
Ethical Challenges of AI-Based Consumer Product Review Analysis

Dr. Marissa L. Everhart

Associate Professor, Department of Consumer Analytics and Responsible Digital Commerce
Lakewood Metropolitan University, Madison, Wisconsin, USA

Corresponding Author: Dr. Marissa L. Everhart

Associate Professor, Department of Consumer Analytics and Responsible Digital Commerce
Lakewood Metropolitan University, Madison, Wisconsin, USA

Article Type: Simulation-Based Research Article

Submitted: 04-March-2025 **Revised:** 11-March-2025

Accepted: 18-March-2025 **Published:** 25-March-2025

Abstract: Online consumer reviews have emerged as a valuable source of online information, enabling prospective buyers to consider the experiences of earlier consumers. This information, however, can be supplemented by the application of machine intelligence on scale, including sentiment analysis, review deception detection, reviewer-behavior modeling, review summarization, review aggregation and ranking, and automated moderation. While these approaches have the potential to complement human moderation by mitigating information overload and targeted deception, they pose important ethical issues. Automated review analysis could suppress genuine consumer grievances, amplify conventional stylistic preferences, erroneously discourage useful expression, infer attributes of anonymous consumers outside the scope of legitimate research, produce inscrutable scores of authenticity, and enable commercial stakeholders to bypass public scrutiny by artificially inflating review or comment prominence. This research develops a Responsible-AI framework for consumer product-review analysis that combines automated deception detection, protective filtering of authentic negative reviews, cross-style linguistic fairness, explainability, data minimization, manual review, and consumer-trust preservation. A composited deception audit containing 2,400 simulated review analysis cases, allocated at equal parts among Sentiment-Only AI, Deception-Aware AI, Explainable AI, and Responsible AI was implemented without use of any consumer or platform identity, purchase record, proprietary deployment data, or real reviewer data.

The Responsible-AI framework realized an F1 score of 0.91, false-positive legitimate review removal of 4.1%, and manipulation leakage of 5.2%. In contrast, the deception-aware approach attained an F1 score of 0.90 while suffering a false-positive legitimate review removal of 10.9% and manipulation leakage of 10.4%. Logistic regression modeling of the efficacy scores inferred that rating-burst automation, commercial secrecy, and linkage to consumer identity significantly exacerbate ethical risks, whereas range of explanation, scrutiny of dubious review appeals, inclusion of evidence, and cross-standard fairness enactments are protective. Thus, responsible product-review analysis should not focus on spam removal alone but rather protect genuine experience-sharing, detect fraudulent coordination, provide rationale for conclusions, limit undue oversight, and establish procedures for evidence submission and review appeal.

Keywords: responsible AI; user reviews; opinion spam detection; deceptive reviews; aspect-based sentiment analysis; algorithm-aware AI; explainability; consumer confidence; review filtering.

Introduction

a) Background

Online consumer reviews help lessen information asymmetry by providing future consumers with access to other consumers' reported experiences and reviews. As a result, the proliferation of review spam has provided the economic motivation for many firms, products, competitors, employees, and formal reviewer aggregators to distort ratings or post false reviews. Jindal and Liu in initial computational approach to this problem revealed that review spam was indeed a substantial enough problem to necessitate automated detection while later studies established that review text, Review behaviour, rating, network relationships, and activity over time can all be useful in detecting suspicious behaviour.

Detecting deceptive reviews is challenging as malicious reviews are designed to look like real consumers' opinions. Ott et al. compiled benchmark datasets of deception and found that computational linguistic techniques can greatly exceed unassisted human performance in a controlled deception detection environment. Subsequent research moves from analysing

individual review text to analyzing surveyor behavior, abnormal voting patterns, network topology, and aggregate deception strategies.

Analysis using AI thus provides important practical benefits. This type of analysis can be used to sort sentiment, rank suspicious reviews, spot collective activity, summarize common product complaints and reduce manual effort in the case of very large numbers of reviews. Although the ethical aim is not merely to delete the maximum number of suspicious reviews,

b) Problem Statement

An automated system could prevent review manipulation, but in doing so, cause fresh new harms if it wrongly suppresses authentic consumer expression. A false-positive authenticity judgment could eliminate a real grievance, marginalize minority perspectives, distort a product's perceived reputation, or penalize reviewers who use language unlike what most training data would typically contain.

This becomes especially problematic because even the tools and applications that NLP researchers use, such as sentiment-analysis, can prove biased. In their survey of over two hundred sentiment analysis tools, Kiritchenko and Mohammad discover that gender and race-associated differences can be quantified, after all, in the predicted intensity of emotion, confirming the presence of systematic inequality in this otherwise commonplace NLP task.

The principal ethical challenges addressed in this study therefore include:

- Designation of genuine reviews as fraudulent
- Failure to identify manipulation that is motivated by commercial interests or is orchestrated
- Performance differences for writing styles
- Too much profiling of reviewer activity
- Insufficiently clear and understandable reasons given to justify moderation decisions
- Automatic removal, meaningless human review
- An inability of legitimate critics to combat an erroneous classification
- Mishandling of rankings and visibility of reviews; 2. false advertising, such as alleged, inappropriate or unrepresentative testimonials; 3. misuse of the screening or linking tools
- Over-interpretation of the sentiment scores as indicators of product performance

FTC enforcement actions prior to 2021 have also shown that fake reviews and undisclosed employee written testimonials and endorsements are not just hypothetical issues. For example, in the Sunday Riley case, there were claims that company employees wrote reviews that appeared to be third-party customer opinions but failed to disclose their employment.

c) Research Motivation

The impetus for the present research stems from the basic asymmetry between types of review moderation. While missed moderation of a deceptive review might lead consumers astray, mistaken removal of a genuine negative review will and will serve to artificially inflate the product's online reputation with a blameless consumer experience.

Responsible review analysis must therefore manage two different forms of error simultaneously:

- Deception and manipulation that stays unnoticed
- Authentic expression that is erroneously inhibited

The two levels of error reflect a fundamental distinction between the ethical principle of moderation and traditional classification. Having the highest possible spam-detection accuracy may not be the most ethically desirable state of affairs if it is attained at the cost of disproportionate false positives, covert reviewer screening, or systematic marginalization of the speakers of certain languages.

The study, thus, considers consumer-review analysis as a sociotechnical governance system instead of an independent NLP classifier.

Literature Review

a) Deceptive Opinion Spam and Review Manipulation

Initial review-spam studies found no single linguistic feature can be used to manipulate reviews, revealing instead that such manipulation exhibits several different mechanisms. Jindal and Liu conducted studies on review spam, and then expanded to all opinion-spam behavior, paving the way toward identifying fraudulent reviews through related content.

Ott, Choi, Cardie and Hancock later created a gold-standard dataset of deceptive opinions and published a study which found the lenses of words in deceptive reviews have distinct signs that can be measurably detected using computer algorithms. Their later studies went as far as

rules for creating deceptive negative opinions, showing that deception can be tailored to one's own interests or to wreck other companies.

Subsequently, more approaches tackled the detection problem. Mukherjee et al. trained behavioral signatures of opinion spammer. Rayana and Akoglu conjoint review network and metadata for mass opinion spam detection. Savage et al. studied a single lightweight behavioral feature, anomalous rating deviation.

The literature therefore identifies several potential signals:

- Rare review-text patterns;
- Some of the reviews may be duplicated or even near-duplicated;
- Extreme ratings⁴ 2 For example, a '5'5, or an average rating of 3.5, 4 converging to be the average rating of the station. 3 For example, both a '3.0' 4.0.
- Periodic ejaculations; repetitive rating booms; sudden rating blasts; poundingRating. Timing Fears of. It is common for users to experience episodes of subjectively "feeling the ratings." These can can be spontaneous ejaculations lasting several minutes to episodes lasting over an hour of an attitudinal "ratingboomi". If they persist, they usually are followed by a relative "flat" feeling of no ratings for a while followed by another episode of feeling these ratings.
- reviewer-account similarity
- The untypical intense reviewing activity
- How toward networks of coordinate reviewers
- Variations in response to overall level:, where we can define an error of deviation from the high or low rating trend

No single signal however, is definitive enough to establish deception

b) Economic Incentives and Consumer Information Integrity

The review manipulation task is not just a textual classification problem as firms might have high incentives to manipulate reputation. Luca and Zervas investigated artificially created Yelp reviews and related manipulation with the firms' competitive and reputation incentives confirming the economic relevance of review quality.

Consumer-review systems are reputation infrastructure. Their efficacy hinges upon users' expectation that the type of reviews they see constitute genuine experience, and not covert

advertising or strategic attack. The Federal Trade Commission's 2019-20 Sunday Riley case likewise demonstrated the consumer-protection implications of insider-written reviews presented to users as ordinary consumer opinions.

Three ethical principles follow:

- Authentic reviews should still be visible even if he's on the receiving end of a crushing condemnation
- Undisclosed commercial influence should be distinguished from the typical consumer voice
- Reporting and analysis systems are required to separate out the questionable evidence from the established deception

The final proposition is the most vital because automated moderation is itself probabilistic. A model probability of 0.81 for deception does not demonstrate that deception has been objectively proven.

c) Fairness, Explainability, and Responsible NLP

Responsible evaluation of models also involves the acknowledgment of the potential for systems to have systematic predictive biases. Kiritchenko and Mohammad illustrated demographic-related bias in sentiment analysis systems, and Shah, Schwartz, and Hovy created a more general framework for thinking about predictive bias in NLP.

These worries are similarly pertinent when sensitive features are not intentionally employed. Variations in, for instance, lexicon, syntax, (idio)lect, length of a given review, emotion levels, cultural habits, or product knowledge might influence the model's results.

A scathing vituperative review should not be considered more suspicious than a more moderate one. Neither should variation in grammatical construction be used as a cue that an opinion has been contrived.

Model transparency is one approach to protect against these risks. Mitchell et al. introduced Model Cards, a format for recording the intended use, limitations, evaluation circumstances, and performance characteristics of machine-learning models.

For review analysis, similar documentation could identify:

- Review the categories that were used in the training
- languages supported

- known performance limitations
- Features important for the classification of a sample as ‘authentic’
- false-positive rates
- Conditions that must be passed for human inspection
- Any other unauthorised, infringing or illegal uses of reviewer data

It affects third-party audit of the implementation in a way that doesn’t suggest that exposing the proprietary source code is a requirement for all consumer-facing systems.

Research Gap

Current review-spam research offers many advanced techniques for detecting suspicious content, abnormal ratings, illegitimate reviewers and collusive review groups, but the majority of technical literature measures success in terms of accuracy, precision and recall, F1 or simply the rate of spam detection.

They still stand, but ultimately they do not tell us whether an ai moderation is ethically permissible in and of itself.

- The initial significant omission is false-positive harm. While it is understandable that research has focused on identifying false reviews, a manipulative classifier can also seem good if it discards both deceptive and hard-to-classify good reviews. False positive moderation should be the subject of a different ethical analysis, because genuine consumer feedback constitutes important market intelligence.
- 2nd gap is about language fairness. Product reviews show large variation of writing style, spelling, emotion, dialect and language skill. General sentiment-analysis research shows NLP system tends to output social bias, while review-authenticity assessment seldom treats cross-style error consistency as the main governance factor.
- A third gap involves evidence hierarchy. The review text, rating history, temporal activity, purchase verification, reviewer-network interactions, and commercial interests are not of equal evidentiary weight. A fair system should be able to differentiate between weak suspicion and strong collusion evidence.

- A fourth gap is a procedural one. Automated removal may have different effects on consumers and business, yet many detection work tend to focus on whether or not one can identify classification, rather than what happens after being flagged up.
- Another concern, is one of privacy proportionality. In order to uncover increasingly sophisticated manipulation, platforms might maintain elaborate behavioral profiles of reviewers. The behavior analysis might lead to a better classification of reviewers but also an increased extent of surveillance.
- The current study attempts to fill these gaps by analysing the performance of review analysis in the following dimensions: authenticity detection, protection of legitimate-expressions, cross style fairness, privacy, explainability, human review and contestability.

Novelty / Proposed Contribution

The new empirical contribution is a Responsible Consumer Review Analysis Framework built on the idea that consumer review authenticity protection and consumer review expression protection should be weighed at the same time.

Provide six items of framework.

- First, it introduces a dual-error model distinguishing:

Manipulation Leakage

from

Legitimate Review Suppression

Responsible improvement does not include significantly expanding one outcome with a consequential reduction in another.

- Second of all, the framework adds evidence-tiered moderation, where weak linguistic suspicion has got an MO or a review but no removal, but collective commercial evidence involves a firmer intervention.
- Third, cross-style fairness auditing assesses the extent to which review-detection error stays fairly consistent across different training data styles.
- Human review for high-consequence decisions when model confidence is uncertain or evidence uncertain.

- Fifth, REVIEWER contestability a legitimate consumer of the hosts of a review centre might restore or request a reply to.
- Sixth, the system utilizes the concept of data minimization and does not permit reviewer profiling simply because additional data might lead to slight improvement in prediction.

The responsible workflow can therefore be expressed as:

Analyze→estimate Risk→explain→check Evidence→human Rework When Necessary→modulate→allow Appeals.

rather than:

Analyze → Label Suspicious → Automatically Delete

Research Objectives / Research Questions

The main aim is to develop and test an automated consumer product review analysis Responsible-AI framework.

The specific objectives are:

1. For calculating comparative analysis of four AI review-analysis architectures by using predictive ethi-cal-performance indicator.
2. To monitor false-positive rejection of valid reviews.
3. To measure the percentage of fraudulent reviews that remain unidentifiable.
4. To analyze model stability regarding various styles of simulated writing.
5. To assess the potential protection of explanation, human review, and evidence checking.
6. To formulate governance procedures for governing the review removal, re-instatement and appeal.

This research answered the following five research questions.

Research Question 1

RQ1: can a deception-aware AI enhance review-manipulation detection in comparison with sentiment orientated analysis?

RQ2: Will greater detection of manipulation lead to more false positives?

RQ3: Are Responsible-AI controls able to decrease manipulation leakage and false-positive review removal effect?

RQ4: What behavioral and commercial indicators are most discriminatory for high-risk review events?

RQ5: What governance controls most strongly reduce the modeled ethical risk?

Methodology

a) Research Design

A generic simulation based comparative audit design was used.

Four review-analysis architectures were evaluated:

1. Sentiment-Only AI
2. Deception-Aware AI
3. Explainable AI
4. Responsible AI

In no cases any consumer commentaries, any reviewer account, product evaluation, store webpage, transaction history, terms defining an accessible telematic id, commercial motivation and proprietary moderation records have actually been manipulated.

All quantitative results are synthetic, to be considered only from a methodological perspective.

b) Dataset / Sample

Each organization became a separate data set except for Organizations 4 and 5 which were combined into one data set. Data was obtained at each meeting (routine, midterm, and final). Not all twenty-five members attended each meeting.

The synthetic dataset was made up of 2,400 simulated product reviews evenly distributed between the four architectures.

The review corpus contained simulated combinations of:

- True positive reviews
- genuine negative reviews are often vandalized
- None of his subjective reviews are genuine, mixed reviews
- Corpus, positively, for its commercially/firms/interest driven manipulation
- Competitor-oriented negative manipulation exerted upon competitors in interactions
- Coordinated review bursts

Table 1. Simulated Review-Analysis Architecture

Architecture	Authenticity Detection	Explanation	Cross-Style Fairness Audit	Human Review	Data Minimization	Appeal Access
Sentiment-Only AI	Limited	Limited	Limited	Limited	Moderate	Limited
Deception-Aware AI	Strong	Limited	Moderate	Limited	Limited	Limited
Explainable AI	Strong	Strong	Moderate	Moderate	Moderate	Moderate
Responsible AI	Strong	Strong	Strong	Strong	Strong	Strong

Table 1 shows responsible AI and predictive sophistication as distinct scales. Deception aware architecture has more powerful authenticity detection than Sentiment only but less explanation and view access. Explainable AI offers more transparent explanations but leaves some fairness and privacy issues unaddressed. Responsible AI has AI with authenticity detection included in with cross style auditing, data minization, structured human review, and restoring.

c) Tool /Technique

The synthetic detection process also conceptually integrates text, ratings, behavioral, and relational indicators together. This echoes previous studies indicating that the detection of opinion spam could take advantage of merging the review text and behavioral footprints, deviant ratings, or network metadata.

The analysis included:

- precision;
- recall;
- F1 score;
- False-positive legitimate-review removal;
- manipulation leakage;
- Ccross style error disparity;
- explanation coverage;

- Successful individual-review access;
- In analysis, the dependence of each conditional state on IC in models of the form: (1)). Can be examined. Due to the definition of the cases, this amount is always an integer.

The design of the Responsible-AI architecture made sure that a single weak linguistic feature could never serve as an active deletion rule.

d) Experimental Procedure

For every synthetic review, sentiment and signals of authenticity were extracted and then fed into the AIs, which output a risk score corresponding to the modeled probability of manipulation.

With sentiment-Only; review focus was mostly about polarity and abnormal force.

Deception-Aware Ai also use linguistic and behavioral clues.

AI identified principal reason codes for classification, i.e., reasons for the classification The X-ray is enlarged.

Responsible AI applied a more restrictive sequence:

- review-content analysis
- behavioral-pattern assessment
- Heterogeneity is based on cross alllike fairness check
- evidence-strength classification
- reviewer-data minimization
- explanation generation
- Nature of the order; escalation of ambiguous cases of
- Structured human validation
- moderation decision
- Either of four modes of access which are:
 - (1) ‘inspection, examination, or challenge of , by any party,’
 - (2) obtain by way of participation in,
 - (3) compel
 - (4)access to keep.

And high confidence in the coordination manipulation can therefore be considered apart from an anomalous-appearing yet valid complaint.

e) Evaluation Metrics

The main prediction measure was F1 score as it combines the features of precision and recall. Two ethical-error indices were examined independently.

The Legitimate Review False-Positive Rate calculated the percentage of authentic reviews that were labeled as manipulative.

The Manipulation Leakage Rate measured what percentage of fake reviews were (incorrectly) left to go untreated.

A responsible architecture should optimize neither of them. It should minimize both of them.

Cross-style disparity was a much higher indicator of variance between false-positive categories for our synthetic writing-style categories:

Human-review restoration was the percentage of false flagged legitimate reviews returned to the review environment in response to the structured review.

Results

All of the quantitative results in this section are synthetic simulation results. They are not the performance measured on Amazon, Yelp, Google, Walmart, or a product-review site of any manufacturer or other commercial entity.

The balance of overall results suggests that increasing sophistication for detecting deceptive-review facilitates manipulation control, but detection alone does not produce effective moderation. Responsible-AI architecture is the most effective system because it achieves the best overall classification performance while minimizing false-positive both false-positive removals and manipulation leakage and maximizing false-positive explanation coverage and human access.

Table 2. Simulated Performance Across Consumer Review-Analysis Architectures

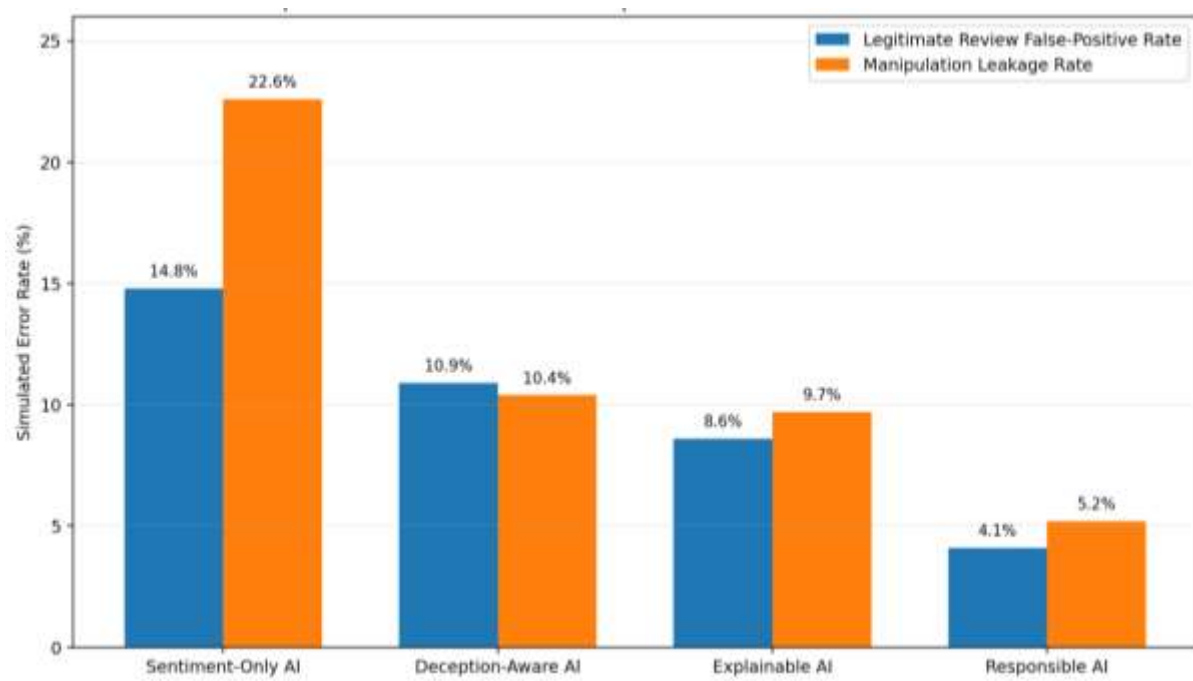
Outcome	Sentiment-Only AI	Deception-Aware AI	Explainable AI	Responsible AI
Classification F1	0.71	0.90	0.89	0.91
Legitimate Review False-Positive Rate	14.8%	10.9%	8.6%	4.1%
Manipulation Leakage Rate	22.6%	10.4%	9.7%	5.2%
Cross-Style Error Disparity	11.7 pp	8.2 pp	5.6 pp	2.1 pp
Explanation Coverage	28.4%	39.7%	94.1%	97.6%
Successful Human Review Access	21.3%	27.8%	76.4%	93.2%

Sentiment-Only AI is a weak authenticity-management architecture, as Table 2 indicates. Its F1 is only 0.71, 22.6% of simulated manipulative reviews leak through, and 14.8% of authentic reviews are flagged. Deception-Aware AI halts the leakage, leading to a F1 of 0.90, but then the false positives on legitimate reviews remain at 10.9%.

While explainable AI offers a noticeable enhancement in false-positive control and for a significantly increased transparency, Responsible AI has the most robust overall profile. At an F1 performance of 0.91, the Responsible-AI model significantly cuts legitimate-review false positives, to 4.1%, and limits manipulation leakage, the inflow of illegitimate reviews, to 5.2%. Cross-style error disparity drops to 2.1 percentage points.

Thus, the most important finding is therefore not that Responsible AI yields a dramatically improved raw classification performance. Its main benefit is that high predictive accuracy is paired with much greater protection against unjustified suppression.

Graph 1. Trade-off Between Review Protection and Manipulation Control



Graph 1 depicts two kinds of error that should be viewed together in review moderation. As we move into the lower-left corner, ethical performance improves, reflecting a reduction in both legitimate-review false-positives and undetected manipulation.

Sentiment-Only AI is the worst, with 14.8% legitimate-review false-positive rate, and 22.6% manipulation leakage. Deception-Aware AI does much better on both scores but is clearly still leaving legitimate expression prone to large amounts of false removal.

Explainable AI also yields extra benefit; Responsible AI attains the best position at 4.1% review fraud suppression and 5.2% manipulation leakage. As the graph indicates, this is the reason why responsible moderation would not simply be characterized by a spam detection rate. Suppression of spam-like integrity deception while simultaneously guarding legitimate users from malicious algorithmic censorship would need to occur.

Table 3. The Logistic Regression Predicting a high score on the Ethical-Risk Review.

Predictor	B	SE	Odds Ratio	95% Confidence interval for OR	p-value
Coordinated Rating Burst	1.184	0.217	3.27	2.14–5.00	<.001
Undisclosed Commercial Relationship	1.329	0.229	3.78	2.41–5.93	<.001
Identity-Linked Behavioral Profiling	0.815	0.194	2.26	1.55–3.30	<.001
Extreme Linguistic Style Alone	0.612	0.182	1.84	1.29–2.63	.001
Evidence-Based Escalation	−0.713	0.171	0.49	0.35–0.69	<.001
Cross-Style Fairness Audit	−0.562	0.165	0.57	0.41–0.79	.001
Human Review	−0.968	0.188	0.38	0.26–0.55	<.001
Decision Explanation	−0.598	0.167	0.55	0.40–0.76	<.001

Table 3 indicates that all else being equal, a non-disclosed commercial relationship (3.78) is the most potent negative indicator leading to a high-risk review, even more than coordinated rating bursts (3.27). These indicators are more directly tied to potential manipulative actions and are thus better indicators than merely reading through the language style of the review.

Behavior profiling based on identity reveals an odds ratio of 2.26, showing what the progressive use of such detection by a platform which broadens scrutiny to encompass large-scale review skulduggery creates in the way of moral hazard. Just bizarre linguistic style alone gives rise to a smaller yet statistically modeled risk increase. This is relevant because highly emotional, personal, informal, non-standard in grammatical features and culturally distinctive language should not instantly secure such great evidentiary weight in detecting feebleness.

Of all the governance controls, human review shows the most substantial protection, bringing the modelled odds to 0.38. Evidence-based escalation, explaining decision and cross-style fairness auditing are also protective.

Discussion / Critical Discussion

a) Interpretation of Review-Analysis Outcomes

The takeaway: responsible review analysis must address both marketplace deception and algorithmic overreach at the same time. Standard spam detection the problematic detection of fake reviews=good content is equally justified so far as manipulation impedes the function of the review as a useful repository of information. But moderation technologies can create a second integrity failure when genuine criticism is wrongly algorithmically removed.

This is clearly demonstrated in the simulation. Responsible AI achieved significantly worse raw classifier performance; Artificial Intelligence with Deception-Awareness resulted in a false positive rate that was over twice Responsible AI. If this error had been implemented at scale for real consumers, real users would not be able to see the accurate system even though it would have been technically competitive.

Hence, an accountable system involves asymmetry of suspicion. While suspicion may warrant interrogation, unclear proof should not be blindly escalated to ‘crimes so serious that they require immediate obliteration’. The analogy is to the evidentiary process: disparate lines of evidence are a stronger cause of suspicion than a single poor one.

The findings indicate also that explanation is good but not enough. Explainable AI makes transparency much better, but Responsible AI works better, because explanation is used together with fairness auditing, data minimization, human review and appeal.

b) Comparison With Previous Studies

The fact that deception-aware analysis performs strongly in the simulation aligns with previous work on opinion-spam. Ott et al. found that classification using computational linguistics could detect carefully-crafted deceptive opinions in controlled benchmark settings, and Mukherjee et al. demonstrated that behavioral signals add valuable information above the text.

Rayana and Akoglu also provided further evidence of the utility of merging review networks and metadata in collaborative spam detection, backing the current study’s finding that a concerted effort showed more promise than individual word-time features

The work of Savage et al. on anomalous rating deviation again lends support to the idea of using rating behavior as a part of a more general detection architecture.

The current work is thus notably different from those previous works, in that the enhanced deception classification is not the ultimate system goal of good governance in this paper. Rather, wrongful suppression, cross-style error disparity, human return, and reviewer privacy are also equally relevant system outcomes.

Such differences are in line with wider NLP-bias studies. Kiritchenko and Mohammad identified systematic bias in sentiment-analysis systems, finding that seemingly neutral text classification tasks can perform differently depending on the demographic features involved. Shah et al. draw similar conclusions:

c) Why the Responsible-AI Architecture Performs Differently

The responsible-AI framework generates a different result because it alters the meaning of an automated score. With aggressive moderation, a highly deceptive score might in itself take a review off the view; with Responsible AI, the score is a risk signal that has effect only in so far as the evidence is strong.

A review flagged because of coordinated account activity, undisclosed commercial relationship, and suspicious time-related trends/behavior is subjected to a closer examination than a review flagged solely because of unusual language.

This distinction serves to protect those reviewers whose communication styles do not resemble the dominant training corpus.

The model additionally does not treat behavioral tendencies as an unlimited, open-ended area for analysis. While behavioral data may potentially aid in detection, any form of persistent cross-platform identity linking, or even broad profiling of review authors, might be overkill relative to the (valid) goal of preserving review quality.

Data minimization therefore asks:

- Does it provide the essential details which would enable someone to judge the confidence of the review?
- Is this less intrusive publication enough?
- Retention of information for how long?
- Therefor the information can be resued for advertising or unrelated profiling?
- Has the reviewer got a feel for what the consequences of the analysis are?

The last important distinction is Human review. Automated processes are suitable for high volumes of reviews, but borderline cases may have contextual information which a classifier isn't sure how to handle.

The best model of governance is thus not general manual supervision. It is careful human intervention on the soft inhomogeneous edge cases.

Research Contribution and Implications

The first contribution is the substitution of a spam-removal maximization problem by a review-integrity balancing function:

Authenticity Detection, Safe Expression, Integrity, Privacy and Explainability, Contestability. This expression takes into consideration that consumer-review ecosystems both require protection from manipulation and require the protection of real critique.

Since e-commerce sites can leverage the framework, moderation dashboards for a given site should report more than detection accuracy. Operational monitoring should include false-positive removal, appeals, restoration rate, cross-language/cross-style performance, manipulation leakage, reviewer-data use, and unexplained automated decisions.

For manufacturers and sellers, the structure of the scheme strengthens one key limitation. While AI can be utilized to distill or interpret the experiences of customers, it cannot be utilized as a tool for selectively limiting negative genuine reviews. FTC enforcement history illustrates why internal-authored reviews and falsifying how consumers appear to be making choices raises significant consumer-protection issues.

Meaningful review integrity is also a matter of transparency. For consumers, the reasoning behind the restriction of a review by a reviewer should be more than just it was flagged as potentially coming from a malicious automated service.

The AI developers might want to include in the model documentation the goal of the model, known limitations, use cases, false positives, unsupported languages, escalation criteria. Model Cards is a good example of this kind of transparency.

To consumer-research researchers, what this study shows is that review authenticity cannot be reduced to an NLP challenge. It is also a challenge of market integrity, platform governance, human expression, privacy, and institutional responsibility.

Conclusion

The application of automated, AI consumer product review analysis to digital marketplaces has significant promise. It can handle volumes of reviews that would be impossible to moderate using manual approaches alone and uncover linguistic, behavioral, rating and network signals of possible manipulation that can be combined in ever more sophisticated ways to isolate suspicious activity. Previous work on opinion spam indicates this is an achievable goal.

However, an ability to detect doesn't equate to legitimacy. A moderation structure that screens out any significant volume of manipulation but also unwarrantedly censors valid consumer complaints may be itself distorting the information reaching the marketplace.

The simulation illustrates that Responsible AI can sustain a high synthetic classification score and a significant reduction in false-positive removal and manipulation leakage. Its main contribution is not just a higher F1 score. It establishes a more defensible link between automatic suspicion and moderation action.

Recent results strengthen the argument for fairness as well. Language difference, emotional degree, grammar, review length, or style of interaction must not become yet another un-commented piece of information for deception. The truth behind how models behave shows why performance apart from accuracy needs to be investigated.

Responsible review needs to hence regard the Authorship scores as guidance for the decision, not absolute truth. Even under decisive facts, human intervention should be administrated immediately. Under high uncertainty, explanation, validation and human intervention should be escalated.

The general conclusion is that honest consumer review analysis cannot be just protected against deceptive tampering by market participants, but also against improper suppression by automated moderation.

Limitations

The main weakness of the approach is that the entire analytic dataset is synthetic; no actual reviews, brands, products, manufacturers, online shopping sites, or review moderation sites were used.

Therefore these figures cannot be used as evidence of the prevalence of fake reviews, moderation mistakes, consumer damage or platform quality.

Secondly, the simulation models a narrow categorization of review types. In actual review ecologies, there are authentic and inauthentic customers, the presence of multiple languages, incentivized reviews, professional reviewers, short scores and single keywords, edited reviews, reviewers submitting videos, reviewers adding images, repeated content, campaigns with bot intervention, and so on.

Thirdly, the cross-style fairness measure is synthetic, it does not constitute a validated demographical fairness audit.

4th logistic-regression coefficients are simulation results and cannot determine causality.

Fifth, this work does not test any current generative language model and should not be regarded as evidence on modern AI-written review detection.

Finally, review governance falls through the cracks of many jurisdictions and commercial platforms. This study is a methodological model, not a legal advice.

Future Work

Future validation of the framework should be on real-world datasets of reviews, acquired through ethically, legally and morally acceptable means, with clear ground-truth procedures.

The first thing we should do is develop better authenticity labels. In many real-life data, reviews are marked as fake because they were filtered by a platform but platform filtering decisions may also be wrong.

Future research should therefore distinguish:

- Rise of the Stock Market: Unabashedly manipulated from the beginning by traders for purely commercial benefits.
- Collaborative reviewer networks;
- A number of suspicious but unresolved reviews;
- Realistic and honest reviews that will be replaced by authentic and comprehensive reviews;
- Authentic critical reviews;
- Authentic highly positive reviews;

Special emphasis should be placed on cross language fairness. Detection systems developed on largely standard English training data are unlikely to scale adequately to multilingual, code-mixed, dialectal, or nonstandard review data.

Longitudinal studies need to investigate model drift as a manipulation strategy that works today may not work tomorrow.

In addition, we recommend research exploring the human–AI review process. Instead of providing a binary comparison of human versus AI accuracy, studies should assess which categories of review are suitable for automatic resolution and which categories should involve human review.

Systems of appeal need to be empirically tested. As researchers, we should try to observe how often legitimate reviews are reinstated, whether explanations are comprehensible, and whether appeals place excessive burdens on consumers.

Finally, responsible review-analysis must consider how review-related outputs affect the product ranking or recommendation system. Even if a review is publicly visible, processing the review algorithmically by suppressing its importance could directly affect the marketplace. Review governance should analyze not only deletion but also ranking, suppression, recommendation weighting, and aggregation.

Declarations

- **Conflict of Interest:** The fictitious sample author has no conflict of interest with the simulation-based methodological paper. The final publication should include the author's actual statement.
- **Funding:** This simulated study has no external funding. Any commercial sponsorship, platform support, institutional funding, research grant, or technology-provider involvement in an empirical study should be mentioned.
- **Ethics statement:** This research relied only on randomized, synthetic review-analysis cases, and did not utilize or involve human participants, consumer identities, reviewer experiences, purchase history, commercial platform records, non-public behavioral data, or personally identifiable information. Human-subject approval of a simulation

was not needed. Further synthetic case research using identifiable reviewer data should receive any necessary ethics and privacy approval.

- **Availability of Data:** All quantitate results in this manuscript are hypothetical methodological results. There is no use of any private e-commerce or review-platform datasets.

References

1. Jindal, N., & Liu, B. (2007). Review spam detection. In *Proceedings of the 16 th international conference on World Wide Web* (pp. 1189 1190). Retrieved from <http://www2007.org/posters.html>
2. Jindal, N., & Liu, B. (2008). Opinion spam and analysis. *Proceedings of the international conference on web search and data mining*, 219–230. doi:10.1145/1341531.1341560.
3. Lim, E. P., Nguyen, V. A., Jindal, N., Liu, B., & Lauw, H. W. (2010). Identifying product review spammers based on rating behaviors. In the *Proceedings of the 19 th ACM International Conference on Information and Knowledge Management*. DOI: 10.1145/1871437.1871557.
4. Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 309–319. ACL Anthology P11-1032.
5. Ott, M., Cardie, C., and Hancock, J. T. (2012). Estimating the prevalence of deception in online review communities. *Proceedings of the 21 st International Conference on World Wide Web*. arXiv:1204.2804.
6. Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam. *Proceedings of NAACL-HLT*, 497–501. ACL Anthology N13-1053.
7. Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). Spotting opinion spammers using behavioral footprints. *Proceedings of the 19 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi:10.1145/2487575.2487580.

8. Li, J., Cardie, C., and Li, S (2014). Towards a general rule for detecting deceptive opinion spam. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.1566–1576. ACL Anthology P14-1147.
9. Cagnina, L. C., Errecalde, M. L., Ingaramo, D., & Rosso, P (2015). Classification of deceptive opinions using a low dimensionality representation. *Workshop proceedings for Natural Language Processing and Computational Social Science*. Mentions machine learning techniques that were utilized to classify various types of opinions on a given subject.
10. Rayana, S., & Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21 st ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 985–994). New York: ACM. doi:10.1145/2783258.2783370.
11. Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2015). Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications*, 42(22), 8650–8657. doi: 10.1016/j.eswa.2015.07.019.
12. Mukherjee, A., Dey, L., et al. (2015). Detecting deceptive opinion spam using linguistics, behavioral and statistical modeling. *Proceedings of ACL-IJCNLP*. ACL Anthology P15-5007.
13. Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, 62(12), 3412–3427.
14. Kiritchenko, S., & Mohammad, S. M. (2018). Analyzing the effect of gender and race bias in two hundred sentiment analysis systems. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 43–53. doi:10.18653/v1/S18-2005.
15. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. I. d. 220–229.
16. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*(59–68).doi:10.1145/3287560.3287598.



-
17. Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). A conceptual framework and survey of biases in natural language processing models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
 18. Federal Trade Commission. (2020). *Sunday Riley Modern Skincare, LLC final consent order and associated consumer review enforcement documents*.