Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Towards Characterizing and Detecting Incentivized Reviews on eCommerce Platforms

Rajvardhan Oak*, **Zubair Shafiq**

University of California, Davis
{rvoak, zubair}@ucdavis.edu

Abstract

Customer reviews play an important role in rankings and visibility on e-commerce sites, and also strongly influence a customer's decision to purchase a product. Motivated by this, malicious sellers engage in incentivized review fraud to inflate their product ratings by providing customers with free products in exchange for five-star reviews, thus compromising review integrity. While there is ample prior work on fake reviews in general, there is limited prior work on incentivized review fraud. In this work, we infiltrate an underground market for fake reviews and implement a custom crawler to collect a dataset of malicious products that seek incentivized reviews. We devise and extract a set of features, and show that these are statistically significant in differentiating between benign and malicious products. Using hypothesis testing, we identify characteristics and trends exhibited by malicious products. While we are unable to achieve a high precision when we train standard machine learning models without compromising on the recall, we propose a lightweight two-phase technique that combines high-precision product classifiers with high-recall review classifiers. This technique allows us to minimize the false positives, with only a slight increase in false negatives. Finally, we also audit the effectiveness of two publicly available tools for incentivized review detection and find that they are not reliable. In summary, we contribute a new high-fidelity dataset, characterize products seeking incentivized reviews, audit existing tools for review analysis, and present a superior method for detecting review fraud. We hope that this research could be useful for e-commerce companies and other entities who have a stake in preserving opinion and review integrity online.

Introduction

Customer reviews play an important role in rankings on e-commerce marketplaces such as Amazon, Walmart, and eBay. Positive customer reviews have been shown to strongly influence a buyer's decision to buy a product (Guo, Wang, and Wu 2020) – 82% customers checked product reviews before buying a product online (Smith and Anderson 2016). Given that there are usually many third-party sellers for any given item, positive customer ratings and reviews play a critical role in the success of a seller on a competitive marketplace. Many sellers, especially with new products without many reviews yet, are willing to pay to obtain positive reviews (Imbler 2018; Nicole Nguyen 2018). E-commerce marketplaces prohibit sellers from soliciting reviews in exchange for free products or other monetary incentives (Chee Chew 2016). Instead, they themselves provide incentivized review services to help new sellers gain initial traction. For example, Amazon Vine[1] is an incentivized review service that enables sellers to send free items to an invite-only group of reviewers that are expected to provide honest and unbiased reviews. Since not every seller gets access to these incentivized review programs and more importantly, get guaranteed positive reviews, an underground ecosystem of incentivized reviews has emerged over the last few years. These underground incentivized review services allow sellers to solicit guaranteed positive reviews from real customers in exchange for free products. These services operate through a complex web of social networks, using Facebook as the primary medium. Sellers reach out to potential reviewers using targeted advertisements and specialized private groups where all forms of engagement (reviews, ratings, likes) are brokered in exchange for cash or free products (Oak and Shafiq 2024). A few examples of group activity can be seen in Fig 1. Fraudulent buyers purchase the product on the platform, an then get reimbursed by the seller via a third party (Venmo, PayPal, Zelle) in exchange for a review. Since e-commerce platforms do not have visibility into transactions on a third party, they are not able to flag such reviews directly. In recent times, such fraudulent incentivized reviews have received significant attention from technology giants and government agencies alike. Since 2022, Amazon has launched a series of lawsuits [2], seeking judicial action against companies and individuals acting as review brokers. In 2023 alone, Amazon removed $250M$ reviews from their platform under suspicion of incentivized review fraud. The Federal Trade Commission (FTC) published new rulemaking that specifically deems undisclosed incentives (such as compensation or refund) in exchange for reviews an act of deception.

*As of this writing, the author was employed by Microsoft. This work is not endorsed by Microsoft in any way.

[1] https://www.amazon.com/vine/about

[2] https://www.aboutamazon.com/news/policy-news-views/amazons-latest-actions-against-fake-review-brokers-2023

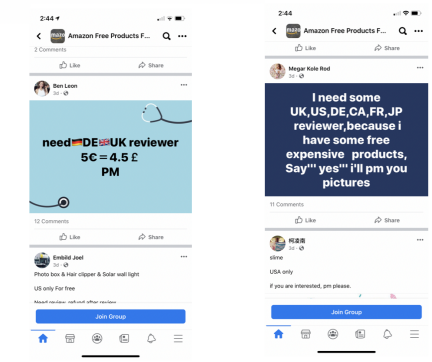


Figure 1: Sample posts in Facebook Review Groups where sellers are asking for buyers and sharing the products that need reviews.

Challenges

Typically, e-commerce marketplaces have strict policies against such reviews and therefore have systems in place to detect and remove them. However, detecting incentivized reviews is a harder problem than detecting traditional fake reviews. Incentivized reviews significantly differ from traditional fake reviews in several key ways as they are not written by bots (Jindal and Liu 2007) or crowdworkers (Fayazi et al. 2015; Rahman et al. 2019). These reviewers are in fact actual customers of the e-commerce marketplaces with a seemingly benign history of making routine purchases. As a result, incentivized reviews do not exhibit any obvious signs of automated behavior (as they are actually *not* automated) such as reviews clustering together in time or location, or reviews being too similar to each other Additionally, a customer who writes incentivized reviews may also be making legitimate purchases on the same platform which are non-incentivized. Therefore, the account activity as a whole is also not indicative of fraudulent behavior, as opposed to bot or crowdworker accounts whose account activity will be focused solely on reviews, and thus entirely fraudulent. These crucial differences make it challenging to reliably detect incentivized reviews.

Motivations

Detecting review fraud is crucial for preserving consumer trust in online platforms, as it ensures that reviews accurately reflect genuine opinions and experiences. Fraudulent reviews may provide malicious sellers with an unfair competitive advantage. According to the latest guidance by the FTC (Commission 2023), undisclosed incentivized reviews are a form of consumer deception and businesses engaging in such review fraud may face penalties. Moreover, such reviews result in counterfeit or harmful products being promoted (Nicole Nguyen 2018). Therefore, understanding and detecting incentivized review fraud is crucial towards improving the integrity of online communities.

Contributions

In this work, we seek to characterize products involved in incentivized review fraud, and then develop methods to detect them. First, we infiltrate an underground market where reviews are purchased, and collect a list of products tat are soliciting incentivized reviews. Then, using our custom crawler, we obtain the product metadata and reviews, from which we extract features for machine learning modeling. We summarize our main contributions as follows:

- We collect and contribute the first known dataset of fraudulent products and incentivized reviews, with labels derived from information from actual incentivized review campaigns.
- We devise a set of 36 features and examine them for statistical significance. We find that most features show statistically significant differences between products with and without incentivized reviews.
- We examine the performance of machine learning algorithms in fraudulent product detection and find that while the performance is reasonable, it is not good enough for use in production systems at the large scales that e-commerce platforms typically operate in.
- We develop a novel, two-stage classification method that combines product and review level classification. Our method tunes the product classifier at high precision and uses review classifier tuned at low recall to correct for the false negatives. We find that this technique allows us to reach high precision with very low drop in recall.
- We audit the effectiveness of two popular tools for incentivized review detection (Fakespot and ReviewMeta) on our dataset of incentivized reviews and find that they are unable to reliably detect incentivized reviews.

Background & Related Work

Fake Reviews

Prior works have detected the existence and impact of fake reviews on websites like TripAdvisor, Expedia (Mayzlin, Dover, and Chevalier 2014). Hu et al. (Hu, Liu, and Sambamurthy 2011) conducted a large-scale analysis on Amazon and Barnes&Noble and found that consumers cannot fully isolate incentivized reviews from genuine reviews and can be easily misled. Fake reviews have evolved over time in terms of the strategy and mechanisms through which they are generated. Traditional fake reviews were generated and posted by automated bots; detection methods for these rely on exploiting features commonly seen in bots such as high burstiness or low entropy in text (Jindal and Liu 2007; Fontanarava, Pasi, and Viviani 2017; Patel and Patel 2018). Later research focused on reviews generated by crowdworkers. Kaghazgaran et al. (Kaghazgaran, Caverlee, and Alfifi 2017) and Fayazi et al. (Fayazi et al. 2015) studied *unverified reviews* [3] written by crowdworkers, and compared it with legitimate reviews. Fornaciari et al. (Fornaciari and Poesio 2014), generated a set of fake book reviews, and

[3] A review being unverified means that the reviewer did not purchase the product on Amazon.

use natural language processing and machine learning classifiers to detect them; their accuracy, however, is a little over 75%. Mukherjee et al. (Mukherjee et al. 2013) presented a model for detecting fake reviews using purely the n-grams extracted from the review text, and obtained an accuracy of 68% which is low, but still indicates that the n-grams convey some signals. Barbado et al. (Barbado, Araque, and Iglesias 2019) scraped user-centric features such as user activity across the platform and other behavior for fake review detection. Similarly, Wang et al. (Wang et al. 2013) developed a detection approach that groups similar user clickstreams into behavioral clusters to identify groups of fake users, an approach that can potentially be used for detecting review fraud via sybil accounts.

Incentivized Reviews

Incentivized reviews have emerged as a popular means to promote a product and create a *buzz* around it. However, typical incentivized reviews (such as sponsored posts) carry a disclosure of sponsorship. Underground review services provide sellers a distinct advantage over the official programs: reviews are guaranteed within a short time-span. Most importantly, they are guaranteed to be five-star. Prior work (Oak and Shafiq 2024) has discovered that incentivized review fraud is a systematic and purposeful manipulation, involving multiple entities (buyers, sellers and agents) and spread across several countries. Based on their findings, the process works as follows. The seller or manufacturer of a product first contacts potential reviewers directly or through a third-party service (typically, via middlemen on social media platforms like Facebook). The reviewer is offered an incentive (generally, cash equal to the value of the product) if they write a five-star review. Once the review is posted and verified, the reviewer receives the promised incentive. Because the incentive is paid out to the reviewer via a third party like PayPal or Zelle, the e-commerce platform has no visibility into it and can therefore not tie the incentivized refund back to a review. While a prior study (He et al. 2022) presents an innovative network-structure approach to detect sellers who buy fake reviews on Amazon, it relies heavily on network features derived from common reviewers among products to identify clusters of fraudulent activity. This approach primarily targets larger-scale operations (i.e. of enough scale to influence network connections) and may miss smaller or more discreet manipulators.

Novelty & Motivation

Our work is distinct from prior work in a number of aspects. First, unlike prior work on fake reviews, the reviewers we consider are not bots (Jindal and Liu 2007) or crowdworkers (Fayazi et al. 2015; Rahman et al. 2019). As a result, their activity does not exhibit typical 'fake' review characteristics like burstiness, IP clustering and high inter-review similarity. Reviewers involved in incentivized review fraud are real amazon customers who buy regularly from amazon, and a part of their purchases is genuine and non-incentivized as well, making it difficult to isolate fraudulent purchase and review activity. Second, while prior work (Oak and Shafiq 2024) has explored the incentivized reviews ecosystem (including trends in review growth, operational characteristics and evasion strategies), it lacks analysis that can illustrate salient characteristics exhibited by products involved in incentivized review fraud. Additionally, while public review fraud detection tools exist, little is known about their accuracy, or the degree to which they produce false positives or negatives. Through our work, we aim to bridge this gap by collecting a new dataset, identifying statistically significant discriminatory patterns between products with and without incentivized reviews, and audit two public tools for incentivized review detection.

Data Collection

Our data collection approach involved two stages: a data curation phase where we identified the list of products to study (along with a corresponding control group), and a web crawler that extracted product and review data from Amazon. We used Amazon as the e-commerce platform to study because of the availability of data, as well as it's popularity. Below, we first explain how we curated our lists, followed by our crawling setup. The end-to-end flow can be seen in Figure 2.

Curating Incentivized Products

Review services operate through social media groups, and involve review agents posting about their products. We identified such groups on Facebook, and leveraged publicly posted information to identify products that were seeking incentivized reviews. We used search terms like *"Free Amazon Products"*, *"free products for reviews"* and their variants to obtain a list of such groups. After ordering the groups by the number of members, we found that the top ranking group had nearly $900k$ members, and was highly active (more than 100 posts a day). In order to advertise their products, agents frequently posted in this group with links to products, or to spreadsheets containing product details. We crawled group posts for a one-week period, and extracted all posts or comments with URLs. Then, we checked whether these URLs resolved to either of the domains we were interested in (Amazon or Google Sheets), and discarded those that did not. The Amazon URLs allowed us to directly obtain product details. For the Google Sheets URLs, we downloaded the sheets, screened them to identify columns referring to product links, and then extracted those URLs. In this way, we were able to collect 1600 product links. Because this list of products was obtained directly through agents, we have high confidence that the reviews of these products have been manipulated through incentivized review services.

Curating Control Products

Third-party sellers, and not first-party, stand to benefit from incentivized reviews. In order to obtain a baseline for trends and features, we curate a list of control products to compare fraudulent products against. We hypothesize that first party products (those which are shipped and sold by Amazon.com as the seller, and of the Amazon Basics brand) will

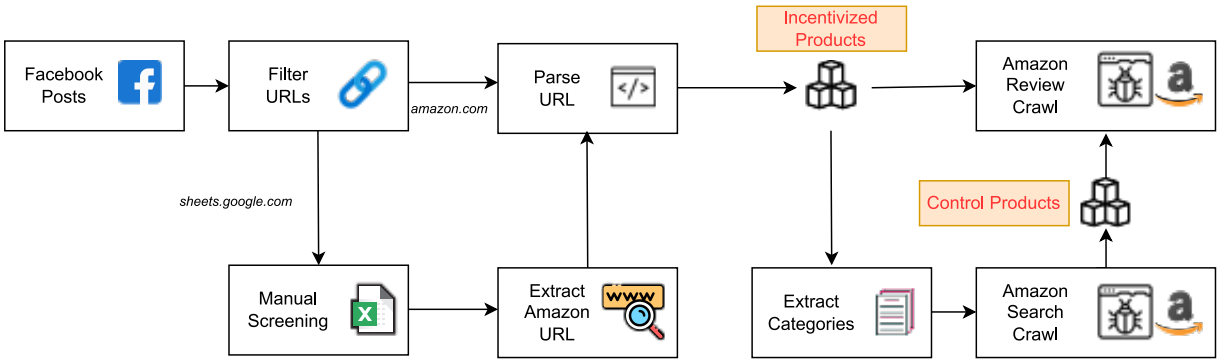


Figure 2: Data Extraction Workflow

be free from manipulation. Our reasoning for this is two-part. First, because of Amazon's prohibition of incentivized reviews, we argue that they will not violate their own guidelines on review integrity. Second, as Amazon itself controls product rankings and appearance in search results, they have no reason to seek incentivized reviews. To identify a control set of products, we first clustered the incentivized products based on their categories (electronics, kitchen, supplements, etc). We then queried Amazon for products belonging to these categories, and chose products proportionately based on category prevalence in the incentivized group. We further verified their legitimacy by confirming that none appeared in review groups during our observation period. Our control group also consisted of 1600 products; because products were broadly similar to those in the incentivized group, this allowed for an apples to apples comparison.

Data Crawling

We developed a custom data crawler to systematically crawl product and review information from Amazon, using only publicly available information. Our crawler operated both for incentivized and control products. The data curation step had already resulted in a list of URLs with their associated fraud labels. Our crawler operated on these URLs. It then visited the product page and systematically crawled product metadata (e.g., product title, description, the day it was added first on Amazon, price, seller, variants) for each product. All of this information is available on the product homepage. Additionally, it crawled reviews for each product including the review text, posted date, additional media attachments, and helpful votes. The product homepage typically shows only the top 20 or so reviews, ordered by recency. To overcome this, we augmented our crawler to implement a click action that would take it to the next page, and continue crawling till reviews were over. Overall, we obtained a list of 3200 products and around $400k$ reviews. Complete dataset statistics can be seen in Table 1Because of the manner in which the products were selected in each category (incentivized and control), we have high confidence that our labels are correct, as they are based on live data from review marketplaces and not based on any heuristic (which may lead to false positives). We will make this gold-standard dataset available for research after the publication of this paper.

Data	Incentivized	Control	Overall
#Products	1600	1600	3200
#Reviews	160,235	248,512	408,747

Table 1: Statistics of Collected Dataset

Feature Extraction

Prior work has employed features such as review burstiness and inter-review similarity to detect fake reviews (Mukherjee et al. 2013)(Jindal and Liu 2007); however, it is unclear whether these incentivized reviews might exhibit similar characteristics. Since the reviews for incentivized products were not organic, they might still show characteristics that differ from the control products. Motivated by this, we construct a set of features both at the product and review level. Features at the review level are further aggregated at the product-level, and therefore become product features.

Rating Features For every product, we extracted the rating displayed on Amazon. This rating is computed using a custom weighting algorithm internal to Amazon, and is not simply an average across all ratings. We also computed simple average over all reviews, as well as the standard deviation in rating across all reviews. Additionally, we consider two more features: the ratio of one-star to five-star reviews, and the ratio of 'bad' (1/2/3-star) to 'good' (4/5-star) reviews. This gives us 5 features at the product level.

Text Features For every review, we computed the review length in terms of characters, words and sentences. This was motivated by the fact that incentivized reviewers put less effort into reviews, and therefore will write reviews that are shorter in length and with fewer words. We also extracted the number of punctuation characters in each review. We hypothesize that the distribution of these features would also differ across the two groups due to review manipulation. Therefore, we considered both the mean and standard deviation in each feature (Sun, Du, and Tian 2016) this leads us to a total of 8 features at the product level.

Semantic Features For every review, we computed the polarity of the text as well as subjectivity of the text, using off-the-shelf text preprocessing modules. Our hypothesis was that since reviews are incentivized, they would be less neutral than organic reviews, and strongly positive. Similar to the above, we considered both the mean and standard deviation in each feature, leading to 4 features.

Content Features For every review, we extract three additional features: the number of photos attached, the number of videos attached, and the number of helpful upvotes that the review received. Because the reviews were incentivized these attributes were subject to manipulation too. Considering both mean and standard deviation, this leads to 6 features.

Five-Star Features Because review manipulation involves adding more five-star positive reviews, we expect to see more significant differences in these features if computed over only five-star reviews for each product (Mukherjee et al. 2013). This gives us 18 additional features (all features replicated except the rating features, which cannot be computed over just 5-star reviews).

Characterizing Products with Incentivized Reviews

We used hypothesis testing in order to examine whether our extracted features showed statistically significant differences between the incentivized and control groups. Our null hypothesis (H_0) for every feature was that the mean value of the feature would be the same in both the incentivized and control groups. The alternative hypothesis (H_a) was that the means would be different. For every feature, we calculated the mean in both groups of products and then compared the means using an unpaired two-tailed *t*-test (Hastie et al. 2009). Below, we present our results and interpret our findings qualitatively.

Product Ratings

Overall Rating We find that products with incentivized reviews, in general, have a higher average review score as compared to those that do not (4.52 vs 4.1). We also discover that there is less variance in ratings in the incentivized product group; as most of the reviews are incentivized, they tend to be rated five-star. Indeed, we found that nearly 23% products in the incentivized group had 0 standard deviation in ratings, as opposed to less than 3% in the control group. On average, the standard deviation in the control group was 49% higher than the incentivized group (1.27 vs 0.85); this difference was found to be statistically significant ($p < 0.01$).

Positive vs Negative Ratings We also examine the distribution of positive (4 or 5-star) versus negative (1, 2 or 3-star) reviews per product. We find that the positive reviews are nearly 5 times the negative ones in the incentivized group, but only 2.5 times in the control group. We further look at the ratio of *extreme positive* (5-star) to *extreme negative* (1-star) reviews. Five-star reviews are 8.3 times more than the one-star reviews in the incentivized group, but only 4 times more in the control group.

Review Content

Review Length We compare review lengths between the two groups in terms of the number of characters, words and sentences. We find that, on average, products in the incentivized group have an average review length of 224 characters, 42 words, and 4 sentences. On the other hand, the products in the control group have an average review length of 303 characters (35.27% more), 57 words(35.71% more) and 5 sentences(25% more). We also find that the standard deviation in the incentivized group is 192 characters, 36 words, and 2.61 sentences, and that in the control group is 285 characters (48.44% more) , 53 words(26.19% more) and 3.8 sentences (45.59% more). These differences were found to be statistically significant ($p < 0.01$). We theorize that incentivized products exhibit shorter reviews on average as a result of users not being able to come up with content, as the review does not reflect an actual user experience.

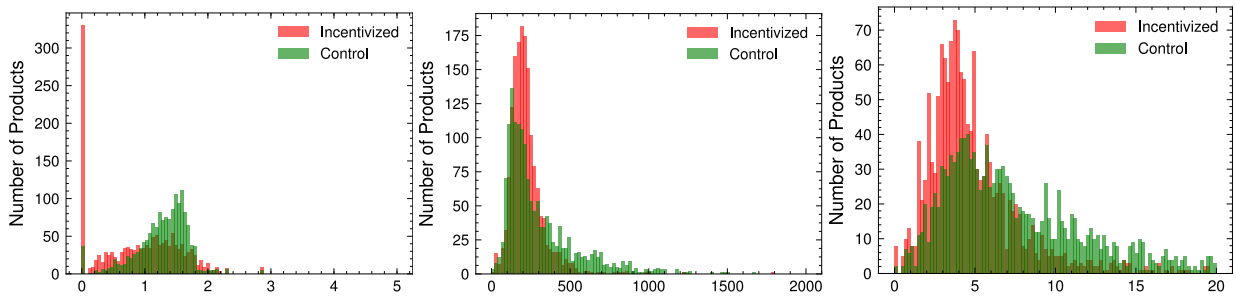
Punctuation We find that for products in the control group, the average review has 8.18 punctuation characters, while the average review in the incentivized group has 5.49. The standard deviation in the number of punctuation are 8.76 and 5.71 in the control and incentivized group, respectively. If we consider only five-star reviews in both groups, we find that the deviation in number of punctuation characters in the control group (8.03) is almost double than the incentivized group (4.49).

Media Attachments We compared the number of photos and videos attached per review between the two groups. We find that for products in the control group, the average review has 0.8 images and 0.3 videos. On the other hand, for products in the incentivized group, the average review has 2.8 images and 1.9 videos. We also find that there is at least one image attached to nearly every five-star review for all incentivized products, while it is not as prevalent in the control products. We qualitatively examined a few randomly sampled products with a high number of images and videos per review, and found that most of them have media attachments when they are not needed: such as video clips showing power banks or nutritional supplements. We hypothesize that media attachments indicate actual use of the product which indicate organic user engagement. Fraudulent reviewers may therefore be overusing them so that their reviews do not get flagged.

Review Semantics

Polarity We find that in general, reviews for incentivized products have higher polarity than those for control products. This trend holds even if we consider only the five-star reviews. This is expected, as the incentivized reviews will be strongly and overwhelmingly positive, while genuine reviews may acknowledge certain shortcomings, even while being positive overall. While the difference in polarity (0.25 in control vs 0.31 in incentivized) is statistically significant, it is not of practical significance because of small magnitude.

Subjectivity We find that subjectivity is marginally higher in the incentivized review group(average of 0.56) compared to the control group (average of 0.54). Similar to polarity,



(a) Std. Deviation in Product Rating (b) Mean Review Length (Chars) (c) Std. Deviation in Punctuation

Figure 3: Distribution histogram plots for some features we examined. Values reported are calculated across all reviews per product.

the difference is statistically significant, but not practically significant enough to be used as a discriminator.

Review Trends

In order to observe the trend in reviews across the two groups, we collected reviews data for all products over a six-week period. We recorded the number of reviews added and removed every week. We describe below our findings from the longitudinal analysis.

Review Growth We observe that incentivized products show abnormal growth in terms of number of reviews added every week. For every product, we compute the review growth based on the number of reviews added as a fraction of the reviews in the first week. We find that incentivized products, on average, grow their reviews by nearly 12% every week; this is in stark contrast to control products which have a growth rate of less than 1%. The maximum growth shown by a product in the incentivized group is 89%, and that in the control group is 8%. Additionally, weekly review trends in the control group are fairly stable with little variance (i.e. the fraction of reviews added remains more or less the same each week). However, incentivized products exhibit *spiky behavior* – we define a spike as a week in which there is a sudden growth in the number of reviews added (at least 10%). We found 43 incentivized products that exhibited at least one spike (an example can be seen in Fig 4). In contrast, no such products were found in the control group. We theorize that sudden spikes are a symptom of an incentivized review campaign running for the product. Sellers or brands interested in boosting their rankings through incentivized reviews will naturally want to do so as quickly as possible, thus resulting in the spiky behavior.

Review Removal While analyzing review growth, we also observed that some products had their reviews *removed*. Our hypothesis is that this is an action taken by Amazon in order to eliminate suspicious reviews from the platform. Out of the 1500 incentivized products we were studying, 235 had at least one review removed. The average product had 8.2 reviews removed, and one product had 1144 reviews removed. In contrast, control products show little to no review removal. Only 21 products had at least one review removed, and the most number of reviews removed for a product were

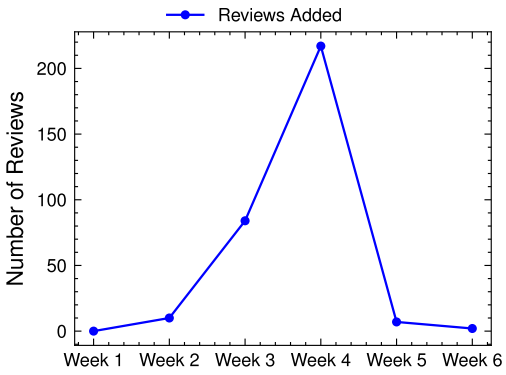


Figure 4: Week-over-Week Review Growth for a product showing spiky behavior

18. We theorize that these removals are a result of one of two things: *(i)* false negatives in fraud detection by Amazon, or *(ii)* collateral damage as a result of the review authors' reviews being removed for suspicion of fraud.

Discussion

We find that several features show statistically significant differences between the incentivized and control group, indicating that they may be useful signals for detecting review fraud using statistical or machine learning methods. Many of these, such as the length and media attachments are supported by empirical findings from prior work (Oak and Shafiq 2024; He, Hollenbeck, and Proserpio 2023). Overall, we see that incentivized products show lower variance in their reviews in terms of the length, rating distributions, indicating that most of the reviews are similar with respect to those features. This indicates that these reviews lack the natural variance that would occur as a result of organic customer purchases. Lower variance indicates that these reviews are curated or influenced in some way. Variance of various parameters across reviews for a product may be indicative of fraud, and a useful signal to leverage for fraud detection. We also find that incentivized products often show abnormal growth in the number of reviews, as well as sudden spikes in reviews, as alluded to empirically in

Model	ACC	PREC	REC	F-1
Logistic Regression	75.00	73.97	76.48	75.20
Decision Tree	76.78	77.10	75.63	76.36
Random Forest	82.14	81.06	83.47	82.25
K-Neighbors	77.20	80.87	70.76	75.48
Naive-Bayes	66.07	59.05	89.83	71.26
MLP	79.72	78.18	81.99	80.04

Table 2: Classification Results at Product Level. ACC, PREC, REC and F-1 refer to Accuracy, Precision, Recall and F-1 score respectively. Values shown are percentages.

prior work (Oak and Shafiq 2024). Control products, exhibit a stable and steady growth in reviews, which is expected because of the products being sold and reviewed organically. Examining for sudden unexplained growth may help identify incentivized review campaigns. These findings offer actionable insights for e-commerce platforms. First, platforms can integrate the identified features (e.g., review growth spikes, media overuse) into existing fraud detection pipelines. While our study focuses on Amazon, the characteristics of incentivized reviews (e.g., shorter length, media inflation, rating homogeneity) are platform-agnostic and likely applicable to other e-commerce ecosystems like Walmart or eBay; sellers on these platforms similarly exploit social media groups to solicit reviews (Oak and Shafiq 2024).

Detecting Review Fraud

In this section, we examine the performance of machine learning models in detecting incentivized reviews. We use six popular, widely used machine learning algorithms for classification: logistic regression, decision trees, random forest, support vector machines (SVM), naive-bayes classifier and a simple feed-forward neural network (multi-layer perceptron or MLP). We first attempt classification at the product and review level individually, and then combine them into a hybrid two-phase strategy.

Incentivized Product Detection

Using the aforementioned statistically significant features, we first trained machine learning classifiers (Fontanarava, Pasi, and Viviani 2017) to detect incentivized products. We reserve 33% of the data for testing, and train on the remaining 67%. We sample randomly but uniformly from both classes to derive our training and testing data. We also verified our results using 5-fold cross-validation. From the results shown in Table 2, we see that the random forest (Ho 1995) is our best performing model with an accuracy of 82.14%. This equates to an error of nearly 20% which we contend is unacceptable for deployment in production systems. By observing the receiver operating characteristic (ROC) (Figure 5) for the random forest, we see that at higher thresholds we can force the model to operate at an extremely high precision. On setting a threshold of 0.9, we see that we obtain a precision of 95.65% on the test set; there are only 7 false positives. However, the recall is low at 32.67% and 318 false negatives.

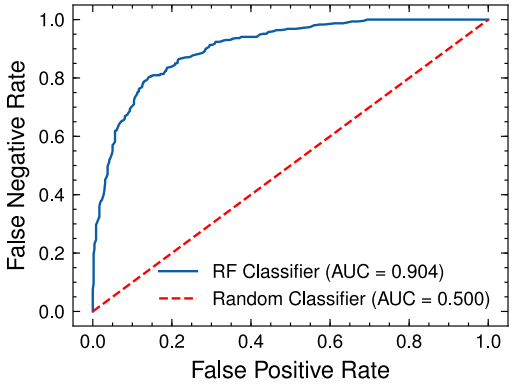


Figure 5: Receiver Operating Characteristic (ROC) curve for Random Forest.

Model	ACC	PREC	REC	F-1
Logistic Regression	83.06	58.59	82.48	68.51
Decision Tree	82.58	62.57	54.90	58.48
Random Forest	86.61	87.35	46.84	60.98
K-Neighbors	78.89	81.97	7.07	13.02
Naive-Bayes	87.83	82.69	57.58	67.89
MLP	87.18	75.20	63.57	68.90

Table 3: Classification Results at Review Level. ACC, PREC, REC and F-1 refer to Accuracy, Precision, Recall and F-1 score respectively. Values shown are percentages.

Review Level Classification

Next, we investigate classification at the review level. We have high-confidence ground-truth about which products have incentivized reviews. If a seller were trying to obtain incentivized reviews, they would want naturally want the reviews to be 5-star. We consider all 5-star reviews of these products to be incentivized. The remaining reviews (i.e. the 1-4 star reviews for incentivized products, and all reviews for control products) are considered to be genuine. Note that we do not have ground truth at the *product-level* and not at the *review-level* (that is, we know which products have incentivized reviews, but we do not know which reviews for each product are incentivized). Therefore, our goal for every review is to predict whether it belongs to a product with review manipulation or not, based only on the review text. Similar to our product-level classifiers, we reserve 33% data for testing and 67% for training. We sample uniformly at random and ensure that there is no product overlap between training and test data. We use a pre-trained BERT (Devlin et al. 2018) model to extract contextual embeddings for each review and use the embeddings as features to off-the-shelf classifiers (Fontanarava, Pasi, and Viviani 2017). As seen from Table 3, the logistic regression is our best performing classifier with an accuracy of just slightly more than 83.06%; however, as discussed above, this error would amount to significant amount of noise and may not be acceptable at the production level.

Precision-Recall Tradeoff

We further study the tradeoff between precision and recall to see if there exists a viable operating point in the receiver operating curve. In the scenario when an e-commerce company like Amazon detects incentivized reviews, they typically do one of two things as mitigation measures: *(i)* remove the product listing from the website, or *(ii)* ban flagged user accounts from contributing to reviews. Note that because these fraudulent reviews are written by real customers, banning user accounts would result in suppressing user opinions and may lead to user churn. Similarly, by de-listing products, the company is also losing revenue and organic impressions from genuine buyers Therefore, any detection of incentivized reviews must operate at a *high precision*; that is, when a product is flagged, the chances of it being a false positive should be extremely low. This will, of course, come at the cost of low recall, which means that there would be some products which have incentivized reviews and would go undetected.

Addressing False Negatives

Our aim is now to use the existing classifier (tuned at a high threshold for high precision) and then apply some methods to reduce the false negatives that the high threshold will introduce. We experimented a variety of methods to correct for the false negatives, as described below.

Label Propagation First, we attempted to correct the false negatives by leveraging labels from similar products. For every product, we obtained the labels of the nearest K neighboring products in the feature space using cosine similarity. We then conducted a majority voting among those labels, and computed the percentage of votes for each class. For the products labeled as non-fraudulent by the base classifier, we switched the labels if at least 60% of the neighbors were marked as fraudulent. We experimented with varying number of neighbors, with values from $K = 3$ to $K = 11$.

Feature Propagation In this approach, we attempted to correct the false negatives by leveraging features from similar products. For every product, we obtained the features of the nearest K neighboring products in the feature space. We then aggregated those features to compute a product vector that incorporates neighboring features along with it's own. This was done by averaging the vectors weighted by the cosine similarity. For the products labeled as non-fraudulent by the base classifier, we then re-scored them using the base classifier and the new feature vector.

Anomaly Detection In this approach, we attempted to correct the false negatives by using one-shot learners trained on the control products to identify anomalies. We trained an unsupervised model (we experimented with Local Outlier Factor (Cheng, Zou, and Dong 2019) and Isolation Forest) on the control products. We then ran inference on this model for the products labeled as non-fraudulent by the base classifier. Products identified as outliers by the anomaly detection model were re-labeled as fraud.

Model	ACC	PREC	REC	F-1
RF @ Thresh=0.5	82.14	81.06	83.47	82.25
RF @ Thresh=0.9	65.86	95.65	32.67	48.65
Two-Stage	88.76	93.54	84.11	88.57

Table 4: Performance Gains with Two-Stage Classification. ACC, PREC, REC and F-1 refer to Accuracy, Precision, Recall and F-1 score respectively. Values shown are percentages.

Results We were not able to obtain meaningful improvements with any of the methods discussed above. Label propagation was not effective, as due to high precision labeling, very few products were newly labeled as fraudulent, resulting in minimal or no elimination of false negatives. Feature propagation resulted in adding several false positives, perhaps due to feature distortion caused by averaging. Anomaly detection did not work, because the dataset was well-balanced, while most anomaly detection methods are designed to detect rare anomalies ($< 2\%$).

Correcting with Reviews

Finally, we investigated whether classification at the review level could help correct false negatives. We use the product classifier at high precision to eliminate false positives, and the review classifier at high recall to correct for the false negatives that the former introduced. First, we trained the Random Forest model on the products and tuned it to a high precision (low, negligible false positives). Then, we trained a logistic regression (Cox 1958) classifier at the review level and tuned it for high recall (low, negligible false negatives).

We first score every product using the base classifier. Because we tuned for high precision, we have high confidence that the products labeled as fraudulent by the classifier are indeed fraudulent, and so we declare that to be the final label for that product. For every product labeled as non-fraudulent by the base (product-level random forest) classifier, we obtain all of it's reviews and score each review using the logistic regression classifier. If a majority voting on the reviews shows that they are incentivized, then we flip the label and declare the product as incentivized. Otherwise, we retain the label assigned by the base classifier.

We see that this approach is highly effective, as evidenced by the metrics in Table 4. The base classifier (at a very high threshold, for high precision) has a precision of 95.65%, but with a very poor recall of 32.67%. However, after correction by our two-phase model, there is large increase in recall (51.44%) at the cost of very small drop in recall (2.9%).

The difference is more obvious when we look at raw numbers as shown in Table 5. The base classifier at the default threshold (0.5) has only 393 true positives, and 93 false positives. Bumping the threshold to a high value of 0.9 drastically reduces the false positives to just 7; but in doing so, it also reduces the true positives by more than half, and raises the false negatives to 318. In comparison, our proposed approach is able to reduce both the false positives and negatives as compared to the base classifier.

Model	#TP	#TN	#FP	#FN
RF @ Thresh=0.5	393	387	93	79
RF @ Thresh=0.9	154	473	7	318
Two-Stage	425	459	21	47

Table 5: Classification Metrics with Two-Stage Classification. #TP, #FP, #TN and #FN represent the true positives, false positives, true negatives and false negatives respectively.

Discussion

Our experiments show that while their exist statistically significant differences between most of the features we computed, the differences are not discriminative enough to build high-precision machine learning classifiers. Low precision is not acceptable at the scale of e-commerce platforms, as it will lead to monetary loss and user churn. We find that our two-step approach (classifying products at high precision, followed by a second high-recall classifier based on individual reviews) is able to optimize jointly for precision and recall, and is able to eliminate false negatives with a minimum increase in false positives. We are able to achieve an F-1 score superior to that of prior work (He et al. 2022) using a lightweight approach that does not rely on network features, which incur heavy computation cost, as well as require frequent recomputation due to the dynamically evolving nature of reviews. It is important to note that our classifiers were based on features derived from publicly available data. In practice, e-commerce platforms like Amazon will have visibility into several other features (such as user journey and historical activity, long-term trends in products) which are likely to improve the performance of machine learning models and result in even higher precision and recall.

Auditing Fake Review Detection Services

There are a number of services that are specifically geared towards detecting fake reviews on Amazon. Operating as web services, mobile apps, or browser extensions, these services score the authenticity of a product's reviews. It is unclear if these services are able to reliably detect incentivized reviews. To bridge this gap, we audit two such services: Fakespot and ReviewMeta. For every product in our dataset, we obtain the review grade for the product by Fakespot and ReviewMeta. Comparing the review grades with our ground truth labels, we are able to evaluate the precision and recall of these services. Note that the total number of products analyzed is less than 1600, as some of the products and/or reviews were removed by Amazon as a part of their cleanup efforts.

Fakespot

Owned by Mozilla, Fakespot[4], is a tool that helps consumers verify review authenticity. According to their website, they use machine learning on a variety of features (product, reviews, seller and user) to produce review grades. Grades A and B mean that a product has reliable reviews, a grade C indicates a mixture of reliable and unreliable reviews, and grades D and F indicate that there are insufficient reliable reviews. A comparison between the grades assigned by FakeSpot and our ground truth labels is shown in Table 6. We see that FakeSpot suffers from both false positives and false negatives. For the products assigned grades A and B by FakeSpot (which indicate that reviews are reliable), 459 (62.48%) and 345 (34.8%) respectively were fraudulent. For the products assigned grades C and D (which indicate unreliable reviews), 265 (67.6%) and 188 (53.41%) were in fact, legitimate products. Overall, we conclude that FakeSpot is unreliable as it suffers from a significant number of false positives and negatives.

Fakespot Grade	Non-Fraud	Fraud	Total
A	275	459	**733**
B	647	345	**992**
C	265	127	**392**
D	188	164	**352**
F	54	84	**138**
N/A	27	250	**276**
Total	**1454**	**1429**	**2883**

Table 6: Fakespot Confusion Matrix

ReviewMeta

ReviewMeta[5] is a website and browser extension that specializes in analyzing product reviews on Amazon. ReviewMeta's algorithm evaluates various factors in Amazon reviews, such as the language used, the timing and frequency of reviews, and the behavior of reviewers. Based on this analysis, ReviewMeta provides users with either a $PASS$, $WARN$ or $FAIL$ rating for a given product. A comparison between the grades assigned by ReviewMeta and our ground truth labels is shown in Table 7. Looking at simply the $PASS$ and $FAIL$ grades, we see that ReviewMeta has an accuracy of around 67%. However, there are a number of false positives and negatives.a significant number of fraudulent products (664) are labeled as non-fraudulent, leading to several false negatives and low recall ($< 21\%$). Additionally, owing to the false positives (166), the precision is also fairly low (around 51%). We also observe that 289 benign products are assigned the $WARN$ label. Reasonably assuming that a user who sees a $WARN$ rating will not trust the product; this thus adds to the false positives even more. Overall, we conclude that ReviewMeta is unreliable as it also suffers from a significant number of false positives and negatives.

[4]https://www.fakespot.com/

[5]http://www.reviewmeta.com/

ReviewMeta Grade	Non-Fraud	Fraud	Total
PASS	971	664	**1635**
FAIL	166	174	**340**
WARN	296	206	**502**
N/A	21	385	**406**
Total	**1454**	**1429**	**2883**

Table 7: ReviewMeta Confusion Matrix

Discussion

Based on our audit of the two fake review detection services, we conclude that they face significant challenges in assessing whether a product has incentivized reviews or not. Both the tools we examined mark several fraudulent products as reliable in terms of reviews. It is important to note that our findings should not be interpreted as criticism of these tools but rather as a demonstration of the inherent challenges posed by the unique characteristics of incentivized reviews. Our measurement represents a snapshot in time; the models used by these tools may learn and improve over time, and produce better results. Additionally, these measurements are over a small sample size, and do not reflect the overall accuracy, precision or recall of these tools. Fakespot and ReviewMeta remain invaluable resources for consumers navigating an increasingly complex review ecosystem. Their limitations in detecting incentivized reviews underscore the need for continuous innovation, not a failure of their current design. By framing these findings as a call for evolution rather than criticism, we emphasize the shared goal of combating fraud while respecting the practical constraints faced by third-party tools.

Conclusion

This paper explores the issue of incentivized review fraud on e-commerce platforms, an emerging issue with serious consequences for the integrity of online platforms as well as consumer safety. We compiled the first known dataset of products soliciting incentivized reviews, enabling detailed analysis of fraudulent practices. Our findings show that products seeking incentivized reviews exhibit certain peculiar characteristics, which are significantly different from products not involved in incentivized review fraud. However, machine learning methods are not able to detect such fraud at high precision without incurring significant loss in recall. Our novel two-phase technique, which combines high-precision product classifiers with high-recall review classifiers, offers a promising solution to this problem. This approach minimizes false positives significantly, which is crucial for maintaining user trust and avoiding undue penalization of legitimate sellers. However, it also results in a slight increase in false negatives, which is an acceptable trade-off in many operational contexts to ensure that only highly probable cases of fraud are acted upon. Our audit of existing tools like Fakespot and ReviewMeta revealed significant limitations in current detection technologies, further underscoring the challenging nature of this problem.

References

Barbado, R.; Araque, O.; and Iglesias, C. A. 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4): 1234–1244.

Chee Chew. 2016. Update on customer reviews, Amazon. https://www.aboutamazon.com/news/innovation-at-amazon/update-on-customer-reviewshttps://www.aboutamazon.com/news/innovation-at-amazon/update-on-customer-reviews.

Cheng, Z.; Zou, C.; and Dong, J. 2019. Outlier detection using isolation forest and local outlier factor. In *Proceedings of the conference on research in adaptive and convergent systems*, 161–168.

Commission, F. T. 2023. Trade Regulation Rule on the Use of Consumer Reviews and Testimonials. https://www.ftc.gov/system/files/ftc_gov/pdf/r311003consumerreviewsandtestimonials_nprm.pdf. Accessed: 2023-09-12.

Cox, D. R. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2): 215–232.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fayazi, A.; Lee, K.; Caverlee, J.; and Squicciarini, A. 2015. Uncovering crowdsourced manipulation of online reviews. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 233–242.

Fontanarava, J.; Pasi, G.; and Viviani, M. 2017. Feature analysis for fake review detection through supervised classification. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 658–666. IEEE.

Fornaciari, T.; and Poesio, M. 2014. Identifying fake Amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 279–287. Association for Computational Linguistics.

Guo, J.; Wang, X.; and Wu, Y. 2020. Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. *Journal of Retailing and Consumer services*, 52: 101891.

Hastie, T.; Tibshirani, R.; Friedman, J. H.; and Friedman, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

He, S.; Hollenbeck, B.; Overgoor, G.; Proserpio, D.; and Tosyali, A. 2022. Detecting fake-review buyers using network structure: Direct evidence from Amazon. *Proceedings of the National Academy of Sciences*, 119(47): e2211932119.

He, S.; Hollenbeck, B.; and Proserpio, D. 2023. Leveraging Social Media to Buy Fake Reviews. *Communications of the ACM*, 66(10): 98–105.

Ho, T. K. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278–282. IEEE.

Hu, N.; Liu, L.; and Sambamurthy, V. 2011. Fraud detection in online consumer reviews. *Decision Support Systems*, 50(3): 614–626.

Imbler, S. 2018. Can You Trust Amazon Vine Reviews? https://www.nytimes.com/wirecutter/blog/amazon-vine-reviews/https://www.nytimes.com/wirecutter/blog/amazon-vine-reviews/. [Accessed 2021-09-04].

Jindal, N.; and Liu, B. 2007. Analyzing and detecting review spam. In *Seventh IEEE international conference on data mining (ICDM 2007)*, 547–552. IEEE.

Kaghazgaran, P.; Caverlee, J.; and Alfifi, M. 2017. Behavioral analysis of review fraud: Linking malicious crowdsourcing to amazon and beyond. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 560–563.

Mayzlin, D.; Dover, Y.; and Chevalier, J. 2014. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8): 2421–55.

Mukherjee, A.; Venkataraman, V.; Liu, B.; Glance, N.; et al. 2013. Fake review detection: Classification and analysis of real and pseudo reviews. *UIC-CS-03-2013. Technical Report*.

Nicole Nguyen. 2018. Inside Amazon's Fake Review Economy. https://www.buzzfeednews.com/article/nicolenguyen/amazon-fake-review-problem. Accessed 2020-09-11.

Oak, R.; and Shafiq, Z. 2024. Understanding Underground Incentivized Review Services. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Patel, N. A.; and Patel, R. 2018. A Survey on Fake Review Detection using Machine Learning Techniques. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 1–6.

Rahman, M.; Hernandez, N.; Recabarren, R.; Ahmed, S. I.; and Carbunar, B. 2019. The art and craft of fraudulent app promotion in google play. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2437–2454.

Smith, A.; and Anderson, M. 2016. Online shopping and e-commerce. *Pew Research Center*, 19.

Sun, C.; Du, Q.; and Tian, G. 2016. Exploiting product related review features for fake review detection. *Mathematical Problems in Engineering*, 2016.

Wang, G.; Konolige, T.; Wilson, C.; Wang, X.; Zheng, H.; and Zhao, B. Y. 2013. You are how you click: Clickstream analysis for sybil detection. In *22nd USENIX Security Symposium (USENIX Security 13)*, 241–256.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Answer
   (g) Did you discuss any potential misuse of your work? No
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? No
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? No
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No. The data, if made available on the Internet may fall

into the hands of Amazon, who may take down the products or reviews from their site, which would result in us being unable to conduct new/follow-up experiments or analyses.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? answerYesYes

(b) Did you mention the license of the assets? No

(c) Did you include any new assets in the supplemental material or as a URL? No

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA, the data does not relate to individuals, but products on e-commerce platforms

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA, the data does not involve people and hence we believe that fairness is not an issue

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? No

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA