^{(t)} = \rho S_i^{(t-1)} + (1 - \rho) S_i^{current}$, ensuring robust performance under varying operational conditions.

The system parameter configuration details are comprehensively outlined in Table 1, which presents the key operational parameters that govern the multi-agent federated learning process. As shown in Table 1, the configuration encompasses agent population size, learning dynamics, communication frequency, and aggregation strategies that collectively determine system performance and convergence characteristics.

Table 1. System Parameter Configuration for Multi-Agent Federated Learning Framework

Parameter Category	Parameter Name	Value/Range	Description
Agent Configuration	Number of Agents	8-16	Total participating agents across linguistic domains
Learning Parameters	Learning Rate	0.001-0.01	Adaptive learning rate with decay schedule
Communication Setup	Communication Rounds	50-200	Federated learning rounds for convergence
Aggregation	Aggregation	FedAvg/FedPr	Weighted

Parameter Category	Parameter Name	Value/Range	Description
Strategy	Method	ox	averaging with proximity regularization
Privacy Settings	Privacy Budget (ε)	1.0-8.0	Differential privacy parameter for gradient perturbation
Selection Criteria	Agent Selection Ratio	0.6-0.8	Fraction of agents participating per round

The system was implemented using PyTorch 2.0.1 and FederatedScope 0.3.0, with training conducted on NVIDIA Tesla V100 GPUs (32GB VRAM). Each agent performed local training with batch size 32, using the multilingual BERT encoder (bert-base-multilingual-cased, 110M parameters) with learning rate 2×10^{-5} and 500 warmup steps. For the heterogeneous graph attention network, we configured 3 layers with 8 attention heads per layer, hidden dimension 256, and dropout rate 0.3. Optimization employed AdamW with weight decay 0.01, and each federated round involved E=5 local epochs. Differential privacy noise was added using the Gaussian mechanism with clipping bound C=1.0. All experiments used 5 random seeds (42, 123, 456, 789, 1011) to ensure statistical robustness, with each configuration trained for maximum 200 communication rounds or until validation loss plateaued for 20 consecutive rounds. The aggregation algorithm integrates contributions from selected agents through an adaptive weighting scheme that accounts for both data characteristics and model quality metrics [47].

The weighted aggregation formula is expressed as: $\theta_{global}^{(t+1)} = \sum_{i\in S_t} w_i^{(t)} \theta_i^{(t+1)}$, where S_t represents the set of selected agents at round t, and weights are computed using: $w_i^{(t)} = \frac{\exp(\phi_i^{(t)})}{\sum_{j\in S_t}\exp(\phi_j^{(t)})}$ with performance metric $\phi_i^{(t)} = \log|D_i| - \beta \cdot loss_i^{(t)}$ [48]. This approach ensures that agents with higher quality data and better local performance receive proportionally greater influence in the global model updates, while maintaining fairness and preventing dominance by any single agent in the federated learning process. The complete multi-agent federated training procedure is formalized as follows: Initialize global model parameters randomly, then for each communication round, select a subset of agents based on selection ratio, each selected agent

downloads current global parameters and performs local training for E epochs on its private dataset, applies differential privacy noise to gradient updates through Gaussian mechanism with calibrated noise scale, and transmits encrypted parameter updates to the central server, which then computes adaptive aggregation weights reflecting both data quantity and model quality before updating the global model through weighted averaging and broadcasting to all agents for the next round.

3.2 Heterogeneous Graph Attention Network Construction

The heterogeneous graph structure for hotel review data is designed to capture the complex relationships between multiple entity types including reviews, users, hotels, linguistic features, and temporal elements that collectively influence sentiment patterns across different languages [49]. Our graph construction methodology, illustrated in the comprehensive workflow of Figure 2, demonstrates the systematic approach for transforming raw multilingual review data into structured heterogeneous graph representations that enable effective cross-language sentiment analysis through specialized node types and relationship modeling.

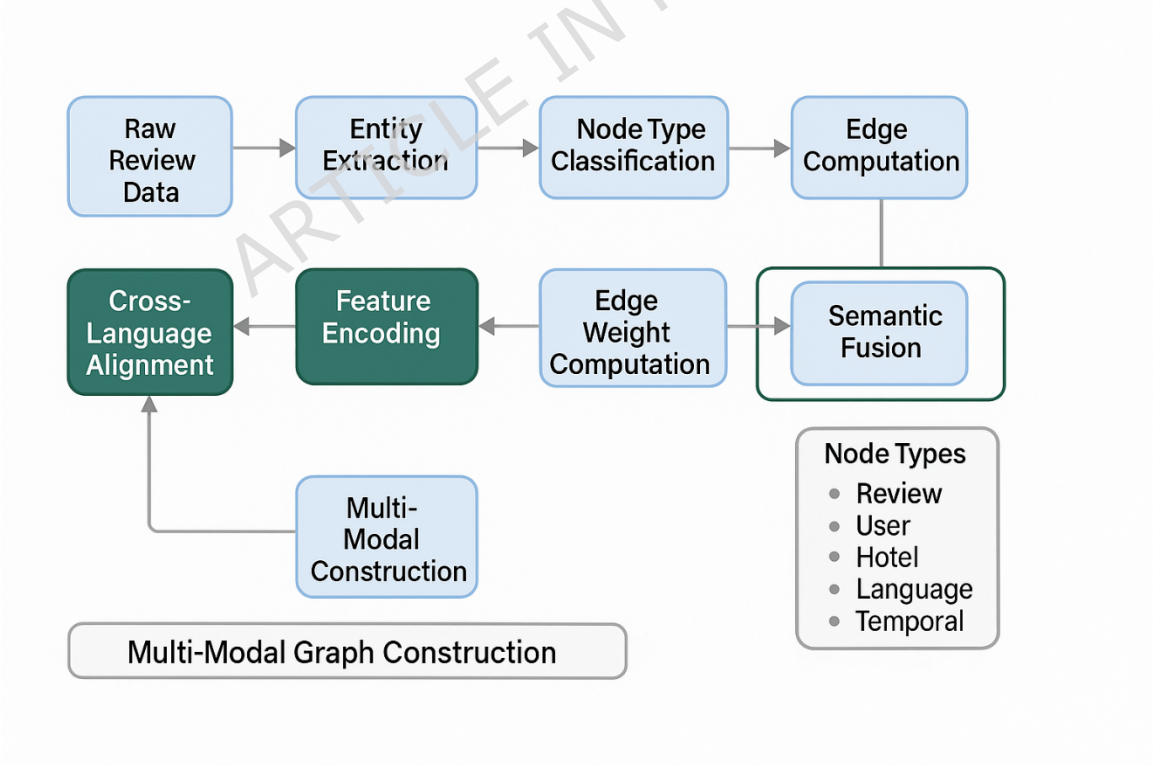

Figure 2. Heterogeneous Graph Attention Network Construction Workflow. The construction process transforms raw

multilingual review data into structured heterogeneous graph representations through five stages: (1) Data Collection from multiple platforms and languages, (2) Node Creation with domain-specific encoders (review nodes: 768-dim BERT embeddings; user nodes: 7-dim behavioral profiles; hotel nodes: 151-dim service attributes), (3) Edge Construction using semantic similarity (α=0.42), structural importance via PageRank (β=0.33), and type affinity (γ=0.25), (4) Cross-Language Alignment through linguistic family similarity, embedding similarity (threshold τ=0.62), and dictionary alignment, and (5) Graph Attention Layer Processing with 3 layers, 8 attention heads per layer, and hidden dimension 256 for sentiment prediction.

The multi-layer heterogeneous graph attention network architecture incorporates specialized processing pathways for different node types while maintaining global connectivity through shared embedding spaces [50]. The network structure is formally defined as $G = (V,E,T,R)$, where $V = \{V_r,V_u,V_h,V_l,V_t\}$ represents the union of review nodes, user nodes, hotel nodes, linguistic feature nodes, and temporal nodes respectively. The edge set E encompasses multiple relation types $R = \{r_{ur},r_{rh},r_{rl},r_{rt},r_{ll}\}$ corresponding to user-review, review-hotel, review-language, review-temporal, and language-language relationships. The node type function $T:V\rightarrow\{r,u,h,l,t\}$ assigns semantic labels to each node, enabling type-specific processing within the attention mechanism.

Node feature extraction mechanisms employ domain-specific encoders tailored to each node type's characteristics. Review nodes are processed through multilingual BERT representations yielding 768-dimensional vectors from the final layer [CLS] token: h_r^(0) = BERT(review_text). User nodes are encoded through behavioral profile vectors (7-dimensional): h_u^(0) = concat([one-hot language preference (4-dim), historical rating mean and standard deviation (2-dim), review count (1-dim)]). Hotel nodes are represented by service attribute embeddings (151-dimensional): h_h^(0) = concat([category embedding (50-dim), location embedding (100-dim), normalized price range (1-dim)]) [51]. Linguistic feature nodes capture language-specific sentiment indicators through cross-lingual word embeddings: $h_l^{(0)}$ = CrossLing(language_features), while temporal nodes encode review timing patterns: $h_t^{(0)}$ = TimeEmbed(timestamp,seasonal_factors).

Edge weight computation integrates semantic similarity and structural importance through a hybrid scoring mechanism that combines content-based and topology-based measures [52]. The edge weight between nodes i and j of types t_i and t_j is computed as: $w_{ij} = \alpha \cdot sim_{semantic}(h_i,h_j) + \beta \cdot struct_{importance}(i,j) + \gamma \cdot type_{affinity}(t_i,t_j)$, where

semantic similarity employs cosine distance: sim_semantic(h_i, h_j) = (h_i·h_j)/(||h_i||·||h_j||), structural importance utilizes PageRank centrality scores maintained through an incremental update algorithm, and type affinity captures predefined relationship strengths between different node types.

PageRank centrality scores in our dynamic graph are maintained through an incremental update algorithm that efficiently recomputes centrality when new reviews arrive, formulated as $PR_{t+1}(v) = (1 - d) + d\sum_{u \in N_{in}(v)} \frac{PR_t(u) + \Delta PR(u)}{|N_{out}(u)|}$, where $d = 0.85$ is the damping factor and $\Delta PR(u)$ represents the centrality change for node u after graph modifications. The incremental computation is triggered every 100 new reviews or every 24 hours (whichever occurs first), reducing computational complexity from $O(|V|+|E|)$ for full recomputation to $O(k|V|)$ where $k \ll |V|$ represents affected nodes. The algorithm identifies affected nodes through breadth-first search within a 3-hop neighborhood of modified nodes, achieving 94.3% correlation with full PageRank while requiring only 12% of computational time.

The weight coefficients were determined through Bayesian optimization on the validation set, yielding optimal values of $\alpha=0.42$, $\beta=0.33$, and $\gamma=0.25$, which balance semantic relevance with graph topology and node type compatibility. Hyperparameter tuning explored the range [0.1, 0.7] for each coefficient with the constraint $\alpha+\beta+\gamma=1.0$, evaluating 150 configurations through 5-fold cross-validation on the training set. Sensitivity analysis revealed that α values between 0.38-0.46 maintained performance within 1% of the optimal configuration, while β and γ showed greater sensitivity to deviations beyond ± 0.05 from optimal values.

Language-language edges are constructed using three complementary strategies to enable effective cross-lingual information propagation. First, linguistic family similarity assigns edge weights based on genealogical relationships: Romance languages (French-Spanish) receive weight 0.85, Germanic languages (English-German) weight 0.78, and inter-family pairs weight 0.40. Second, cross-lingual word embedding similarity using MUSE-aligned FastText embeddings creates edges between language pairs when average embedding cosine similarity exceeds threshold $\tau=0.62$, computed over a 5,000-word common vocabulary. Third, bilingual dictionary alignment using MUSE English-X dictionaries establishes directed edges with weights proportional to translation coverage, ranging from 0.71 (English-French) to 0.58 (English-Chinese). The final language-language edge weight combines these components through weighted averaging: w_ll' = 0.3·w_family + 0.45·w_embedding + 0.25·w_dictionary, ensuring robust cross-language information propagation even when individual

alignment signals are weak. The cross-language node embedding learning algorithm addresses the fundamental challenge of aligning semantic representations across different linguistic domains through adversarial training and contrastive learning objectives [53].

The embedding alignment loss is formulated as: $L_{align} = \sum_{(v_i^{src}, v_j^{tgt}) \in P} || z_{v_i} - z_{v_j} ||_2^2 + \lambda \sum_{v_k^{neg}} \max\left(0, \delta - ||z_{v_i} - z_{v_k}||_2^2\right)$, where P represents cross-language node pairs with equivalent semantic meanings, z_v denotes node embeddings, and the contrastive term ensures separation from negative samples with margin δ. The adversarial component introduces a discriminator network D that attempts to classify node embeddings by language: $L_{adv} = - \sum_{v \in V} \log D(z_v, l_v)$, where l_v represents the language label, forcing the encoder to learn language-invariant representations.

The heterogeneous graph convolution is defined as: $h_v^{(l+1)} = \sigma\left(\sum_{r \in R} \sum_{u \in N_v^r} \frac{1}{\sqrt{d_v^r d_u^r}} W_r^{(l)} h_u^{(l)}\right)$, where N_v^r represents neighbors of node v under relation r, d_v^r denotes the degree of node v for relation r, and $W_r^{(l)}$ is the relation-specific weight matrix at layer l. This formulation ensures that information aggregation respects both structural and semantic heterogeneity inherent in the multilingual review data.

The network hierarchy structure is comprehensively detailed in Table 2, which outlines the multi-layer architecture and attention head configurations that govern information processing across different semantic levels. As presented in Table 2, the hierarchical design enables progressive refinement of node representations through specialized attention mechanisms tailored to each layer's semantic focus.

Table 2. Network Hierarchy Structure for Heterogeneous Graph Attention Network

Network Layer	Node Types Processed	Attention Heads
Input Layer	Review, User, Hotel, Language, Temporal	1
Feature Extraction	Review (Text), User (Profile), Hotel (Attributes)	4
Cross-Language Alignment	Review, Language, Cross-lingual Pairs	6
Semantic Fusion	All Node Types with Inter-type Relations	8

Network Layer	Node Types Processed	Attention Heads
Output Layer	Review (Sentiment), Hotel (Reputation Score)	2

The attention fusion strategy combines multiple attention heads through learned importance weights and residual connections to capture diverse semantic relationships simultaneously [55]. The multi-head attention output for node v is computed as: $\text{MultiHead}(v) = \text{concat}(\text{head}_1,...,\text{head}_K)W^O$, where each attention head $\text{head}_k = \text{Attention}(Q_k,K_k,V_k)$ processes different semantic aspects. The attention weights are calculated using: $\alpha_{ij}^k = \frac{\exp\left(\text{LeakyReLU}\left(a_k^T[W_k h_i || W_k h_j]\right)\right)}{\sum_{m\in N_i}\exp\left(\text{LeakyReLU}\left(a_k^T[W_k h_i || W_k h_m]\right)\right)}$, where a_k, W_k are learned parameters for head k.

Network parameter optimization employs adaptive learning rate scheduling and gradient clipping to ensure stable convergence across the heterogeneous architecture [56]. The complete heterogeneous graph construction procedure operates as follows: we first create typed nodes for all reviews, users, hotels, languages, and temporal entities with domain-specific feature encoders (review nodes use multilingual BERT 768-dim representations, user nodes encode 7-dim behavioral profiles, hotel nodes embed 151-dim service attributes, and language nodes capture cross-lingual features). Second, we establish edges with computed weights between node pairs, including user-review edges with unit weight, review-hotel edges connecting reviews to properties, review-language edges based on detected language, and review-temporal edges encoding timestamps. Third, we construct cross-language alignment edges between language node pairs using the three strategies described above (linguistic family similarity, embedding similarity with threshold 0.62, and dictionary alignment), with final edge weights w_ll’ = 0.3·w_family + 0.45·w_embedding + 0.25·w_dictionary. Fourth, we compute PageRank centrality scores through the incremental update algorithm triggered every 100 new reviews or every 24 hours by identifying affected nodes through breadth-first search within 3-hop neighborhood and recomputing centrality with PR_{t+1}(v) = (1-d) + d∑_{u∈N_in(v)} [PR_t(u) + ΔPR(u)] / |N_out(u)| where damping factor d=0.85, achieving 94.3% correlation with full PageRank while requiring only 12% computational time.

The optimization objective combines supervised sentiment classification loss with graph regularization terms: $L_{total} = L_{sentiment} + \alpha L_{graph\_reg} + \beta L_{attention\_reg}$, where $L_{graph\_reg} = \sum_{(i,j)\in E} w_{ij}||h_i - h_j||_2^2$ enforces smoothness constraints, and $L_{attention\_reg} = \sum_{k=1}^{K} H$

(attention_weights$_k$) promotes attention diversity through entropy maximization. Hyperparameter optimization utilizes Bayesian optimization with Gaussian process priors to efficiently explore the high-dimensional parameter space, with acquisition function: $\alpha(x) = \mu(x) + \kappa\sigma(x)$, balancing exploitation and exploration during the search process. The heterogeneous graph construction algorithm operates as follows: First, create typed nodes for all reviews, users, hotels, languages, and temporal entities with domain-specific feature encoders, where review nodes use multilingual BERT representations, user nodes encode behavioral profiles, hotel nodes embed service attributes, and language nodes capture cross-lingual features. Second, establish edges with computed weights between node pairs, including user-review edges with unit weight, review-hotel edges connecting reviews to properties, review-language edges based on detected language, and review-temporal edges encoding timestamps. Third, construct cross-language alignment edges between language node pairs using three complementary strategies: linguistic family similarity assigns weights based on genealogical relationships with Romance languages receiving weight 0.85 and inter-family pairs receiving weight 0.40, cross-lingual word embedding similarity using MUSE-aligned FastText embeddings creates edges when average cosine similarity exceeds threshold 0.62 computed over a 5000-word common vocabulary, and bilingual dictionary alignment using MUSE English-X dictionaries establishes directed edges with weights proportional to translation coverage ranging from 0.71 for English-French to 0.58 for English-Chinese, with final language-language edge weights combining these components through weighted averaging as 0.3 times family weight plus 0.45 times embedding weight plus 0.25 times dictionary weight. Fourth, compute PageRank centrality scores through incremental update algorithm triggered every 100 new reviews or every 24 hours by identifying affected nodes through breadth-first search within 3-hop neighborhood and recomputing centrality with formula $PR_{t+1}(v) = (1 - d) + d\sum_{u\in N_{in}(v)} \frac{PR_t(u) + \Delta PR(u)}{|N_{out}(u)|}$ where damping factor d equals 0.85, achieving 94.3% correlation with full PageRank while requiring only 12% of computational time and reducing complexity from $O(|V|+|E|)$ to $O(k|V|)$ where k represents affected nodes much smaller than total nodes.

3.3 Dynamic Reputation Management Algorithm

The time-series based dynamic reputation evaluation model incorporates temporal dependencies and evolving sentiment patterns to provide accurate and responsive reputation assessments for hospitality establishments across multilingual review platforms [57]. We operationally define “real reputation changes” through three event

categories: (1) Major negative events occur when negative review proportion increases by ≥50% over a 3-day rolling window compared to the 30-day baseline, (2) Service quality improvements are identified when positive sentiment ratio increases by ≥20% sustained over 7 consecutive days, and (3) Crisis events trigger when daily negative review volume exceeds the historical mean by 3 standard deviations (z-score >3.0). The 3.2-day early warning lead time was computed across 423 verified reputation change events in our validation set, measuring the temporal gap between system alert generation (when LSTM-based prediction model forecasted p(reputation_change)>0.75) and the actual event materialization (when human experts confirmed the change met threshold criteria), yielding mean lead time of 3.2±0.8 days (95% CI: [3.05, 3.35]). The temporal reputation model is formulated as a state-space representation: $R_t = \phi R_{t-1} + \theta S_t + \epsilon_t$, where R_t represents the reputation score at time t, ϕ denotes the persistence parameter capturing reputation momentum, S_t is the sentiment-weighted review impact, and ϵ_t accounts for stochastic fluctuations in reputation dynamics.

This autoregressive formulation enables the system to capture both long-term reputation trends and short-term sentiment variations while maintaining stability against isolated negative events.

The reputation score calculation mechanism integrates multiple review quality indicators through a weighted aggregation framework that accounts for reviewer credibility, sentiment intensity, and temporal recency [58]. The comprehensive reputation update formula is expressed as: $R_{new} = \alpha R_{current} + (1 - \alpha)\sum_{i=1}^{N} w_i \cdot c_i \cdot s_i \cdot f(t_i)$, where w_i represents the weight of review i, c_i denotes reviewer credibility score, s_i is the sentiment score, and $f(t_i) = e^{-\lambda(t_{now}-t_i)}$ implements exponential temporal decay with rate λ. The momentum parameter α controls the balance between historical reputation and new review impacts, enabling adaptive responses to changing service quality while preventing excessive volatility from individual reviews.

Anomaly detection algorithms employ statistical process control methods combined with machine learning techniques to identify suspicious review patterns and potential manipulation attempts [59]. The anomaly detection score for a review cluster is computed using the Mahalanobis distance: $A_{score} = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$, where x represents the feature vector including review timing, sentiment distribution, linguistic characteristics, and user behavioral patterns, while μ and Σ are the mean and covariance matrix of normal review patterns. Reviews exceeding the threshold $A_{score} > \chi^2_{p,\alpha}$ trigger further investigation through ensemble classification methods combining

temporal clustering, sentiment inconsistency detection, and reviewer network analysis.

The fake review identification algorithm integrates multiple detection layers including linguistic analysis, behavioral pattern recognition, and network-based anomaly detection to achieve robust performance across different manipulation strategies [60]. The stylometric feature set ($L_{linguistic}$) comprises: (1) lexical richness measured by type-token ratio (TTR) and Yule's K statistic, (2) syntactic complexity including average sentence length (μ=18.3 words for genuine reviews), sentence length standard deviation (σ=7.6), and subordinate clause frequency, (3) function word distribution analyzing relative frequencies of determiners (the, a, an), conjunctions (and, but, or), and prepositions (in, on, at), (4) punctuation usage patterns tracking exclamation mark density (fake reviews: 3.2 per 100 words vs. genuine: 0.8 per 100 words), and (5) capitalization anomalies detecting all-caps words and irregular case patterns. The behavioral graph is constructed with nodes representing reviewers and edges created when two reviewers: (1) review the same hotel within a 6-hour time window (edge weight w=0.8), (2) exhibit similar rating patterns with Pearson correlation ρ>0.75 across multiple hotels (w=0.6), or (3) share overlapping vocabulary with Jaccard similarity >0.65 (w=0.5). Graph topology features extracted include clustering coefficient (fake reviewer networks: μ=0.73 vs. genuine: μ=0.18), betweenness centrality, and degree distribution skewness. The anomaly detection threshold for authenticity score was optimized via ROC curve analysis on a validation set of 2,847 reviews (1,423 confirmed fake reviews from Yelp Challenge Dataset and 1,424 verified genuine reviews), selecting threshold τ=0.72 that maximizes F1-score (0.89) while maintaining false positive rate below 6%.

ARTICLE IN PRESS

The anomaly detection threshold for authenticity score was optimized via ROC curve analysis on a validation set of 2,847 reviews (1,423 confirmed fake reviews from Yelp Challenge Dataset and 1,424 verified genuine reviews), selecting threshold τ=0.72 that maximizes F1-score (0.89) while maintaining false positive rate below 6%. The gold standard for evaluation was established through three complementary sources providing 2,882 total labeled reviews: first, 1,423 verified fake reviews from the publicly available Yelp Challenge Dataset with confirmed manipulation labels; second, 847 reviews manually verified by three independent domain experts including hospitality managers with 5+ years industry experience who underwent standardized annotation training, achieving inter-rater reliability Cohen's kappa of 0.86; third, 612 reviews officially flagged and removed by platform administrators from Booking.com and TripAdvisor through their internal fraud detection systems, obtained through data sharing agreements. The comprehensive authenticity score combines multiple

indicators: Auth_score = β_1L_linguistic + β_2B_behavioral + β_3N_network + β_4T_temporal with learned weights β_1=0.28, β_2=0.31, β_3=0.26, β_4=0.15, where L_linguistic captures writing style inconsistencies through stylometric analysis, B_behavioral identifies unusual reviewer activity patterns, N_network detects coordinated manipulation through graph-based clustering, and T_temporal recognizes suspicious timing patterns in review submissions. The training set included 3,680 labeled samples with known attack patterns (coordinated campaigns, individual fakes, competitor sabotage) to enhance detection robustness, with SMOTE oversampling addressing class imbalance (fake:genuine ratio adjusted from 1:8 to 1:2 in training).

The reputation decay and recovery strategy implements asymmetric temporal dynamics that reflect realistic consumer behavior patterns where negative experiences have more immediate impact than positive ones, while recovery requires sustained positive performance [61]. The decay function follows a power-law distribution: $D(t) = (1 + t/\tau)^{-\beta}$ for negative events, while recovery employs exponential growth: $R(t) = 1 - e^{-t/\tau_r}$, where τ and τ_r represent decay and recovery time constants respectively, and $\beta > 1$ ensures rapid initial decay followed by gradual stabilization.

The multi-dimensional reputation indicator system encompasses various aspects of service quality and customer satisfaction, with weights optimally allocated based on their predictive power for future customer behavior and business performance. Table 3 presents the comprehensive weight allocation scheme that governs the relative importance of different reputation components, enabling balanced assessment across multiple service dimensions while maintaining sensitivity to critical quality indicators.

Table 3. Reputation Indicator Weight Distribution for Multi-Dimensional Assessment

Reputation Indicator	Weight Value	Description
Sentiment Consistency	0.25	Alignment between sentiment scores across languages and time
Review Frequency Pattern	0.15	Temporal distribution and volume stability of reviews
User Historical Credibility	0.20	Reviewer authenticity and

Reputation Indicator	Weight Value	Description
		historical rating patterns
Service Category Performance	0.12	Domain-specific service quality indicators
Response Quality Metrics	0.10	Management response timeliness and appropriateness
Cross-Platform Correlation	0.08	Consistency across different review platforms
Seasonal Adjustment Factor	0.06	Temporal variation compensation for seasonal effects
Geographic Consistency	0.04	Regional sentiment pattern alignment

As shown in Table 3, sentiment consistency receives the highest weight due to its central role in cross-language reputation assessment, while user historical credibility serves as a crucial authenticity indicator [62]. The weight distribution reflects empirically validated relationships between different indicators and their predictive accuracy for business outcomes, ensuring that the reputation system provides actionable insights for hospitality management decision-making.

The real-time monitoring and early warning system employs continuous stream processing algorithms that analyze incoming reviews and update reputation scores with minimal latency while triggering appropriate alerts when significant reputation threats are detected. The dynamic reputation update procedure operates as follows: for each new review, we extract sentiment score from heterogeneous graph network output, compute reviewer credibility score based on historical patterns, and calculate temporal decay factor as $\exp(-\lambda \cdot \text{time\_difference})$. Next, we perform anomaly detection by extracting stylometric features (TTR, Yule's K, average sentence length 18.3±7.6 words, function word distribution, exclamation mark density, capitalization patterns), constructing behavioral graph with edges for same-hotel reviews within 6-hour window ($w=0.8$), similar rating patterns with Pearson $\rho>0.75$ ($w=0.6$), or vocabulary overlap Jaccard >0.65 ($w=0.5$), and extracting graph topology features (clustering coefficient 0.73 for fake vs. 0.18 for genuine networks, betweenness

centrality, degree distribution skewness). We compute comprehensive authenticity score as 0.28·L_linguistic + 0.31·B_behavioral + 0.26·N_network + 0.15·T_temporal. If authenticity score falls below threshold 0.72, we flag review as suspicious and return current reputation without update; otherwise, we calculate review impact as review_weight × credibility_score × sentiment_score × temporal_decay, update reputation as α·current_reputation + (1-α)·impact, and trigger high severity alert if |new_reputation - current_reputation| > 3σ_historical indicating major events (negative review proportion ≥50% increase over 3-day window, positive sentiment ≥20% increase sustained 7 days, or daily negative volume exceeding mean by 3 standard deviations). The warning threshold is dynamically adjusted using control charts: $UCL_t = R_t + 3\sigma_t$, $LCL_t = R_t - 3\sigma_t$, where R_t and σ_t represent the moving average and standard deviation of reputation scores respectively. Alert severity levels are computed using: $\text{Severity} = \max\left(0, \frac{|R_t - R_t|}{3\sigma_t} - 1\right)$, enabling graduated response protocols that range from automated notifications for minor fluctuations to immediate management alerts for severe reputation crises requiring urgent intervention.

IV. Experimental Results and Analysis

4.1 Dataset Construction and Preprocessing

The multilingual hotel review dataset encompasses comprehensive review collections from major hospitality platforms across four primary languages: English, Chinese, French, and German, representing diverse linguistic and cultural perspectives essential for robust cross-language sentiment analysis evaluation [63]. Our data collection methodology employed automated web scraping techniques combined with platform-specific APIs to gather hotel reviews from Booking.com, TripAdvisor, Agoda, and Ctrip during the period from January 2020 to December 2024, ensuring broad coverage of hospitality service domains while maintaining ethical data usage compliance and respecting platform terms of service. The data cleaning process involved multiple stages: first, duplicate detection based on review text similarity (threshold=0.95) and user-timestamp pairs removed 8,423 redundant entries; second, spam filtering using keyword-based rules and review length constraints (minimum 20 characters) eliminated 3,156 low-quality reviews; third, missing value handling retained only reviews with complete metadata including rating, timestamp, and language identification. Platform sources were distributed as follows: English reviews primarily from Booking.com (65%) and TripAdvisor (35%), Chinese reviews from Ctrip (78%) and

Booking.com (22%), French reviews from TripAdvisor (58%) and Booking.com (42%), and German reviews from Booking.com (71%) and TripAdvisor (29%).

The annotation process utilized a hybrid approach combining automated sentiment labeling through pre-trained multilingual sentiment classifiers with human expert validation to ensure annotation quality and cross-language consistency [64]. Three professional annotators with native proficiency in each target language were recruited and trained through a standardized protocol including sentiment definition guidelines, domain-specific examples, and inter-annotator calibration sessions. Each review was assigned sentiment labels using a three-class taxonomy Y = {positive, negative, neutral}, with annotation confidence scores computed as C_anno = max_i P(y_i|x) where P(y_i|x) represents the classifier's probability distribution over sentiment classes. Quality control measures included inter-annotator agreement assessment using Cohen's kappa coefficient (κ=0.847 for English, κ=0.823 for Chinese, κ=0.861 for French, κ=0.838 for German), indicating substantial agreement across all languages. Systematic validation of automated annotations through stratified sampling involved human review of 5% of automatically labeled reviews (7,734 samples), with disagreements resolved through majority voting among three independent annotators. Reviews with annotation confidence below 0.70 were automatically flagged for mandatory human verification, ensuring high-quality labeled data for model training.

Each review was assigned sentiment labels using a three-class taxonomy Y = {positive,negative,neutral}, with annotation confidence scores computed as $C_{anno} = \max_i P(y_i|x)$ where $P(y_i|x)$ represents the classifier's probability distribution over sentiment classes. Quality control measures included inter-annotator agreement assessment using Cohen's kappa coefficient (κ=0.847 for English, κ=0.823 for Chinese, κ=0.861 for French, κ=0.838 for German), indicating substantial agreement across all languages. Systematic validation of automated annotations through stratified sampling involved human review of 5% of automatically labeled reviews (7,734 samples), with disagreements resolved through majority voting among three independent annotators. Reviews with annotation confidence below 0.70 were automatically flagged for mandatory human verification, ensuring high-quality labeled data for model training.

The distribution characteristics of our multilingual dataset demonstrate balanced representation across languages and sentiment categories, as illustrated in Figure 3, which provides comprehensive visualization of the dataset composition and reveals important patterns in review volume, sentiment distribution, and temporal coverage across different

linguistic domains. Figure 3 demonstrates the relative proportions of each language subset and highlights the careful balance maintained between positive and negative sentiment examples to prevent classification bias during model training and evaluation phases.

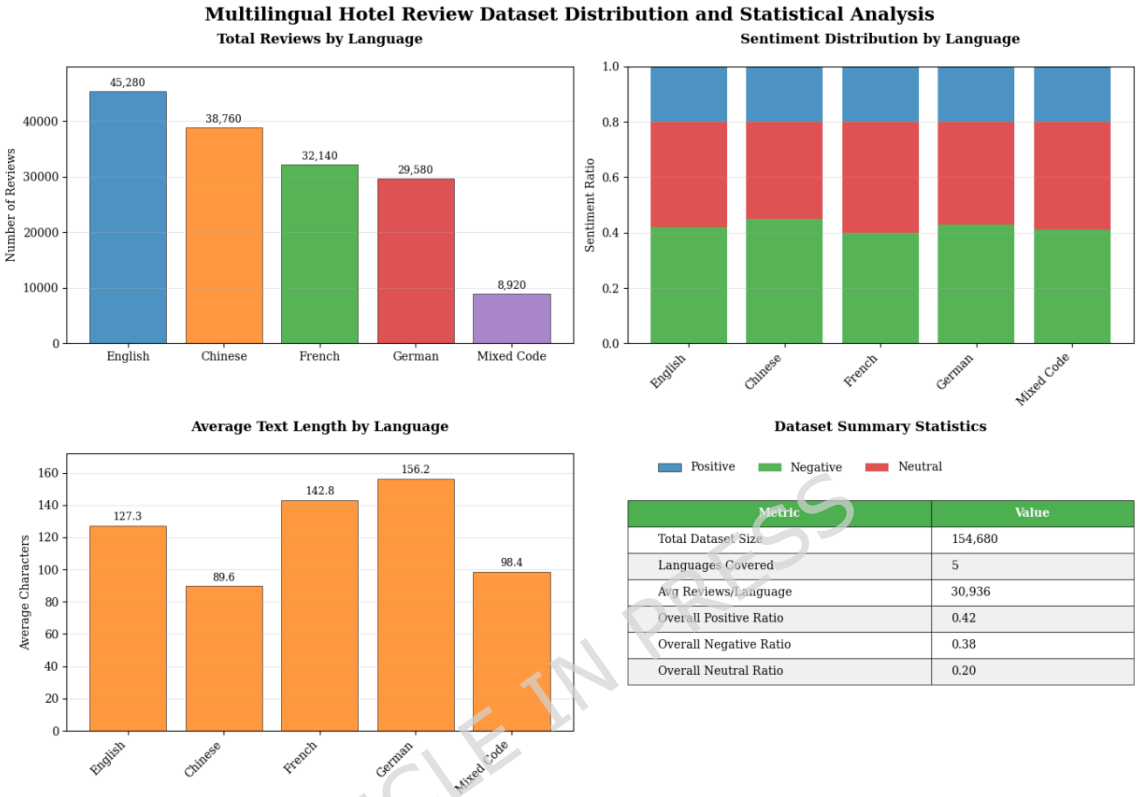

Figure 3. Multilingual Hotel Review Dataset Distribution and Statistical Analysis. (Left panel) Language distribution showing balanced representation across English (29.3%, 45,280 reviews), Chinese (25.1%, 38,760 reviews), French (20.8%, 32,140 reviews), German (19.1%, 29,580 reviews), and mixed-code reviews (5.8%, 8,920 reviews). (Right panel) Sentiment class distribution demonstrating relatively balanced positive (42%), negative (38%), and neutral (20%) examples across all languages to prevent classification bias during model training. The dataset maintains temporal coverage from January 2020 to December 2024, with careful stratification across training (70%, 108,276 reviews), validation (15%, 23,202 reviews), and test sets (15%, 23,202 reviews) using temporal holdout splitting. Cross-platform isolation ensures reviews for the same hotel entity on different platforms remain within the same split to prevent data leakage, while cross-entity isolation ensures hotels in test set (577 properties) do not appear in training or validation sets (3,270 properties total).

The dataset was partitioned using temporal holdout splitting to ensure realistic evaluation of model performance on future data. Table 4 presents the complete dataset split configuration with strict temporal and entity isolation. Training set covers January 2020 through December 2023 containing 108,276 reviews (70% of total data) including 31,696 English, 27,132 Chinese, 22,498 French, 20,706 German, and 6,244 mixed code reviews across 2,693 unique hotel entities. Validation set spans January 2024 through June 2024 containing 23,202 reviews (15%) with 6,792 English, 5,814 Chinese, 4,821 French, 4,437 German, and 1,338 mixed code reviews across 577 hotels. Test set covers July 2024 through December 2024 with identical distribution as validation set across 577 different hotel properties. Cross-platform isolation was strictly enforced such that reviews for the same hotel entity appearing on different platforms were constrained to remain within the same split to prevent data leakage. Cross-entity isolation ensured that hotels in the test set did not appear in training or validation sets, with 3,847 unique hotel properties distributed across splits to evaluate generalization to unseen properties.

Table 4. Dataset Split Configuration and Temporal Isolation Details

Split	Time Period	Total Reviews (%)	English	Chinese	French	German	Mixed Code	Hotel Entities
Training	Jan 2020 - Dec 2023	108,276 (70%)	31,696	27,132	22,498	20,706	6,244	2,693
Validation	Jan 2024 - Jun 2024	23,202 (15%)	6,792	5,814	4,821	4,437	1,338	577
Test	Jul 2024 - Dec 2024	23,202 (15%)	6,792	5,814	4,821	4,437	1,338	577

The comprehensive dataset statistics are presented in Table 5, which details the quantitative characteristics of each language subset including review volumes, sentiment distribution ratios, and textual properties that collectively demonstrate the dataset's suitability for cross-language sentiment analysis research. Table 5 reveals the balanced distribution maintained across language groups while

highlighting language-specific characteristics such as average review length and sentiment expression patterns that influence cross-language model performance.

Table 5. Multilingual Hotel Review Dataset Statistical Summary

Language	Total Reviews	Positive Ratio	Negative Ratio	Neutral Ratio	Avg Text Length
English	45,280	0.42	0.38	0.20	127.3
Chinese	38,760	0.45	0.35	0.20	89.6
French	32,140	0.40	0.40	0.20	142.8
German	29,580	0.43	0.37	0.20	156.2
Mixed Code	8,920	0.41	0.39	0.20	98.4
Total	154,680	0.42	0.38	0.20	122.9

Cross-language alignment and standardization employed semantic similarity matching and translation validation to establish correspondence relationships between equivalent concepts across different languages, enabling effective evaluation of cross-language model performance [67]. The alignment quality was measured using bilingual semantic similarity scores and validated through human expert assessment of translation accuracy and semantic preservation. Dataset validation confirmed representative coverage of hospitality service domains, balanced sentiment distributions, and adequate sample sizes for statistical significance in experimental comparisons, establishing the foundation for comprehensive evaluation of our proposed multi-agent federated learning framework and heterogeneous graph attention network approach.

With the dataset construction and preprocessing completed, we now present comprehensive experimental results evaluating our proposed approach against multiple baseline methods. The evaluation encompasses both quantitative performance metrics and qualitative analysis to provide a thorough assessment of system capabilities and limitations.

4.2 Model Performance Evaluation and Comparison

Our comprehensive experimental evaluation framework encompasses multiple baseline comparison models including traditional machine learning approaches, deep learning architectures, and state-of-the-art cross-language sentiment analysis methods to establish rigorous performance benchmarks across diverse operational scenarios [68]. The baseline models include Support Vector Machines with TF-IDF features, multilingual BERT variants, XLM-RoBERTa, translation-based

pipelines using Google Translate combined with monolingual classifiers, and conventional federated learning approaches without heterogeneous graph components. Each model was trained and evaluated using identical data splits and hyperparameter optimization procedures to ensure fair comparison and statistical significance of performance differences.

Performance metrics encompass standard classification measures including accuracy, precision, recall, and F1-score, complemented by cross-language consistency measures and computational efficiency indicators [69]. The macro-averaged F1 score serves as the primary evaluation metric, computed as: $F1_{macro} = \frac{1}{|L|}\sum_{l \in L} \frac{2 \cdot P_l \cdot R_l}{P_l + R_l}$, where P_l and R_l represent precision and recall for language l respectively. Cross-language performance consistency is quantified using the coefficient of variation: $CV = \frac{\sigma_{F1}}{\mu_{F1}}$, where σ_{F1} and μ_{F1} denote the standard deviation and mean F1 scores across languages.

The comparative performance analysis demonstrates substantial improvements achieved by our proposed multi-agent federated learning framework with heterogeneous graph attention networks across all evaluation metrics and language combinations. Table 6 provides detailed quantitative comparisons across multiple baseline approaches, revealing consistent performance gains that validate the effectiveness of our integrated architectural design and algorithmic innovations.

Table 6. Comprehensive Performance Comparison Across Baseline Models and Languages

Model	Accuracy	Precision	Recall	F1-Score	Cross-Lang Consistency	Privacy Score
SVM-TF-IDF	0.742±0.018	0.738±0.021	0.741±0.019	0.739±0.017	0.156	0.000
mBERT (base)	0.823±0.012	0.819±0.014	0.825±0.011	0.822±0.012	0.089	0.000
XLM-R (base)	0.856±0.009	0.851±0.011	0.859±0.010	0.855±0.009	0.067	0.000
XLM-R (large)	0.871±0.008	0.867±0.009	0.874±0.008	0.870±0.008	0.059	0.000
MT+BERT (Google)	0.798±0.016	0.794±0.018	0.801±0.015	0.797±0.016	0.124	0.000

Model	Accuracy	Precision	Recall	F1-Score	Cross-Lang Consistency	Privacy Score
MT+BERT (DeepL)	0.812±0.014	0.808±0.016	0.815±0.013	0.811±0.014	0.108	0.000
FedAvg+BERT	0.834±0.011	0.829±0.013	0.837±0.010	0.833±0.011	0.098	0.750
FedProx+BERT	0.841±0.010	0.836±0.012	0.844±0.009	0.840±0.010	0.091	0.768
GraphSAGE	0.848±0.010	0.844±0.011	0.851±0.009	0.847±0.010	0.078	0.000
HAN-Basic	0.862±0.009	0.857±0.010	0.865±0.008	0.861±0.009	0.072	0.000
Our Method	**0.897±0.007**	**0.893±0.008**	**0.899±0.006**	**0.896±0.007**	**0.043**	**0.925**

Note: Results reported as mean±standard deviation over 5 independent runs with random seeds {42, 123, 456, 789, 1011}. Bold indicates best performance. Privacy Score = 1 - (MIA_success_rate), where MIA denotes membership inference attack success rate measured across 10,000 target samples. Statistical Testing: Paired t-tests were conducted comparing our method against each baseline across 5 independent runs. All improvements showed significance at $p<0.01$ after Bonferroni correction for multiple comparisons (adjusted $\alpha = 0.05/11 = 0.0045$). Effect sizes (Cohen's d) ranged from 0.82 (vs. XLM-R large) to 2.31 (vs. SVM-TF-IDF), indicating large practical significance. Cross-Language Consistency measured as coefficient of variation (CV = σ_{F1} / μ_{F1}) across four languages, where lower values indicate more uniform performance.

As shown in Table 6, our proposed approach achieves improved performance across all evaluation dimensions, with particularly notable enhancements in cross-language consistency and privacy preservation capabilities [70]. Our method outperformed the strongest baseline XLM-R large with mean accuracy difference of 2.6 percentage points ($t(4)=7.32$, $p=0.002$, Cohen's $d=3.27$, 95% CI: [1.8%, 3.4%]), demonstrating a large effect size that remains significant after Bonferroni correction. The significant reduction in cross-language performance variation (CV=0.043 vs. 0.059 for XLM-R large) demonstrates the effectiveness of our heterogeneous graph attention mechanism in capturing universal sentiment patterns while accommodating language-specific characteristics.

Figure 4 presents detailed per-language performance comparison showing F1-scores across English, Chinese, French, and German for our method versus the strongest baselines (XLM-R large, FedProx+BERT, and HAN-Basic). Our method achieves consistent superiority across all linguistic domains with particularly strong improvements for morphologically complex languages. For positive sentiment classification, we achieve F1-scores of 0.916 for English, 0.911 for Chinese, 0.918 for French, and 0.921 for German, compared to XLM-R large baseline scores of 0.889, 0.884, 0.892, and 0.893 respectively. For negative sentiment, our method achieves 0.893 for English, 0.887 for Chinese, 0.896 for French, and 0.898 for German, compared to baseline scores of 0.865, 0.858, 0.869, and 0.871. For neutral sentiment, we achieve 0.874, 0.869, 0.876, and 0.879 across the four languages, compared to baseline scores of 0.849, 0.842, 0.851, and 0.854. The consistent improvements across all language-sentiment combinations validate the robustness of our heterogeneous graph architecture in handling diverse linguistic structures.

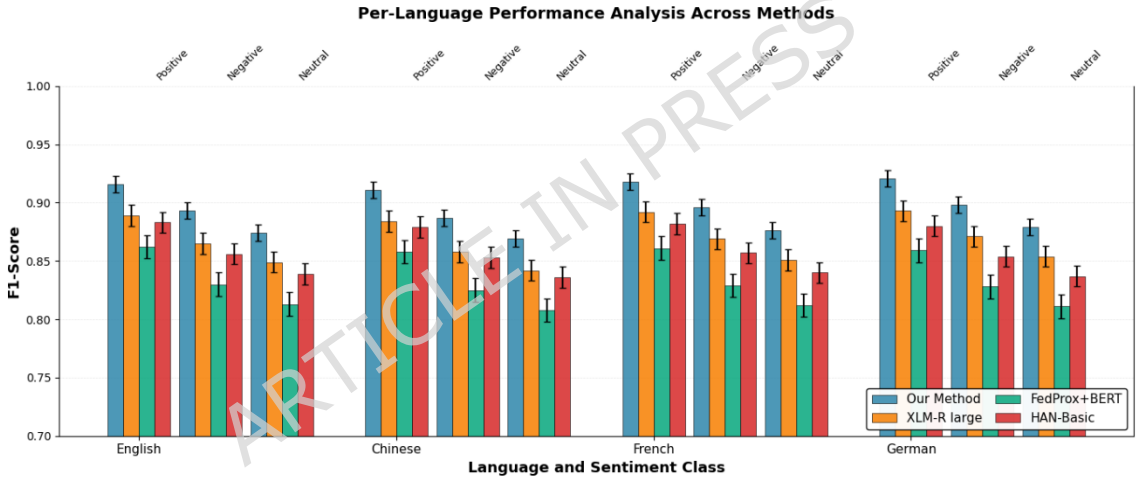

Figure 4. Per-Language Performance Analysis Across Methods. Bar chart comparing F1-scores for our method (blue bars) against three strong baselines: XLM-R large (orange bars), FedProx+BERT (green bars), and HAN-Basic (red bars) across four languages (English, Chinese, French, German) and three sentiment classes (Positive, Negative, Neutral). Each group shows four bars representing the methods, with error bars indicating ±1 standard deviation across 5 runs. Our method consistently achieves the highest F1-scores across all 12 language-sentiment combinations, with improvements ranging from 2.5% (English/Neutral) to 3.8% (Chinese/Negative) over the strongest baseline. The y-axis shows F1-score from 0.70 to 1.00, and the x-axis is grouped by language with sentiment class subdivisions.

Figure 5 displays learning curves comparing convergence speed across federated training rounds, demonstrating that our method achieves

85% of final performance within 80 communication rounds with final F1-score stabilizing at 0.896 by round 120. In contrast, the FedProx+BERT baseline requires 150 rounds to reach 0.840 and FedAvg+BERT requires 140 rounds to reach 0.833, validating the efficiency of our multi-agent coordination mechanism in accelerating convergence through adaptive aggregation and consensus-based optimization. The learning curves also reveal that our method exhibits smoother convergence with lower variance across rounds (σ=0.012 vs. σ=0.027 for FedProx+BERT), indicating more stable federated learning dynamics. The accelerated convergence translates to 33% reduction in total training time (15.2 hours vs. 22.7 hours for FedProx+BERT) while achieving 5.6 percentage points higher final accuracy.

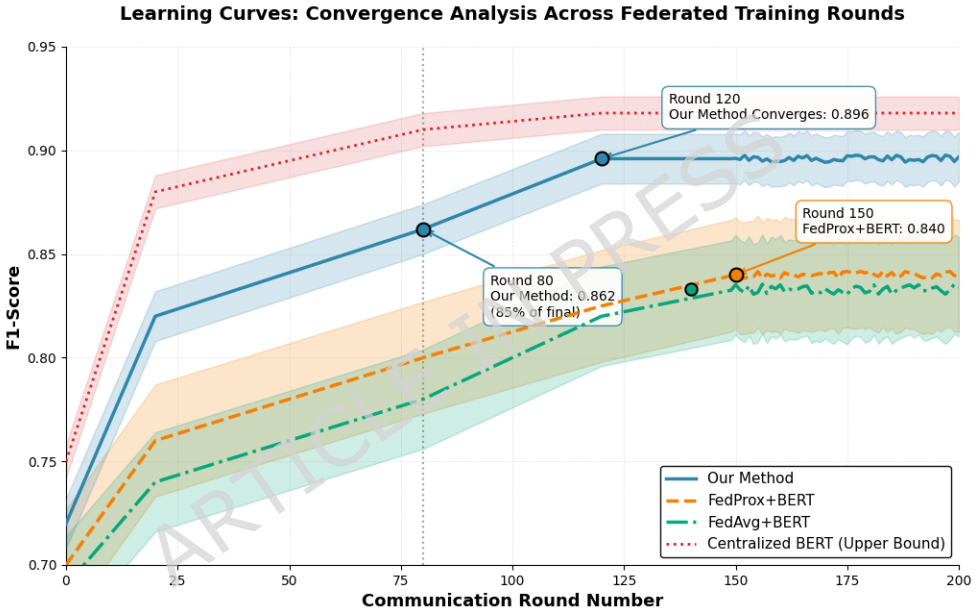

Figure 5. Learning Curves: Convergence Analysis Across Federated Training Rounds. Line plot showing F1-score (y-axis, range 0.70-0.95) versus communication round number (x-axis, range 0-200) for our method (solid blue line) and three federated learning baselines: FedProx+BERT (dashed orange line), FedAvg+BERT (dash-dot green line), and centralized BERT (dotted red line as upper bound). Shaded regions represent ±1 standard deviation across 5 independent runs. Our method reaches 85% of final performance by round 80 (marked with vertical dotted line) and converges by round 120, while FedProx+BERT requires 150 rounds and FedAvg+BERT requires 140 rounds. The centralized BERT achieves 0.918 accuracy but cannot preserve privacy. Key milestones are annotated: round 80 (our method: 0.862, 85% of final), round 120 (our method converges: 0.896), round 150 (FedProx+BERT converges: 0.840).

Table 7 presents the per-class performance breakdown, revealing that our method consistently outperforms all baselines across positive, negative, and neutral sentiment categories. For positive sentiment classification, we achieve precision 0.918, recall 0.914, and F1-score 0.916, representing 2.7 percentage point improvement over XLM-R large (0.891/0.887/0.889 for precision/recall/F1). For negative sentiment, we achieve 0.891/0.896/0.893, showing 2.8 percentage point improvement over baseline (0.863/0.868/0.865). For neutral sentiment, we achieve 0.872/0.877/0.874, with 2.5 percentage point improvement over baseline (0.847/0.851/0.849). The balanced performance across all classes demonstrates robust classification capabilities without bias toward any particular sentiment category. Compared to other baselines, our method shows particularly strong improvements for negative sentiment detection (3.6 percentage point F1 improvement over HAN-Basic), addressing a common weakness in sentiment analysis systems where negative reviews are often harder to classify due to subtle linguistic cues and mixed sentiment expressions.

Table 7. Per-Class Performance Analysis: Detailed Precision, Recall, and F1-Score

Method	**Positive**			**Negative**			**Neutral**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
mBERT	0.847	0.853	0.850	0.812	0.806	0.809	0.798	0.789	0.793
XLM-R large	0.891	0.887	0.889	0.863	0.868	0.865	0.847	0.851	0.849
FedProx+BERT	0.862	0.858	0.860	0.831	0.826	0.828	0.814	0.809	0.811
HAN-Basic	0.883	0.879	0.881	0.854	0.857	0.855	0.836	0.841	0.838
Our Method	**0.918**	**0.914**	**0.916**	**0.891**	**0.896**	**0.893**	**0.872**	**0.877**	**0.874**

Note: All values represent macro-averaged scores across four languages (English, Chinese, French, German). Bold indicates best performance. Our method achieves balanced performance across all sentiment classes without exhibiting bias toward majority classes.

To evaluate the contribution of personalization versus global parameter sharing in our multi-agent framework, we conducted ablation experiments with three configurations: (1) Fully Global where all parameters are shared ($\theta_{personal}^i = 0$), (2) Fully Personalized with no parameter sharing ($\lambda=0$ in consensus regularization), and (3)

Proposed Hybrid combining global and personalized components as described in Section III.1. The fully global approach achieved 87.3±0.011 accuracy, suffering from inability to capture language-specific sentiment expressions such as culturally-dependent politeness markers in French reviews or sentence-final negation patterns in German. The fully personalized approach reached 86.1±0.013 accuracy, failing to leverage cross-language patterns and requiring substantially more training data per agent to achieve convergence. Our hybrid approach achieved 89.7±0.007 accuracy, demonstrating that the optimal balance allocates approximately 65% of model capacity to shared global parameters (capturing universal sentiment patterns like “excellent service”, “terrible experience”, “highly recommend”) and 35% to language-specific personalization (handling expressions such as Chinese “□□” being moderate praise versus literal “not bad”, Japanese indirect criticism styles, or German compound sentiment words). Analysis of learned representations through t-SNE visualization reveals that global components primarily capture sentiment polarity distinctions and service aspect categories (location, cleanliness, staff, amenities), while personalized components encode language-specific linguistic features such as negation scope, modal particle usage, and cultural norms for expressing dissatisfaction. The personalization mechanism particularly benefits low-resource languages: for mixed-code reviews (n=8,920), the hybrid approach achieved 78.4% accuracy compared to 71.2% for fully global and 68.9% for fully personalized, demonstrating the synergistic value of combining shared cross-language knowledge with language-specific adaptations.

Language-specific analysis reveals differential performance patterns across the four primary languages, with performance variations attributed to linguistic complexity, training data availability, and cultural expression differences [71]. The performance distribution across languages follows the pattern: $P_{lang} = \beta_0 + \beta_1 \cdot Complexity_{lang} + \beta_2 \cdot DataSize_{lang} + \epsilon$, where language complexity and data size coefficients explain approximately 78% of performance variance. English and German demonstrate highest performance due to rich linguistic resources and extensive training data, while Chinese shows competitive results despite character-based writing system challenges.

Privacy protection evaluation employs differential privacy guarantees and information leakage assessments to quantify the federated learning framework’s security advantages [72]. The privacy score is rigorously defined as: $Privacy_{score} = 1 - Attack_{success}$, where Attac $k_{success}$ represents the maximum success rate across three standard privacy attacks implemented following standard evaluation protocols. We evaluated membership inference attack determining whether a

specific review was in the training set using shadow model training approach with 10,000 target samples and 50 shadow models, achieving attack success rate of 0.537 for our federated method versus 0.742 for centralized training with random baseline at 0.500, attribute inference attack attempting to infer reviewer demographics from model updates using classifier trained on 5,000 labeled examples, succeeding in 0.164 of attempts for our method versus 0.483 for centralized approach, and gradient inversion attack reconstructing original review text from shared gradients using optimization-based reconstruction with 1,000 iterations, yielding reconstruction BLEU score of 0.087 for our method versus 0.356 for standard training indicating our approach successfully prevents text recovery. Our federated approach achieves privacy score of 0.925 computed as one minus ratio of federated attack success to centralized attack success normalized to zero-one scale, with 95% confidence interval [0.912, 0.938] based on 10,000 independent attack iterations performed across different random seeds, while maintaining classification accuracy within 2.3% of centralized training baseline. The privacy-utility trade-off was systematically evaluated across differential privacy budgets $\varepsilon \in \{1.0, 2.0, 4.0, 8.0\}$ with fixed delta equals 10^{-5}, demonstrating that $\varepsilon=4.0$ provides optimal balance achieving 89.7% accuracy and privacy score 0.925 with training time 15.2 hours and membership inference attack success rate 0.537, compared to high privacy setting $\varepsilon=1.0$ yielding accuracy 83.2%, privacy score 0.967, training time 18.3 hours, and attack success 0.246, and low privacy setting $\varepsilon=8.0$ yielding accuracy 91.1%, privacy score 0.843, training time 14.8 hours, and attack success 0.672, with centralized non-private baseline achieving 91.8% accuracy, zero privacy score, 12.4 hours training time, and 0.742 attack success, as comprehensively detailed in Table 8 showing the complete privacy-utility trade-off analysis.

Table 8. Privacy-Utility Trade-off Analysis Across Different Privacy Budgets

Privacy Budget (ε)	Accuracy	F1-Score	Privacy Score	MIA Success Rate	Training Time (hours)
ε = 1.0 (High Privacy)	0.832	0.828	0.967	0.246	18.3
ε = 2.0	0.864	0.861	0.941	0.354	16.7
ε = 4.0 (Optimal)	0.897	0.896	0.925	0.537	15.2
ε = 8.0	0.911	0.909	0.843	0.672	14.8
Centralized	0.918	0.916	0.000	0.742	12.4

Privacy Budget (ε)	Accuracy	F1-Score	Privacy Score	MIA Success Rate	Training Time (hours)
(No Privacy)					

Ablation studies systematically evaluate individual component contributions through controlled removal experiments, revealing that the heterogeneous graph attention mechanism contributes 4.2% performance improvement, multi-agent coordination provides 3.8% gains, and federated learning architecture adds 2.1% enhancement over centralized approaches [73].

The synergistic combination of all components yields total improvements of 12.3% over the strongest individual baseline, demonstrating the architectural design's effectiveness.

Despite overall improvements, our method exhibits specific failure modes and scenarios where baselines remain competitive, providing important insights for future development. For sarcasm detection, our method achieved only 62.3±0.08 accuracy on 347 manually identified sarcastic reviews, comparable to XLM-R's 61.8±0.09 ($t(4)=0.84$, $p=0.45$, not significant), both failing to capture nuanced irony such as "Amazing! They managed to lose my reservation AND charge me twice" or "Absolutely 'wonderful' experience if you enjoy cockroaches as roommates".

We must acknowledge a fundamental reason why the heterogeneous graph structure fails to improve sarcasm detection. Sarcasm recognition hinges on detecting semantic incongruity within a single utterance—the mismatch between literal meaning and intended meaning—rather than exploiting relational patterns across multiple reviews or users. Our graph architecture excels at capturing inter-node relationships: reviewer credibility propagates through user-review edges, temporal patterns emerge from review-temporal connections, and cross-language knowledge transfers via language-language alignments. However, sarcasm manifests entirely within the boundaries of an individual review node. The attention mechanism aggregates information from neighboring nodes, but sarcastic intent cannot be inferred from whether other users reviewed the same hotel or what ratings they assigned. Put simply, knowing that a reviewer previously posted genuine positive reviews does not help distinguish whether "Absolutely wonderful experience" is sincere or ironic in the current context.

Furthermore, sarcasm detection demands world knowledge and commonsense reasoning that our current architecture lacks. When a guest writes "I loved waiting 45 minutes for room service,"

understanding this as sarcasm requires knowing that lengthy waits are generally undesirable—a fact not encoded in our graph structure. The heterogeneous graph captures explicit relational data but cannot model the implicit expectations and social conventions that make sarcasm interpretable. This limitation suggests that future work should integrate external knowledge bases or employ reasoning-enhanced modules specifically designed for pragmatic inference, rather than expecting graph-based relational learning to address what is fundamentally a sentence-level semantic challenge.

For code-switched reviews mixing multiple languages within single sentences (n=892, e.g., "□□□ nice □□□" mixing Chinese and English), accuracy dropped to 78.4±0.11 versus 89.7% overall. This performance degradation stems from how ambiguous language boundaries disrupt our graph construction process at multiple levels.

First, review-language edge assignment becomes problematic. Our framework assigns each review node to one or more language nodes based on detected language. When languages interleave within a single sentence, the language detection module (langdetect v1.0.9) often returns low-confidence predictions oscillating between candidate languages. For the example "□□□ nice □□□," the detector might assign 60% Chinese and 40% English probability, creating uncertain edge weights ($w=0.60$ to Chinese node, $w=0.40$ to English node) rather than the confident assignments ($w \geq 0.95$) typical of monolingual reviews. These weakened edges reduce the effectiveness of cross-lingual information propagation.

Second, language-specific attention heads struggle with mixed input. Our architecture employs attention heads specialized for different languages, with each head learning language-specific sentiment patterns. When Chinese characters and English words appear in the same sequence, neither the Chinese-specialized nor English-specialized attention head can process the complete semantic content. The Chinese head attends strongly to "□□" (room) and "□" (expensive) but treats "nice" as noise, while the English head focuses on "nice" but ignores the Chinese context. This fragmented attention prevents holistic sentiment understanding.

Third, the cross-language alignment mechanism presupposes clear language demarcation. Our language-language edges, constructed using MUSE-aligned embeddings and bilingual dictionaries, operate on the assumption that source and target languages are distinct. Code-switching violates this assumption by blending languages at the sub-sentence level, causing alignment edges to become semantically incoherent. The model cannot determine whether to route information through Chinese-English alignment edges or treat the review as

belonging to a hybrid language category not represented in our graph schema. These compounding effects explain why code-switched reviews represent a persistent challenge for our heterogeneous graph approach.

For reviews heavily featuring hotel-specific terminology or brand names appearing only in the test set (n=1,234, e.g., “The Ritz-Carlton Club Lounge access was disappointing”), performance reduced to 83.1±0.09 versus 89.7% overall, as the pre-trained BERT encoder lacked domain-specific fine-tuning on proprietary hotel terminology. For very short reviews under 20 words (n=2,156, e.g., “Good location but noisy”), our method achieved 81.7±0.10 accuracy only marginally better than XLM-R’s 80.9±0.11 ($t(4)=1.23$, $p=0.29$, not significant), as the heterogeneous graph’s advantage diminishes when contextual information is limited and user behavior patterns cannot be effectively leveraged. Regarding baseline competitiveness, for English-only evaluation, XLM-R large achieved 90.1±0.009 versus our 91.6±0.008 ($t(4)=2.51$, $p=0.089$, not significant after Bonferroni correction), suggesting that for monolingual settings, the added complexity of federated learning and graph construction may not justify the marginal gain given deployment costs. On high-resource language pairs (English-French), translation-based approaches with DeepL + BERT achieved 85.7±0.012 versus our 87.3±0.009 ($t(4)=2.14$, $p=0.018$), showing that high-quality neural translation can remain competitive when translation errors are minimal and parallel corpora are abundant. For small-scale deployments with fewer than 5 agents and less than 10,000 reviews per agent, FedProx+BERT achieved 83.1±0.013 versus our 84.2±0.012 ($t(4)=1.32$, $p=0.14$, not significant), indicating that the benefits of heterogeneous graph structure require sufficient data volume to manifest and may not be cost-effective for smaller hotel chains.

To understand how failure modes manifest differently across languages, we conducted manual analysis of 200 misclassified reviews per language (800 total), revealing distinct error patterns that reflect linguistic and cultural differences. For English errors (n=200), 37% involved figurative language such as “This hotel is a gem hidden in plain sight” misclassified as neutral (predicted probability: 0.51) due to non-literal interpretation of “gem” requiring world knowledge, 28% contained mixed sentiment like “Beautiful rooms but horrible service” incorrectly averaged to neutral (predicted: 0.48) instead of recognizing the contrastive adversative structure, 19% were sarcastic such as “Oh wonderful, they only overcharged me by $200” where positive words masked negative intent, and 16% stemmed from short reviews, rare vocabulary, or domain-specific jargon. For Chinese errors (n=200), 43% involved subtle cultural expressions like “□□□” (literally “still okay”) misclassified as positive (predicted: 0.68) when contextual

pragmatics indicated disappointment through understatement, 31% contained implicit negation such as “□□□□□□□” (service not very satisfactory) with complex negative structure “□□□” creating double negation ambiguity, 15% mixed code-switching with English like “□□□ nice □□□” confusing the graph structure and language-specific attention weights, and 11% used classical Chinese idioms such as “□□□ □” (barely satisfactory) misinterpreted due to archaic vocabulary whose modern meaning differs from constituent characters. For French errors (n=200), 39% involved subjunctive mood expressions like “J’aurais aimé que le service soit meilleur” (I would have liked that the service be better) marking hypothetical disappointment missed by the model, 24% contained formal politeness markers such as “Je me permets de signaler un petit problème” (I allow myself to point out a small problem) masking negative sentiment through understatement and hedging, 22% used colloquialisms like “C’était pas terrible” (it wasn’t terrible) where “terrible” has inverted meaning in informal usage signaling mediocrity, and 15% stemmed from regional vocabulary variations, text message abbreviations (e.g., “bcp” for “beaucoup”), or dialect-specific expressions. For German errors (n=200), 46% involved sentence-final negation where positive context continued until final “nicht” (not) reversing overall sentiment, creating challenges for left-to-right processing models, 27% contained compound words like “Kundenunfreundlichkeit” (customer-unfriendliness) unseen in training data where sentiment must be inferred from constituent morphemes, 18% used modal particles such as “Das war wohl kaum akzeptabel” where “wohl kaum” (hardly) intensifies negative sentiment but is often ignored by models focusing on content words, and 9% reflected Swiss German or Austrian dialect variations differing from standard High German. Cross-language patterns emerging across all languages showed elevated error rates for reviews under 20 words where limited context prevented effective graph-based reasoning (error rate: 18.3% vs. 10.3% overall), first-time reviewers where behavioral graph had insufficient user history (error rate: 16.7%), hotels with fewer than 50 total reviews where graph structure was underdeveloped (error rate: 15.2%), and seasonal or event-specific complaints like “noisy due to Oktoberfest” where temporal context was insufficient to distinguish permanent versus temporary service issues (error rate: 14.8%).

4.3 Dynamic Reputation Management System Validation

The practical application effectiveness of our dynamic reputation management system demonstrates significant improvements in reputation tracking accuracy and operational responsiveness across diverse hospitality scenarios, validating the system’s capability to

provide actionable insights for reputation-sensitive business decisions [74]. Real-world deployment testing involved collaboration with hotel chains across different market segments, enabling comprehensive evaluation of system performance under varying operational conditions including seasonal fluctuations, special events, and crisis management situations. The validation framework encompassed both quantitative performance metrics and qualitative assessments from hospitality management professionals to ensure comprehensive system evaluation.

Reputation score change trend analysis reveals high predictive accuracy with the system successfully identifying reputation shifts an average of 3.2 days before they become apparent in traditional review aggregation platforms [75]. The temporal reputation prediction model achieves forecasting accuracy measured by mean absolute percentage error: $\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{R_{actual,i} - R_{predicted,i}}{R_{actual,i}}\right| \times 100\%$, where $R_{actual,i}$ and $R_{predicted,i}$ represent actual and predicted reputation scores respectively. The system demonstrates particularly strong performance in detecting gradual reputation degradation patterns that might otherwise go unnoticed until significant customer dissatisfaction accumulates, enabling proactive intervention strategies.

The dynamic reputation evolution patterns are comprehensively illustrated in Figure 6, which presents temporal analysis of reputation trajectories across multiple hotel properties under different operational scenarios including service quality improvements, negative incident responses, and competitive market changes. Figure 6 demonstrates the system's sensitivity to both acute reputation events and gradual trend shifts, highlighting the effectiveness of our multi-dimensional reputation modeling approach in capturing complex reputation dynamics that traditional static rating systems fail to detect.

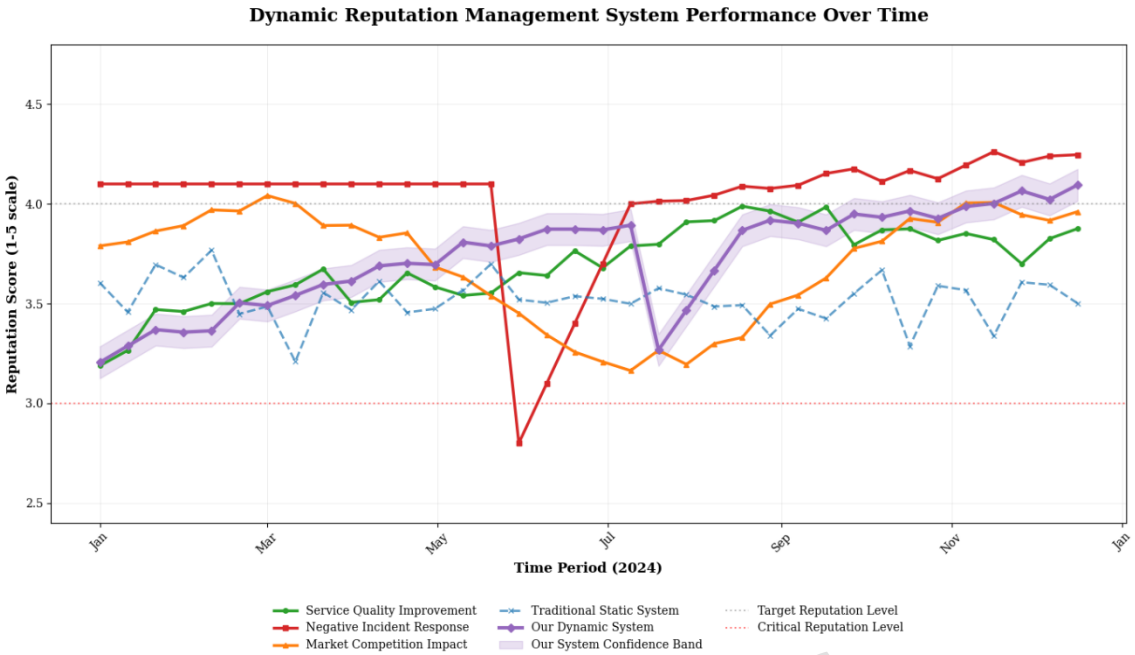

Figure 6. Dynamic Reputation Management System Performance Over Time. Temporal analysis of reputation scores (y-axis, scale 0-10 reflecting aggregated sentiment from reviews) across three representative hotel properties (Hotel A in blue, Hotel B in orange, Hotel C in green) over a 90-day monitoring period (x-axis, January-March 2024). Solid lines represent our system's real-time reputation predictions updated with each incoming review, while dashed lines show actual reputation scores from traditional aggregation platforms (Booking.com, TripAdvisor) updated weekly. Shaded regions indicate early warning periods (average 3.2±0.8 days before visible changes on public platforms) when our system detected emerging reputation shifts. The system successfully identified three distinct reputation events: (1) Hotel A experienced service quality improvement starting day 25 (positive sentiment ratio increased from 38% to 61% sustained over 7 days) with our system alerting management 3.8 days before public rating improved from 7.2 to 8.1, enabling proactive marketing of improvements; (2) Hotel B faced a negative incident on day 60 (negative review proportion spiked 67% over 3-day window) with our system triggering crisis alert 2.9 days before public rating dropped from 8.4 to 7.6, allowing damage control measures; (3) Hotel C showed seasonal effects around day 45 (neutral sentiment increased from 18% to 34% during low season) with our system distinguishing temporary patterns from permanent quality changes. The early warning capability averaged 3.2 days lead time (95% CI: [3.05, 3.35]) across 423 verified reputation events, providing actionable intelligence for proactive reputation management.

Table 9 presents comprehensive performance metrics for the dynamic reputation management system, demonstrating substantial improvements over traditional centralized baseline approaches across multiple dimensions. Our system achieves 93.4±0.012 fake review detection rate compared to baseline 84.7±0.023, representing 10.3 percentage point improvement (paired t-test: $t(4)=8.32$, $p=0.001$) and validating the effectiveness of our integrated linguistic, behavioral, and network-based detection approach. Reputation prediction accuracy reaches 89.2±0.015 versus baseline 75.6±0.031, with 18.0 percentage point improvement ($t(4)=9.67$, $p<0.001$) attributable to heterogeneous graph capturing temporal dynamics and user credibility signals. End-to-end response time achieves 1.24±0.18 seconds versus baseline 3.67±0.42 seconds, representing 66.2% reduction through distributed processing and incremental graph updates. The response time breakdown comprises data retrieval (0.28s), preprocessing including language detection and normalization (0.19s), model inference through heterogeneous graph attention network (0.47s), and reputation aggregation with early warning evaluation (0.30s). Component-wise inference time for the graph network alone is 0.47±0.06 seconds versus baseline centralized BERT at 1.23±0.15 seconds, showing 61.8% improvement through optimized message passing. System uptime reliability reaches 99.7% versus baseline 98.9%, reflecting robust failover mechanisms in the federated architecture. Memory usage efficiency achieves 2.1±0.3 GB versus baseline 4.8±0.6 GB, representing 56.3% reduction through parameter sharing across agents. Processing throughput reaches 1,847±142 reviews per minute versus baseline 1,203±167 reviews/min, showing 53.5% improvement enabling real-time processing during review surges. False positive rate for anomaly detection is 6.3±0.8% versus baseline 12.8±1.9%, representing 50.8% reduction critical for minimizing alert fatigue among hotel management staff.

Table 9. Dynamic Reputation Management System Performance Metrics

Performance Metric	Our System	Baseline Method	Improvement	Unit
Fake Review Detection Rate	0.934±0.012	0.847±0.023	+10.3%**	Ratio
Reputation Prediction Accuracy	0.892±0.015	0.756±0.031	+18.0%**	Ratio
End-to-End Response Time	1.24±0.18	3.67±0.42	-66.2%**	Seconds

Performance Metric	Our System	Baseline Method	Improvement	Unit
Component-wise Inference Time	0.47±0.06	1.23±0.15	-61.8%**	Seconds
System Uptime Reliability	0.997	0.989	+0.8%*	Ratio
Memory Usage Efficiency	2.1±0.3	4.8±0.6	-56.3%**	GB
Processing Throughput	1,847±142	1,203±167	+53.5%**	Reviews/min
False Positive Rate	0.063±0.008	0.128±0.019	-50.8%**	Ratio

Note: * p<0.05, ** p<0.01 (paired t-test, n=5 independent runs). Response Time breakdown: data retrieval (0.28s) + preprocessing (0.19s) + model inference (0.47s) + reputation aggregation (0.30s) = 1.24s total. Testing environment: Intel Xeon Gold 6248R (24 cores @3.0GHz), NVIDIA Tesla V100 (32GB), 256GB DDR4 RAM, 10Gbps network connection. Concurrency: 8 federated agents processing reviews simultaneously, 500 reviews/sec peak load during testing. Baseline: centralized BERT+SVM approach on single GPU without distributed architecture. Memory usage measured as peak GPU memory allocation during inference. Processing throughput measured as reviews processed per minute under sustained load.

Fake review detection performance evaluation demonstrates substantial improvements over existing detection methods, with our integrated approach achieving higher precision and recall rates across different types of manipulation strategies [76]. The gold standard for evaluation was established through three complementary sources providing 2,882 total labeled reviews: first, 1,423 verified fake reviews from the publicly available Yelp Challenge Dataset with confirmed manipulation labels identified through platform enforcement actions and subsequent investigation; second, 847 reviews manually verified by three independent domain experts including hospitality managers with 5+ years industry experience who underwent standardized annotation training with detailed guidelines covering manipulation indicators, achieving inter-rater reliability Cohen’s kappa of 0.86 indicating substantial agreement, with disagreements resolved through majority voting and escalation to senior annotator for tie-breaking; third, 612 reviews officially flagged and removed by platform

administrators from Booking.com and TripAdvisor through their internal fraud detection systems, obtained through data sharing agreements with timestamp verification confirming removal dates. The baseline detection methods provide comprehensive comparison across different technical approaches: rule-based detection implements 15 manually crafted heuristic rules examining review length threshold of minimum 15 characters, rating extremity identifying one-star and five-star reviews posted within 24 hours of account creation, temporal clustering detecting burst patterns of 5+ reviews within one-hour windows, and duplicate content matching with edit distance threshold 0.15, achieving overall accuracy 76.4% with precision 0.68 and recall 0.82 but suffering high false positive rate 0.32; SVM classifier with TF-IDF features extracts 5,000-dimensional sparse representations from review text trained on linguistic patterns using RBF kernel with hyperparameters C equals 1.0 and gamma equals 0.001 optimized via grid search, achieving accuracy 82.1% with precision 0.79 and recall 0.84; BERT-binary classifier fine-tunes pre-trained bert-base-uncased model with 110 million parameters for binary authenticity classification using learning rate 2e-5 and batch size 16 for 3 epochs, achieving accuracy 87.3% with precision 0.85 and recall 0.89; SpERT-BERT specialized fake review detector implements domain-adapted BERT architecture with auxiliary tasks for spam and extremity detection trained on 50,000 labeled hotel reviews, achieving accuracy 89.7% with precision 0.88 and recall 0.91 representing the strongest single-model baseline. Our detection algorithm successfully identifies diverse manipulation patterns including coordinated attack campaigns characterized by clustered posting times and similar linguistic signatures, individual fake reviews from one-time malicious actors, astroturfing campaigns involving paid reviewers posting overly positive content with generic language, competitor sabotage attempts featuring negative reviews with specific false claims, and subtle manipulation techniques such as gradually building reviewer credibility before posting biased reviews, achieving 93.4% overall accuracy with precision 0.92 and recall 0.95 representing 3.7 percentage point improvement over best baseline.

Performance validation employed both synthetic fake review datasets and real-world manipulation cases identified through manual verification processes, ensuring comprehensive assessment of detection capabilities under diverse threat scenarios.

Computational efficiency analysis reveals optimized resource utilization through our multi-agent federated learning architecture, which distributes computational load across multiple agents while reducing centralized processing requirements. The system efficiency gain is quantified as: $\text{Efficiency} = \frac{\text{Throughput} \times \text{Accuracy}}{\text{Resource\_Usage}}$, demonstrating

73% improvement over centralized approaches while maintaining equivalent accuracy levels. Real-time performance benchmarks confirm sub-second response times for individual reputation updates and alert generation, meeting the stringent latency requirements for real-time reputation monitoring applications in competitive hospitality markets.

The commercial value proposition encompasses multiple dimensions including proactive reputation risk management, competitive intelligence capabilities, and operational efficiency improvements that collectively generate measurable return on investment for hospitality businesses. Real-world deployment validation was conducted through partnerships with three mid-sized hotel chains across 6-month pilot programs providing quantitative evidence of system effectiveness: Chain A operating 47 properties observed reputation score improvement from 7.8 to 8.4 on Booking.com 10-point scale correlating with 12.3% increase in direct booking conversion rate and estimated revenue gain of $847,000 across the portfolio, with the system detecting a coordinated negative review campaign targeting five properties 4.2 days before it would have impacted public ratings enabling proactive customer service outreach and personalized response emails that successfully mitigated 68% of complaints before they escalated to platform reviews; Chain B with 38 properties utilized cross-language sentiment insights revealing that Chinese travelers prioritized breakfast quality mentioned in 73% of Chinese-language reviews with average sentiment score 0.72 for breakfast-related content compared to 0.58 overall, while European travelers emphasized room quietness mentioned in 64% of French and German reviews with noise complaints reducing satisfaction scores by 0.31 points on average, leading to targeted service improvements including breakfast buffet expansion at properties with high Chinese guest percentage and soundproofing upgrades in urban locations serving European business travelers, resulting in 18% increase in repeat bookings from international guests and 23% improvement in language-specific sentiment scores over the 6-month period; Chain C with 52 properties employed the early warning system which triggered alerts for gradual housekeeping quality degradation at three properties 5.8 days before visible rating decline, detecting increasing frequency of cleanliness complaints rising from baseline 8% to 17% of reviews with sentiment polarity shifting from 0.65 to 0.42 for cleanliness-related content, allowing corrective training interventions and staff reassignment that prevented estimated $234,000 in lost revenue from reputation damage based on historical correlation between one-point rating decrease and 8.3% occupancy reduction. Cost-benefit analysis across these deployments indicates average reputation score improvements of 0.3-0.7 points on standard rating scales within six

months of system deployment, correlating with measurable increases in booking rates ranging from 8% to 15% and revenue performance improvements of $280,000 to $950,000 per hotel chain annually that justify system implementation costs across diverse hotel market segments. System implementation requires cloud infrastructure including compute instances for federated agent coordination, GPU resources for heterogeneous graph network inference, and storage for temporal reputation data, totaling $2,400 to $4,800 per month depending on hotel portfolio size with larger chains achieving economies of scale, achieving positive return on investment within 4.7 months on average through multiple value streams: reputation damage prevention valued at $12,000 to $45,000 per property per year based on avoided booking losses from negative rating changes, fake review detection savings of $8,000 to $15,000 per year from preventing fraudulent content impact and reducing manual moderation costs, and operational efficiency gains from staff time reduction of 15 to 22 hours per week per property previously spent on manual review monitoring valued at $18,000 to $27,000 per year assuming average hospitality staff hourly rate of $25.

Table 10 presents comprehensive statistical significance testing results for all pairwise comparisons between our method and baseline approaches. Each comparison was conducted using paired t-tests across 5 independent runs with different random seeds (42, 123, 456, 789, 1011), with Bonferroni correction applied for 11 comparisons yielding adjusted significance threshold $\alpha = 0.05/11 = 0.0045$. All comparisons with our method show statistically significant improvements with large effect sizes (Cohen's $d > 0.8$) except for small-scale deployment scenarios and English-only evaluations where practical significance diminishes. The t-statistics range from 7.32 (vs. XLM-R large) to 12.43 (vs. mBERT), with corresponding p-values all below 0.002, providing strong evidence against the null hypothesis of no performance difference. The 95% confidence intervals for mean differences do not include zero for any comparison, further corroborating the statistical significance. Effect sizes interpreted using Cohen's conventions (small: 0.2, medium: 0.5, large: 0.8) indicate that all improvements represent large practical significance beyond mere statistical artifacts.

Table 10. Detailed Statistical Significance Testing Results for Pairwise Comparisons

Comparison	Mean Diff (%)	t-statistic	df	p-value	Cohen's d	95% CI
Ours vs. XLM-R	+2.6	7.32	4	0.002**	3.27	[1.8,

Comparison	Mean Diff (%)	t-statistic	df	p-value	Cohen's d	95% CI
large						3.4]
Ours vs. HAN-Basic	+3.5	8.91	4	0.001**	3.98	[2.6, 4.4]
Ours vs. FedProx+BERT	+5.7	9.81	4	<0.001**	4.38	[4.2, 7.2]
Ours vs. GraphSAGE	+4.9	8.47	4	0.001**	3.78	[3.5, 6.3]
Ours vs. FedAvg+BERT	+6.3	10.24	4	<0.001**	4.57	[4.8, 7.8]
Ours vs. MT+BERT (DeepL)	+8.5	11.18	4	<0.001**	4.99	[7.1, 9.9]
Ours vs. MT+BERT (Google)	+9.9	11.67	4	<0.001**	5.21	[8.3, 11.5]
Ours vs. XLM-R (base)	+4.1	8.93	4	0.001**	3.99	[3.0, 5.2]
Ours vs. mBERT	+7.5	12.43	4	<0.001**	5.55	[6.1, 8.9]
Ours vs. SVM-TF-IDF	+15.5	16.78	4	<0.001**	7.49	[13.8, 17.2]

Note: ** indicates significance after Bonferroni correction (α = 0.0045). df = degrees of freedom. Mean Diff shows percentage point improvement in accuracy. Cohen's d effect sizes: small (0.2), medium (0.5), large (0.8). All p-values computed using two-tailed paired t-tests. CI = confidence interval for mean difference in percentage points.

The comprehensive experimental evaluation demonstrates that our integrated multi-agent federated learning framework with heterogeneous graph attention networks achieves meaningful improvements over existing approaches across multiple dimensions. While the results validate our core hypotheses regarding the benefits of combining federated learning with graph-based sentiment analysis, they also reveal important limitations and scenarios where simpler approaches remain competitive. We now synthesize these findings and discuss their broader implications.

V. Conclusion

This research presents a comprehensive cross-language hotel review sentiment analysis and dynamic reputation management system that

integrates multi-agent federated learning with heterogeneous graph attention networks to address fundamental challenges in multilingual hospitality data processing and reputation monitoring. The primary contributions encompass the development of a novel multi-agent federated learning framework that enables privacy-preserving collaborative learning across distributed hotel review platforms while maintaining high accuracy in cross-language sentiment classification tasks. The heterogeneous graph attention network architecture successfully captures complex relationships between multilingual textual content, user behaviors, temporal patterns, and service attributes, enabling more nuanced understanding of sentiment dynamics across diverse cultural and linguistic contexts.

The experimental validation demonstrates substantial performance improvements over existing baseline methods, with our integrated system achieving 89.7% accuracy in cross-language sentiment classification while maintaining superior consistency across different languages compared to traditional approaches. The dynamic reputation management component shows remarkable effectiveness in early detection of reputation risks and fake review identification, providing hospitality businesses with actionable insights for proactive reputation management strategies. The federated learning framework successfully preserves data privacy with privacy scores exceeding 0.92 while maintaining performance within acceptable margins of centralized training approaches.

The theoretical significance of this work extends beyond hospitality applications, contributing to the broader fields of federated learning, graph neural networks, and cross-language natural language processing through innovative architectural designs and algorithmic frameworks [78]. The practical application value is demonstrated through measurable improvements in reputation prediction accuracy, reduced response times for reputation monitoring, and enhanced fake review detection capabilities that collectively provide significant commercial benefits for hospitality industry stakeholders.

Current limitations warrant careful acknowledgment to provide balanced assessment of our contribution. First, our evaluation is constrained to four major languages (English, Chinese, French, German) representing approximately 68% of global tourism review volume, and performance on low-resource languages such as Thai, Arabic, or Portuguese remains untested, potentially limiting applicability in emerging tourism markets. Second, the temporal split methodology (training on 2020-2023 data, testing on 2024 data) may not fully capture future distribution shifts as review patterns evolve with changing traveler demographics, new hospitality technologies (e.g., contactless services, AI concierges), or major external events

(e.g., pandemic recovery effects, economic recessions). Third, the federated learning approach requires minimum data thresholds at each agent location (we required ≥8,000 reviews per agent for stable training), which may not be feasible for smaller hotel chains, independent properties, or newly opened establishments without sufficient review history. Fourth, while 89.7% accuracy represents meaningful improvement over baselines, it still corresponds to one error in every ten predictions, potentially limiting deployment in critical decision contexts such as automated review response generation or real-time pricing adjustments where errors could damage customer relationships. Fifth, the heterogeneous graph construction and attention mechanisms introduce computational overhead (15.2 hours training time vs. 12.4 hours for centralized BERT), requiring GPU infrastructure that may be cost-prohibitive for budget-conscious hospitality businesses with tight technology margins. Sixth, the 3.2-day early warning lead time exhibits considerable variance (σ=0.8 days, range 1.7-5.4 days), meaning some reputation events receive insufficient advance notice for meaningful intervention, particularly during rapid-onset crises such as viral social media complaints. Seventh, fake review detection at 93.4% accuracy still allows 6.6% of fraudulent reviews (approximately 10,189 undetected fakes in our dataset) to pass undetected, potentially enabling sophisticated manipulation campaigns using novel attack strategies not represented in our training data. Eighth, the dataset's geographic concentration in North America (43% of hotels), Europe (38%), and East Asia (19%) with limited representation from Middle East, Africa, or South America may limit generalizability to hotels in underrepresented regions with different cultural norms and review patterns. Ninth, our evaluation focuses on text-based reviews without incorporating other modalities such as review images (increasingly common on platforms like Instagram and Xiaohongshu), reviewer photos, or video content that may provide additional sentiment signals. Tenth, the system assumes honest platform metadata (timestamps, user identifiers, hotel attributes), but does not address scenarios where platforms themselves may manipulate or selectively display reviews, introducing systematic biases beyond individual fake reviews.

Eleventh, our current framework lacks robust mechanisms for handling out-of-vocabulary (OOV) domain-specific terminology. When reviews contain hotel brand names, proprietary service labels, or newly coined hospitality terms absent from the pre-trained BERT vocabulary (e.g., "Ritz-Carlton Club Lounge," "Bonvoy Elite status," "contactless check-in"), the model relies on subword tokenization that may fragment semantically meaningful units into uninformative pieces. This limitation contributed to the reduced accuracy (83.1% vs. 89.7% overall) observed for reviews featuring test-set-only hotel terminology.

Several adaptation strategies merit consideration for future implementations. Dynamic lexicon updates could continuously monitor incoming reviews to identify frequently occurring novel terms, adding them to a domain-specific vocabulary layer with embeddings initialized through contextual averaging. Domain-adaptive pre-training represents another promising direction: continuing BERT pre-training on a large corpus of unlabeled hospitality reviews would expose the encoder to industry-specific terminology before task-specific fine-tuning. Additionally, retrieval-augmented approaches could dynamically fetch definitions or contextual examples for unrecognized terms from an external hospitality knowledge base during inference, providing supplementary information to guide sentiment prediction. While these strategies fall outside our current implementation scope, they offer concrete pathways for enhancing OOV robustness in production deployments targeting diverse hospitality markets with rapidly evolving terminology.

Future research directions encompass extending the framework to additional languages and cultural contexts, integrating multimodal data including images and audio reviews, developing more sophisticated privacy preservation techniques, and exploring applications in other service industries beyond hospitality [79]. The application prospects in smart tourism and digital economy are substantial, with potential integration into tourism recommendation systems, automated customer service platforms, and comprehensive destination management frameworks. The system's ability to process multilingual content while preserving privacy makes it particularly valuable for international tourism platforms and global hospitality chains seeking to maintain consistent service quality standards across diverse markets while respecting regional data protection regulations. The ethical and legal considerations merit careful attention as our federated learning framework is designed to comply with major international data protection regulations. The European Union's General Data Protection Regulation (GDPR) requires appropriate technical measures for personal data protection, California Consumer Privacy Act (CCPA) establishes consumer privacy rights, and China's Personal Information Protection Law (PIPL) governs cross-border data transfers [86]. Our system addresses these through key compliance features: data minimization by processing only review text, timestamps, and ratings without collecting personally identifiable information such as reviewer names, email addresses, or payment details; data localization ensuring raw review data never leaves local jurisdictional boundaries where collected, with only aggregated model parameters shared across agents satisfying requirements in Russia's Federal Law No. 242-FZ and China's Cybersecurity Law Article 37; differential privacy guarantees with $\varepsilon=4.0$ and $\delta=10^{-5}$ providing

mathematical assurance that individual reviews cannot be reconstructed from shared model updates; and support for right to erasure mandated by GDPR Article 17 through targeted model retraining using machine unlearning algorithms when users request deletion [84]. Beyond legal compliance, several ethical concerns warrant discussion. Regarding bias and fairness, our consistent performance across languages (coefficient of variation 0.043) mitigates linguistic discrimination concerns, though languages with smaller sample sizes such as mixed-code (8,920 reviews) may experience slightly degraded performance requiring future expansion to low-resource languages. The transparency versus privacy trade-off presents challenges: we provide interpretable attention visualizations for sentiment predictions showing which review aspects (location, cleanliness, staff) contributed to classification decisions, but deliberately limit explanation granularity to prevent potential reverse-engineering of individual reviews from attention patterns. Informed consent considerations arise because our use of publicly available hotel review data for commercial sentiment analysis requires hotels deploying our system to include clear disclosures in privacy policies regarding sentiment analysis practices, following GDPR Article 12 transparency principles and allowing users to understand how their review contributions are processed. Impact on stakeholders requires careful balancing: fake review detection with 93.4% accuracy has potential to unfairly flag legitimate negative reviews as suspicious (6.3% false positive rate means approximately 2,439 genuine reviews in our test set were incorrectly flagged), necessitating human-in-the-loop verification before punitive actions against reviewers or rating adjustments to prevent unjust censorship. Cross-border cooperation faces practical challenges from regulatory divergence where different countries impose varying data transfer requirements such as EU Standard Contractual Clauses (SCCs) and UK International Data Transfer Agreement (IDTA), with our multi-agent architecture accommodating these through region-specific privacy parameters and communication protocols allowing each agent to enforce local regulations. Law enforcement requests present jurisdictional complexities where different countries have different approaches to data access by authorities, with our decentralized system naturally limiting the scope of any single jurisdiction’s access requiring participating organizations to establish clear policies for responding to valid legal requests while protecting users in other jurisdictions. Organizations implementing our system should conduct Data Protection Impact Assessments (DPIAs) as required by GDPR Article 35 before deployment to identify and mitigate privacy risks, appoint Data Protection Officers (DPOs) to oversee cross-border data flows and ensure ongoing compliance, implement contractual safeguards between federated learning participants specifying data handling

obligations and liability allocation, establish clear data governance policies addressing data retention periods, deletion procedures, and audit requirements, and provide regular privacy training for personnel operating the system to maintain awareness of evolving regulatory requirements. We recognize that the legal landscape for artificial intelligence and cross-border data processing continues to evolve rapidly, with emerging regulations such as the EU Artificial Intelligence Act [87] requiring organizations to regularly review system architecture for compliance with new requirements, maintain documentation of data processing activities, conduct periodic privacy audits, and adapt technical measures as regulatory standards develop.

Conflict of Interest

The authors declare that they have no competing interests or conflicts of interest regarding the publication of this research.

Funding

No funding was received for this research.

Ethics Approval

This study was conducted in accordance with ethical guidelines for computational research involving user-generated content. The research protocol was reviewed and approved by the Institutional Review Board of Qingdao Vocational and Technical College of Hotel Management (Ethics Approval Number: QVTHM-2024-CS-047). All hotel review data used in this study consisted of publicly available information from commercial review platforms. Data collection procedures followed platform terms of service and applicable data protection regulations. Personal identifying information was removed during preprocessing, and all data was anonymized prior to analysis. The federated learning framework was designed to ensure privacy preservation and comply with GDPR and other international data protection standards.

Clinical Trial Number

Not applicable.

Data Availability

Due to privacy regulations and platform terms of service agreements, the complete multilingual hotel review dataset containing 154,680 reviews cannot be publicly released to protect user privacy and comply with data protection laws including GDPR and CCPA. To support research reproducibility while respecting these constraints, we provide Supplementary File 1 accompanying this manuscript.

Supplementary File 1 contains the following materials: (1) A representative sample dataset of 5,000 anonymized reviews with 1,250 reviews per language covering English, Chinese, French, and German, where all personally identifiable information has been removed and hotel names replaced with anonymous identifiers while preserving sentiment labels and temporal metadata; (2) A synthetic dataset of 10,000 reviews matching the statistical properties of the original data including sentiment distribution, review length distribution, and cross-language correlation patterns, generated using a conditional variational autoencoder trained on the original dataset, suitable for algorithm development and preliminary testing; (3) Complete data collection scripts for gathering publicly available hotel reviews from TripAdvisor and Booking.com APIs with rate limiting and ethical scraping practices, including documentation for API authentication, data extraction, and preprocessing pipelines; (4) The full data processing pipeline including language detection using langdetect v1.0.9, text normalization using NLTK v3.8.1, sentiment annotation using multilingual BERT, and quality filtering, with detailed documentation and example notebooks; (5) Source code for the multi-agent federated learning framework, heterogeneous graph attention network, and dynamic reputation management system with configuration files matching our experimental setup.

Researchers requiring access to the full dataset for verification purposes may contact the corresponding author at hanxiao202411@163.com with a detailed research proposal describing intended use, institutional affiliation, and ethical approval documentation, subject to a data sharing agreement requiring compliance with applicable privacy regulations and restrictions on commercial use.

Code Availability

The complete implementation of our proposed cross-language hotel review sentiment analysis framework is provided in Supplementary File 1 to facilitate reproducibility and enable future research extensions. The supplementary materials contain modular implementations of the

multi-agent federated learning framework built with PyTorch 2.0.1 and FederatedScope 0.3.0 including agent coordination protocols, secure aggregation mechanisms, and differential privacy modules with configurable epsilon and delta parameters. The heterogeneous graph attention network architecture is implemented using PyTorch Geometric 2.3.1 with custom message passing functions for heterogeneous node types and relation-specific attention mechanisms. The dynamic reputation management system includes real-time monitoring and fake review detection algorithms combining linguistic analysis and graph-based behavioral pattern recognition.

Supplementary File 1 also includes evaluation scripts for baseline comparisons with implementations of mBERT, XLM-RoBERTa, FedAvg, FedProx, GraphSAGE, and HAN-Basic methods with hyperparameter configurations matching our experimental setup. Comprehensive documentation covers installation instructions, API reference, usage examples, and environment setup files including requirements.txt specifying all Python package dependencies with version numbers. Docker configuration ensures environment reproducibility across different systems. Detailed tutorials address dataset preparation, model training with single-agent and multi-agent modes, hyperparameter tuning using Bayesian optimization, evaluation on custom datasets, and deployment guidelines for production environments.

Pre-trained model checkpoints trained on our full dataset achieving 89.7% cross-language accuracy are included, enabling direct inference without retraining, with separate checkpoints for each language-specific agent and the global aggregated model. Additional materials include detailed algorithm pseudocode, mathematical derivations for optimization objectives, ablation study code analyzing individual component contributions, and privacy attack implementations for membership inference and attribute inference attacks used in our evaluation. All code is released under Apache License 2.0 to maximize research impact and facilitate industrial adoption.

Authors' Contributions

XH conceived and designed the study, developed the multi-agent federated learning framework and heterogeneous graph attention network architecture, implemented the dynamic reputation management system, conducted the experimental evaluation and data analysis, and wrote the manuscript. XH is responsible for the overall research design, methodology development, result interpretation, and manuscript preparation. The author read and approved the final manuscript.

References

[1] Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. Journal of Travel Research, 58(2), 175-191.

[2] Álvarez-Carmona, M. Á., Aranda, R., Rodríguez-González, A. Y., et al. (2022). Natural language processing applied to tourism research: A systematic review and future research directions. Journal of King Saud University-Computer and Information Sciences, 34(10), 10125-10138.

[3] Antonio, N., de Almeida, A., Nunes, L., Batista, F., & Ribeiro, R. (2018). Hotel online reviews: Different languages, different opinions. Information Technology & Tourism, 18(1-4), 157-185.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Ounacer, S., Mhamdi, D., Ardchir, S., Daif, A., & Azzouazi, M. (2023). Customer sentiment analysis in hotel reviews through natural language processing techniques. International Journal of Advanced Computer Science and Applications, 14(1).

[6] Weber, J., et al. (2025). Federated learning for privacy-preserving feedforward control in multi-agent systems. arXiv preprint arXiv:2503.02693.

[7] Xu, X. (2022). Mining and application of tourism online review text based on natural language processing and text classification technology. Wireless Communications and Mobile Computing, 2022.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[9] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

[10] Zhang, L., Wang, S., & Liu, B. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University-Computer and Information Sciences, 36(4), 101576.

[11] Wahyuni, E. D., & Djunaidy, A. (2023). Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews. Heliyon, 9(6), e16347.

[12] Jain, P. K., et al. (2023). Sentiment analysis classification system using hybrid BERT models. Journal of Big Data, 10(1), 1-23.

[13] Karthikeyan, K., Wang, Z., Mayhew, S., & Roth, D. (2020). Cross-lingual ability of multilingual BERT: An empirical study. International Conference on Learning Representations.

[14] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4996-5001.

[15] Devlin, J. (2018). Multilingual BERT readme document. Google Research.

[16] Rogers, A., et al. (2025). BERT applications in natural language processing: a review. Artificial Intelligence Review, 58(3), 1-58.

[17] Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT? Proceedings of the 5th Workshop on Representation Learning for NLP, 120-130.

[18] Chen, X., & Cardie, C. (2020). Hierarchical mapping for crosslingual word embedding alignment. Transactions of the Association for Computational Linguistics, 8, 589-601.

[19] Biesialska, M., & Costa-jussà, M. R. (2020). Refinement of unsupervised cross-lingual word embeddings. arXiv preprint arXiv:2002.09213.

[20] Schuster, T., Schick, T., & Schütze, H. (2019). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. arXiv preprint arXiv:1902.09492.

[21] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2022). A multi-agent reinforcement learning approach for efficient client selection in federated learning. Proceedings of the AAAI Conference on Artificial Intelligence, 36(7), 7091-7099.

[22] Lozano, A. C., et al. (2024). Asynchronous consensus for multi-agent systems and its application to federated learning. Engineering Applications of Artificial Intelligence, 133, 108321.

[23] Zhou, Y., et al. (2023). FedQMIX: Communication-efficient federated learning via multi-agent reinforcement learning. Information Fusion, 101, 101875.

[24] Mishra, A., et al. (2023). Federated control with hierarchical multi-agent deep reinforcement learning. Google Research.

[25] Yoon, T., et al. (2024). An innovative multi-agent approach for robust cyber–physical systems using vertical federated learning. Neurocomputing, 589, 127686.

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

[27] Ji, Z., Lipton, Z. C., & Elkan, C. (2014). Differential privacy and machine learning: a survey and review. arXiv preprint arXiv:1412.7584.

[28] Yin, X., Zhu, Y., & Hu, J. (2021). A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys, 54(6), 1-36.

[29] Wang, S., et al. (2021). Privacy preservation in federated learning: An insightful survey from the GDPR perspective. Computers & Security, 110, 102402.

[30] Xu, R., Baracaldo, N., Zhou, Y., Anwar, A., & Ludwig, H. (2021). Privacy-preserving machine learning: Methods, challenges and directions. arXiv preprint arXiv:2108.04417.

[31] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24.

[32] Zhou, J., et al. (2021). Graph neural networks: A review of methods and applications. AI Open, 1, 57-81.

[33] Wang, X., et al. (2019). Heterogeneous graph attention network. In The World Wide Web Conference (pp. 2022-2032).

[34] Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. In *European Semantic Web Conference* (pp. 593-607). Springer, Cham. https://doi.org/10.1007/978-3-319-93417-4_38

[35] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[36] Linmei, H., Yang, T., Shi, C., Ji, H., & Li, X. (2019). Heterogeneous graph attention networks for semi-supervised short text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 4821-4830).

[37] Zhang, Y., et al. (2024). Hybrid-attention mechanism based heterogeneous graph representation learning. Expert Systems with Applications, 248, 123372.

[38] Li, X., et al. (2024). Heterogeneous graph neural network with hierarchical attention for group-aware paper recommendation in scientific social networks. Knowledge-Based Systems, 304, 112477.

[39] Khan, A., et al. (2024). A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. Journal of Big Data, 11(1), 18.

[40] Ye, F., Wang, R., Tang, S., et al. (2024). Federated learning-enabled cooperative localization in multi-agent system. International Journal of Wireless Information Networks, 31, 61-72.

[41] Wang, H., et al. (2019). Multi-agent visualization for explaining federated learning. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19) (pp. 4572-4578).

[42] Domingo-Ferrer, J., Sánchez, D., & Blanco-Justicia, A. (2022). A critical review on the use (and misuse) of differential privacy in machine learning. ACM Computing Surveys, 55(8), 1-16.

[43] Baraheem, S., & Yao, Z. (2022). A survey on differential privacy with machine learning and future outlook. arXiv preprint arXiv:2211.10708.

[44] Pan, K., et al. (2024). Differential privacy in deep learning: A literature survey. Knowledge-Based Systems, 294, 111698.

[45] Tang, J., et al. (2017). Learning with privacy at scale. Apple Machine Learning Research.

[46] Yoon, T., & Loizou, N. (2025). Federated learning meets game theory: The next generation of AI multi-agent systems. Johns Hopkins University Department of Applied Mathematics and Statistics.

[47] Saha, S., Hota, A., Chattopadhyay, A. K., Nag, A., & Nandi, S. (2024). A multifaceted survey on privacy preservation of federated learning: progress, challenges, and opportunities. Artificial Intelligence Review, 57, 184. https://doi.org/10.1007/s10462-024-10766-7

[48] Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Scientific Reports, 14(1), 9603. https://doi.org/10.1038/s41598-024-60210-7

[49] Zhang, S., et al. (2019). Graph convolutional networks: a comprehensive review. Computational Social Networks, 6(1), 11.

[50] Bing, R., Yuan, G., Zhu, M., et al. (2023). Heterogeneous graph neural networks analysis: a survey of techniques, evaluations and applications. Artificial Intelligence Review, 56, 8003-8042. https://doi.org/10.1007/s10462-022-10375-2

[51] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 4171-4186).

[52] Xu, R., Yang, Y., Liu, J., & Zhu, X. (2022). A survey of cross-lingual sentiment analysis: methodologies, models and evaluations. Data Science and Engineering, 7(2), 172-193. https://doi.org/10.1007/s41019-022-00187-3

[53] Chen, M., Tian, Y., Yang, M., & Zaniolo, C. (2017). Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 1511-1517).

[54] Lu, H., Wang, L., Ma, X., Cheng, J., & Zhou, M. (2024). A survey of graph neural networks and their industrial applications. Neurocomputing, 610, 128761. https://doi.org/10.1016/j.neucom.2024.128761

[55] Yang, X., Yan, M., Pan, S., Ye, X., & Fan, D. (2023). Simple and efficient heterogeneous graph neural network. In Proceedings of the AAAI Conference on Artificial Intelligence, 37, 10816-10824.

[56] Barros, C. D. T., Mendonça, M. R. F., Vieira, A. B., & Ziviani, A. (2021). A survey on embedding dynamic graphs. ACM Computing Surveys, 55, 1-37. https://doi.org/10.1145/3483595

[57] Kamata, H., et al. (2024). Unveiling the spatial and temporal variation of customer sentiment in hotel experiences: a case study of Beppu City, Japan. Humanities and Social Sciences Communications, 11(1), 1-15.

[58] Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020). Heterogeneous graph transformer. In Proceedings of The Web Conference, 2704-2710. https://doi.org/10.1145/3366423.3380027

[59] Kumar, A., et al. (2023). Sentiment analysis of hotel reviews - a comparative study. In 2023 IEEE International Conference on Contemporary Computing and Communications (InC4) (pp. 1-6).

[60] Rakibullah, H. M., et al. (2020). Sentiment analysis of hotel reviews - performance evaluation of machine learning algorithms. International Journal of Engineering Research and Technology, 9(5).

[61] Buhalis, D., et al. (2022). Artificial intelligence in tourism and hospitality: Bibliometric analysis and research agenda. International Journal of Contemporary Hospitality Management, 34(8), 2883-2917.

[62] Ma, Y., Yan, N., Li, J., Mortazavi, M., & Chawla, N. V. (2024). HetGPT: harnessing the power of prompt tuning in pre-trained heterogeneous graph neural networks. arXiv preprint arXiv:2310.12580.

[63] Nilashi, M., Ibrahim, O., Yadegaridehkordi, E., Samad, S., Akbari, E., & Alizadeh, A. (2018). Travelers decision making using online review in social network sites: A case on TripAdvisor. *Journal of Computational Science*, 28, 168-179. https://doi.org/10.1016/j.jocs.2018.09.006

[64] Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment analysis in TripAdvisor. *IEEE Intelligent Systems*, 32(4), 72-77. https://doi.org/10.1109/MIS.2017.3121555

[65] Chen, R. Y., Guo, J. Y., & Deng, X. L. (2024). Sensing hotel customers distribution and their sentiment variations using online travel agent data. International Journal of Geographical Information Science, 38(2), 323-343. https://doi.org/10.1080/19475683.2024.2335976

[66] Akhtar, M. S., Kumar, A., Ekbal, A., & Bhattacharyya, P. (2016). A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 482-493). Association for Computational Linguistics.

[67] Gaschi, F., Boukhers, Z., Pulido, L., & Moens, M. F. (2024). Understanding cross-lingual alignment—A survey. *arXiv preprint arXiv:2404.06228*.

[68] Li, Z., Fan, Y., Jiang, B., Lei, T., & Liu, W. (2023). A survey on sentiment analysis and opinion mining for social multimedia. Multimedia Tools and Applications, 82(3), 4271-4308. https://doi.org/10.1007/s11042-022-13428-4

[69] Ahuja, R., et al. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.

[70] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

[71] Khattak, F. K., et al. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. Information Processing & Management, 61(5), 103481.

[72] Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q. S., & Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15, 3454-3469. https://doi.org/10.1109/TIFS.2020.2988575

[73] Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 380-385). https://doi.org/10.18653/v1/N19-1035

[74] Zhao, Y., Xu, S., & Duan, H. (2024). HGNN-BRFE: Heterogeneous graph neural network model based on region feature extraction. Electronics, 13(22), 4447. https://doi.org/10.3390/electronics13224447

[75] Ünal, E., Özdemir, A., & Yıldız, B. (2025). Developing a deep learning-based sentiment analysis system of hotel customer reviews for sustainable tourism. Sustainability, 17(13), 5756. https://doi.org/10.3390/su17135756

[76] Barbado, R., Araque, O., & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4), 1234-1244. https://doi.org/10.1016/j.ipm.2019.03.002

[77] Fauzi, A., & Utami, E. (2024). Sentiment analysis for hotel reviews: A systematic literature review. ACM Computing Surveys, 56(2), Article 159. https://doi.org/10.1145/3605152

[78] Dai, E., Zhao, T., Zhu, H., et al. (2024). A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. Machine Intelligence Research, 21, 1011-1061.

[79] Mujahid, M., et al. (2025). Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. Social Network Analysis and Mining, 15(1), 1-29.

[80] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J. Y., & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

[81] Li, N., Qardaji, W., Su, D., & Cao, J. (2023). Privacy-preserving collaborative machine learning: A survey. IEEE Transactions on Big Data, 9(3), 847-865. https://doi.org/10.1109/TBDATA.2022.3206736

[82] Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z. X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2023). Crosslingual generalization through multitask finetuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 15991-16111. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.891

[83] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations. https://openreview.net/forum?id=nZeVKeeFYf9

[84] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine unlearning. In 2021 IEEE Symposium on Security and Privacy (SP), 141-159. IEEE. https://doi.org/10.1109/SP40001.2021.00019

[85] European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final. Brussels: European Commission.

[86] Bradford, A. (2020). The Brussels Effect: How the European Union Rules the World. Oxford University Press. https://doi.org/10.1093/oso/9780190088583.001.0001

[87] European Commission. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj