MGLS: Malicious User Detection on Movie Review Platforms with Multi-Channel Graphs and LLM-assisted Semantic Guidance

Jianwei Xu, Haizhou Wang*
School of Cyber Science and Engineering, Sichuan University, Chengdu, China
2022141410250@stu.scu.edu.cn, whzh.nc@scu.edu.cn

***Abstract*—As online movie review platforms play an increasingly important role in users' viewing decisions and movie marketing, rating manipulation and spam flooding driven by commercial or adversarial motives have become increasingly severe, making it urgent to identify malicious users in the complex user–movie–review interaction network. However, existing studies in the movie review domain focus mainly on individual reviews and have not yet addressed malicious user detection at the user level. To fill this gap, we construct a user-level malicious user dataset based on data from the Douban movie platform and propose a malicious user detection model that uses multi-channel graphs and LLM-assisted semantic guidance (MGLS). Specifically, we first build user representations by integrating BERT-based review embeddings with sentiment intensity and off-topic probability. From four perspectives—sentiment, content, temporal patterns, and behavior—we construct multi-channel user relation graphs and use multiple GCNs to learn channel-specific representations. Finally, we design an LLM-assisted semantic guidance module and a dual-stream attention fusion mechanism that adaptively fuses GCN-learned channel weights with movie-level semantic channel weights produced by the LLM, enabling collaborative modeling of multi-channel information and robust malicious user detection. Experimental results show that our model is more effective than state-of-the-art baselines with an F1-score of 87.16% for malicious user detection in movie reviews and provides a useful reference for future work on malicious user detection in movie review platforms.**

***Index Terms*—malicious user detection; graph neural networks; multi-channel graph representations; large language model; movie review platforms**

I. Introduction

With the rapid development of the global digital entertainment industry, online movie review platforms have become key infrastructure for audiences to obtain movie information, share viewing experiences, and shape public opinion. Studies show that users commonly consult other users' ratings and reviews before watching a movie to judge its quality [1]. Since these systems heavily rely on user-generated content, the authenticity and fairness of the review ecosystem not only affect viewers' decisions and local box-office performance [2], [3], but also influence cross-regional and cross-lingual distribution patterns and the platform's brand credibility. However, driven by commercial interests or malicious attacks, some users manipulate online word-of-mouth by, among other means, giving extreme ratings in a concentrated manner and repeatedly posting templated or content-irrelevant comments under multiple films, thereby disrupting normal recommendation and decision-making processes. Therefore, accurately identifying malicious users on movie review platforms has become an important and urgent task.

Currently, malicious user detection on movie review platforms faces three main challenges:

First, there is no labeled dataset specifically targeting malicious users on movie review platforms at present.

Second, it is difficult to extract and recognize features of malicious behavior. In movie review scenarios, abnormalities often appear simultaneously in multiple dimensions such as rating bias, extreme sentiment and off-topic flooding. A single perspective cannot accurately capture users' overall behavioral patterns and easily confuse occasional anomalous reviews with persistent malicious behaviors.

Third, existing malicious user detection models are hard to transfer. Most existing methods are designed for generic social platforms, emphasizing social features like posting frequency and retweet relationships, whereas the core structure of movie review platforms is the user–movie–review triad, which relies more on information such as rating behaviors and textual content. The two differ significantly in structure and feature space, and directly applying existing models yields limited performance, making it necessary to design a new malicious user detection model for movie review platforms.

To address the challenges mentioned above, our contributions are as follows:

- **We construct the first malicious-user dataset for movie review platforms.** To the best of our knowledge, DMU-Dataset is the first publicly available user-level dataset for malicious user detection on movie review platforms, with a tailored labeling strategy.
- **We propose a user-level representation learning framework based on multi-channel graphs.** Considering that abnormal behaviors in movie review scenarios often manifest simultaneously in multiple dimensions such as rating bias, extreme sentiment, off-topic spam, and cross-movie behavioral consistency, we construct user relationship graphs from four dimensions: sentiment, content, temporal patterns, and behavior, to jointly model various user characteristics. Multi-channel GCNs are then used to learn comprehensive user representations across

*Corresponding author.

multiple movies, enabling a more complete characterization of complex malicious behavior patterns.

- **We design a malicious user detection model for movie review platforms with multi-channel graphs and LLM-assisted semantic guidance.** We introduce a dual-stream attention fusion mechanism: on one side, GCNs adaptively learn the importance of each channel from data; on the other side, based on the movie synopsis and representative critic reviews, the LLM outputs, for each movie, semantic weights and confidence scores for the four graph channels, as well as a movie-level gate. These outputs semantically evaluate the information in each channel and thus guide channel fusion. In the fusion stage, we combine the GCN-learned channel weights with the LLM-provided semantic channel weights and use a consistency regularizer to balance data-driven graph structural information and LLM-assisted semantic guidance. As far as we know, MGLS is the first framework for malicious user detection on movie review platforms that deeply integrates multi-channel graph neural networks with LLM-assisted semantic guidance.

II. RELATED WORK

Existing studies related to our task mainly fall into three lines. First, in movie review analysis, early work mainly relies on traditional machine-learning methods, while more recent studies widely adopt deep neural networks and pre-trained language models to better capture sentiment polarity, aspect-level opinions, and semantic information in movie reviews [4]–[7]. These studies provide useful techniques for understanding emotional and textual patterns in review content. Second, malicious review detection mainly focuses on identifying spam or deceptive reviews at the review level. Representative methods combine textual features with graph structures or feature interaction mechanisms to model explicit and implicit relations among reviews, thereby improving the detection of anomalous review content [8]–[11]. However, such methods generally stop at classifying individual reviews and do not explicitly model malicious users across multiple movies. Third, malicious user detection has been extensively studied on general social platforms using traditional machine learning, graph neural networks, heterogeneous graph models, and graph Transformers [12]–[24]. Nevertheless, these methods are mostly designed for social bots or spam accounts and usually rely on social relations such as following, reposting, and interaction networks. In contrast, malicious behavior on movie review platforms is more closely reflected in rating bias, review semantics, temporal patterns, and cross-movie behavioral consistency. Therefore, existing methods cannot be directly transferred to user-level malicious account detection in movie review scenarios.

III. METHOD

A. Dataset

As far as we know, there is currently no publicly available dataset for malicious user detection on movie review platforms. Therefore, we construct a user-level dataset for this task, denoted as DMU-Dataset[1]. (Douban Movie Malicious User Dataset), based on the Chinese Douban movie platform.

We collect review texts, ratings, timestamps, interaction statistics, and user/movie metadata from 434 representative movies on Douban. After data cleaning and deduplication, we aggregate the original review-level records into user-level samples by selecting users with multiple reviews across different movies.

To assign user-level labels, we jointly consider multiple dimensions, including abnormal rating patterns, repetitive or templated review text, off-topic content, concentrated temporal activity, cross-movie behavioral consistency, and interaction/activity statistics. Users with persistent anomalies across multiple dimensions are labeled as malicious, while users with mostly regular behavior are labeled as benign. Ambiguous cases are further manually inspected based on their full cross-movie histories.

The final DMU-Dataset contains 59,161 review records, 14,795 users, and 434 movies, including 9,686 benign users (38,889 reviews) and 5,109 malicious users (20,272 reviews). On average, each user has about four reviews across different movies.

B. Detection Model

Our model consists of four modules: *Representation Extraction Module*, *Multi-Channel Graph Representation Module*, *LLM-assisted Semantic Guidance Module*, and *Multi-Channel Fusion and Classification Module*. The model structure is shown in Fig. 1.

1) *Representation Extraction Module:* For each user review under movie m, we use Chinese BERT to obtain a semantic vector $h_{u,m}$, an MLP to predict the sentiment score $s^{\text{sent}}_{u,m}$, and an LLM-based relevance module to estimate the off-topic probability $q_{u,m} \in [0, 1]$. These features are combined with user behavioral statistics, such as rating deviation and interaction features, to form the user node representation.

2) *Multi-Channel Graph Representation Module:* After completing representation extraction, we construct a multi-channel user relationship graph from four perspectives: sentiment, content, time sequence, and behavior, using users as nodes. Each channel is encoded using a graph convolutional network to obtain a channel-specific user representation.

Sentiment graphs are used to characterize the similarity of users' emotional attitudes towards the same movie. Based on the sentiment score $s^{\text{sent}}_{u,m}$ obtained by the representation extraction module, we divide users into three sentiment nodes: "positive / neutral / negative" according to a set threshold. Within the same movie, we connect users with the same sentiment category and similar sentiment intensity. The smaller the sentiment difference, the larger the edge weight, thereby highlighting suspicious sentiment patterns.

Content graphs are used to characterize the semantic similarity of user comment texts. We utilize comment embeddings

[1]https://github.com/yiyepianzhounc/MGLS

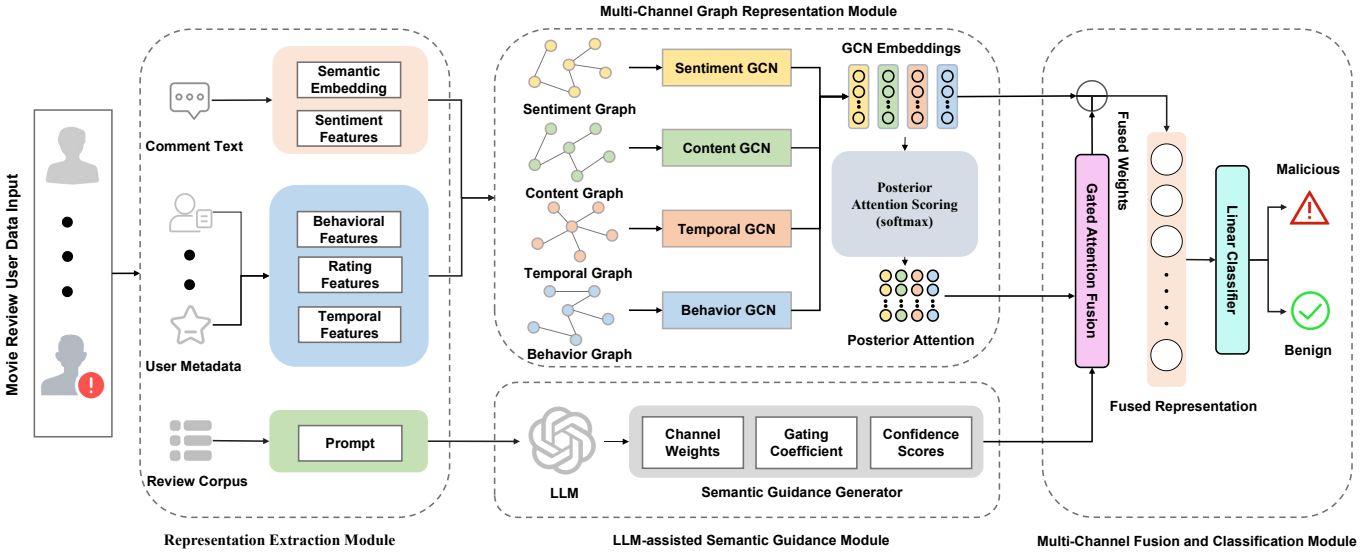

Fig. 1. Framework of MGLS: Malicious User Detection on Movie Review Platforms with Multi-Channel Graphs and LLM-assisted semantic guidance

$h_{u,m}$ to calculate the semantic similarity of users (u, v) within the same movie, and incorporate off-topic probabilities to suppress the impact of irrelevant comments:

$$s_{uv}^{c} = \cos\left(h_{u,m}, h_{v,m}\right) \tag{1}$$

$$w_{uv}^{c} = s_{uv}^{c} \cdot \left(1 - \max\left(q_{u,m}, q_{v,m}\right)\right) \tag{2}$$

When two comments are semantically similar and highly relevant to the film's content, w_{uv}^{c} is larger; if either comment is significantly off-topic, the corresponding edge weight will be significantly weakened.

Time sequence graphs are used to describe the similarity of users' commenting behavior over time, depicting a pattern of "concentrated action within the same time window". Let $t_{u,m}$, $t_{v,m}$ be the time period at which user u,v comments on movie m, and let the time sequence edge weights be defined as follows:

$$w_{uv}^{t} = \exp\left(-\frac{|t_{u,m} - t_{v,m}|}{\tau}\right) \tag{3}$$

where τ is the temporal window hyperparameter that controls how fast the weight decays with the time difference. The closer the comment times are, the closer w_{uv}^{t} is to 1; the further apart they are, the more it decays toward 0, thus emphasizing suspicious review manipulation related to the release date and time synchronization.

Behavioral graphs characterize the similarity between users based on long-term user behavior statistics. We use behavioral vectors b_u to calculate the similarity between any two users and establish behavioral edges when the similarity exceeds a threshold. The weight of the edges increases with the increase in behavioral similarity. This allows us to characterize normal, highly active user groups and clusters of accounts with similar malicious behaviors.

Each channel is encoded by a two-layer GCN to obtain channel-specific user embeddings, which are then fed into the fusion module.

3) *LLM-assisted Semantic Guidance Module:* To inject higher-level semantic information into multi-channel graph modeling, we use a large language model to evaluate the importance of each channel for every movie and treat these evaluations as semantic guidance for channel fusion. Intuitively, we hope that the LLM can answer "For this film, which type of information is more critical: emotional relationships, content similarity, temporal patterns, or behavioral characteristics?" based on the film content and representative critic reviews, and then translate this judgment into channel weights and confidence scores.

Concretely, for each movie m, we construct a prompt that includes the movie title, its synopsis d_m, and several representative professional critic reviews, and feed it to the LLM to score the four dimensions (sentiment/content/temporal/behavior). Let the raw score vector be $\tilde{r}_m \in \mathbb{R}^4$. We apply softmax to obtain a movie-level semantic channel weight vector

$$p_m = \mathrm{softmax}(\tilde{r}_m) \in \mathbb{R}^4, \tag{4}$$

where the four elements of p_m correspond to the four channels and sum to 1; a larger value means the LLM considers that channel more important for movie m. At the same time, the LLM outputs a confidence score for each channel, which we map to a confidence vector $\sigma_m \in \mathbb{R}^4$. We then rescale p_m with σ_m and renormalize it to obtain

$$\tilde{p}_m = \frac{p_m \odot \sigma_m}{\sum_{k=1}^{4} p_m^{(k)} \sigma_m^{(k)}}. \tag{5}$$

where $\odot$ denotes element-wise multiplication. In addition, the LLM outputs a movie-level gate $g_m \in [0, 1]$, which is used

TABLE I
EXPERIMENTAL RESULTS COMPARISON ON DMU-DATASET

Category	Models	Accuracy	Malicious			Benign		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Graph Learning	GAT [25]	0.6908	0.6157	0.3040	0.4070	0.7064	0.8982	0.7909
	RGCN [26]	0.8248	0.7727	0.7057	0.7377	0.8492	0.8886	0.8685
	HGT [27]	0.8263	0.7824	0.6960	0.7367	0.8461	0.8962	0.8704
	RGT [28]	0.8302	0.7686	0.7348	0.7513	0.8610	0.8814	0.8711
	S-HGN [29]	0.8410	0.7786	0.7609	0.7696	0.8733	0.8840	0.8786
Malicious User	BotRGCN [30]	0.8238	0.7702	0.7057	0.7365	0.8489	0.8871	0.8676
	RF-GNN [31]	0.8041	0.7365	0.6816	0.7068	0.8375	0.8693	0.8528
	BotDGT [22]	0.8575	0.8108	0.7792	0.7947	0.8815	**0.9004**	0.8909
	BotStip [23]	0.8115	0.7308	0.7403	0.7355	0.8566	0.8505	0.8536
	AdaBot [24]	0.8070	0.7611	0.6776	0.7169	0.8288	0.8800	0.8536
LLM	GPT-4	0.5032	0.3520	0.5080	0.4159	0.6550	0.4960	0.5645
	Gemini-3.0	0.5155	0.3787	0.6465	0.4776	0.7083	0.4472	0.5482
Malicious Movie User	**MGLS (ours)**	**0.9027**	**0.8793**	**0.8612**	**0.8716**	**0.9037**	0.8862	**0.8934**

to balance graph-learned and LLM-guided channel weights during fusion.

4) *Multi-Channel Fusion and Classification Module:* The goal of this module is to adaptively calculate the weights of the four channels for each user, fuse these channel representations, and complete malicious user classification.

First, we obtain channel scores based on the initial node features x_i and channel embeddings $z_i^{(k)}$ using a lightweight attention scoring network. We then perform softmax along the channel dimension to obtain the posterior channel weights $\alpha_i^{(k)}$ learned by the GCN, reflecting the relative importance of channel k for user i when learning purely from graph structure.

Simultaneously, for each movie m, we rescale p_m with σ_m and renormalize it to obtain $\tilde{p}_m$. We then introduce a learnable channel gating vector $\lambda \in (0,1)^4$, which balances "trusting the GCN" and "trusting the LLM". The final channel weights used for fusion are

$$\beta_i = \mathrm{norm}\big(\lambda \odot (g_m \alpha_i) + (1-\lambda) \odot ((1-g_m)\,\tilde{p}_m)\big) \quad (6)$$

where $\odot$ denotes element-wise multiplication and $\mathrm{norm}(\cdot)$ denotes normalization along the channel dimension, such that $\sum_k \beta_i^{(k)} = 1$.

To align graph-learned and LLM-guided channel weights, we introduce a consistency loss $\mathcal{L}_{\mathrm{cons}}$. The fused user representation is

$$z_i^{\mathrm{fusion}} = \sum_{k \in \{s,c,t,b\}} \beta_i^{(k)} z_i^{(k)} \quad (7)$$

A linear classifier is then trained with a binary cross-entropy classification loss $\mathcal{L}_{\mathrm{cls}}$, and the final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \gamma\, \mathcal{L}_{\mathrm{cons}}, \quad (8)$$

where γ is a hyperparameter corresponding to the consistency loss weight in the code.

IV. EXPERIMENT

This section provides a systematic evaluation of the proposed MGLS model. First, we describe the experimental configuration. Then, we evaluate the performance of MGLS through comparative, ablation and robustness experiments. Finally, the parameter robustness of MGLS is further evaluated through sensitivity analysis of key hyperparameters.

A. Experimental Configuration

In our work, all experiments are carried out on a workstation equipped with NVIDIA GeForce RTX 4090 GPU and 24GB of memory. Considering Chinese language comprehension and instruction following capabilities, as well as the manageable efficiency and cost in a single-card environment, we use Qwen2.5-7B-Instruct for both the off-topic detection and the LLM-assisted semantic guidance module. All the experiments are evaluated on DMU-Dataset, which is divided into 60% for training, 20% for validation, and 20% for test.

B. Comparative Experiment with Baselines

To evaluate the effectiveness of our model and verify its advantages over existing methods, we compare it against several baselines from three categories: general graph learning models, malicious user detection models, and large language models.

Experimental results in Table I show that MGLS significantly outperforms all comparable methods in overall *Accuracy*, as well as *Precision*, *Recall* and *F1-score* for both malicious and benign user classes. Compared to simple graph learning models and traditional malicious user detection models, MGLS can more accurately depict the complete behavioral trajectory of users on movie review platforms, thus achieving better performance. Furthermore, compared to directly relying on LLMs such as GPT-4 and Gemini-3.0 to detect malicious

users through simple inference, deeply fusing large models as semantic guidance signals with multi-channel graph neural networks achieves better performance on this task.

C. Ablation Experiment

To evaluate the contribution of key components, we conducted ablation experiments. F represents the Multi-Channel Fusion and Attention Module, S represents the LLM-assisted semantic guidance Module, and '/' represents 'none'.

TABLE II
DESCRIPTION OF FEATURE SETS

Feature set	Feature categories
MGLS	Multi-Channel Graph Representation Module, Multi-Channel Fusion and Attention Module, LLM-assisted semantic guidance Module
MGLS/S	Multi-Channel Graph Representation Module, Multi-Channel Fusion and Attention Module
MGLS/F	Multi-Channel Graph Representation Module, LLM-assisted semantic guidance Module
MGLS/S&F	Multi-Channel Graph Representation Module

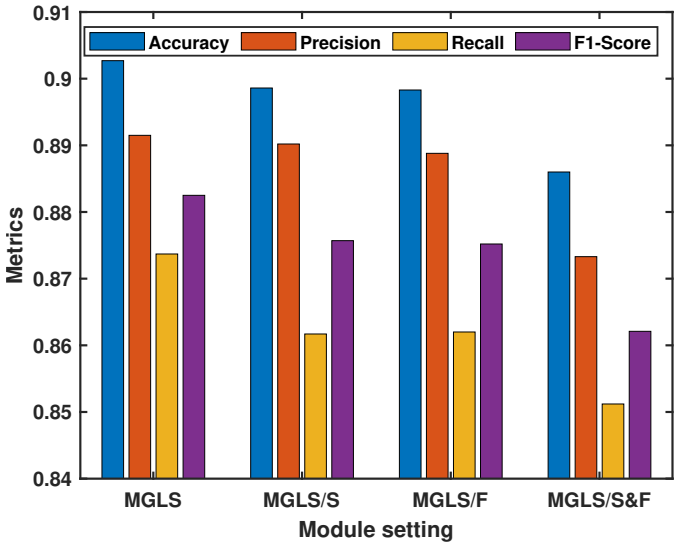

Fig. 2. Results for Ablation Experiment

Fig. 2 shows that Multi-Channel Graph Representation Module provides the core structural representation of the model. On this basis, the introduction of Multi-Channel Fusion and Attention Module as well as LLM-assisted semantic guidance Module can further significantly improve the model performance. The two form a good complementary effect on multi-channel graph representation, which significantly improves the performance of malicious user detection.

D. Robustness Experiment

We randomly label the training set incorrectly according to a certain ratio (5%-45%), and train our model and the baseline models to test the robustness under different noise levels. The experimental results are shown in Fig. 3. The results show that as the noise rate continues to increase, MGLS consistently performs better than the baselines under different noise levels. The F1-score curve of MGLS is smoother and has a smaller slope in terms of different proportions, which means MGLS has excellent robustness.

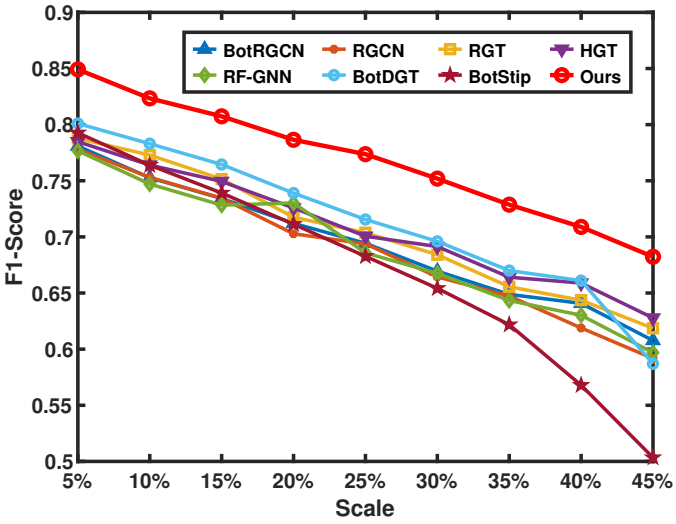

Fig. 3. Results for Robustness Experiment

E. Parameter Sensitivity Experiment

On the DMU-Dataset, we further analyze the sensitivity of MGLS to two key hyperparameters: the consistency loss weight γ, which controls the strength of the KL regularizer between the GCN-learned and LLM-guided channel weights, and the temporal window τ, which controls temporal edge construction in the time-sequence graph; the results are shown in Fig. 4.

As γ increases from 0 to 0.10, both F1-score and accuracy improve and peak at $\gamma = 0.10$, then slightly decline at 0.15 and 0.20, indicating that a moderate consistency constraint best balances GCN and LLM signals. For τ, the best performance is obtained at 14 days, while shorter or longer windows degrade performance. Consequently, we adopt $\gamma = 0.10$ and $\tau = 14$ days in all experiments.

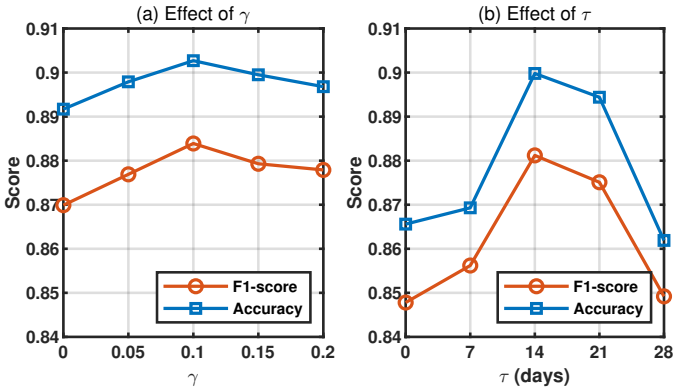

Fig. 4. Results for Parameter Sensitivity Experiment

V. CONCLUSION

This paper studies malicious user detection on movie review platforms. We construct a user-level malicious behavior dataset and propose MGLS, a model that integrates multi-channel graph representation with LLM-assisted semantic guidance. Specifically, we build user representations from textual semantics, sentiment, off-topic information, and behavioral statistics, and construct multi-channel user graphs from sentiment, content, temporal, and behavioral perspectives. Channel-specific representations are learned by multiple two-layer GCNs, and a dual-stream attention fusion mechanism is introduced to combine GCN-based channel weights with LLM-provided semantic weights.

Experiments on DMU-Dataset show that MGLS consistently outperforms strong baselines. Ablation, robustness, and sensitivity studies further confirm the importance of multi-channel modeling and semantic weighting.

In future work, we will extend MGLS to other review platforms, explore its cross-platform and cross-lingual transferability, investigate labeling bias and its impact on fairness, and develop more interpretable explanation mechanisms for practical deployment.

REFERENCES

[1] H. Ma, J. M. Kim, and E. Lee, "Analyzing dynamic review manipulation and its impact on movie box office revenue," *Electronic Commerce Research and Applications*, vol. 35, p. 100840, 2019.

[2] P. K. Chintagunta, S. Gopinath, and S. Venkataraman, "The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets," *Marketing Science*, vol. 29, pp. 944–957, 2010.

[3] Y. Wu, E. W. Ngai, P. Wu, and C. Wu, "Fake online reviews: Literature review, synthesis, and directions for future research," *Decision Support Systems*, vol. 132, p. 113280, 2020.

[4] A. Rahman and M. S. Hossen, "Sentiment Analysis on Movie Review Data Using Machine Learning Approach," in *Proceedings of 2nd IEEE International Conference on Speech and Language Processing (ICBSLP)*, 2019, pp. 1–4.

[5] V. U. Ramya and K. T. Rao, "Sentiment Analysis of Movie Review using Machine Learning Techniques," *International Journal of Engineering & Technology*, vol. 3, p. 676, 2018.

[6] M. M. Danyal, S. S. Khan, M. Khan, S. Ullah, F. Mehmood, and I. Ali, "Proposing sentiment analysis model based on BERT and XLNet for movie reviews," *Multimedia Tools and Applications*, vol. 83, pp. 64 315–64 339, 2024.

[7] G. Yang, Y. Xu, and L. Tu, "An intelligent box office predictor based on aspect-level sentiment analysis of movie review," *Wireless Networks*, vol. 29, pp. 3039–3049, 2023.

[8] Y. Gao, M. Gong, Y. Xie, and A. K. Qin, "An Attention-Based Unsupervised Adversarial Model for Movie Review Spam Detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 784–796, 2020.

[9] L. Zhang, X. Song, X. Zhao, Y. Fang, D. Li, and H. Wang, "GAIM: Graph-aware Feature Interactional Model for Spam Movie Review Detection," in *Proceedings of 26th IEEE International Conference on Pattern Recognition (ICPR)*, 2022, pp. 621–628.

[10] H. Cao, H. Li, Y. He, X. Yan, F. Yang, and H. Wang, "GCMK: Detecting Spam Movie Review Based on Graph Convolutional Network Embedding Movie Background Knowledge," in *Proceedings of 31st International Conference on Artificial Neural Networks (ICANN)*, 2022, pp. 494–505.

[11] Y. Cai, H. Wang, H. Cao, W. Wang, L. Zhang, and X. Chen, "Detecting Spam Movie Review Under Coordinated Attack With Multi-View Explicit and Implicit Relations Semantics Fusion," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 7588–7603, 2024.

[12] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting Burstiness in Reviews for Review Spammer Detection," in *Proceedings of 17th international AAAI conference on Web and Social Media (ICWSM)*, 2013, pp. 175–184.

[13] X. Kan, Y. Fan, J. Zheng, A. Kudreyko, C.-h. Chi, W. Song, and A. Tregubova, "User-level malicious behavior analysis model based on the NMF-GMM algorithm and ensemble strategy," *Nonlinear Dynamics*, vol. 111, pp. 21 391–21 408, 2023.

[14] D. Koggalahewa, Y. Xu, and E. Foo, "An unsupervised method for social network spammer detection based on user information interests," *Journal of Big Data*, vol. 9, p. 7, 2022.

[15] P. Agarwal, M. Srivastava, V. Singh, and C. Rosenberg, "Modeling User Behavior With Interaction Networks for Spam Detection," in *Proceedings of 45th international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2022, pp. 2437–2442.

[16] L. Deng, C. Wu, D. Lian, Y. Wu, and E. Chen, "Markov-Driven Graph Convolutional Networks for Social Spammer Detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 12 310–12 322, 2022.

[17] L. Lin, Y. Huang, L. Xu, and S.-Y. Hsieh, "Better Adaptive Malicious Users Detection Algorithm in Human Contact Networks," *IEEE Transactions on Computers*, vol. 71, pp. 2968–2981, 2022.

[18] K. Gu, D. Yang, W. Zhao, and X. Li, "Overlapping community-based malicious user detection scheme in social networks," *Knowledge-Based Systems*, vol. 312, p. 113139, 2025.

[19] M. Song, Z. Hua, Y. Zheng, R. Lan, Q. Liao, and G. Xu, "Enabling Reliable and Anonymous Data Collection for Fog-Assisted Mobile Crowdsensing With Malicious User Detection," *IEEE Transactions on Mobile Computing*, vol. 25, pp. 1414–1430, 2025.

[20] K. Lu, H. Zhang, Y. Yang, and B. Fang, "Multi-Relation Aware Heterogeneous Graph Transformer for Robust Spammer Detection via Contrastive Learning in Social Networks," in *Proceedings of 68th IEEE International Conference on Communications (ICC)*, 2025, pp. 4683–4688.

[21] F. Zhang, C. Huo, R. Ma, and J. Chao, "Local interpretable spammer detection model with multi-head graph channel attention network," *Neural Networks*, vol. 184, p. 107069, 2025.

[22] B. He, Y. Yang, Q. Wu, H. Liu, R. Yang, H. Peng, X. Wang, Y. Liao, and P. Zhou, "BotDGT: Dynamicity-aware Social Bot Detection with Dynamic Graph Transformers," *arXiv preprint arXiv:2404.15070*, 2024.

[23] F. Liu and R. Ma, "From the past to the present: A social bot detection method based on spatio-temporal interactive perception," *Knowledge-Based Systems*, p. 113712, 2025.

[24] H. Sun, H. Peng, Y. Cao, Q. Dai, and X. Bai, "Adaptive Social Bot Detection through Bridging the Feature Bias Between Source and Target Users," in *Proceedings of 33rd International Conference on Multimedia Retrieval (ICMR)*, 2025, pp. 1228–1236.

[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[26] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks," in *Proceedings of 15th European Semantic Web Conference (ESWC)*, 2018, pp. 593–607.

[27] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proceedings of 29th ACM International Conference on World Wide Web (WWW)*, 2020, pp. 2704–2710.

[28] S. Feng, Z. Tan, R. Li, and M. Luo, "Heterogeneity-Aware Twitter Bot Detection with Relational Graph Transformers," in *Proceedings of 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022, pp. 3977–3985.

[29] Q. Lv, M. Ding, Q. Liu, Y. Chen, W. Feng, S. He, C. Zhou, J. Jiang, Y. Dong, and J. Tang, "Are we really making much progress?: Revisiting, benchmarking and refining heterogeneous graph neural networks," in *Proceedings of 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, 2021, pp. 1150–1160.

[30] S. Feng, H. Wan, N. Wang, and M. Luo, "BotRGCN: Twitter bot detection with relational graph convolutional networks," in *Proceedings of 13rd IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2021, pp. 236–239.

[31] S. Shi, K. Qiao, J. Yang, B. Song, J. Chen, and B. Yan, "RF-GNN: Random Forest Boosted Graph Neural Network for Social Bot Detection," *arXiv preprint arXiv:2304.08239*, 2023.