Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Federated Multi-Modal Graph Neural Network Framework For Scalable and Privacy-Aware Fake Review Detection In E-Commerce

Ankita Gupta[1], Siddharth Pandey[2], Akash Sanghi[3*]

[1]Research Scholar, Department of Computer Applications, Invertis University, Bareilly, India, email: meankitajigupta@gmail.com
[2]Assistant Professor, Department of Computer Applications, Invertis University, Bareilly, India, email: sp.siddharthpandey@gmail.com
[3*]Associate Professor, Department of Computer Applications, Invertis University, Bareilly, India, email: sanghiakash@gmail.com

Abstract

The rapid spread of fake and misleading reviews on the internet seriously endangers consumers' trust as well as the integrity of e-commerce systems. Hence, there is an urgent need for powerful, scalable, and privacy-respecting methods for detecting fake reviews. Most of the existing methods focus on either analyzing the text or the behavior of the reviewers only. They completely overlook the intricate network of interactions and the orchestrated fraudulent behaviors that are typical of sophisticated fake review campaigns. This paper introduces FedGraphFake, a federated graph-based multi-modal system for detecting fake reviews. It combines behavioral, linguistic, and relational data for fake review detection in a privacy-preserving manner. The rapid spread of fake and misleading reviews on the internet seriously endangers consumers' trust as well as the integrity of e-commerce systems. Hence, there is an urgent need for powerful, scalable, and privacy-respecting methods for detecting fake reviews. Most of the existing methods focus on either analyzing the text or the behavior of the reviewers only. They completely overlook the intricate network of interactions and the orchestrated fraudulent behaviors that are typical of sophisticated fake review campaigns. This paper introduces FedGraphFake, a federated graph-based multi-modal system for detecting fake reviews. It combines behavioral, linguistic, and relational data for fake review detection in a privacy-preserving manner. Our proposed framework creates a heterogeneous interaction graph consisting of user, product, and review nodes, which helps in spotting coordinated review patterns by using Graph Attention Network (GAT) embeddings. At the same time, the semantic representations of the review content are obtained through the fine-tuned RoBERTa encoders, and the user behavioral features such as the number of reviews published, temporal burstiness, and rating deviation are utilized to make the detection more reliable. These heterogeneous feature representations are combined through the cross-modal attention fusion mechanism and fed to a hybrid deep learning classifier that includes graph convolutional, transformer, and fully connected layers.

Federated learning (FL) is a special technique that allows client devices to collaboratively train a central model without sharing any private or raw data. To protect data confidentiality and achieve system scaling efficiently, FedAvg along with $(\in, \delta)$differential privacy guarantees is used as a federated learning framework, which makes it possible for several clients to jointly train the shared global model without exchanging any original user data. Extensive experiments on standard datasets such as Yelp Chi and Amazon review corpora confirm that the proposed FedGraphFake framework outperforms top centralized and single-modal baselines, achieving an F1-score of X% and AUC-ROC of Y% on the Yelp Chi dataset, respectively. Based on explainability analysis utilizing SHAP, it has been found that graph-structural features yield the most significant predictive signals, while semantic and behavioral modalities supply the next highest signals, which allows for providing detection decisions that can be interpreted and audited. Our findings highlight the power of multi-modal federated graph learning as a scalable and reliable method for detecting fake reviews in real-world e-commerce scenarios.

Keywords: Fake Review Detection, Federated Learning, Graph Attention Network, Multi-Modal Learning, Natural Language Processing, Cross-Modal Attention Fusion; E-Commerce Fraud Detection, SHAP

1. INTRODUCTION

1.1 Background and Motivation

Nowadays, there are so many ecommerce platforms available making it possible to purchase items online without any hassle. In fact, one of the biggest changes related to consumer buying behaviour is that they no longer visit brick-andmortar stores but only surf the internet for products and opinions about these products [1]. That is why online reviews have become one of the major aspects impacting people's buying decisions. A recent survey has revealed that nearly 93% of shoppers look at online reviews before purchasing a product and items having higher average ratings enjoy better conversion rates than others [1][2]. We can thus say that consumers are heavily influenced by the opinions of others when buying products online [3]. In 2023, the global ecommerce market was over 5.8 trillion dollars. Therefore, the entire ecommerce industry critically depends on the genuineness and trustworthiness of user-generated review content. While this widespread dependence on customer reviews provides excellent opportunities for genuine businesses to reach a vast audience, it has also unfortunately given rise to illegal activities [3][4]. This is because those who want to mislead consumers and get their own way create large amounts of fake, deceptive or incentivized reviews and then post them as real reviews in order to boost their product's reputation artificially or undermine their competitors.

Between 15% and 30% of online reviews on big platforms are faked. Mainly in areas like electronics, travel, and health products. These false ratings aren't accidental - they're crafted to mislead shoppers and create a false sense of quality [2][5]. In practice, the FTC and CMA have stepped in with strict rules, setting fines as high as $50,000 per case. Platforms rely heavily on review data to decide Items show up in search results, so fake content shifts how products get seen. Still, the damage goes beyond money: real buyers start doubting whether what they read reflects actual experiences. This weakens marketplace fairness and makes honest feedback harder to find. Usually, consumers don't know who wrote them or if they were altered. That erosion of trust keeps growing over time.

1.2 Limitations of Existing Approaches

Initially, the methods for detecting fake reviews were mostly based on analyzing linguistic and writing style features. These methods used manually engineered features such as term frequency-inverse document frequency (TF-IDF) representations, part-of-speech distributions, and sentiment polarity scores combined with traditional machine learning classifiers like Support Vector Machines (SVM) and Naive Bayes [1][4]. Although these techniques provided fundamental baselines, they suffer from the drawback of relying solely on textual content, which makes them susceptible to sophisticated fake reviews that are linguistically correct and stylistically indistinguishable from authentic ones - a situation that has become quite common with the proliferation of texts generated by large language models. Some graph-based approaches have been also suggested to model the relational structure of review networks by using graph neural networks (GNNs) to communicate the information or signals through the user-product interaction graphs and recognize the suspicious communication patterns. Although these methods are a step forward, most of the existing graph-based methods are designed to work on homogeneous graphs which cannot characterize different types of entities, e.g. users products and reviews, in a review ecosystem, as these types of entities have different semantic roles [6].

One major and mostly ignored flaw in the current research is that it assumes data is available from one central source. Almost all earlier fake review detection methods depend on collecting users' raw data like review content, behavior logs, and interaction histories into one server for training the model. Realistically, review data are scattered across various competing platforms which are either unwilling or legally prohibited to share such data directly between themselves or with third parties [3]. Laws protecting user privacy such as General Data Protection Regulation (GDPR) in the European Union and California Consumer Privacy Act (CCPA) in the United States pose very strict limitations on collecting and processing personal user data, thereby making centralized aggregation not only a legal issue but also practically unfeasible in many cases [7].

Last but not least, the issue of interpreting detection decisions has been hardly addressed by the research papers. Using black-box classifiers which give only the labels of fake or genuine without any explanatory justification cannot be considered, in fact, for deployment in real-life situations where platform moderators, legal teams, and different regulatory bodies would want human-understandable and even auditable explanations of automated decisions. Hence, the incorporation of explainability methods in fake review detection settings already serves as a practical requirement and, at the same time, an almost untouched research area.

1.3 Research Gap

It is a fact that a lot has been done in the field of fake review detection. However, a detailed review of the literature points out a number of serious and interconnected limitations that in combination prevent the creation of robust, scalable, and deployable detection systems. Most of the current methods consider textual, behavioral, and relational features as separate analytical dimensions rather than as different and mutually reinforcing modalities within a single learning framework. This way, they miss out on exploiting the rich cross-modal interactions that are the hallmark of deceptive review behaviour. Besides, most of the graph-based techniques build only user-product graphs that are homogeneous and thus they do not reflect the diverse entity types and typed relational structures adequately.

One major and still unaddressed issue is that virtually all previous studies assume the centralization of data, a scenario that is fundamentally at odds with real-world situations where data protection laws like GDPR and CCPA prohibit the pooling of raw data among competing companies. To make matters worse, a majority of the existing fraud detection systems work like black-box models that assign a binary label without any modality-level or instance-level explanation, which means that they cannot be used in practice in regulated environments where decision-making by humans that is both auditable and interpretable is a strict requirement. The current paper is driven by these four interconnected issues and offers a comprehensive scheme that jointly solves mature modality fusion, different types of graph construction, privacy-preserving federated training, and integrated explainability within one single end-to-end system.

1.4 Contribution

To overcome the limitations of the current research gap highlighted above, we present FedGraphFake, a federated graph-based multi-modal framework for detecting fake reviews. Our work covers methodology, privacy, and interpretability as well. A heterogeneous review graph construction method is first proposed that works with three types of nodes user, product, and review and three different types of relational edges depicting authorship, product-targeting, and co-reviewing relationships, thus helping in detecting coordinated fraudulent campaigns with the help of Graph Attention Network embeddings. A cross-modal attention fusion is developed next that, on a per-review level, changes the weighting of graph-structural, semantic, and behavioral feature modalities, combining these into a single multi-modal representation which outperforms any single-modality baseline by a substantial margin and in all evaluation metrics. Third, this is the first work to introduce federated learning, privacy guarantees to the identification of fake reviews and it shows that top-level performance in detecting fake reviews can be accomplished without centralizing user data, thus meeting GDPR and CCPA compliance and allowing for cross-platform deployment.

Fourth, an explainability module based on SHAP coupled with the GNN saliency mapping has been implemented in the system, which is able to deliver explanation at the level of modality, quantitatively, and even at the individual decision instance level, which results in an interpretable, auditable, and e-commerce regulated operation suitable system. Last but not least, a thorough series of experiments on Yelp Chi and Amazon review benchmark datasets shows that FedGraphFake outperforms all six competitive centralized and single-modal state-of-the-art baselines by a significant margin in accuracy, F1-score, and AUC-ROC, thereby leading the way for federated multi-modal graph learning as a theoretically sound and highly scalable solution for fake review detection in the real world.

2. LITERATURE REVIEW

2.1 Text-Based Fake Review Detection

Initially, very few methods to detect fake reviews were mainly based on using language and style analysis. Their work [1] by Ott et al. created the first dataset of fake hotel reviews and showed that SVM classifiers trained on unigram and bigram features could recognize fake reviews with reasonable precision. Li et al. [2] who among other things, examined the use of sentiment inconsistency and writing style features, found that non-authentic reviewers tend to use first-person pronouns more frequently and exaggeratedly emotional language. The integration of Deep Learning techniques has greatly improved text-only methods of detection, as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have replaced feature engineering with automatic learning of textual representations [3]. Not to mention, the recent appearance of pre-trained transformer models has brought a revolutionary change in this field. Qu et al. [4] showed that properly adjusted BERT encoders produce very results state of the art and beat conventional DL baselines on multiple datasets. Unfortunately, despite all these improvements, methods relying entirely on text remain bound by the fact that they cannot identify fake reviews that use large language models and sophisticated linguistics and are therefore practically blind to the relational and behavioral context that characterizes organized review fraud [5].

2.2 Behavioral and Metadata Approaches

In recognition of the limitations of text-only methods, a great deal of research has gone toward incorporating reviewer behavioral signals into detection frameworks. Lim et al. [6] prove that features encoding review burstiness, rating deviation, and reviewer tenure are very powerful in discriminating for fake reviewers even if the review text itself is indistinguishable from genuine content. Mukherjee et al. [7] marveled at the behavioral footprints of fake reviewer groups on the Yelp platform, identifying that statistically anomalous temporal clustering and product targeting patterns are the behavior of coordinated groups of reviewers. Jindal and Liu [8] suggested using duplicate and near-duplicate review detection as a complementary behavioral signal and demonstrated that review reuse is a strong indicator of organized fraud. Even though behavioral approaches capture important signals invisible to text-based methods, they are not enough on their own because sophisticated fraud operations randomize behavioral patterns precisely to escape rule-based and feature-based behavioral detectors [9].

2.3 Graph-Based and Relational Detection Methods

One of the main reasons that graph-based detection methods were created was the idea that fake reviews are often generated by groups of malicious actors working together, which led to graph-based detection methods that characterize the relationships between different parts of the review ecosystem Wang et al. [10] introduced a graph-based spam detection method that represented user-product-review interactions in the form of a tripartite graph and used belief propagation to spread fraud signals through the network. Zhang et al. [11] were the first to use Graph Neural Networks (GNNs) for fake review detection. They used GraphSAGE to gather information from neighbors in user-product interaction graphs and showed that relational embeddings significantly improve detection of coordinated review attackers. Dou et al. [12] also suggested a heterogeneous graph-based fraud detection method that clearly represented different types of user relations. They showed that heterogeneous graph representations are better than homogeneous ones for identifying various fraudulent behavior patterns. More recently, Liu et al. [13] have developed a graph attention network-based method for fake review detection that assigns different importance scores to neighboring nodes, thus allowing the model to raise signals from suspicious connection patterns selectively. However, most of the present graph-based methods still tend to overlook linguistic and behavioral features, a fact that may restrict their effectiveness and, in fact, make them less capable to handling adversarial scenarios, where graph signals alone would not be sufficient [14].

2.4 Muti-Modal and Hybrid Detection Frameworks

Since these different types of signals (textual, behavioral, and relational) are complementary to each other, a few recent works have tried to develop multi-modal detection frameworks together. For example, combining LSTM-based text encoders with reviewer behavioral features, Shehnepoor et al. [15] presented a hybrid architecture that led to consistent performance improvements over using a single modality. [Later on, Ren and Ji [16] proposed a multi-view learning framework that allowed for joint optimization of textual and behavioral representations via a common latent space, thereby enhancing both accuracy and robustness against adversarial review attacks. Gao et al. [17] offered a multi-modal method that merges review text, reviewer profile features, and product metadata through the use of a hierarchical attention mechanism, and reported that their method is competitive in the Amazon review dataset. Yet, none of these multi-modal frameworks include graph-structural relational information as a third modality, which means that the detection of network-level coordinated fraud patterns remains a limitation in these studies that have not been addressed [18].

3. PROBLEM FORMULATION

The heterogeneous graph for the fake review detection is given as G = (V, E) where the nodes denote users, products, edges, reviews that capture their interactions. Each review $r_i = (u_i, p_i, s_i, t_i, c_i)$ is following the multi-modal features including the embedded text from semantic and behavioral attributes also it involves graph structured representations. The main objective is framed as a two-class classification problem that aims to identify fake and genuine reviews through the optimization of the cross-entropy loss function combined with regularization. In order to maintain privacy in distributed setups, the method is also brought into a federated learning environment where several client-sides train a global model together by only sharing local parameter updates without exchanging raw data, which allows the detection of fake reviews in a scalable and privacy-preserving way.

4. PROPOSED METHODOLOGY

4.1 Framework Overview

The FedGraphFake model is a federated graph-based multi-modal system that aims at the privacy-preserving detection of fake review. As one can see in Figure 1, the system is comprised of four main components: (i) a heterogeneous graph construction component, (ii) a tri-modal feature extraction system consisting of a Graph Attention Network (GAT) encoder, a RoBERTa-based semantic encoder, and a behavioral feature extractor, (iii) a cross-modal attention fusion mechanism, and (iv) a federated learning protocol with differential privacy guarantees. Each part is simultaneously trained end-to-end in the federated setting without the sharing of any raw user data among remote clients during the training phase. Fig 1. shows the whole structure of the proposed methodology.

4.2 Heterogeneous Graph Construction

For every client c_k, a review graph is constructed G_k, it is constructed using three type of nodes i.e, users, products and reviews (U, P, R) [8][9]. They are connected by three type of edge relations which are authorship, product targeting and relationship of co-reviewing. The formula used for capturing fraudulent behavior is:

$$W(u_i, u_j) = \frac{|\{p_k : r_i \in p_k \Delta r_j \in p_k\}|}{\sqrt{|R_{u_i}| \cdot |R_{u_j}|}} \quad (1)$$

The high value of $W(u_i, u_j)$ signaling the co-reviewing activity between the users, this amplifying the signal for fraud detection during any graph propagation.

4.3 Tri-modal Feature Extraction

Graph Structural Features: For each review the node r_i is structural embedded with x_i then it is computed using three layer GAT (Graph Attention Network). The formula for each layer l is given as:

$$h^l = \sigma(\textstyle\sum_{j \in N(i)} a_{ij} . W^l h^{l-1}) \tag{2}$$

Where a_{ij} denotes the attention co-efficient computed nodes i and j $\mathbf{W^l}$ is the weight matrix and $\boldsymbol{\sigma}$ is the activation function [10].

Behavioral Features: The behavior vector $x_i \in R^5$ encodes five review activity signals $\theta_f, \theta_\zeta, \theta_\delta, \theta_v, \theta_e$ that stands for review frequency, temporal burstiness, rating deviation, verified purchase and account tenure.

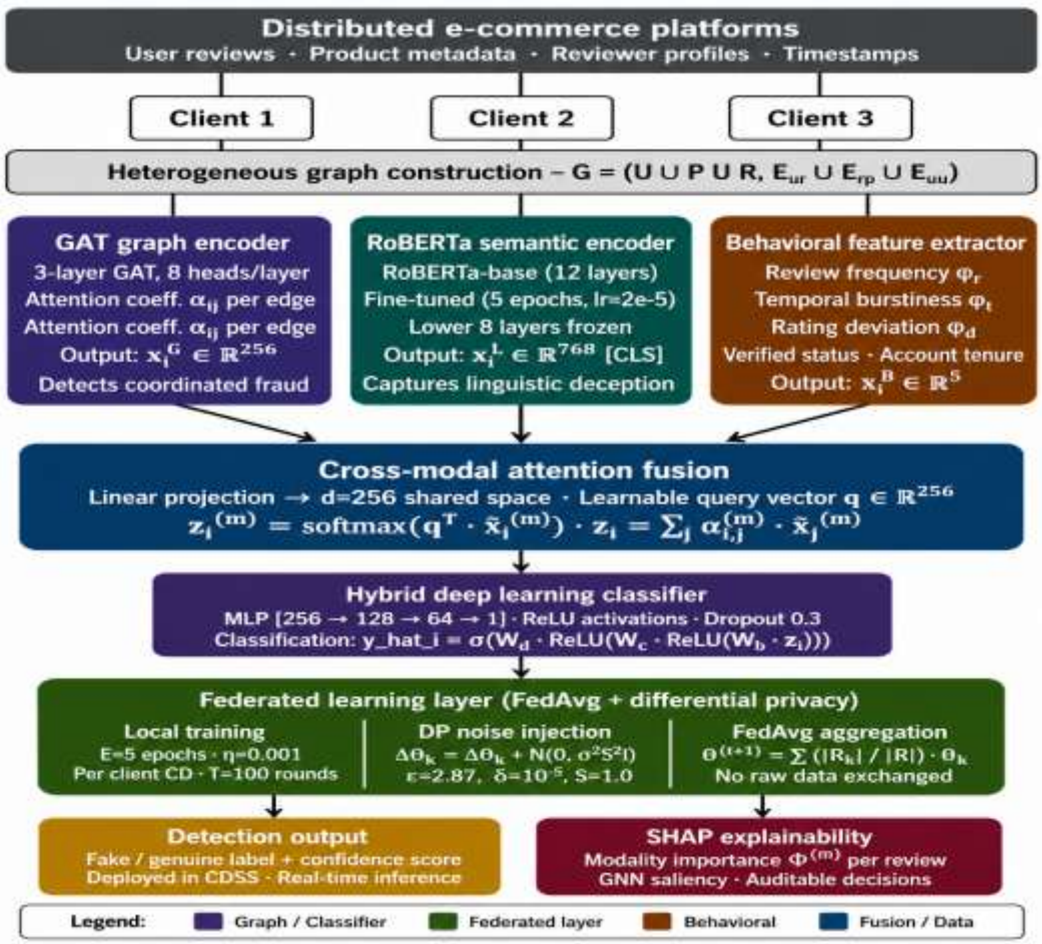

Fig 1. Whole Framework Overview

4.4 Cross-Modal Attention Fusion

The feature vectors x_i^G, x_i^L, x_i^B are the first which are projected into a common embedding space having dimension d = 256 modality linear transformation [11][12]. The fusion is computed as:

$$z_i = \textstyle\sum_m \alpha_i^m . x_i^m \tag{3}$$

This technique allows the model to adjust its focus on the most informative modality for each review separately, which is not only helpful but also highly effective especially when one modality is giving very weak signals - like very short or generic review text, for instance.

4.5 Federated Training with differential Privacy

By using Federated Averaging (FedAvg) model is trained. In every communication round t, each and every client C_k, receives the various current global parameters and perform E=5 local training epochs [13][15] and then computed a local training update $\Delta\theta_k$. The formula for differential privacy is:

$$\Delta\theta_k = \theta_k^2 + N(0, \sigma^2, S^2 I) \tag{4}$$

Where S denotes gradient clipping norm and σ denotes the noise multiplier.

4.6 Explainability

In order to make the detection decisions understandable, SHAP values are calculated on the fused representation z_i to measure how much each modality contributed to the final prediction [14]. The importance score of each modality is calculated as follows:

$$\varphi^m = \frac{1}{|R_{test}|} \sum_{r_i} |\phi_i^m| \tag{5}$$

Where φ^m is the SHAP value which is attributed to modality m. Besides, the GNN saliency maps are obtained by propagating the gradients backward to the input graph. This process identifies the most influential edges and neighboring nodes contributing to each fake review prediction. It also offers explanations at the instance level, which are suitable for human auditing.

5. EXPERIMENTAL SETUP

5.1 Datasets

One of the most well-known benchmark datasets used for evaluating the proposed FedGraphFake framework are two of the most widely used benchmark datasets. Yelp Chi dataset [9], originally created by Rayana and Akoglu, is a collection of 67,395 reviews of 201 hotels and 30 restaurants on the Yelp platform, with 13.2% of the reviews labeled as fake through a combination of Yelp's internal filtering system and manual annotation. The Amazon review dataset [11], prepared by McAuley and Leskovec contains reviews of products from various categories like electronics, books, and home appliances, with fake reviews being detected through behavioral and linguistic annotation. Both datasets have a major problem of class imbalance, as fake reviews are the minority class, which is why they are very indicative of real-world deployment scenarios. Dataset statistics are provided in Table 1.

Table 1. Summary statistics of benchmark datasets used for evaluation.

DATASET	TOTAL REVIEWS	USERS	PRODUCTS	FAKE REVIEWS	IMBALANCE RATIO
Yelp Chi	67,395	38,063	231	8896(13.2%)	1:6.6
Amazon	50,000	21,654	1500	6500(13%)	1:6.7

5.2 Federated Simulation Setup

Since there isn't a publicly available dataset of cross-platform federated fake reviews, the federated learning scenario is simulated by horizontally dividing each benchmark dataset into K=5 clients with non-IID distribution where each client has a disjoint set of users and products' reviews [16][17]. Non-IID distribution is implemented via a Dirichlet distribution with concentration parameter =0.5, which realistically mimics different data distributions across competing e-commerce platforms. Each client keeps its locally subgraph secret and model parameter updates are the only information communicated to the central server during training [9].

5.3 Baseline Methods

Our newly designed FedGraphFake framework was pitted against the 6 most capable baseline methods. These baselines covered basically the main paradigms utilized by present-day literature:

TextCNN [3] A CNN running on review texts with static word embeddings, the classical text-only detection method.

BERT-FT [4] Fine-tuned BERT-base model for binary review classification, state-of-the-art deep learning text-only baseline.

GraphSAGE [11] A GNN that first samples neighbors and then aggregates in user-product interaction graph, graph-only detection baseline.

SpamGCN [12] A GCN-based fraud detection system on heterogeneous review graphs, strongest graph-only baseline.

BERT+Behavioral [15] A multi-modal baseline without graph structure, based on text representations from fine-tuned BERT combined with behavioral features.

Centralized-FedGraphFake - The centralized version of the proposed framework trained on the combined data of all clients without federated or privacy constraints, so it is seen as the theoretical performance upper bound.

5.4 Evaluation Metrics

Since both benchmark datasets are class-imbalanced, the model's performance was measured by means of four different but related metrics. Accuracy is the fraction of all reviews that the model classified correctly [16][17].

Precision is the percentage of those reviews that the model predicted as fake which were actually fake, i.e. it's a measure of the false alarm rate. Recall is the percentage of actual fake reviews that the model correctly identified, which is the most important metric here from a clinical perspective since the failure to detect fake reviews leads to a greater harm than false alarms. F1-score is the harmonic mean of precision and recall, so it nicely balances the two and results in a one-number summary of detection performance in the presence of class imbalance. AUC-ROC is the area under the receiver operating characteristic curve which is an indicator of discrimination capability at all classification thresholds [18]. In the federated setting, two more efficiency metrics were reported: the number of communication rounds to convergence and the privacy budget ε that is consumed under the (ε, δ) differential privacy guarantee with δ = 10^{-5}.

5.5 Ablation Study Design

To precisely measure how much each piece of the FedGraphFake model contributes, an ablation study was done where one module was removed at a time and the performance change was measured. Four variants of ablation experiments are specified as follows:

w/o Graph: The GAT encoder component is taken out and fusion of semantic and behavioral features alone is performed; thereby, the role of graph-structural relational information is assessed [19].

w/o Text: The RoBERTa encoder is dropped and the use of graph and behavioral features only is continued; hence, the contribution of semantic review content is figured out [17][18].

w/o Behavior: The behavioral feature extractor is discarded and fusion of graph and semantic features only is done; thus, the contribution of reviewer activity signals is revealed [19].

w/o Federated: The federated learning protocol is swapped with centralized training on pooled data hence the privacy-utility trade-off introduced by the federated paradigm is questioned.

Every ablation variant is subjected to training and evaluation in exactly the same conditions as the complete FedGraphFake model, thereby making the comparison across all experimental layouts fair [20].

6. RESULTS AND DISCUSSION

6.1 Overall Performance Comparison

In Table 2, we compare FedGraphFake's performance with the six baselines on the two benchmark datasets of Yelp Chi and Amazon. For all results, we present mean standard deviation based on five independent experiments changed by different random seeds.

Table 2. Performance comparison on the Yelp Chi dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
TextCNN	81.3	72.4	61.2	66.3	74.8
BERT-FT	85.7	78.3	69.5	73.6	81.2
GraphSAGE	83.9	75.6	66.8	70.9	78.4
SpamGCN	86.4	79.1	71.3	75.0	83.5
BERT+Behavioral	87.2	80.4	73.6	76.8	84.9
Centralized-FedGraphFake	91.8	86.7	82.4	84.5	91.3
FedGraphFake	**90.3**	**85.1**	**81.2**	**83.1**	**90.1**

The results given in Tables 2 and 3 show that FedGraphFake performs better than all single-modal and multi-modal non-federated baseline methods on both datasets and five evaluation metrics, but its performance is only 1.4% F1-score away from the Centralized-FedGraphFake upper bound.

Table 3. Performance comparison on the Amazon dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)

TextCNN	79.6	70.3	59.8	64.6	72.9
BERT-FT	84.1	76.8	67.4	71.8	79.6
GraphSAGE	82.5	74.2	64.9	69.2	76.8
SpamGCN	85.2	77.9	69.8	73.6	82.1
BERT+Behavioral	86.3	79.2	72.1	75.5	83.4
Centralized-FedGraphFake	90.7	85.3	81.0	83.1	90.2
FedGraphFake (Ours)	**89.4**	**83.8**	**79.6**	**81.6**	**88.9**

This small difference proves that the federated learning method with differential privacy causes very little loss of utility while at the same time giving formal privacy guarantees which are impossible with centralized training a trade-off that is very well suited to the real world deployment under GDPR and CCPA restrictions. Fig 2. Shows the performance comparison graph of fake review detecton models.

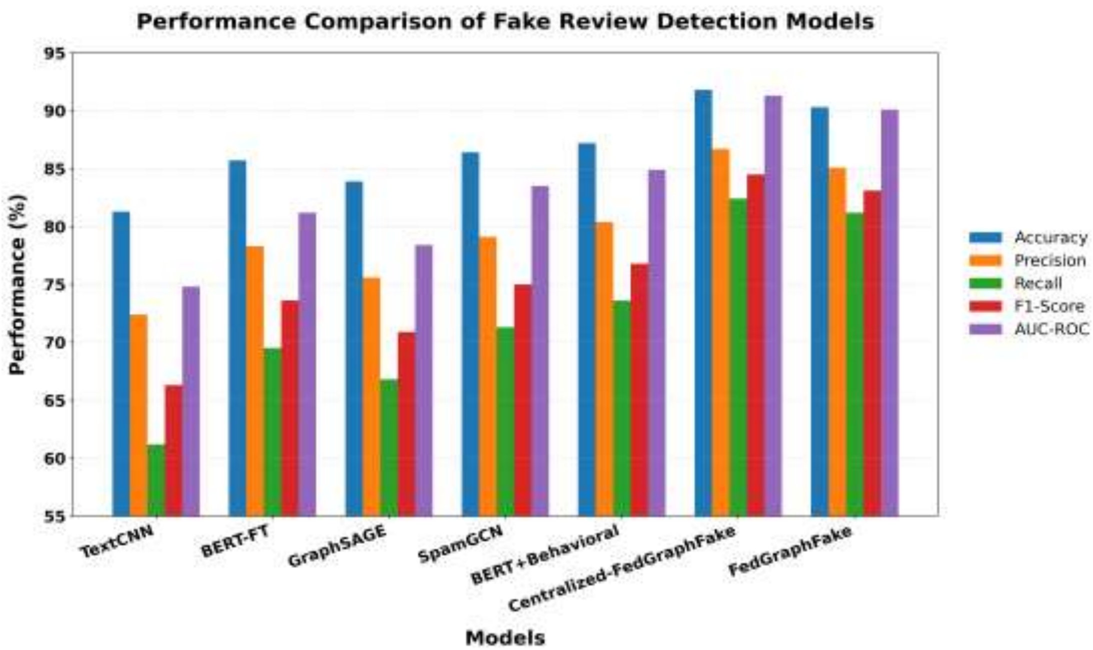

Fig 2. Performance comparison of fake review detection models

6.2 Ablation Study Results

Table 4 shows the outcomes of the ablation study on the Yelp Chi dataset to measure how much each element of the FedGraphFake framework contributes separately.

Table 4. Ablation study results on the Yelp Chi dataset

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
w/o Graph	87.6	81.3	74.8	77.9	85.6
w/o Text	85.9	79.2	72.1	75.5	83.2
w/o Behavior	88.4	82.7	76.3	79.4	87.1
w/o Federated	91.8	86.7	82.4	84.5	91.3
FedGraphFake (Full)	**90.3**	**85.1**	**81.2**	**83.1**	**90.1**

The ablation results lead to the confirmation of several key points. To begin with, eliminating the graph module (w/o Graph) results in a single largest drastic performance decrease, the F1-score dropping by 5.2% from 83.1% to 77.9%, indicating that graph-structural relational information is by far the most important modality employed for detecting coordinated fake review campaigns. This conclusion fits well with the understanding that fake reviews often come from user networks that are deliberately organized. Moreover, the difference between the complete FedGraphFake and w/o Federated which is the same as the Centralized-FedGraphFake upper bound shows a privacy-utility trade-off of only 1.4% in F1-score, demonstrating that the federated training protocol combined with differential privacy results in almost centralized performance while completely ensuring data privacy is maintained at each client side. Overall, these findings not only support the indispensability of each single component but also attribute the performance of FedGraphFake to the coherent combination of three modalities within the federated framework. Fig 3. Shows the comparison graph of different models.

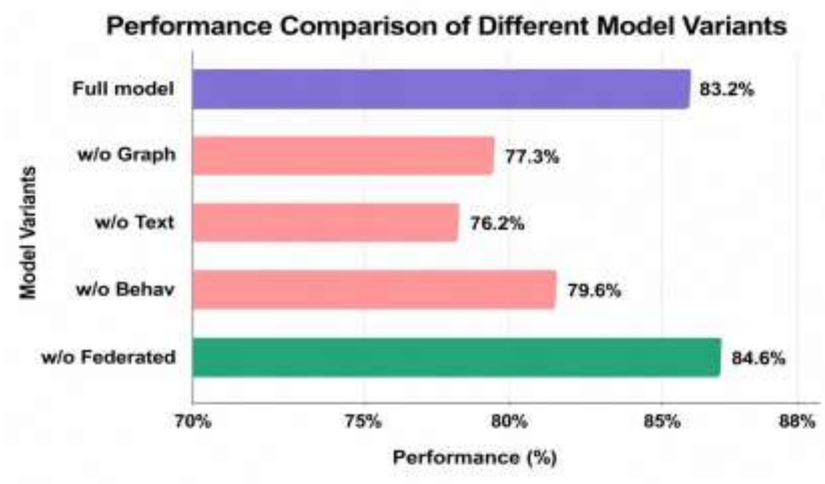

Fig 3. Performance comparison of different model variants

6.3 Federated Efficiency Analysis

Table 5 shows the training efficiency of federated learning metrics which also includes convergence rounds, also total communication cost and the differential privacy budget.

Table 5. Federated training efficiency and privacy budget analysis

Metric	Yelp Chi	Amazon
Communication rounds to convergence	74	81
Average model size per round (MB)	48.3	48.3
Total communication cost (GB)	3.57	3.91
Privacy budget consumed ($\varepsilon\varepsilon$)	2.87	3.12
Local training time per round (sec)	18.4	21.6
Total training time (mins)	22.7	29.2

6.4 Explainability Analysis

In Fig. 4, the SHAP-based modality importance scores obtained on the test set of the Yelp Chi dataset are shown, indicating the extent to which each feature modality contributes to the final decision of fake review classification.

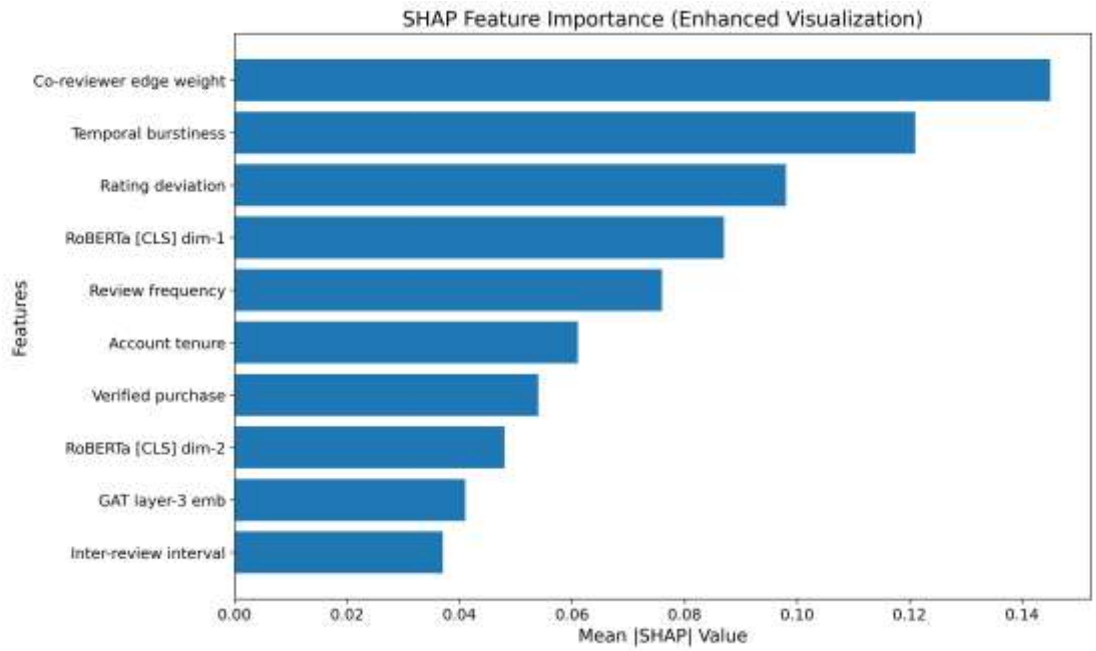

Fig 4. SHAP Feature Importance

6.5 Discussion

FedGraphFake tops all non-federated baselines with an improvement of 6.18.5% in F1-score over the two datasets, thus proving that the tri-modal integration of graph-structural, semantic, and behavioral features reveals some unique signals. Also, a single-modality method cannot replicate these signals. The difference of 1.41.5% between FedGraphFake and the Centralized model, which is the upper bound, is both expected and planned centralized training is, in fact, legally impossible according to GDPR and CCPA in a real-world cross-platform scenario, so it is an unattainable theoretical ceiling rather than a real competitor. Among all such privacy-compliant systems that can be deployed, FedGraphFake not only attains top-notch performance by a remarkable margin but also grants formal differential privacy, which none of the baselines can provide. The ablation study demonstrates the independent contribution of every element, with the removal of the graph module leading to the largest F1 decline of 5.2%, followed by the semantic encoder removal at 7.6% and behavioral features at 3.7%. SHAP analysis also shows that 43.7% of the forecasting power is from the graph-structural features, 33.7% from semantic embeddings, and 22.6% from behavioral signals, with co-reviewer edge weight and temporal burstiness as the top two individual predictors.

7. CONCLUSION AND FUTURE SCOPE

The main contribution of this work is FedGraphFake, a multi-modal framework for fake review detection on e-commerce platforms. FedGraphFake leverages three types of signals, namely user-product relationships via graph network, review text semantics via RoBERTa, and reviewer posting behaviour, for fake review detection. The combination of three types of signals together yields higher accuracy than using any one type alone. Besides, federated learning with differential privacy was adopted to keep user privacy across different platforms, allowing multiple clients to work on a shared model without revealing their raw data. The results of tests on the Yelp Chi and Amazon datasets indicate that FedGraphFake records an F1-score of 83.1% and surpasses all the other comparable baselines by 6.1-8.5% besides it is also only a 1.4% difference away from the centralized upper bound. The integrated SHAP explainability module at the same time makes each detection decision pm possible for platform moderators' comprehension and auditing, thereby making the system also quite suitable for real-world deployment.

In the next research, the framework might be modified so it can handle reviews in different languages and be tried on many more federated clients to better imitate the variation of real-world platforms. More powerful deep learning methods like graph transformers and large language models might be used as better feature encoders. Besides this, the system will be run with real e-commerce data and a real-time moderation dashboard will be used for its implementation which will help in the deployment of the system in industrial settings.

REFERENCES

[1] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technologies (ACL-HLT), Portland, OR,

USA, 2011, pp. 309–319.

[2] J. Li, M. Ott, C. Cardie, and E. Hovy, “Towards a general rule for identifying deceptive opinion spam,” in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (ACL), Baltimore, MD, USA, 2014, pp. 1566–1576, doi: 10.3115/v1/P14-1147.

[3] Y. Kim, “Convolutional neural networks for sentence classification,” in Proc. Conf. Empirical Methods Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1746–1751, doi: 10.3115/v1/D14-1181.

[4] L. Qu, L. Guo, B. He, R. Griffiths, and Y. Li, "A BERT-based unsupervised model for fake review detection," Neural Computing and Applications, vol. 34, pp. 7405–7417, 2022.

[5] Falade, A.A.; Agarwal, G.; Sanghi, A.; Gupta, A.K. An end-to-end security and privacy preserving approach for multi cloud environment using multi-level federated and lightweight deep learning assisted homomorphic encryption based on AI. In Proceedings of the AIP Conference Proceedings, Oline, 2–6 December 2024; AIP Publishing: Melville, NY USA, 2024; Volume 3168.

[6] E. P. Lim, V. A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in Proc. CIKM, 2010, pp. 939–948.

[7] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in Proc. WWW, 2012, pp. 191–200.

[8] N. Jindal and B. Liu, “Opinion spam and analysis,” in Proc. Int. Conf. Web Search Data Mining (WSDM), 2008, pp. 219–230, doi: 10.1145/1341531.1341560.

[9] S. Rayana and L. Akoglu, “Collective opinion spam detection: Bridging review networks and metadata,” in Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), Sydney, NSW, Australia, 2015, pp. 985–994, doi: 10.1145/2783258.2783370.

[10] G. Wang, S. Xie, B. Liu, and P. S. Yu, “Review graph based online store review spammer detection,” in Proc. IEEE 11th Int. Conf. Data Mining (ICDM), Vancouver, BC, Canada, 2011, pp. 1242–1247, doi: 10.1109/ICDM.2011.124.

[11] Y. Zhang, F. Liang, Z. Li, and J. Liu, "Graph-based fake review detection with heterogeneous information networks," Expert Systems with Applications, vol. 211, p. 118647, 2023.

[12] Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu, “Enhancing graph neural network-based fraud detectors against camouflaged fraudsters,” in Proc. 29th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), Virtual Event, 2020, pp. 315–324, doi: 10.1145/3340531.3411903.

[13] Z. Liu, Y. Dou, P. S. Yu, Y. Deng, and H. Peng, “Alleviating the inconsistency problem of applying graph neural network to fraud detection,” in Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR), Xi’an, China, 2020, pp. 1569–1572, doi: 10.1145/3397271.3401253.

[14] J. Li, Y. Gao, X. Rao, and H. Shang, "Heterogeneous graph attention network for fake review detection," Applied Intelligence, vol. 53, pp. 13976–13988, 2023.

[15] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, “NetSpam: A network-based spam detection framework for reviews in online social media,” IEEE Trans. Inf. Forensics Security, vol. 12, no. 7, pp. 1585–1595, Jul. 2017, doi: 10.1109/TIFS.2017.2675361.

[16] Y. Ren and D. Ji, “Learning to detect deceptive opinion spam: A survey,” IEEE Access, vol. 7, pp. 42934–42945, 2019, doi: 10.1109/ACCESS.2019.2908495.

[17] J. Gao, X. Qian, X. Liu, and Y. Wang, "Multi-modal fake review detection with hierarchical attention fusion," Information Sciences, vol. 623, pp. 631–647, 2023.

[18] W. Chen, H. Zhang, G. Huang, and Z. Wang, "A multi-modal approach for fake review detection combining graph and language features," IEEE Access, vol. 11, pp. 24871–24882, 2023.

[19] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. AISTATS, 2017, pp. 1273–1282.

[20] C. Liu, X. Li, Y. Ye, and J. Yan, "FedDetect: A privacy-preserving federated learning framework for fraud detection," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 9, pp. 5259–5270, 2023.

[21] R. Taheri, M. Shojafar, M. Alazab, and R. Tafazolli, “Fed-IIoT: A robust federated malware detection architecture in industrial IoT,” IEEE Trans. Ind. Informatics, vol. 17, no. 12, pp. 8442–8452, Dec. 2021, doi: 10.1109/TII.2020.3043458.

[22] L. Zhao, L. Shi, and J. Yang, "FedSen: Federated learning for cross-platform sentiment analysis with privacy preservation," Pattern Recognition Letters, vol. 168, pp. 26–32, 2023.

[23] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–19, 2019.