Towards Amazon Fake Reviewers Detection: The Effect of Bulk Users

Youssef Esseddiq Ouatiti
youssefesseddiq_ouatiti@um5.ac.ma
ENSIAS, Mohammed V University
Rabat, Morocco

Noureddine Kerzazi
n.kerzazi@um5s.net.ma
ENSIAS, Mohammed V University
Rabat, Morocco

ABSTRACT

Online marketplaces such as Amazon allow people to share their experiences about purchased products using textual comments known as product reviews. These reviews have become a common tool that users rely on to get insights on the quality and functionality of products and services from online consumers. However, like any other online information, reviewers raise serious questions concerning the credibility and reliability, since anyone can post reviews, which might impact the reliability of the information. This paper tackles the phenomenon of Bulk reviewers. We first analyze a large dataset of reviews from Amazon aiming to spot bulk reviewers according to their behavior. We then apply a what-if analysis to assess the effect of bulk reviews on the online marketplaces using a metric called Net Promoter Score to measure the willingness of users to recommend products. Our Results reveal that bulk users (i.e., users that review multiple times) have same distribution of ratings as non-bulk users indicating that a bulk reviewer is not automatically a fake reviewer. Yet, we discover that bulk users do inflate NPS metric and thus contribute to overestimate the level of customer satisfaction.

CCS CONCEPTS

• **Computing methodologies** → **Anomaly detection**.

KEYWORDS

Reviews; Net Promoter Score; Recommendations.

ACM Reference Format:
Youssef Esseddiq Ouatiti and Noureddine Kerzazi. 2020. Towards Amazon Fake Reviewers Detection: The Effect of Bulk Users. In *13th International Conference on Intelligent Systems: Theories and Applications (SITA'20), September 23–24, 2020, Rabat, Morocco.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3419604.3419800

1 INTRODUCTION

Online marketing has allowed people to share their experiences about their purchased products using textual comments known as products' reviews [7]. Online reviews provided by the previous consumers are key information source for both consumers and marketers [5, 17]. These reviews aim to help people making informed decisions before buying products. However, many online companies recruit either influential reviewers or fake ones to provide a non-truthful experience of inferior products [3].

Empirical evidence for the existence of bulk reviewers and fake reviews and associations between paid reviews and the opinion about online products could help marketers and users make more informed choices. Like a lot of Amazon customers, we often read reviews before deciding to buy products [5]. For many, it's a critical part of product research because it can help to identify potential issues that one can encounter while using that product. We continue to treat reviews as credible evidence that proves reliability and quality of products on Amazon. The question is: **should we do that?**

To answer this question, we need to spot fake reviews and find a true opinion of real users' experiences. Obtaining such evidence, however, is a challenging task. Considering a large number of fake five-star reviews, for example, we need to know where they come from. To boost the popularity of their products, sellers often hire people to act as "shoppers" and write positive reviews. For instance, according to Amazon's blog[1], there were about 100 of paid groups on Facebook divided by product categories and geographic region. One of these groups had more than 50,000 members engaged to make fraudulent reviews.

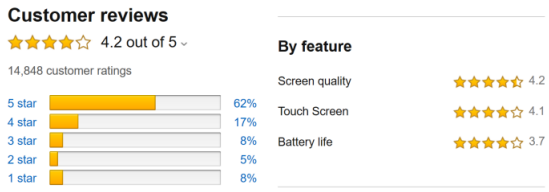

Figure 1: Feature Example of an Amazon's 'Request a Review'

Our goal is to identify bulk reviewers on Amazon marketplace and understand their role in order to develop effective countermeasures. More specifically, this paper will answer the following research questions:

- **RQ1. Are Amazon bulk reviewers necessary Fake ones?**

[1]https://sellercentral.amazon.com/forums/t/how-merchants-use-facebook-to-flood-amazon-with-fake-reviews/398434

SITA'20, September 23–24, 2020, Rabat, Morocco

ACM ISBN 978-1-4503-7733-1/20/09…$15.00
https://doi.org/10.1145/3419604.3419800

We start our analysis with a general characterization of reviewers' behavior. Through a statistical study, we try to spot 'bulk users' and compare their rating behavior to 'non-bulk users'. *RQ1* leads to a related question.

- **RQ2. How can we measure clients' global loyalty to a brand of product?**
  We calculate the NPS for Amazon and compare it with other major companies around the world. We then select a product within the products available on our dataset (e.g., Fire TV) and calculate its Net Promoter Score (NPS), and make a what-if-analysis to understand the effect of bulk users on this measure.

Paper organization: The remainder of this paper is organized as follows. Section II surveys related works on Amazon reviews analysis. Section III presents the applied methodology to answer our research questions along with threats to validity. Section IV highlights our findings. Finally, section V draws conclusions and enlightens future work.

2 MOTIVATION & RELATED WORK

In this section, we discuss prior research with regard to reviews/ratings analysis.

2.1 Motivation

Having easy access to the web has radically changed the way people shop for almost everything today [3, 5, 10, 11]. Amazon as a pioneer of e-commerce has used user online reviews/ratings in order to optimize customers' experience. Although, these reviews are posted by people nothing ensure their authenticity [7, 8]. That said, an illegal market for fake app reviews has emerged, where we provide[2] services to help vendors improve their ratings and ranking in app stores. Real users are supposed to report feedbacks on their satisfaction or dissatisfaction of consuming a product. Fake reviewers, however, get paid to submit reviews aiming to blunt sales. They might or might not be real consumers of the product.

This in fact led us to introduce a key concept of this work, which is 'bulk users'; defined as users that have rated a product several times. In addition, beyond plain summaries on the ratings (e.g, # 5 stars, # 4 stars) we aim to measure willingness of a user/client to recommend a given product. In other words, we want to measure the loyalty of that client to a given company. To do so, we build on a metric called Net Promoter Score (NPS) [7], which provides an indicator for assessing and managing client satisfaction.

2.2 Related Work

There are growing interests in understanding how online customers make purchase decisions [1, 2, 5, 14–17].

[2] An example of a Web site providing fake reviews: https://blog.reviews.io/thinking-of-buying-google-reviews-think-again

Fraud Reviews - Duang et al. [5] studied the word-of-mouth (WOM) generation on the Internet and how it influences consumers' purchase decisions at retail outlets. Authors applied their approach to the movie industry and found that both a movie's box office revenue and WOM valence significantly influence WOM volume. Ying et al. replicate the study on Amazon book review focusing on the impact of fraud reviews and found that fraudulent reviews have a significant impact on consumer purchasing decisions. Attributes such as more number, higher proportion, longer word count and higher promotion of rating characterize fraud reviews lead to higher sales. We leverage on these indicators among others to identify bulk users on Amazon. Ruiz et al. [14] have analyzed the mobile app store reviews and how such ratings influence a customer's decision to acquire a mobile App. Authors pointed out the fact that in contrast with products an app can be updated in a very short time period. Authors concluded that there is a need for a careful rethinking of the review and rating system.

Sentiment Analysis Facet - Other Amazon reviews analysis research focus on the Sentiment analysis facet [1–3, 5]. Tkachenko et al. [15] try to build a model that can detect comparison between products from online reviews. Kuppili et al. [10] have propose combining both text reviews (using sentiment analysis) and ratings (using clustering) to provide recommendations that are more accurate. Lima et al. discuss the impact of specificity in a review on its helpfulness for other users. Almjawel et al. [1] work consists on a statistical analysis on Amazon product ratings/reviews in order to produce dashboards visualizing the results of their review sentiment analysis. Yet, to the extent of our knowledge, no previous works have attempted to discuss the 'bulk users' concept, nor customer satisfaction in e-commerce using NPS metric.

Predicting the helpfulness of online reviews - Zhang et al. [17] studied semantic and structural features of reviews to predict the helpfulness of online reviews automatically. Authors used a machine learning approach on a dataset of reviews from Amazon and proposed a model with an accuracy of 68.7% on predicting and classifying the helpfulness of online reviews for products. They found that a reviewer's reputation is the major determinant in predicting the helpfulness of online reviews.

3 METHODOLOGY

Here, we describe the Dataset that we used, and the metrics and analysis methods we applied to answer our research questions. The goal of our work is as follows:

Goal We seek to address the need for more effective fraud reviews on online markets such as Amazon, with the ultimate goal to identify bulk users and mitigate their influences on consumers' purchase decisions.

3.1 Dataset Description

We use an online available dataset[3] of over 34,000 Amazon products' reviews. In this dataset, we identify 21 attributes from which we extract 7 useful features as presented in Table 1. All duplicated reviews are filtered out. This dataset has a large enough data ensuring an empirical study.

Features	Description
Id	Product indentifier
Name	Product name
Reviews.date	Review submission date
Reviews.do.Recommend	Recommended or not (0,1).
Reviews.rating	Rating of the product
Reviews.text	Text of the review
Reviews.username	Unique username of the reviewer

Table 1: Selected features description

3.2 Identifying Bulk Reviewers and Fake Reviews

There are a lot of factors that can dictate whether a review is genuine or not [4], among which we mention: length of the review (Short reviews are generally more suspicious), reviews coming from users that don't have previous reviewing history or those that rate a lot of products in a short time frame [9], and reviews from users who did not buy the product [4].

Based on the data available, we are interested in the "'bulk users"' as an explanation for fake reviews. Bulk reviewing is a concept manifested in users making several ratings, this in fact can be discussed in two setups: 1) within the same product; 2) and cross-products. For the first setup, a number of reviews greater than one is suspicious, although in the second setup we have to fix a threshold for the accepted number of reviews per user. Therefore, fake reviews (if they exist) would likely be the ones published by bulk users. To avoid any misconception, we will call reviews published by 'bulk user' and explore their genuineness late no.

3.3 NPS Metric Analysis

In the contexts of Amazon, we discuss the level of satisfaction and loyalty of users to the company/brand using a measure called: Net Promoter Score (NPS) [13]. NPS metric is an index ranging from [-100 to 100] that measures the willingness of customers to recommend a company's products or services to others. Specifically, NPS is used as an indicator for assessing the customer's overall satisfaction with a company's product or service. To calculate this measure, users are classified into three categories see Table 2 [6]: Detractors, Passives and Promoters.

[3]https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products

Class of clients	Description
Detractors (D)	Software analysis & Intelligence laboratory
Passives (Ps)	Users that are somewhat satisfied with the product. Yet, they can switch to another brand given better offer. (i.e. ratings=3).
Promoters (P)	Satisfied users that are loyal to the brand and likely to recommend it to others. (i.e. ratings >3)

Table 2: Client category

3.4 Threats to validity

Our approach to identifying bulk users is somehow soft even it is based on previous studies. This is a crucial thread to validity that will be investigated in future research.

Recent studies have investigated the accuracy of the NPS measure. Kristensen et al. [6, 9] attempted to demonstrate that the NPS is an inferior measure compared to standard measures of loyalty. We are aware that such a suspicious claim might affect our analysis. Due to time constraints, we postpone this investigation for further analysis.

4 RESULTS

In the following, we present our findings with respect to our two stated research questions.

RQ1. Are Amazon bulk reviewers necessary Fake ones?

Motivation - Basically, we are interested to answer the question: Is a bulk reviewer necessarily a fake reviewer? There are growing interests in understanding how fake reviews are generated and how it influences consumers' purchase decisions at Amazon [1, 2, 5]. The review mechanism, generated by online consumers, aims to help people making informed decisions before buying a given product [7, 8, 10]. However, many online companies recruit either influential reviewers or fake ones to boost their inferior products. These recruits generally don't settle for one occasional review of a certain product, instead they are very active and provide many reviews (i.e., bulk reviewers). In the next section, we share our findings related to the Amazon review activity carried out by bulk users.

Approach - In order to explain fake reviews, we start by investigating and spotting 'bulk users', which are users that reviewed multiple times a product or a group of products. We are then interested in the distribution of ratings for this kind of users and the volume of their activity and also uncover specific characteristics of their reviews, which can help build a model to predict fake reviewers.

Results - We start by exploring the results of our preliminary statistical analysis. Our analysis in the cross-product context shows that 69% of online customers review and rate only once. Over 83% never rate/review more than three times

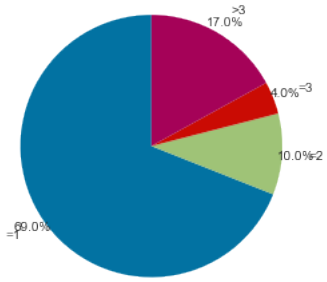

Figure 2: Number of reviews per user (cross product setup)

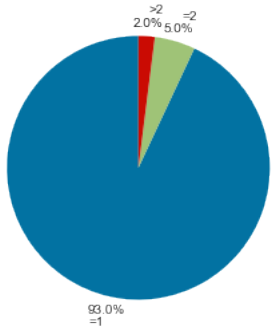

Figure 3: Number of reviews per user (within product setup)

which we consider to be the standard behavior, as depicted in Figure 2. This pattern is even more present within product context as more than 93% of the users never rate more than once as presented in Figure 3.

We found a proportion of 0.54% of bulk users in the cross-product reviews and 7.7% in the within-product reviews. It's worth noticing that bulk users in the within product context are more present than in the context of cross-product. Now, the number of reviews made by bulk reviewers for the cross-product setup reaches 9.11% (3160 out of 34660 reviews). This is somehow a small portion in general, but it is only provided by less than 0.55% of the users in our dataset. We hypothesis that this community of users that generate bulk reviews are very active. This gives us strong clues for our hypothesis of these reviews being fake ones (or at least suspicious). A minority generating this great deal of reviews is unlikely to be random.

In the context of cross-product setup, we found that the number of products in which bulk users are involved ranges from 1 to 13 as shown in Figure 4. With a normal distribution centered around 6-7 products per Bulk user as depicted in Figure 5.

We also aim to explore the distribution of the 'bulk user' vs. 'non-bulk users' seeking to confirm our suspiciousness around the genuineness of bulk user reviews. Specifically,

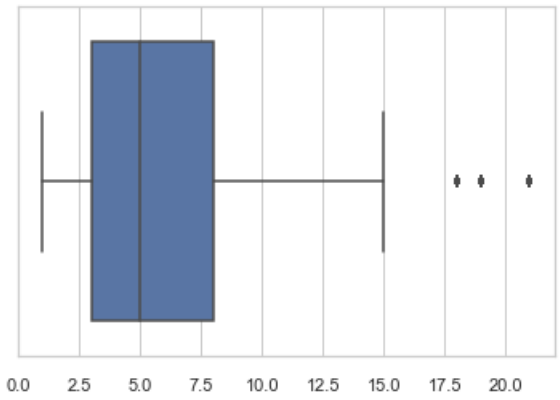

Figure 4: Product count per bulk user.

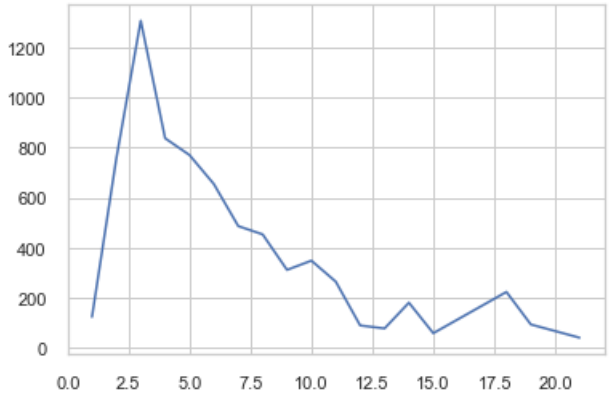

Figure 5: Product count per bulk user distribution.

we seek to know how does bulk users rate products? We hypothesis that fake reviews have a skewed distribution either towards high ratings (paid reviews or gifted products in exchange for stunning reviews) or towards low ratings (Intent to damage the reputation of a product). We found that 93.3% of the bulk users rate highly (4-5) in which over 68% of them vote a full 'five stars' rating. Meanwhile, only 2.4% of the ratings are low which suggests that bulk users in this dataset do not try to damage public opinion on a given product, or just that they do not express their unhappiness with products as often as their satisfaction with them see Figure 6. If we compare this distribution of bulk users to the rating distribution of non-Bulk users. Surprisingly, we found that the two distributions show a quite similar shape, which discarded our assumptions. In other words, the number of reviews of a user is not enough to unveil that this user is a fake reviewer. Further investigation is required in order to identify fake reviewers according to their activity.

In order to further investigate the impact of bulk users on predicting fake reviews, we leveraged Martens et al. work [4]

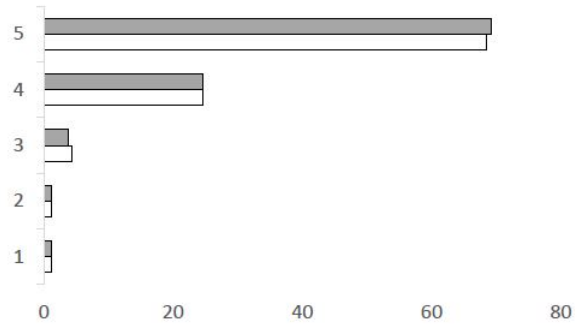

Figure 6: Bulk Vs. Non-bulk Users Rating Distributions

models. The authors used features based on the product, the reviewer and the review itself, to build multiple classification models for App store applications reviews genuineness. We calculate the same metrics see Table 3 and apply the models to our dataset. We found that only 8% of the reviews made by bulk reviewers are classified as fake based on Martens et al. models [4]. Yet, Martens's models consider the 'number of reviews' made by a reviewer as one of the most important feature in predicting fake reviews, this further confirms that high volume of reviewing does not necessary imply fake behavior.

The most important feature for detecting fake log levels according to Martens et al. models is the number of reviews for the product being reviewed, we found that the median number of reviews per product for the fake bulk users is larger than the median for non-fake bulk users see Figure 7, which we explain by the fact that very active reviewers are not automatically fake reviewers but when their reviews become targeted towards more popular products they are more likely to be fake.

Features	Description
ReviewerReviews	Number of reviews made by the reviewer
U%Ri (i=1,2,..5)	Percentage of i stars reviews
Frequency of rating	Number of reviews per day.
ProductReviews	Number of reviews for the product.
P%Ri (i=1,2,..,5)	Percentage of i star reviews
Length of review	Text of the review

Table 3: Features used for clustering the reviews.

RQ2. How can we measure clients' global loyalty to a brand of product?

Motivation - One or two fake reviews have not a big impact. However, lots of them end up with an artificially inflated product rating [12, 16]. Online customers could encounter a situation where every single review was a fake one [12]. Consequently, there is a need to discuss the sensitivity of fraud reviews and the effect of them on consumer behavior

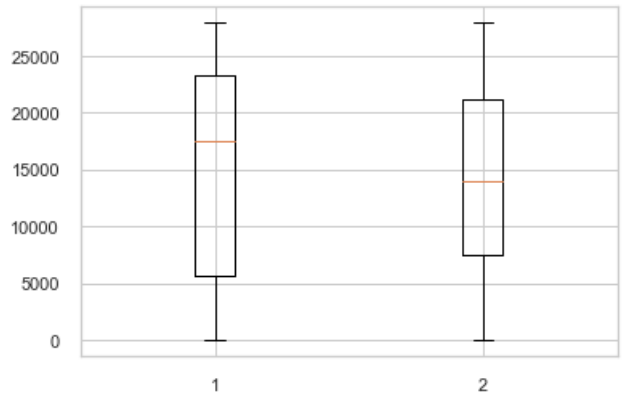

Figure 7: Product number of reviews for bulk fake users Vs. Bulk non-fake reviews.

through empirical research [12, 14].

Approach - We discuss customers' satisfaction through NPS metric and investigate the effect of bulk users on the NPS. We apply 'What-if' analysis to Fire TV products, showing that removing Bulk users significantly influences the NPS which means that these users cause overestimating the customers' satisfaction level for a given product.

Results - First, we discuss the NPS score of Amazon in comparison to other major companies and drive some insights from it. We start by calculating the NPS metric for Amazon and it turns out that its value is 61.99. High value, which means that Amazon customers are quite satisfied with the products and are likely to recommend products to others. Figure 8 shows a comparative study (3) of NPS between major companies.

The results show that apart from Starbucks company, the leading companies are tech companies, this might be due to the fact that they are accessible for more users and thus having more promoters. In addition, except for some outliers (e.g., Nescafe, Facebook), we can clearly see that the fancier and luxurious the product is, the lower NPS score it gets (e.g., Prada, Porsche, etc.). This, in fact, might be explained by the culture in itself of those companies; as their products are dedicated to specific consumers and do not care much about detractors who probably might be just complaining about pricing.

Finally, it is important to note that a low NPS score does not define the success of a company. Yet, it is an important measure to work on in order to maximize clients' satisfaction and guarantee an increasing customer base. We will now look further into one of Amazon's products, which is Amazon Kindle Paperwhite - eBook reader. This product has 3176 reviews made by 2866 consumers. The NPS score for this product is over the NPS of Amazon and is 77.08. We plot

the ratings distribution for this product, which has the same shape as the entire dataset (i.e. skewed towards high ratings).

We now move on to retrieve bulk users from the 2866 that reviewed this product. We found that 186 (6.48%) out of them are bulk and they are contributing to 496 (15.61%) of the reviews. In addition, their ratings distribution is similar to non-bulk users which is in line with what we discovered in *RQ1*. Our aim now is to produce results for the what-if-analysis in which we remove bulk users and re-measure the NPS metric. Upon omitting the Bulk users from the calculation of the NPS, it decreases from 77.08 to 64.82a decrease of over 9% caused by only removing the bulk users. This in fact means that while bulk reviewing might not be the root cause for amazon fake reviews but it surely contributes to inflating the NPS score and consequently provide a misleading idea about clients' satisfaction level for the product.

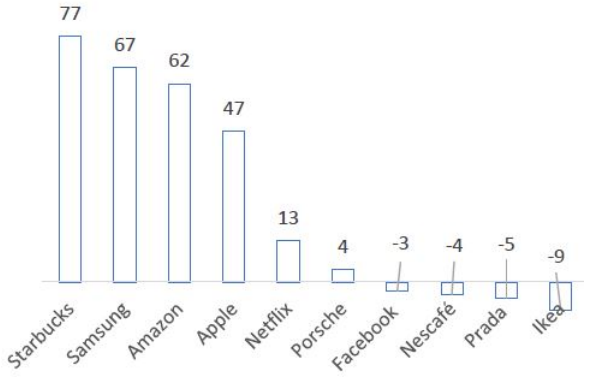

Figure 8: NPS metric comparative study

5 CONCLUSION

In recent years, the use of online reviews has been largely exploited in order to improve customer's experience and improve loyalty. In this paper, we went through a statistical analysis with the objective of exploring 'bulk users' concept and its relation with review authenticity. We also assessed the impact of Bulk users rating on the overall client satisfaction through the NPS metric, and we compared this value for several major companies.

Our results suggest that bulk reviews are not paid/fake reviewers, yet these bulk users are not totally harmless, as they tend to overestimate client satisfaction measures such as NPS. Some improvements can be added for future works that would include a more in deep study of reasons that make reviews fake and through other datasets that provide more information about users. We can also enforce our client satisfaction study via introducing other metrics and combine them to describe client satisfaction widely.

REFERENCES

[1] A. Almjawel, S. Bayoumi, D. Alshehri, S. Alzahrani, and M. Alotaibi. Sentiment analysis and visualization of amazon books' reviews. In *2019 2nd International Conference on Computer Applications Information Security (ICCAIS)*, pages 1–6.

[2] S. AlZu'bi, A. Alsmadiv, S. AlQatawneh, M. Al-Ayyoub, B. Hawashin, and Y. Jararweh. A brief analysis of amazon online reviews. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 555–560.

[3] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, WWW 09, page 141150, New York, NY, USA, 2009. Association for Computing Machinery.

[4] M. Daniel and M. Walid. Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering*, 24(6):3316–3355, 2019.

[5] W. Duan, B. Gu, and A. Whinston. The dynamics of online word-of-mouth and product sales-an empirical investigation of the movie industry. *Journal of Retailing*, 84(2):233–242, 2008.

[6] J. K. Eskildsen and K. Kristensen. The accuracy of the net promoter score under different distributional assumptions. In *2011 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*, pages 964–969.

[7] C. Fry and S. Manna. Can we group similar amazon reviews: A case study with different clustering algorithms. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 374–377.

[8] J. Kikuchi and V. Klyuev. Gathering user reviews for an opinion dictionary. In *2016 18th International Conference on Advanced Communication Technology (ICACT)*, pages 566–569.

[9] K. Kristensen and J. Eskildsen. Is the net promoter score a reliable performance measure? In *2011 IEEE International Conference on Quality and Reliability*, pages 249–253.

[10] V. Kuppili, D. Kumar, G. P. Kudchadker, and A. Arora. Variance based product recommendation using clustering and sentiment analysis. In *2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI)*, pages 1–5.

[11] Y. Liu, X. Huang, A. An, and X. Yu. Arsa: A sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 607–614.

[12] J. Mackiewicz, D. Yeats, and T. Thornton. The impact of review environment on review credibility. *IEEE Transactions on Professional Communication*, 59(2):71–88, 2016.

[13] Z. W. Ras, K. A. Tarnowska, J. Kuang, L. Daniel, and D. Fowler. User friendly nps-based recommender system for driving business revenue. In L. Polkowski, Y. Yao, P. Artiemjew, D. Ciucci, D. Liu, D. Ślęzak, and B. Zielosko, editors, *Rough Sets*, pages 34–48. Springer International Publishing.

[14] I. J. M. Ruiz, M. Nagappan, B. Adams, T. Berger, S. Dienst, and A. E. Hassan. Examining the rating system used in mobile-app stores. *IEEE Software*, 33(6):86–92, 2016.

[15] M. Tkachenko and H. W. Lauw. Comparative relation generative model. *IEEE Transactions on Knowledge and Data Engineering*, 29(4):771–783, 2017.

[16] T. Yin, W. Wang, and W. Shi. A study on fraud reviews: Incentives to manipulate and effect on sales. *China Communications*, 16(3):165–178, 2019.

[17] Y. Zhang and D. Zhang. Automatically predicting the helpfulness of online reviews. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration*, pages 662–668, 2014.