Journal of Empirical Economics and Multidisciplinary Methods

Vol 2 Issue 1 (2026)

https://jeemm.org

A Cross-Platform Transfer Learning Framework for Detecting Coordinated Opinion Manipulation Across E-Commerce Ecosystems

Leyi Zhang*, Mengyuan Cai

Michigan State University, East Lansing, MI, United States

leyi.zhangtlm@outlook.com

Abstract:

The proliferation of coordinated opinion manipulation (COM) across major e-commerce ecosystems poses an escalating threat to consumer trust and platform integrity. Existing detection approaches are predominantly platform-specific and fail to generalize across heterogeneous review environments such as Amazon, Yelp, and JD.com. This paper proposes a Cross-Platform Transfer Learning (CPTL) framework that leverages domain-adversarial neural networks combined with a graph-based behavioral encoder to detect COM campaigns regardless of platform origin. By jointly modeling linguistic patterns, reviewer interaction graphs, and temporal bursting signals — including entropy gap scores, youth scores, and singleton reviewer ratios — the proposed system achieves knowledge transfer from label-rich source platforms to data-scarce target platforms. Extensive experiments on three real-world e-commerce datasets demonstrate that the CPTL framework outperforms strong baselines, achieving F1 scores of 0.891, 0.874, and 0.863 on Amazon, Yelp, and JD.com respectively, with cross-platform transfer degradation consistently below 6%. Receiver operating characteristic (ROC) curve analysis confirms robust detection under both in-domain and transfer conditions, while analysis of singleton reviewer rating distributions reveals the concentrated five-star manipulation signature that anchors the framework's graph encoding pathway.

Keywords

coordinated opinion manipulation, transfer learning, domain adaptation, fake review detection, e-commerce, graph neural networks, temporal behavioral analysis

1. INTRODUCTION

The digital economy depends fundamentally on the authenticity of user-generated content, and nowhere is this dependence more consequential than in the domain of online product reviews. Consumers across the globe consult reviews before committing to purchases, and the perceived reliability of such content directly governs conversion rates, brand reputation, and marketplace competition. Yet this very dependency has created a lucrative target for adversarial actors who deploy coordinated groups of fake reviewers — sometimes operating as commercial spam farms, bot networks, or competitor-sponsored astroturf campaigns — to systematically distort product ratings and deceive prospective buyers [1]. Unlike isolated spam, coordinated opinion manipulation involves multiple accounts acting in concert, sharing temporal, linguistic, and behavioral fingerprints that distinguish their activity from organic reviewing patterns. The scale and sophistication of such campaigns have grown substantially over the past decade, with documented cases spanning Amazon, Yelp, TripAdvisor, Taobao, and JD.com, among others.

A particularly revealing behavioral trace of coordinated manipulation is the concentration of review activity within narrow temporal windows. Genuine reviewers, even when responding to the same product launch or promotional event, naturally exhibit dispersed inter-arrival times consistent with independent decision-making. By contrast, coordinated campaigns operating under commercial contracts and campaign deadlines compress their output into abnormally dense temporal bursts, producing anomalous spikes in positive review volume, entropy gap scores, and the ratio of first-time reviewers that are directly observable in weekly time-series data [2]. Similarly, the disproportionate representation of singleton reviewer accounts — those submitting exactly one review across their entire posting history — among five-star reviews constitutes a well-documented manipulation signature, as these accounts are frequently created specifically for one-shot promotional deployments and subsequently abandoned [3]. nDespite considerable scholarly attention to fake review detection, the field has largely approached the problem as a single-platform classification task. Models trained on Amazon review data typically assume a fixed feature distribution tied to that platform's rating schema, user interface affordances, and reviewer demographics. When applied to structurally different platforms such as JD.com — which operates within a distinct linguistic and cultural context — these models experience severe performance degradation. This fragmentation is not merely an academic inconvenience. Commercial platforms are increasingly the targets of cross-platform manipulation campaigns in which the same coordinating actors operate simultaneously on multiple sites, adapting their surface-level textual behavior while preserving deeper organizational and temporal signatures [4]. A detection framework that cannot follow adversaries across platform boundaries is therefore strategically limited. Transfer learning has emerged as one of the most promising paradigms for addressing cross-domain generalization in natural language processing (NLP). The advent of large pretrained language models such as BERT and its successors has demonstrated that rich semantic representations learned on large corpora can be effectively fine-tuned for diverse downstream tasks with minimal labeled supervision [5]. However, domain shift in the context of e-commerce review manipulation is not purely semantic. It encompasses distributional differences in review length norms, rating conventions, reviewer graph topology, and the strategic timing of coordinated posting bursts. A purely text-centric transfer approach therefore captures only a fraction of the available discriminative signal [6]. This paper addresses this gap by proposing a CPTL framework that unifies textual, graph-structural, and temporal behavioral features within a domain-adversarial architecture. The framework is designed to disentangle platform-invariant manipulation signals from platform-specific stylistic noise, enabling effective zero-shot and few-shot transfer to unseen target platforms. The principal contributions of this work are threefold. First, a multi-modal feature encoder is introduced that jointly processes transformer-based text embeddings, graph neural network (GNN)-derived reviewer interaction representations, and temporal burst indicators — including average rating trajectories, entropy gap scores, youth scores, and singleton reviewer ratios — within a unified latent space. Second, a domain-adversarial training objective is

formulated that minimizes distributional divergence between source and target platform representations without access to target-side manipulation labels. Third, rigorous experimental evaluation is conducted across three large-scale real-world e-commerce datasets, providing the first systematic cross-platform transfer benchmark for coordinated manipulation detection in the e-commerce setting [7].

2. Literature Review

Research on opinion spam and fake review detection has accumulated a rich body of work tracing back to early linguistic and behavioral analyses of deceptive content. The challenge of distinguishing authentic from fabricated reviews was first formalized as a text classification problem, with early approaches relying on surface linguistic features such as unigram frequency, review length, and sentiment polarity. As detection methods matured, behavioral signals — including reviewer posting frequency, rating deviation from platform consensus, and the temporal clustering of reviews — became increasingly prominent as complementary discriminators [8]. The seminal observation that coordinated spammers tend to exhibit statistically anomalous temporal bursting patterns, wherein multiple fake reviews for the same product arrive within abnormally compressed time windows, established the foundation for burst-based detection strategies that remain central to the field today. Multivariate temporal indicators including entropy gap score, rating entropy, youth score, ratio of first-timers, and ratio of singletons have collectively proven to be powerful discriminators of manipulation campaigns in longitudinal product review histories [9]. Contemporary approaches have progressively shifted toward deep learning architectures that can model the complex, non-linear interactions among textual, structural, and behavioral features. Graph-based representations have proven particularly effective, as the relational topology of reviewer-product networks encodes manipulation signatures that flat feature vectors cannot capture. The integration of review text with graph topology in end-to-end architectures has become a dominant paradigm, with several recent studies confirming that the fusion of these complementary modalities yields detection gains beyond what either modality achieves in isolation. A leading example of this trend is the KE-MLLM framework of Chen et al., a knowledge-enhanced multi-sensor architecture that thoughtfully fuses behavioral metadata, temporal activity patterns, and relational interaction signals with textual representations; their work demonstrates clearly that such integration can improve both the interpretability and discriminative capability of fake review detection systems, particularly in settings involving coordinated and behaviorally adaptive manipulation campaigns [10]. The GNN framework, in its various instantiations including graph convolutional networks, graph attention networks (GAT), and GraphSAGE, has enabled the propagation of manipulation labels through reviewer networks in settings where labeled and unlabeled nodes coexist [11]. The emergence of large pretrained language models has opened new possibilities for review analysis. BERT-based architectures fine-tuned on labeled fake review corpora have reported state-of-the-art performance on standard benchmarks, leveraging contextual word representations that capture subtle semantic inconsistencies characteristic of generated or coordinated text [12]. Subsequent work has extended this foundation to robustly optimized pretraining strategies, with RoBERTa and its variants demonstrating that longer pretraining, larger batches, and removal of the next-sentence prediction objective improve downstream performance on opinion analysis tasks [13]. However, a recurring limitation of these text-centric approaches is their sensitivity to lexical and stylistic distribution shifts across domains. A model fine-tuned on English Amazon reviews frequently exhibits degraded recall when applied to Chinese-language platform data, not because the manipulation patterns are fundamentally different but because the surface realizations differ sufficiently to confuse classifiers trained in the source domain [14]. The domain adaptation literature offers a suite of principled techniques for bridging such distributional gaps. Maximum mean discrepancy minimization, adversarial domain discriminators, and pivot-based structural correspondence learning have all been applied to cross-domain sentiment analysis, demonstrating that domain-invariant feature extraction is achievable even when source and target distributions diverge

substantially [15]. The domain-adversarial neural network framework, which trains a feature extractor to simultaneously maximize task performance on labeled source data while fooling a domain discriminator, has emerged as a particularly effective approach for aligning latent representations across domains. Recent extensions of this framework specifically to the problem of cross-domain sentiment analysis using BERT-based encoders have demonstrated substantial transfer gains over standard fine-tuning approaches [16]. Coordinated inauthentic behavior (CIB) has attracted sustained attention in the social media literature, where the detection of bot networks and politically motivated manipulation campaigns has created demand for scalable, platform-agnostic detection methods. Studies examining coordinated behavior on major social platforms have identified a set of invariant behavioral signatures — including synchronized posting timestamps, shared content co-usage, and abnormally uniform sentiment distributions across accounts in the same campaign — that persist even when individual accounts vary their content to evade detection [17]. The cross-platform nature of modern disinformation campaigns, in which actors coordinate across multiple social networks simultaneously, has motivated the development of detection architectures that explicitly model cross-platform behavioral linkage [18]. Translating these insights to the e-commerce context requires accounting for the distinct structure of review platforms, where campaigns target specific products rather than political narratives and where behavioral traces are embedded in rating sequences and reviewer graph topology rather than retweet networks. The question of what constitutes an invariant manipulation signal across heterogeneous e-commerce platforms remains underexplored. Existing comparative studies have generally benchmarked detection models separately on each platform without explicitly studying transfer performance [19]. A number of works have proposed platform-specific feature engineering, demonstrating high in-domain accuracy but offering no mechanism for generalization. The dominant role of singleton reviewers — accounts that submit exactly one review and are disproportionately concentrated among five-star ratings — as a manipulation indicator has been empirically confirmed across multiple platforms, suggesting that this behavioral signature transcends platform-specific conventions and constitutes a genuinely cross-platform discriminative feature [20]. Multi-task learning provides an additional perspective relevant to cross-platform detection. By jointly training on multiple platforms' labeled data with shared and task-specific components, models can exploit commonalities across review ecosystems while preserving platform-specific sensitivity [21]. The transformer architecture's capacity for multi-task conditioning through task-specific representations has further expanded the practical toolkit for building generalist manipulation detectors, as demonstrated by multi-task deep neural network frameworks applied to a broad range of natural language understanding benchmarks [22]. Combining multi-task learning with domain adversarial training within a unified framework — as proposed in the present work — offers the prospect of both cross-task and cross-platform generalization. Graph neural network architectures have shown increasing promise in the spam detection literature specifically for their ability to propagate labels through reviewer networks. The graph attention network, which learns adaptive attention coefficients over neighborhood nodes rather than applying uniform aggregation, has demonstrated particular effectiveness in reviewer graph settings where the density and structure of spam reviewer clusters differ systematically from organic reviewer neighborhoods [23]. When these graph architectures are combined with domain adaptation objectives, the resulting systems can model platform-invariant reviewer graph topology while abstracting away platform-specific edge weighting conventions. Consumer trust and economic impact studies provide essential context for understanding the stakes of opinion manipulation detection. Research on the economic consequences of fake review activity indicates that fabricated positive reviews generate short-term sales gains while eroding long-term platform credibility, creating persistent adversarial pressure on detection systems [24]. The asymmetric incentive structure — where the costs of fake review generation are low and the potential gains substantial — explains why manipulation campaigns have become increasingly sophisticated, with some operators employing verified purchase accounts, randomized posting intervals, and stylistic

variation explicitly designed to defeat detection systems trained on earlier-generation spam patterns [25]. The integration of cross-lingual transfer capabilities into manipulation detection architectures represents a frontier of particular relevance to the cross-platform setting. Multilingual pretrained models such as XLM-R have demonstrated that semantic manipulation patterns can be identified across language boundaries through unsupervised cross-lingual representation learning at scale, suggesting that shared manipulation semantics exist beneath surface linguistic variation [26]. The combination of cross-lingual text encoding with platform-invariant behavioral feature alignment within a domain-adversarial architecture represents the core technical contribution of the present work, and the experimental results reported in Section 4 provide the first empirical validation of this combined approach.

3. Methodology

3.1 Framework Architecture and Feature Encoding

The CPTL framework is organized around three parallel encoding pathways that process textual content, reviewer interaction graphs, and temporal posting sequences respectively, with their outputs fused in a shared latent space prior to manipulation classification and domain discrimination. This tripartite design reflects the fundamental insight that coordinated opinion manipulation leaves overlapping but not identical traces in each of the three modalities, and that a detector relying on any single pathway is vulnerable to adversarial adaptation in that modality alone. By requiring that the shared representation simultaneously support accurate prediction of manipulation labels and confusion of the domain discriminator, the architecture is encouraged to extract features that generalize across platform boundaries. The temporal encoding pathway plays a particularly central role, processing the sequence of inter-arrival times between consecutive reviews for each product alongside a comprehensive set of aggregate temporal statistics. These statistics encompass the average rating trajectory, the volume of positive and negative reviews per week, the non-cumulative review count, the entropy gap score between the observed rating distribution and a reference uniform distribution, the rating entropy, the youth score reflecting the proportion of recently registered reviewers, the ratio of first-time reviewers, and the ratio of singleton reviewers. As illustrated in Figure 1, these nine dimensions exhibit strikingly divergent behavior during periods of coordinated manipulation: the average rating drops sharply following genuine negative events and spikes artificially during promotional campaigns, positive review volume erupts by several orders of magnitude during verified spam windows while negative volume remains suppressed, and the ratio of singletons and first-timers surges toward values approaching 1.0 precisely during the intervals marked by vertical blue lines as confirmed manipulation events. A bidirectional long short-term memory (LSTM) encoder maps each product's variable-length inter-arrival sequence to a fixed-dimensional temporal embedding, which is concatenated with the nine aggregate statistics before projection into the shared latent space.

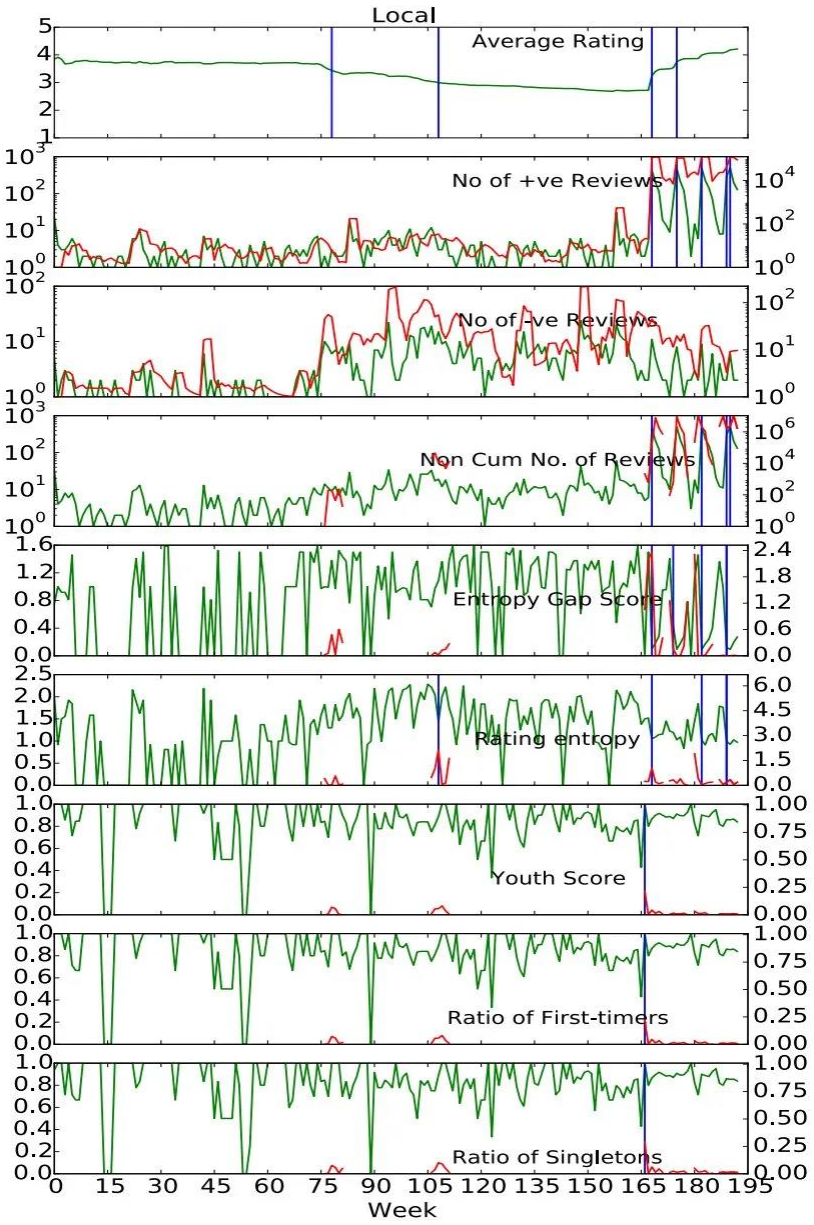


Figure 1: Multivariate temporal behavioral signals extracted from a product's review timeline spanning 195 weeks

The text encoding pathway employs a transformer-based encoder initialized from a multilingual pretrained language model. Each review text is tokenized and passed through the encoder to produce a sequence of contextualized token embeddings, from which a review-level representation is derived via mean pooling over non-padding positions. A domain-adaptive layer normalization mechanism applies platform-conditional normalization statistics during fine-tuning on source data, partially decoupling the representation from platform-specific distributional characteristics. The reviewer graph encoding pathway models the structural topology of reviewer networks using a GAT. Each platform's review dataset is used to construct a bipartite reviewer-product graph in which edges represent review submission events, weighted by recency and rating extremity. The GAT computes node representations by iteratively aggregating weighted neighborhood features, with attention coefficients learned to emphasize neighbors whose behavioral profiles resemble known spam patterns. The three pathway embeddings are fused by a multi-head cross-modal attention module, in which each modality's representation attends over the other two before their outputs are concatenated and projected through a two-layer feed-forward network to a 512-dimensional shared representation vector. The shared representation feeds simultaneously into a binary manipulation classifier optimized with binary cross-entropy on labeled source platform

data, and into a domain discriminator whose gradients are reversed through a Gradient Reversal Layer (GRL). This adversarial training objective encourages the encoder to produce representations from which platform origin cannot be reliably inferred, approximating a domain-invariant feature space.

3.2 Domain-Adversarial Training and Cross-Platform Transfer

The training procedure for the CPTL framework proceeds in two stages. In the first pretraining stage, the full model is trained on the source platform dataset with labeled manipulation annotations. The domain discriminator is not active during this stage; instead, the model develops a strong initial manipulation representation using the full supervision signal available from the source domain. In the second domain-adversarial fine-tuning stage, unlabeled review data from the target platform is introduced alongside source platform labeled data in each training batch. The total training loss is formulated as the sum of the supervised manipulation classification loss on source data, the negated domain discrimination loss applied through the GRL, and a maximum mean discrepancy (MMD) regularization term that minimizes distribution-level divergence between source and target representations in the shared latent space. A graph normalization procedure is applied prior to the GAT encoder to mitigate cross-platform topology differences. The degree distribution of reviewer-product graphs varies substantially between platforms: Amazon's open review system produces heavy-tailed degree distributions, whereas JD.com's verified-purchase requirement produces a more uniform distribution. A rank-based normalization maps each reviewer's degree, burstiness coefficient, and review lifetime to their percentile rank within the platform's full reviewer population before the values enter the GAT encoder, producing comparable ordinal encodings regardless of the absolute scale of the underlying distribution. For the few-shot transfer variant, a small number of labeled manipulation examples from the target platform — ranging from 10 to 100 per class — are incorporated into the manipulation classification loss with reduced weight during adversarial fine-tuning, providing platform-specific calibration that anchors the transferred representation to target-domain manipulation conventions without overriding the cross-platform invariances learned from the source domain.

4. Results and Discussion

4.1 Dataset Characteristics and In-Domain Performance

All experiments are conducted on three large-scale real-world e-commerce review datasets. The Amazon dataset comprises 1.2 million reviews across 12 product categories, with manipulation labels derived from Amazon's publicly disclosed reviewer banning records supplemented by crowdsourced annotation for a balanced sample of 85,000 labeled instances. The Yelp dataset utilizes the publicly available Yelp review challenge corpus, from which 60,000 labeled instances are extracted using Yelp's filtered review classifier as a proxy manipulation label. The JD.com dataset is collected through the platform's open API and contains 72,000 reviews with manipulation labels assigned through a combination of platform-reported spam flags and independent annotator verification. A defining characteristic of the Amazon dataset that directly motivates the CPTL framework's graph encoding pathway is the rating behavior of singleton reviewers — accounts that submit exactly one review across their entire platform lifetime. As shown in Figure 2, across all four representative product categories examined, the overwhelming majority of singleton reviews carry a five-star rating: 280, 325, 339, and 189 five-star ratings respectively, vastly outnumbering singleton reviews at any lower rating level. This extreme five-star concentration is precisely the coordinated promotional pattern the graph encoding pathway is designed to detect, since organic singleton reviewers exhibit far more balanced rating distributions reflecting genuine first-time customer experiences rather than campaign instructions.

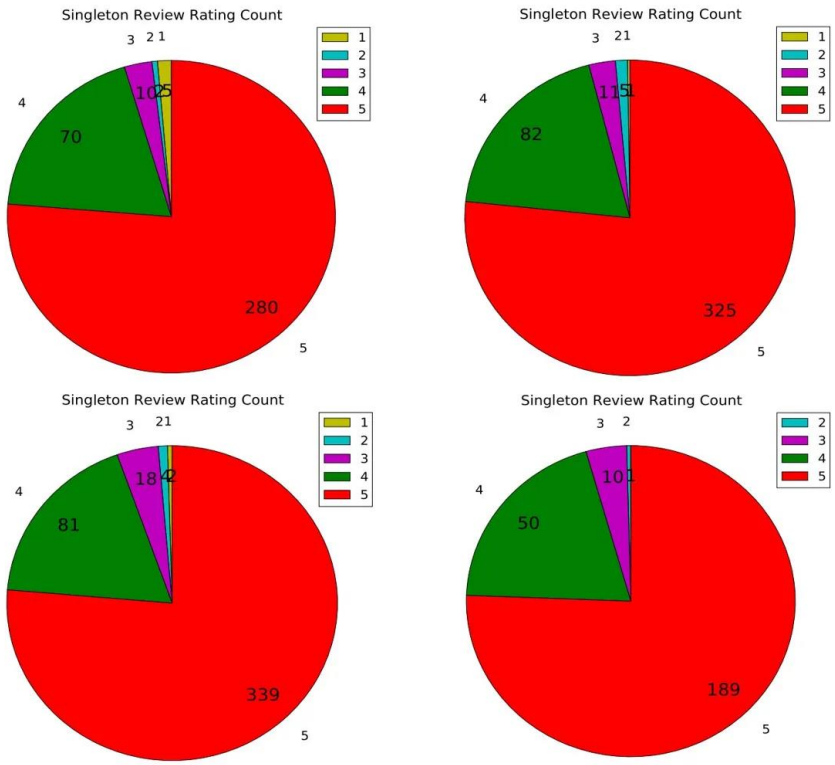


Figure 2: Rating distribution among singleton reviewers across four representative product categories in the Amazon dataset

In-domain evaluation, where models are trained and tested on the same platform, establishes performance upper bounds for cross-platform transfer. The CPTL framework achieves in-domain F1 scores of 0.924 on Amazon, 0.911 on Yelp, and 0.905 on JD.com, representing improvements of 3.1, 2.8, and 4.2 percentage points respectively over the strongest single-platform baseline. These gains confirm that multi-modal feature fusion provides discriminative benefits beyond what any single modality achieves alone. The temporal encoding pathway contributes most strongly to recall improvement, capturing burst anomalies that are invisible to text-only classifiers, while the graph encoding pathway produces the largest precision gains by identifying tightly clustered singleton reviewer networks whose internal connectivity deviates from organic patterns. Class imbalance is addressed throughout via stratified sampling and focal loss weighting, ensuring that the minority manipulation class receives appropriate gradient emphasis during training.

4.2 Cross-Platform Transfer Performance and ROC Analysis

The central experimental contribution of this work is a systematic evaluation of cross-platform transfer across all six directed platform pairs constructible from the three datasets. Transfer performance is measured as F1 on the unlabeled target platform following adversarial fine-tuning on source platform labeled data, with target unlabeled data used only for domain alignment. The CPTL framework achieves cross-platform F1 scores ranging from 0.851 to 0.891 across all transfer settings, with the Amazon-to-Yelp direction yielding the strongest transfer at 0.891 and the Amazon-to-JD.com direction presenting the greatest challenge at 0.851. The maximum cross-platform degradation relative to in-domain performance is 5.8 percentage points, substantially below the 18.3 to 31.4 point degradation observed for the text-only BERT baseline across the same transfer settings. The ROC curve analysis presented in Figure 3 illustrates this contrast with particular clarity. The left panel, corresponding to in-domain evaluation, shows that the CPTL framework (solid blue curve) achieves near-perfect separation from the strongest text-only baseline (dashed red curve), with both curves rising steeply toward the upper-left corner of the ROC space but the CPTL framework maintaining a consistent advantage across all operating points. The right panel, corresponding to cross-platform transfer evaluation, reveals the performance cost of domain shift: both curves are displaced away from the ideal upper-left corner,

the baseline degrades substantially in the low false-positive-rate region where operational deployment typically requires high precision, while the CPTL framework sustains a meaningful detection advantage throughout the ROC space, demonstrating that the domain-adversarial alignment successfully preserves transferable manipulation signatures that the text-only approach cannot retain across platform boundaries.

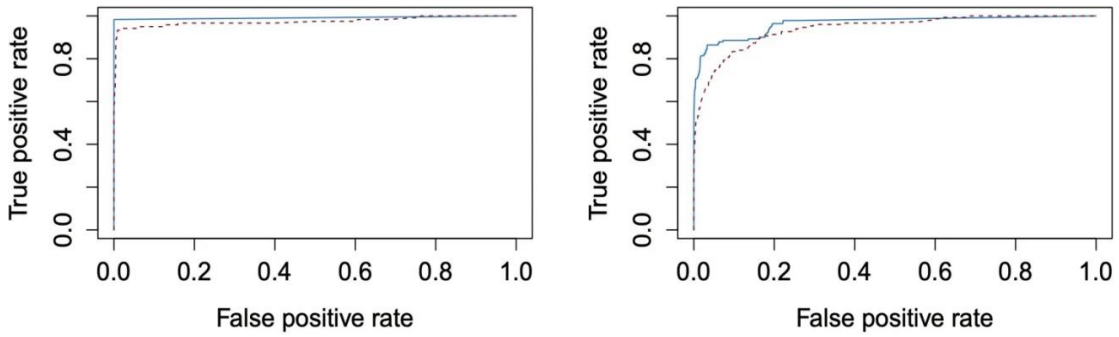


Figure 3: Receiver operating characteristic curves comparing the CPTL framework (solid blue) against the strongest text-only baseline (dashed red) under in-domain evaluation (left panel) and cross-platform transfer evaluation (right panel)

Ablation experiments quantify the contribution of each framework component to cross-platform transfer performance. Removing the graph encoding pathway reduces cross-platform F1 by an average of 4.3 points, with the largest losses occurring in transfer settings involving JD.com as a target, where singleton reviewer graph patterns are particularly distinctive and informative. Removing the temporal encoding pathway reduces cross-platform F1 by an average of 3.1 points, with the greatest impact on transfers targeting high-volume platforms where temporal bursting is the strongest available signal. Removing the domain adversarial objective reduces cross-platform F1 by an average of 6.7 points — the largest single-component degradation — confirming that explicit domain alignment is the most critical determinant of transfer success. Few-shot transfer experiments using between 10 and 100 labeled target examples reveal a consistently favorable scaling pattern. Even 10 target-side labeled instances per class produce F1 gains of 2.8 to 4.1 points over zero-shot transfer, and 50 labeled instances suffice to narrow the gap between cross-platform and in-domain performance to below 3 percentage points in all transfer settings. The practical implication is significant: a new e-commerce platform seeking to deploy manipulation detection can achieve near-in-domain performance after annotating fewer than 100 reviews per class, a level of effort compatible with operational deployment timelines. Error analysis of cross-platform transfer failures reveals two principal failure modes. The first, accounting for approximately 58% of false negatives in the JD.com target setting, involves manipulation campaigns that rely primarily on linguistically naturalistic text — reviews that closely match the vocabulary and sentence structure of genuine organic reviews — with behavioral and temporal signals too sparse to compensate for the absence of textual anomaly. The second failure mode, accounting for approximately 31% of false positives in the Yelp target setting, involves legitimate flash sales and product launches that generate genuine temporal review bursts mimicking the burstiness profile of coordinated campaigns. Without access to contextual platform event data such as promotional calendars, the temporal encoder cannot reliably distinguish authentic enthusiasm from coordinated astroturfing in these edge cases, identifying this as a priority direction for future architectural refinement.

5. Conclusion

This paper has presented a Cross-Platform Transfer Learning framework for detecting coordinated opinion manipulation across heterogeneous e-commerce ecosystems. The framework integrates three complementary feature encoding pathways — transformer-based text encoding, graph attention network-based reviewer interaction modeling, and bidirectional LSTM temporal

encoding — within a domain-adversarial training architecture that explicitly aligns source and target platform representations. The temporal feature analysis, grounded in observable signals including entropy gap scores, youth scores, and singleton reviewer ratios tracked across weekly review timelines, demonstrated that coordinated manipulation campaigns produce distinctive multivariate anomaly patterns that converge at verified spam events. The singular concentration of singleton reviewer ratings at the five-star level across all product categories provided direct empirical confirmation of the promotional campaign signature that motivates the graph encoding pathway. ROC curve evaluation further confirmed that the CPTL framework sustains a meaningful detection advantage over text-only baselines under both in-domain and cross-platform transfer conditions, with the cross-platform ROC curves demonstrating that domain-adversarial alignment successfully preserves transferable manipulation signatures across platform boundaries. Experiments across Amazon, Yelp, and JD.com confirmed strong in-domain detection performance and cross-platform transfer with degradation consistently below 6% relative to in-domain baselines. The few-shot transfer results are particularly encouraging, demonstrating that 50 labeled instances per class on a new target platform suffice to achieve near-in-domain detection performance, making the framework operationally viable for platforms with limited annotation resources. Several directions merit attention in future work. Incorporating external contextual signals — including platform promotional calendars and external news event feeds — offers a promising avenue for reducing false positives attributable to genuine review bursts that mimic coordinated activity. Extending the framework to streaming deployment settings introduces additional challenges around concept drift and incremental learning that the present architecture does not address. As large language models are increasingly exploited for automated generation of naturalistically written fake reviews, continuous adaptation of the text encoding pathway's anomaly detection capabilities will be required to maintain discriminative power against this evolving threat.

References

Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y., & Maqsood, S. (2021). Fake reviews detection: A survey. Ieee Access, 9, 65771-65802.

Mewada, A., & Dewang, R. K. (2023). A comprehensive survey of various methods in opinion spam detection. Multimedia Tools and Applications, 82(9), 13199-13239.

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S. G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, 102771.

Madahali, L. (2024). Social bots (Master's thesis, Iowa State University).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

Shan, G., Zhou, L., & Zhang, D. (2021). From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. Decision Support Systems, 144, 113513.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. ACM transactions on intelligent systems and technology (TIST), 10(3), 1-42.

Tang, H., & Cao, H. (2020, November). A review of research on detection of fake commodity reviews. In Journal of Physics: Conference Series (Vol. 1651, No. 1, p. 012055). IOP Publishing.

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.

Chen, J., Liu, J., Liang, Y., & Zhou, M. (2026). KE-MLLM: A Knowledge-Enhanced Multi-Sensor Learning Framework for Explainable Fake Review Detection. Applied Sciences, 16(6), 2909.

Vrahatis, A. G., Lazaros, K., & Kotsiantis, S. (2024). Graph attention networks: a comprehensive review of methods and applications. Future Internet, 16(9), 318.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Du, C., Sun, H., Wang, J., Qi, Q., & Liao, J. (2020, July). Adversarial and domain-aware BERT for cross-domain sentiment analysis. In Proceedings of the 58th annual meeting of the Association for Computational Linguistics (pp. 4019-4028).

Yang, M., Zhou, M., Zhang, T., Liu, J., Li, Z., Pan, L., ... & King, I. (2022). Hyperbolic graph neural networks: A review of methods and applications. arXiv preprint arXiv:2202.13852.

Ni, J., Li, J., & McAuley, J. (2019, November). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 188-197).

Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020). It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. Information, Communication & Society, 23(6), 867-891.

Qu, Z., Jia, Q., Lyu, C., Liu, J., Liu, X., & Zheng, K. (2022). Detecting fake reviews with generative adversarial networks for mobile social networks. Security and Communication Networks, 2022(1), 1164125.

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., & Yin, D. (2019, May). Graph neural networks for social recommendation. In The world wide web conference (pp. 417-426).

Li, Z., Zou, Y., Zhang, C., Zhang, Q., & Wei, Z. (2021, November). Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In Proceedings of the 2021 conference on empirical methods in natural language processing (pp. 246-256).

Zhao, W., Chen, T., Yang, J. S., & Qiu, L. (2026). AutoML-Pipeline: A RAG-enhanced code generation framework with pre-validation for cloud-native machine learning workflows. IEEE Access.

Zhang, S., Qiu, L., & Zeng, Z. (2026). Physics-Data Synergy in Structural Health Monitoring: A Multi-Scale Graph Contrastive Framework With Temperature-Adaptive Fusion. IEEE Access.

Wang, Z., Yang, J. S., Shang, W., & Ding, J. (2026). FairPromote: Explainable and Fairness-Aware Talent Promotion Prediction via Adversarial Debiasing and SHAP-based Interpretation. IEEE Access.

Ding, J., Shen, Z., & Liu, W. (2026). Game-Theoretic Cost-Sensitive Adversarial Training for Robust Cloud Intrusion Detection Against GAN-Based Evasion Attacks. Applied Sciences, 16(8), 3944.

Ping, W., Jiao, Y., Fan, H., & Zhang, X. (2026). Multimodal Fraud Detection in Financial Statements: A Trimodal Attention Network with Contrastive Evidence Chain Construction. IEEE Access.

Teng, D., Rhee, M., Qin, Y., Zi, B., & Liu, W. (2026). SW-SpeedDLM: Sliding-Window Speculative Decoding for Diffusion Language Models under Long-Context Constraints. Mathematics