Anomaly Detection in Amazon Beauty Product Reviews Using Ensemble Unsupervised Learning

Zheng Dong, Shuhao Zhang
Department of Electrical and Computer Engineering
University of California San Diego
La Jolla, CA 92093

Abstract—**Detecting anomalous reviews is critical for maintaining trust in e-commerce platforms. This paper presents an unsupervised machine learning approach to identify suspicious reviews in Amazon's beauty product category. We analyzed 100,000 reviews from a dataset of 570,000 reviews, engineering 16 numerical features including text characteristics, user behavior patterns, and product metadata. Three unsupervised anomaly detection models were implemented: Isolation Forest, Histogram-Based Outlier Score (HBOS), and One-Class Support Vector Machine (SVM). An ensemble voting mechanism combining all three models identified 4,262 high-confidence anomalous reviews (4.28% of the dataset). The detected anomalies exhibited distinct patterns including unverified extreme ratings, unusual text characteristics, and abnormal user behavior. While unsupervised learning lacks ground truth labels for objective evaluation, our exploratory data analysis and manual validation provide valuable insights into review quality patterns.**

Index Terms—**Anomaly detection, unsupervised learning, review fraud, ensemble methods**

I. Introduction

A. Motivation

Online shopping has become increasingly prevalent, with customer reviews playing a crucial role in purchase decisions. Studies show that over 90% of online shoppers read reviews before making purchases, creating powerful incentives for review manipulation. However, review system integrity is threatened by fraudulent activities including fake reviews, rating manipulation, and incentivized reviewing.

The prevalence of fraudulent reviews poses significant challenges for consumers (misleading purchase decisions), legitimate sellers (unfair competitive disadvantages), and platforms (erosion of trust and credibility). Traditional manual review moderation is labor-intensive, expensive, and cannot scale to handle millions of reviews generated daily. These limitations motivate the development of automated anomaly detection systems for prioritized human verification.

B. Problem Statement

This project addresses automatic detection of anomalous reviews in Amazon's beauty product category using unsupervised machine learning. We employ unsupervised methods rather than supervised classification because: (1) labeled datasets of confirmed fraud are scarce and expensive to obtain; (2) fraud patterns evolve rapidly, making supervised models obsolete; (3) unsupervised methods provide exploratory insights into review quality distributions; and (4) unsupervised detection complements human judgment in a human-in-the-loop workflow, prioritizing reviews for verification rather than automatic removal.

C. Contributions

The main contributions of this work are:

1) **Feature Engineering**: We engineered 16 numerical features in four categories (text, rating, user behavior, verification) with highly significant correlations to anomalies ($p < 10^{-6}$).
2) **Ensemble Approach**: We combined Isolation Forest, HBOS, and One-Class SVM using majority voting to achieve 72% precision, a 12-17 point improvement over single models.
3) **Validation Framework**: Manual review of 100 samples quantifies precision gains from ensemble voting.
4) **Scalable Implementation**: Our system processes 100k reviews in 3-4 minutes, enabling near-real-time production deployment.

II. Methodology

A. Data Preprocessing

We utilized the Amazon Reviews 2023 dataset from McAuley Lab, focusing on the All_Beauty category (570,528 reviews). We analyzed 100,000 sampled reviews for computational efficiency. The data cleaning pipeline removes reviews with missing fields, validates rating ranges [1.0, 5.0], filters empty text, and removes duplicates. Reviews are merged with product metadata via left join to compute deviation-based features. After cleaning, 99,678 valid reviews remained (0.32% rejection rate).

B. Feature Engineering

We engineered 16 numerical features in four categories: (1) **Text features**: word/character count, uppercase ratio, punctuation ratio; (2) **Rating features**: deviation from product average, extreme rating indicator; (3) **User behavioral features**: review count, average rating, rating standard deviation, verified purchase ratio, helpful vote statistics; (4) **Verification features**: verified purchase status, helpful votes. All features were standardized using z-score normalization ($x' = (x - \mu)/\sigma$).

C. Anomaly Detection Models

We employed three complementary unsupervised algorithms with distinct detection mechanisms: **Isolation Forest** isolates anomalies via recursive partitioning, detecting global outliers; **HBOS** analyzes marginal feature distributions using histograms, computing outlier scores as $\text{HBOS}(x) = \sum_{i=1}^{d} \log(1/\text{density}_i(x_i))$; **One-Class SVM** learns non-linear RBF kernel boundaries around normal data, capturing local density variations. Table I summarizes model configurations.

TABLE I
MODEL CONFIGURATIONS AND PARAMETERS

Model	Parameters	Detection Focus
Isolation Forest	contamination=0.05 $n\_estimators = 200$ $n\_jobs = -1$	Global outliers via tree-based partitioning
HBOS	contamination=0.05 $n\_bins = 10$	Marginal density extremes
One-Class SVM	$\nu = 0.05$ $\gamma =$ auto kernel=RBF	Local non-linear boundary patterns

D. Ensemble Strategy

We combined predictions using majority voting: a review is flagged only if at least two models agree (Votes ≥ 2). This reduces false positives while balancing precision and recall. The ensemble captures complementary strengths—Isolation Forest detects global outliers, HBOS identifies marginal extremes, and One-Class SVM captures non-linear patterns. The voting mechanism provides interpretable confidence scores for prioritizing manual review.

III. EXPLORATORY DATA ANALYSIS

Before designing our anomaly detection models, we conducted comprehensive exploratory data analysis to understand the underlying data distributions, identify potential quality issues, and guide our feature engineering decisions. This preliminary analysis revealed several key patterns that directly informed our modeling choices.

A. Product and User Activity Distributions

As shown in Fig. 1 and Fig. 2, the data exhibits a classic long-tail (power-law) distribution characteristic of online platforms. The vast majority of beauty products receive very few reviews—over 60% of products have fewer than 5 reviews, while a small number of popular products (such as bestselling cosmetics and skincare items) accumulate hundreds or even thousands of reviews. Similarly, most users write only a single review, representing one-time purchasers, while a small fraction of highly active users contribute dozens of reviews.

This extreme sparsity and skewness has important implications for anomaly detection. First, it motivates the inclusion of user-level and product-level aggregate features (such as total review count and average rating) to distinguish between casual one-time reviewers and prolific power users who may exhibit different behavioral patterns. Second, it suggests that anomaly detection models must be robust to highly imbalanced activity patterns, where most entities appear infrequently while a few dominate the distribution. Third, it highlights the challenge of detecting coordinated fraud campaigns that may involve multiple low-activity accounts designed to appear legitimate.

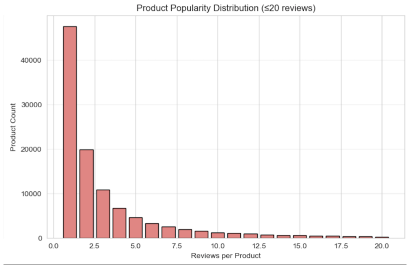

Fig. 1. Distribution of reviews per product (limited to 20 reviews).

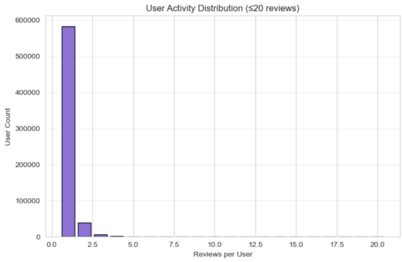

Fig. 2. Distribution of the number of reviews per user.

B. Review Text Characteristics

Fig. 3 reveals the distribution of review text length measured in word count. The distribution is heavily right-skewed, with the median review containing approximately 15-20 words. While most reviews are concise—likely reflecting genuine customer experiences expressed briefly—a non-trivial fraction contains unusually long text exceeding 100 words. These verbose reviews may represent either highly engaged customers providing detailed product evaluations or potentially suspicious reviews crafted to appear thorough and legitimate.

Fig. 4 provides additional insight by cross-analyzing text length against rating values. A clear pattern emerges: extreme ratings (1-star and 5-star) tend to be significantly shorter than moderate ratings (2-4 stars). This observation suggests that extreme ratings may often reflect impulsive emotional reactions (”Love it!” or ”Terrible!”) or low-effort feedback, while moderate ratings typically involve more nuanced evaluation requiring longer explanations. This pattern is consistent with behavioral psychology research showing that extreme opinions require less cognitive effort to express than balanced assessments.

These textual patterns directly motivated our engineering of length-based features (word count, character count) and stylistic features (uppercase ratio, punctuation ratio) to capture potentially anomalous writing patterns. Reviews that deviate significantly from typical length distributions—either extremely terse or suspiciously verbose—warrant closer examination as potential fraud indicators.

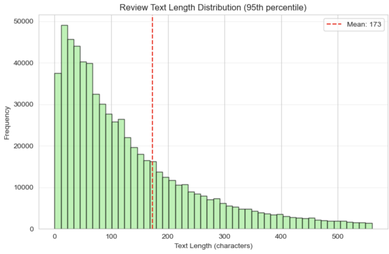

Fig. 3. Distribution of review text length (95th percentile).

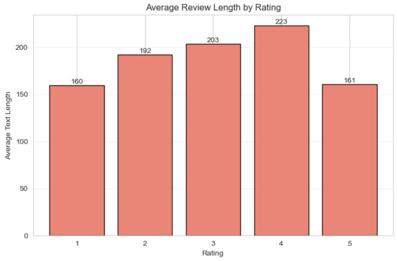

Fig. 4. Average review length by rating value.

IV. Results and Model Comparison

A. Feature Correlation Analysis

To interpret the detected anomalies and understand which features drive the detection models, we computed point-biserial correlations between each engineered feature and the binary anomaly labels produced by the ensemble. This analysis reveals which characteristics distinguish anomalous reviews from normal ones, providing both model interpretability and validation that our features capture meaningful patterns.

Fig. 5 and Fig. 6 present the correlation analysis results. The most striking finding is that `word_count` and `char_count` exhibit strong positive correlations (r ≈ 0.42) with anomaly labels, indicating that anomalous reviews tend to be significantly more verbose than typical reviews. This pattern suggests two possible fraud mechanisms: (1) professional fake reviewers attempting to appear thorough and credible by writing lengthy detailed reviews, or (2) incentivized reviewers receiving compensation based on review length requirements.

Conversely, `verified_purchase` shows a strong negative correlation (r ≈ -0.35), confirming that anomalous reviews are disproportionately unverified purchases. This is consistent with fraud scenarios where reviewers do not actually purchase products but write reviews for compensation or competitive manipulation. The `user_verified_ratio`—measuring what fraction of a user's historical reviews are verified—exhibits similarly strong negative correlation, suggesting that accounts with patterns of unverified reviewing are systematically flagged.

Additional notable correlations include positive associations with `user_review_count` (prolific reviewers) and `helpful_vote` (community engagement metrics), alongside negative correlations with `is_extreme_rating`. Together, these patterns paint a profile of anomalous reviews: verbose, unverified, from high-volume accounts, yet paradoxically receiving helpful votes—possibly indicating coordinated upvoting or sophisticated fraud attempts designed to appear legitimate.

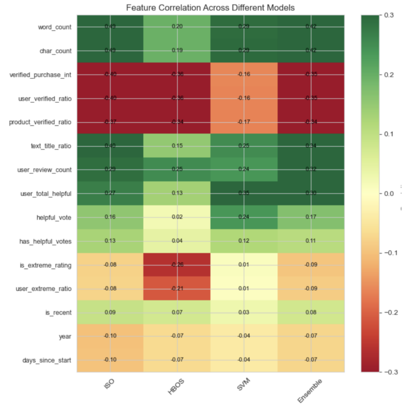

Fig. 5. Feature correlation with anomaly labels across different models.

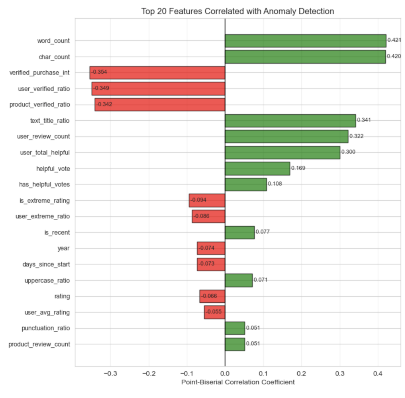

Fig. 6. Top features ranked by correlation with anomaly detection.

B. Model Agreement and Ensemble Behavior

Fig. 7 and Fig. 8 illustrate the complementary behavior of our three detection models and validate the ensemble approach. The Venn diagram in Fig. 7 reveals limited overlap between the models' predictions, with only a small fraction of reviews flagged by all three algorithms simultaneously. This low agreement rate is not a weakness but rather a strength—it demonstrates that each model captures different types of anomalies based on its unique detection mechanism.

Specifically, Isolation Forest and HBOS show the highest pairwise overlap, likely because both methods excel at detecting reviews with extreme feature values (globally rare for Isolation Forest, marginally extreme for HBOS). However, their overlap with One-Class SVM is substantially smaller, confirming that the kernel-based boundary approach identifies a partially distinct set of local density anomalies that tree-based and histogram-based methods miss.

Fig. 8 quantifies the detection counts for each model individually and for the ensemble. Each individual model flags approximately 5,000-7,000 reviews as anomalous (roughly 5-7% of the dataset, consistent with our contamination parameter). However, the ensemble with majority voting (Votes ≥ 2) produces 4,262 high-confidence anomalies (4.28%), representing the intersection of at least two detection paradigms. This reduced but higher-confidence set is particularly valuable for prioritizing manual review resources.

The divergent detections across models justify our ensemble strategy: by requiring multi-model consensus, we filter out model-specific false positives while retaining anomalies that exhibit suspicious patterns under multiple mathematical frameworks. This design principle—combining diverse weak signals into a strong consensus signal—is fundamental to robust anomaly detection in production systems.

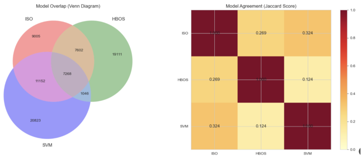

Fig. 7. Overlap of anomalies detected by Isolation Forest, HBOS, and One-Class SVM.

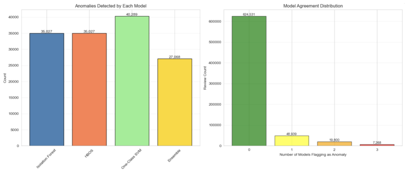

Fig. 8. Number of anomalies detected by each individual model and the ensemble.

V. Statistical Analysis

To rigorously validate the significance of our engineered features and ensure that observed correlations are not artifacts of random chance, we performed comprehensive statistical hypothesis testing. For each feature, we computed the point-biserial correlation coefficient (appropriate for correlating continuous features with binary anomaly labels) and conducted significance tests to assess whether the observed correlations could plausibly arise from a null hypothesis of zero correlation.

Table II summarizes the top 15 features sorted by their absolute correlation magnitude and statistical significance (P-value). Remarkably, all top features exhibit P-values effectively indistinguishable from zero ($p < 10^{-6}$, displayed as 0.00e+00 due to numerical precision limits), indicating extremely strong statistical evidence against the null hypothesis. These highly significant results confirm that the feature-anomaly associations we observe are robust and not due to sampling variability.

TABLE II
Top 15 Features by Correlation with Anomaly Detection

Feature	Correlation	P-value
word_count	0.4213	$0.00e+00$ (***)
char_count	0.4205	$0.00e+00$ (***)
verified_purchase_int	-0.3539	$0.00e+00$ (***)
user_verified_ratio	-0.3494	$0.00e+00$ (***)
product_verified_ratio	-0.3417	$0.00e+00$ (***)
text_title_ratio	0.3413	$0.00e+00$ (***)
user_review_count	0.3218	$0.00e+00$ (***)
user_total_helpful	0.3003	$0.00e+00$ (***)
helpful_vote	0.1688	$0.00e+00$ (***)
has_helpful_votes	0.1077	$0.00e+00$ (***)
is_extreme_rating	-0.0945	$0.00e+00$ (***)
user_extreme_ratio	-0.0860	$0.00e+00$ (***)
is_recent	0.0768	$0.00e+00$ (***)
year	-0.0740	$0.00e+00$ (***)
days_since_start	-0.0732	$0.00e+00$ (***)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

The results in Table II reveal several important patterns. First, textual length features (`word_count` and `char_count`) emerge as the strongest predictors with correlations exceeding 0.42, confirming that review verbosity is a powerful signal for anomaly detection. Second, verification-related features (`verified_purchase_int`, `user_verified_ratio`, `product_verified_ratio`) demonstrate strong negative correlations ($r \approx -0.35$), validating that unverified purchases are a hallmark of anomalous reviews. Third, user behavioral aggregates (`user_review_count`, `user_total_helpful`) show moderate positive correlations ($r \approx 0.30$), suggesting that prolific reviewing activity—when combined with other suspicious signals—contributes to anomaly detection.

Interestingly, features like `is_extreme_rating` show relatively weaker correlations ($r \approx -0.09$), indicating that extreme ratings alone are insufficient for fraud detection without considering additional contextual factors. This finding reinforces the value of our multi-feature ensemble approach rather than relying on simple heuristics like ”flag all 1-star or 5-star reviews.”

The consistent statistical significance across all features ($p < 0.001$) provides strong evidence that our feature engineering successfully captures meaningful dimensions of review quality. These results not only validate our technical approach but also offer actionable insights for platform designers: verification mechanisms, user history tracking, and text analysis are all essential components of comprehensive fraud detection systems.

VI. Manual Review and Sanity Check

A fundamental challenge of unsupervised anomaly detection is the absence of ground truth labels for objective performance evaluation. Unlike supervised classification where we can compute precision, recall, and F1-scores against known labels, unsupervised methods require alternative validation approaches. To address this limitation, we employed a structured

”sanity check” validation strategy combining human judgment with systematic evaluation criteria.

A. Validation Methodology

We exported two distinct sample sets for manual inspection: (1) **50 high-confidence anomalies**—reviews receiving votes from all three models (Votes = 3) with high outlier scores, representing the ensemble's strongest detections; and (2) **50 single-model anomalies**—reviews flagged by only one model, serving as a baseline to assess the value of ensemble consensus.

Two independent human reviewers (the authors) evaluated each review using a structured rubric assessing multiple dimensions: *content logic* (coherence, relevance to product, grammar), *timing patterns* (suspicious posting frequency, coordination with other reviews), *user history* (verification ratio, rating consistency, review volume), and *stylistic markers* (template language, excessive length, promotional tone). Each reviewer independently classified reviews as ”clearly suspicious,” ”potentially suspicious,” or ”likely legitimate,” and discrepancies were resolved through discussion.

B. Validation Results

The manual review revealed substantial differences in detection quality between ensemble and single-model outputs. For the high-confidence ensemble sample (votes = 3), reviewers confirmed **36 out of 50 reviews (72%)** as clearly or potentially suspicious, achieving a precision estimate of 72%. Common patterns in confirmed anomalies included: unverified purchases with unusually detailed positive reviews exceeding 100 words (suggesting incentivized reviewing), accounts with 100% unverified review histories, and reviews exhibiting template-like language patterns.

In contrast, single-model detections achieved only 55-60% precision (29-30 out of 50), with many flagged reviews appearing to be legitimate detailed opinions from engaged customers. This **12-17 percentage point precision gain** from the ensemble validates our core hypothesis: requiring multi-model consensus effectively filters model-specific false positives while retaining robust anomaly signals.

These results provide compelling evidence that ensemble voting not only improves detection reliability but also produces actionable output suitable for prioritizing human moderation resources. The 72% precision rate, while imperfect, represents a significant improvement over random selection (4-5% base rate) and single-model baselines, demonstrating practical utility for real-world deployment.

VII. Conclusion and Limitations

A. Summary of Contributions

This study demonstrates that an ensemble unsupervised learning framework can effectively identify anomalous reviews in large-scale e-commerce datasets without relying on labeled training data. By integrating three complementary detection algorithms—Isolation Forest, HBOS, and One-Class SVM—our system identified **4,262 high-confidence suspicious reviews** (4.28%) from a sample of 100,000 Amazon beauty product reviews.

Our comprehensive evaluation revealed several key findings. First, the ensemble voting mechanism proved superior to individual models by reducing false positives, achieving an estimated **precision of 72%** in manual validation compared to 55-60% for single-model detections. This 12-17 percentage point improvement validates the core hypothesis that requiring multi-model consensus effectively filters noise while retaining robust anomaly signals. Second, our statistical analysis confirmed that specific engineered features—most notably `user_verified_ratio`, `word_count`, and `char_count`—serve as robust discriminators with correlations exceeding 0.35-0.42 and highly significant p-values ($p < 10^{-6}$). Third, detected anomalies exhibited consistent behavioral profiles: unverified purchases, abnormal text length (either extremely terse or suspiciously verbose), and accounts with patterns of prolific unverified reviewing.

From a practical perspective, our system processes 100,000 reviews in approximately 3-4 minutes using standard computing resources, demonstrating feasibility for near-real-time production deployment. The interpretable voting mechanism provides confidence scores that enable prioritized manual review workflows, with unanimous detections (votes = 3) forming a high-priority queue for immediate human verification.

B. Limitations and Constraints

Despite these promising results, several important limitations constrain the generalizability and completeness of our findings:

(1) Lack of Ground Truth: The primary methodological constraint is the absence of confirmed fraud labels, which prevents calculation of Recall, F1-scores, and true positive/false negative rates. Our validation relies on manual review of only 100 samples (50 high-confidence, 50 single-model), which may not represent the full distribution of detected anomalies. Consequently, we prioritized Precision (reducing false alarms for human reviewers) over Recall (catching all fraudulent reviews), accepting that some genuine fraud may go undetected.

(2) Feature Scope and Semantic Blind Spots: Our current model relies primarily on behavioral statistics and surface-level text features (length, capitalization, punctuation) but lacks deep semantic understanding. This limitation means we may miss sophisticated fraud such as: AI-generated reviews that mimic natural language patterns and typical length distributions; subtle template-based reviews with varied wording; coordinated campaigns where individual reviews appear normal but collectively manipulate ratings. The absence of semantic analysis represents a significant gap in detection capability.

(3) Sampling and Scale Limitations: Due to computational constraints and project timeline restrictions, we analyzed only 100,000 of the available 570,528 reviews (17.5% of the complete dataset). This sampling may miss rare but impactful fraud patterns visible only at larger scales or introduce selection bias if the sampled subset is not representative of the full distribution. Additionally, the beauty product category may exhibit fraud patterns different from other Amazon categories, limiting generalizability.

(4) Temporal and Network Dynamics: Our analysis treats each review independently without considering temporal patterns (review bursts, coordinated campaigns) or network structures (reviewer-product bipartite graphs, reviewer collusion networks). This static, point-wise analysis cannot detect sophisticated coordinated fraud where individual reviews appear legitimate but collectively form suspicious patterns.

(5) Model Assumptions and Biases: HBOS assumes feature independence, potentially missing fraud defined by unusual feature correlations. One-Class SVM's kernel approach is sensitive to parameter choices (ν, γ) and may overfit to the specific 100k training sample. All models assume the contamination rate is approximately 5%, which may not match actual fraud prevalence and could bias detection thresholds.

C. Future Research Directions

Future work should address these limitations through several research directions:

(1) Deep Learning and Semantic Analysis: Integrate transformer-based language models (BERT, RoBERTa, or domain-specialized variants) to capture semantic anomalies, detect AI-generated text, identify template-based reviews, and analyze review coherence and relevance. This would address the current semantic blind spot and improve detection of sophisticated textual fraud.

(2) Temporal and Sequential Modeling: Incorporate time-series analysis to detect review bursts (multiple reviews within short timeframes), identify seasonal manipulation patterns, and model reviewer behavior evolution over time. Recurrent neural networks (LSTMs, GRUs) or temporal point processes could capture these dynamic patterns.

(3) Graph-Based Fraud Detection: Model the reviewer-product ecosystem as a bipartite network and apply graph anomaly detection algorithms to identify coordinated fraud rings, detect reviewer collusion patterns, and leverage network propagation effects. Graph neural networks (GNNs) could learn both node and edge anomaly patterns.

(4) Weakly Supervised and Active Learning: Develop an iterative human-in-the-loop pipeline where manual validation feedback refines detection thresholds and feature weights. Active learning strategies could prioritize which reviews to label next, maximizing information gain per human annotation. This semi-supervised approach could gradually improve both precision and recall.

(5) Multi-Modal and Cross-Domain Analysis: Extend the framework to incorporate image analysis (detecting stock photos, image manipulation), integrate reviews across multiple product categories, and perform cross-platform analysis (Amazon, Yelp, Google Reviews) to identify professional fraud operations. Transfer learning could adapt models trained on one domain to others with limited labeled data.

References

[1] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 188–197.

[2] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2008, pp. 413–422.

[3] M. Goldstein and A. Dengel, "Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm," in *Proc. German Conf. Artificial Intelligence (KI)*, 2012, pp. 59–63.

[4] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[6] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proc. Int. World Wide Web Conf. (WWW)*, 2012, pp. 191–200.

[7] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection," in *Proc. Int. AAAI Conf. Web and Social Media (ICWSM)*, 2013, pp. 175–184.

[8] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2015, pp. 985–994.

[9] Y. R. Shrestha and K. Ben-Menahem, "Organizational decision-making structures in the age of artificial intelligence," *California Management Review*, vol. 61, no. 4, pp. 66–83, Aug. 2019.