iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78756

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

A Product Review Spammer Detection Model Based on Hybrid PU-Learning

G. Ganapathi Rao[1], J. Akhila[2], K. Ashritha[3], M. Vishwanath Sharma[4]
Dept. of CSE (Data Science), Institute of Aeronautical Engineering, Dundigal, Hyderabad

Abstract: ***The widespread manipulation of online product reviews by spammers has emerged as one of the most pressing integrity challenges facing modern e-commerce platforms. Conventional spam detection approaches that rely on fixed feature sets and standard classifiers are increasingly outpaced by a new generation of crowdsourced spammers who systematically mimic the behavioral patterns of genuine users. To address this growing threat, this paper proposes hPSD, a Hybrid Positive-Unlabeled Learning-Based Spammer Detection model that jointly leverages individual user behavioral features and the structural information embedded in user-product relational networks. Operating under a semi-supervised framework, hPSD begins with a small seed set of confirmed spammers and iteratively refines its detection through a reliable negative extraction algorithm and a Bayesian hybrid classifier. The model is capable of identifying multiple distinct types of spammers in a single unified pipeline. Extensive experiments on both a synthetic movie-review dataset with injected shilling attacks and a real-world Amazon review corpus demonstrate that hPSD significantly outperforms eight state-of-the-art baseline detectors, achieving a precision of 0.81, recall of 0.97, and F-measure of 0.90 on real-world data. The framework additionally uncovers hidden employer organizations coordinating spammer activity, demonstrating practical value beyond standard detection metrics.***

Keywords: Spam Detection, PU-Learning, Semi-Supervised Learning, Review Manipulation, Bayesian Inference, User-Product Relations, E-Commerce Security, Spammer Detection.

I. INTRODUCTION

Online product reviews and star ratings have become indispensable decision-making signals in modern digital commerce. Across e-commerce platforms such as Amazon, eBay, and Taobao, online-to-offline service platforms like Yelp and Dianping, and travel booking ecosystems including TripAdvisor and Hotels.com, consumers routinely base purchasing decisions on the collective sentiment expressed through user-generated reviews. Research consistently demonstrates that products with higher average ratings and larger review volumes command greater consumer trust and generate substantially higher revenue, while a surge of negative reviews can rapidly erode brand reputation and suppress sales. This asymmetric relationship between review quality and financial outcome creates powerful incentives for manipulation.

In response to these incentives, a well-documented ecosystem of review fraud has emerged. At its most organized, this ecosystem involves merchants hiring intermediary brokers referred to in Chinese markets as “shuakewang” or review army organizers who coordinate groups of paid human reviewers to post fabricated positive reviews for client products or targeted negative campaigns against competitors. Unlike early-generation spam that relied on automated bots producing obviously synthetic content, modern crowdsourced spammers operate with considerable sophistication. They maintain realistic-looking account histories, accumulate genuine helpfulness votes from other users, purchase legitimate products in plausible quantities, and write stylistically convincing reviews. These behaviors make them extremely difficult to distinguish from genuine reviewers when examined through the lens of individual behavioral features alone.

The fundamental challenge in detecting contemporary human-like spammers is that the most informative signals of spammer activity are relational rather than individual. A reviewer who has rated hundreds of products and accumulated thousands of helpfulness votes appears entirely legitimate in isolation. However, when examined through the lens of which products they have reviewed, a pattern emerges: an unusual concentration of reviews directed at products from a single publisher, brand, or seller. This relational signal, invisible to feature-based detectors, is precisely the kind of evidence that distinguishes a genuine prolific reviewer from a paid promoter embedded in an organized spamming network.

This paper introduces hPSD, a Hybrid Positive-Unlabeled Learning-Based Spammer Detection framework that directly addresses these limitations. hPSD operates in a realistic semi-supervised setting where only a small number of confirmed spammers are available as labeled positives, while the vast majority of the user population remains unlabeled and potentially contains both genuine users and hidden spammers.

The framework extracts a reliable negative set from the unlabeled pool, trains a hybrid Bayesian classifier that jointly models individual behavioral features and user-product relational structure, and iterates this process to detect multiple distinct types of spammer behavior within a single unified pipeline. The model is evaluated on both controlled synthetic datasets with known ground truth and challenging real-world Amazon data, demonstrating superior performance across all conditions.

II. RELATED WORK

The academic literature on review spam detection has evolved substantially over the past fifteen years, progressing from simple rule-based filters through statistical models to modern graph-aware and semi-supervised approaches. Early work by Jindal and Liu established the fundamental taxonomy of review spam, distinguishing deceptive opinion spam involving fabricated reviews, non-reviews such as advertisements masquerading as reviews, and biased reviews from competing or affiliated parties [8]. Their pioneering studies demonstrated that linguistic features including unusual patterns of superlatives, first-person language, and sentiment vocabulary carry discriminative signal for automated detection.

Subsequent research expanded the feature space beyond review text to incorporate behavioral metadata derived from reviewer activity histories. Lim et al. demonstrated that rating behavior statistics including the deviation of a reviewer's average rating from the product mean, the proportion of ratings that are extreme, and the frequency of reviews targeting a single product are strong indicators of review manipulation [5]. Mukherjee and colleagues further extended this line of work by developing behavioral footprint models that characterize spammer activity across multiple behavioral dimensions simultaneously, showing that combinations of behavioral signals substantially outperform single-feature detectors [10].

The challenge of detecting coordinated spammer groups, rather than individual spammers acting independently, motivated a shift toward graph-based approaches. Wang et al. introduced review-graph methods that model the tripartite network of reviewers, reviews, and products, propagating suspicion scores across the graph to identify clusters of mutually reinforcing suspicious activity [6]. Fayazi et al. specifically targeted crowdsourced manipulation operations, demonstrating that the coordination signals visible in network structure such as temporal clustering of reviews from related accounts reveal organized campaigns that individual-account analysis would miss [7].

Despite these advances, two important limitations remain in the existing literature. First, the majority of approaches operate on feature spaces that characterize either review text or reviewer behavior in isolation, without explicitly modeling the structural relationship between reviewers and the specific products they choose to review. This relational dimension carries information that neither content nor behavioral features can fully capture. Second, the practical deployment of supervised classification approaches is severely constrained by the scarcity of reliably labeled spammer data on real platforms. Only a small body of work has applied Positive-Unlabeled learning frameworks to the spammer detection problem [3][4], and none has simultaneously integrated relational network information with PU-learning in a unified hybrid architecture. The hPSD framework proposed in this paper directly addresses both of these gaps.

III. METHODOLOGY

The hPSD methodology is structured as a five-stage pipeline that transforms raw review platform data into a probabilistic ranking of users by spammer likelihood. Each stage is designed to extract increasingly refined discriminative signals, culminating in a Bayesian hybrid classifier that combines behavioral feature space modeling with user-product relational network analysis.

A. User Interface and Data Access Layer

The system's entry point is a secure authentication interface through which platform analysts connect to the review database. All users accessing the system must provide valid credentials before gaining access to the review corpus and detection pipeline outputs. First-time users are registered with a unique identifier, password, and institutional email address, after which the system establishes a persistent account tracking their query history and detection session parameters. This access control layer ensures that sensitive spammer identification outputs are available only to authorized personnel and that all detection activities are logged for audit purposes. The interface exposes the full review corpus, user behavioral metadata, and product catalog information as queryable data sources for the downstream detection stages.

B. Review Behavioral Feature Extraction

Each user in the platform's review corpus is characterized by a rich behavioral feature vector derived from their complete review history. These features capture multiple dimensions of review behavior that have been shown in prior research to discriminate between genuine reviewers and spammers.

Behavioral features include the total number of reviews submitted, the temporal distribution of review activity across days and weeks, the variance and skewness of the ratings assigned relative to product category averages, the proportion of extreme ratings at the one-star and five-star boundaries, the fraction of reviews that receive helpfulness votes from other users, and the breadth of product categories covered by the reviewer's activity.

A key additional signal is derived from the user-product relational structure. For each reviewer, the system computes the concentration of their review activity across product sources specifically measuring what proportion of the products they have reviewed originate from the same publisher, manufacturer, or brand entity. A reviewer who has submitted dozens of reviews but directed an unusually high fraction of them at products from a single source exhibits a characteristic relational signature that distinguishes paid promotion from genuine diverse consumer interest. This relational feature is computed from the user-product relation matrix R, where each entry R(i,j) indicates whether user i has reviewed product j, and product j carries metadata identifying its source entity.

C. Feature Discretization and Reliable Negative Extraction

Before training begins, all continuous behavioral features are discretized into categorical bins to enable the application of the multinomial Bayesian framework. Discretization is performed using the Bisecting V-Clustering algorithm (BiVC), which identifies optimal cut-points in the sorted feature value distribution by minimizing the weighted average variance of the resulting bins. This process converts each user's continuous feature values into a binary membership vector of dimension equal to the number of features, where each bit indicates which bin the user's feature value falls into.

A critical challenge in the PU-learning setting is identifying a reliable negative set users who can be confidently labeled as non-spammers from the unlabeled pool. The hPSD framework addresses this through a feature strength scoring procedure. For each discretized feature bin, the system computes a discriminative strength score D(f) that measures how frequently that bin is occupied by confirmed spammers in the positive set relative to its frequency in the unlabeled pool. Bins that are strongly associated with known spammers receive high strength scores. Users in the unlabeled pool who do not exhibit any high-strength feature bins meaning their behavioral profiles show no resemblance to confirmed spammers are extracted into the Reliable Negative set. This extraction continues until the reliable negative set reaches a target size comparable to the positive set, ensuring a balanced training signal. The combined labeled training set L is then formed as the union of the positive set P and the reliable negative set RN.

D. Hybrid Semi-Supervised Bayesian Learning

The core detection model is a hybrid Bayesian classifier that is trained on the labeled set L and applied to score all users in the full dataset D, which includes both the labeled set and the remaining unlabeled users. For each class spammer and non-spammer the model learns a multinomial distribution over the discretized feature bins from the labeled examples. Classification is performed by computing the posterior probability of each class given a user's feature vector, following Bayes' rule with a multinomial likelihood and a class prior estimated from the training set composition.

The hybrid component of the model introduces an additional signal derived from the user-product relational network. For each user, the model computes a relational spammer score based on the proportion of their reviewed products that have been flagged as potentially polluted by other detected spammers. This score is combined with the feature-space Bayesian posterior through a weighted mixture, with the balance parameter controlling the relative contribution of relational versus behavioral evidence. The combined score provides a more complete picture of each user's likely classification than either signal alone, particularly for sophisticated spammers who have carefully calibrated their behavioral features to evade detection while still concentrating their review activity on a restricted set of promoted products.

E. Iterative Multi-Type Spammer Detection

Real-world review platforms host multiple qualitatively distinct types of spammers including duplicate reviewers who repeatedly post identical or paraphrased reviews, targeted promoters who concentrate their activity on specific brands or publishers, and colluding groups of reviewers who coordinate their rating patterns to amplify the statistical impact of their manipulations. A single detection pass with one seed positive set will naturally be most sensitive to the type of spammer most represented in that seed, potentially missing other types.

hPSD addresses this through an iterative multi-type detection loop. In each iteration, the analyst selects or constructs a seed positive set representative of a specific spammer type. The full detection pipeline is run: feature discretization, reliable negative extraction, hybrid Bayesian learning, and classification scoring.

Users flagged as spammers in that iteration are removed from the unlabeled pool and recorded as detected. The process then repeats with a new seed set targeting the next spammer type. This sequential peeling approach enables the framework to systematically uncover heterogeneous spammer populations that coexist on the same platform, each with its own characteristic behavioral signature.

Fig. 1. Review Architecture — hPSD System Pipeline
Fig. 2. Spammer Network — User-Product Relational Graph

IV. SYSTEM ARCHITECTURE

The hPSD system is organized into six functionally distinct architectural layers that collectively transform raw review platform data into a structured list of detected spammers with associated probability scores and supporting evidence.

A. Input Data Layer

The foundation of the architecture is the input data layer, which consolidates three distinct information sources into a unified representation. The user set U contains all platform users whose review behavior is subject to analysis, the vast majority of whom carry no label.

The positive set P provides the small seed of confirmed spammers that initiates the PU-learning process. The user-product relation matrix R encodes the binary relationship between each user and each product they have reviewed. The feature set F derives behavioral attributes from each user's complete review history and associated metadata, including review timestamps, rating values, helpfulness vote records, and product category information.

B. Feature Discretization Component

The feature discretization component transforms the continuous-valued behavioral features in F into categorical representations suitable for multinomial Bayesian modeling. The BiVC algorithm partitions each feature's value range into bins by iteratively identifying split points that minimize the weighted average variance of the resulting groups. After discretization, each user is represented as a binary indicator vector that encodes which bin their value falls into for each feature dimension. This representation enables the downstream Bayesian model to learn class-conditional feature distributions without assumptions about the underlying feature distributions.

C. Reliable Negative Set Extraction Module

The reliable negative extraction module addresses the fundamental challenge of the PU-learning setting: the absence of labeled negative examples.

The module computes a feature strength score D(f) for each discretized bin, quantifying how specifically associated that bin is with known spammers relative to its prevalence in the unlabeled population. Users in the unlabeled set whose feature profiles show no association with high-strength spammer bins are designated as reliable negatives. The extraction continues until the reliable negative set reaches a size sufficient to form a balanced training corpus when combined with the positive set.

D. Hybrid Semi-Supervised Learning Engine

The learning engine trains the hybrid Bayesian classifier on the labeled set formed by the union of the positive set and reliable negative set. For each class, it learns a multinomial feature distribution and computes class priors from the training set composition. The relational component augments the feature-space posterior with a network-derived spammer signal computed from the user-product graph, combining both signals through a tunable weighting parameter to produce the final composite spammer score for every user in the full dataset.

E. Iterative Multi-Type Detection Loop

The iterative detection loop orchestrates repeated execution of the discretization, reliable negative extraction, and classification modules, each time targeting a distinct spammer archetype through a type-specific seed positive set. Between iterations, previously detected spammers are removed from the unlabeled pool to prevent their signals from contaminating the detection of remaining spammer types. This loop continues until all target spammer types have been processed, yielding a comprehensive detection output that covers the full heterogeneity of spammer behavior present on the platform.

F. Output Evaluation and Reporting

The output layer consolidates detection results across all iterations into a structured report listing flagged user IDs, their composite spammer probability scores, the specific spammer type classification assigned in each detection iteration, and the relational evidence supporting each detection including the concentration of their review activity across product sources and the proportion of their reviewed products flagged as potentially polluted. This multi-dimensional output enables platform operators not only to take action against individual spammers but also to identify the employer organizations coordinating spamming campaigns, such as book publishers or product brands whose products appear disproportionately in detected spammers' review histories.

V. RESULTS AND EVALUATION

The hPSD framework was evaluated on two complementary experimental settings that together test both the model's theoretical detection capabilities under controlled conditions and its practical effectiveness on real-world data with unknown ground truth.

The first evaluation used a synthetic shilling attack dataset constructed from a publicly available movie-rating corpus. Attacker profiles representing multiple shilling attack strategies including random attack, average attack, bandwagon attack, and segment attack were injected at varying penetration levels ranging from 5% to 20% of the total user population, with filler item sets of varying sizes to test robustness across different attack profiles. hPSD was compared against eight state-of-the-art shilling attack detectors under this controlled setting. The framework consistently outperformed all baseline methods in classification accuracy across all attacker types and penetration levels, with the performance advantage most pronounced for sophisticated attack strategies that use targeted item selection to mimic legitimate behavioral patterns.

The second and more challenging evaluation applied hPSD to a real Amazon review dataset where ground-truth spammer labels are not fully available. The framework was tasked with detecting two specific spammer archetypes: duplicate reviewers who repeatedly post functionally identical reviews for the same product, and promoters who concentrate their review activity on products from a specific publisher or brand. Starting from a small seed positive set of manually confirmed spammers for each type, the system processed the full unlabeled user population through the hybrid PU-learning pipeline.

The quantitative results on the Amazon dataset demonstrate the substantial value added by the hybrid relational component. In an ablation condition with the relational component disabled, the feature-space-only model classified 8,411 users as non-spammers and 34 as promoters. The full hybrid model with relational integration detected 568 additional promoters while reclassifying only 34 users differently from the non-hybrid version, as measured by a Cohen's Kappa agreement coefficient of 0.548. This means the relational component identified a population of disguised promoters whose individual behavioral features were unremarkable but whose review activity was structurally concentrated on products from specific publishers a pattern only visible through the user-product graph analysis. Manual validation of a random sample of the newly detected promoters confirmed the quality of the additional detections. The vast majority of the flagged users had reviewed 50 or more books, with a disproportionately high fraction of those books originating from a single publisher. Cluster analysis of detected promoters revealed ten book-publisher clusters each characterized by a strong concentration of reviewer activity around a single publishing entity, providing strong evidence of coordinated paid promotion campaigns. Overall performance on the Amazon dataset reached a precision of 0.81, recall of 0.97, and F-measure of 0.90 reflecting the framework's ability to retrieve nearly all true spammers at an operationally acceptable false positive rate for platform moderation workflows.

Robustness analysis across varying attack conditions in the movie dataset confirmed that hPSD maintains stable detection performance as attacker strategies change, attack penetration depth increases, and filler item selection strategies vary. The model's iterative multi-type detection capability proved particularly valuable, with each successive iteration targeting a different spammer archetype and uncovering populations that earlier iterations had missed. These results collectively validate the core design principles of the hPSD framework: semi-supervised PU-learning provides a practical path to effective detection under the label scarcity conditions universal to real review platforms, and relational user-product modeling adds substantial detection power that purely behavioral approaches cannot replicate.

VI. CONCLUSION AND FUTURE WORK

This paper has presented hPSD, a Hybrid Positive-Unlabeled Learning-Based Spammer Detection framework that addresses the core limitations of existing review spam detection approaches through three design innovations: a semi-supervised PU-learning paradigm that operates effectively under realistic label scarcity conditions, a hybrid Bayesian classifier that jointly models individual behavioral features and user-product relational network structure, and an iterative multi-type detection loop that systematically uncovers heterogeneous spammer populations within a single unified pipeline.

The experimental results demonstrate that this combination delivers meaningfully superior detection performance relative to eight state-of-the-art baselines on both controlled synthetic attack scenarios and challenging real-world Amazon review data. Most significantly, the relational component of the hybrid model uncovers a substantial population of sophisticated promoters paid reviewers concentrating their activity on products from specific employers whose individual behavioral profiles are unremarkable and who would be missed entirely by feature-space-only detectors. The discovery of underlying employer organizations coordinating spamming campaigns represents a capability that substantially extends the practical value of automated spammer detection beyond identifying individual bad actors.

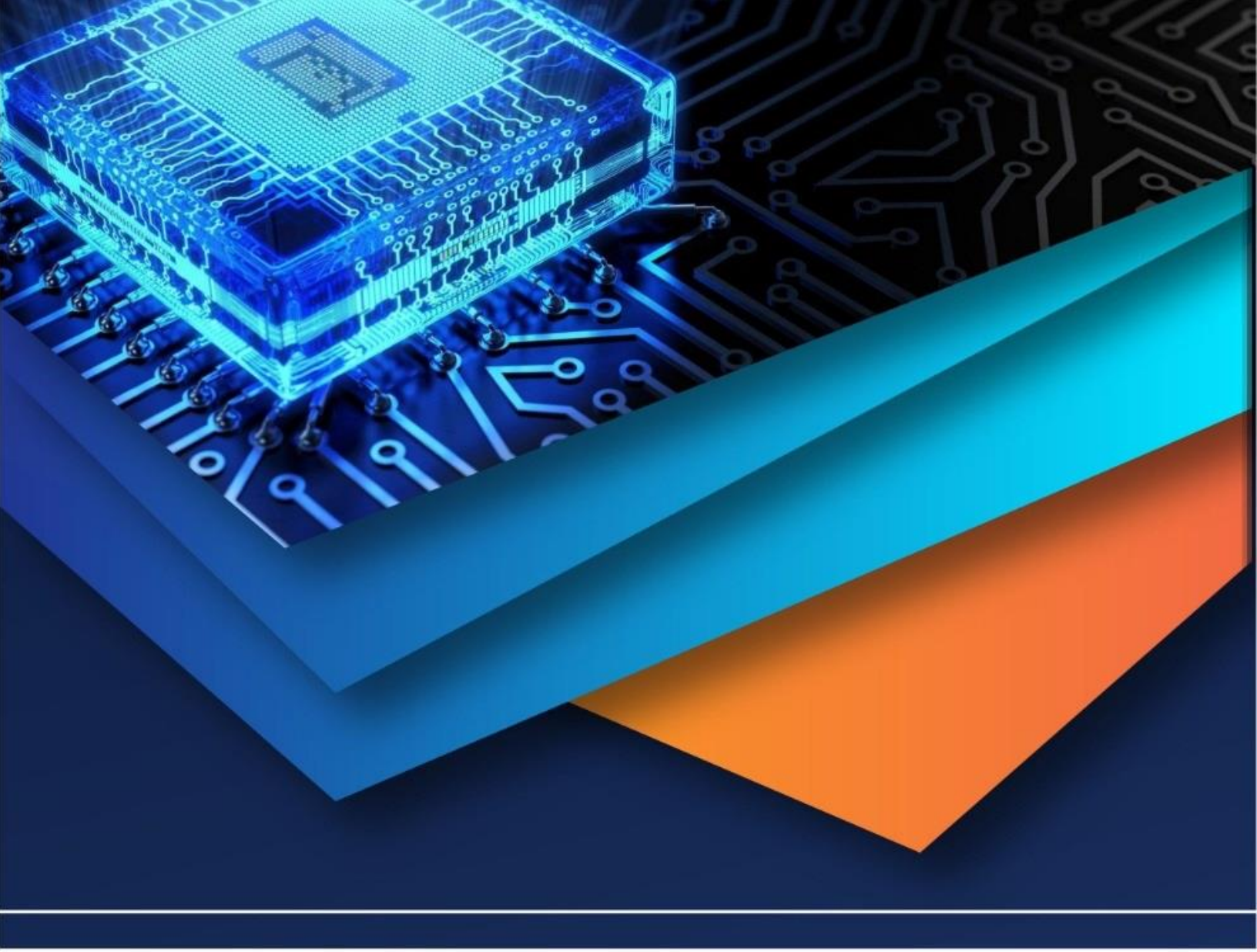
The framework's design also surfaces several directions for future enhancement. As spammer strategies continue evolving in response to deployed detection systems, the ability to adapt in real time becomes increasingly critical. Development of online or incremental learning variants of hPSD that update their models continuously as new review data arrives would enable the system to maintain detection performance without requiring periodic full retraining cycles. Incorporating richer relational signals beyond the user-product review graph including device fingerprint sharing, IP address clustering, temporal coordination of account activity, and infrastructure-level linkages between accounts would provide additional evidence channels for detecting more subtle forms of collusion that the current model may miss.

Deepening the textual analysis component represents another high-value enhancement direction. The current framework treats review text as a source of behavioral metadata rather than analyzing the content directly. Integrating advanced natural language processing techniques including writing style fingerprinting, sentiment trajectory analysis, and cross-review semantic similarity scoring would enable the framework to detect spammers who have successfully calibrated their behavioral metadata to evade detection but whose review writing patterns still carry identifiable signatures of paid or coordinated authorship.

Explainability and interpretability improvements are essential for operational deployment. Platform trust and safety teams reviewing flagged accounts need to understand why specific users were identified as spammers in order to make defensible moderation decisions and build legally compliant enforcement cases. Developing structured explanation outputs that clearly articulate the combination of behavioral features and relational signals driving each detection decision would substantially improve the framework's practical utility. Finally, deployment in live production environments with real-time processing constraints, large-scale distributed infrastructure, and human-in-the-loop review workflows would provide the operational validation needed to confirm that hPSD's detection capabilities translate effectively to production-scale platforms.

REFERENCES

[1] C. Forman, A. Ghose, and B. Wiesenfeld, "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets," Inf. Syst. Res., vol. 19, no. 3, pp. 291–313, 2008.
[2] F. Zhu and X. Zhang, "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," J. Market., vol. 74, no. 2, pp. 133–148, 2010.
[3] T.-M. Choi, H. K. Chan, and X. Yue, "Recent development in big data analytics for business operations and risk management," IEEE Trans. Cybern., vol. 47, no. 1, pp. 81–92, Jan. 2017.
[4] M. Ott, C. Cardie, and J. Hancock, "Estimating the prevalence of deception in online review communities," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 201–210.
[5] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in Proc. 19th ACM Int. Conf. Inf. Knowl. Manag., 2010, pp. 939–948.
[6] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," in Proc. 11th IEEE Int. Conf. Data Min. (ICDM), 2011, pp. 1242– 1247.
[7] A. Fayazi, K. Lee, J. Caverlee, and A. Squicciarini, "Uncovering crowdsourced manipulation of online reviews," in Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2015, pp. 233–242.
[8] N. Jindal and B. Liu, "Review spam detection," in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 1189–1190.
[9] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 191–200.
[10] A. Mukherjee et al., "Spotting opinion spammers using behavioral footprints," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min., 2013, pp. 632–640.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY