Sorting Out the Lemons – Identifying Product Failures in Online Reviews and their Relationship with Sales

Dominik Gutt[1]

[1] Paderborn University, Business Administration and Economics, Paderborn, Germany
{dominik.gutt}@uni-paderborn.com

Abstract. It is well-established that both average online ratings and the number of ratings positively impact product sales. Yet, the economic implications of the information contained in the online review texts are not that well understood. In this study, we contribute to the understanding of online review texts and their economic implications by conducting and validating an unsupervised machine learning algorithm, the latent dirichlet allocation, to identify online reviews that mention product failures. We show that the textual information on product failures are associated with lower product sales. Our results help online review system designers identifying these reviews and to make them easily accessible to potential customers to support their purchasing decisions. Academics can build on our results by applying our validated topic identification strategy and by linking reviews mentioning product failure to a range of different outcomes.

Keywords: Latent dirichlet allocation, text mining, online reviews

1 Introduction

User-generated social media, in particular online reviews, are a key success factor for businesses because they help erode the information asymmetry between sellers and buyers prior to purchase. Unsurprisingly, the positive economic impact of several metrics of online ratings, such as the average online rating score of a business [1] or the sheer number of online reviews [2, 3], is increasingly well understood. However, online reviews do not only consist of a rating score, commonly on a scale from 1 to 5, but usually also comprise a review text written by the reviewer. Currently, little is known about the economic impact of online review texts. Yet, review text can contain highly valuable information to the customer that probably play a role in her purchasing decisions.

Reviews that contain information about product failures (e.g., malfunctioning of electronic devices) are of particular interest for a number of reasons. For instance, (i) because they might affect economic success negatively [4], (ii) because firms can use them to improve their own product [5], or (iii) because they the assist future customers in making a purchase decision [6]. However, while customers can easily compare the average rating and the number of reviews of a seller or a business, review texts are much more difficult and time-consuming to assess, analyze, and compare. Thus, this

Multikonferenz Wirtschaftsinformatik 2018,
March 06-09, 2018, Lüneburg, Germany

study attempts to narrow the research gap on impacts of online review texts on economic outcomes by proposing and validating a machine learning approach to identify online reviews that contain information on product failures. In particular, we would like to answer the following two research questions: 1. *Can we automatically identify online reviews that focus on product failure?* 2. *What is the relationship between online reviews that focus on product failure and economic success?*

To answer these two questions, first, this research in progress applies an unsupervised machine learning approach – the latent dirichlet allocation (LDA, [7]) – to automatically identify the online reviews that mention product failure (*failure reviews*) in a large data set of online reviews for digital cameras from amazon.com. Second, we validate the identification of the failure reviews that resulted from our LDA using manual coders. Third, we conduct regression analyses to examine the association between product failures and sales performance of digital cameras. Our preliminary results indicate that the automatic identification approach is well capable of reliably capturing the failure reviews in our dataset and that failure reviews are, as expected, negatively associated with sales performance. To the best of our knowledge, we are the first to automatically identify and manually validate mentions of product failures in online reviews, as well as quantifying the relationship of these reviews with sales. Based on our results, online rating websites could use this approach to identify and tag failure reviews so website visitors can easily identify these reviews to make better purchasing decisions. This identification can help sellers filtering out failure reviews from all the reviews they receive to either respond to these reviews or further improve their product.

2 Related Literature

A large body of literature has delivered empirical evidence collated from across different industries supporting the claim that there is a positive effect of numerical online ratings. These include the effect of average ratings [1, 2, 8], and the number of ratings [2, 3] on online and offline sales of, for example, movies [9], books [2], and restaurants [1].

A nascent stream of research investigated possible ways to leverage the information contained in online review texts. Studies have demonstrated that the textual information in online reviews can be used to predict the pricing power of products [10] or that review texts can be used for the inference and surveillance of market structure [11]. There are two studies pursing a similar goal to our study, but they differ substantially with regards to the applied methodology and the aim of their research. One study leverages human coders to build a logistic regression framework to detect defects in cars and consumer electronics [12]. Another study identifies potentially dangerous toys from online reviews by using a catchword list obtained through a sampling approach [13]. Our study contributes to both of these streams by proposing and validating a quick, automated machine learning approach to identify one specific topic in online reviews, namely product failures, and by relating this topic to economic outcomes.

3 Topic Modelling and Empirical Analysis

We obtained a data set from amazon.com containing 29,332 single online reviews of 1,146 digital cameras [14] that were collected in July 2014 and contain all reviews from May 1996 until the time of collection. We chose digital cameras for our analysis because they can exhibit product failure in a straightforward way but also comprise substantial share of experience attributes that can be described in the reviews (e.g., usability, picture quality, etc.) Based on the online reviews, we computed the average ratings (AVG_RATING), the number of ratings (NUM_REVIEWS), and the variance of ratings (VARIANCE), as depicted in Table 1.

Table 1. Descriptive Statistics

Variable	N	Mean	Std. Dev.	Min	Max
SALES_RANK	1,128	10,930.72	10,086.21	12	155,504
NUM_FAILS	494	2.014	1.887	1	21
PRICE (in US Dollar)	1,051	141.71	169.489	0.01	899
AVG_RATING	1,146	4.067	0.491	1	5
NUM_REVIEWS	1,146	25.595	33.774	5	280
VARIANCE	1,146	1.210	0.683	0	3.551

3.1 Identification of Product Failure Reviews

To identify failure reviews in our large data set of online reviews, we employ probabilistic topic modeling based on LDA. LDA is a widely-used unsupervised machine learning method that can identify topics in large collections of documents, in our case online reviews, with written text. The essential idea behind LDA, according to Blei et. al [7], is that the authors compose documents *D* by first deciding about a discrete distribution of topics *T* to write about, relying on words *W* from a discrete distribution of words that are typical for the chosen topic. Put differently, a document is defined by a probability distribution over a fixed set of topics and each topic is defined by a probability distribution over a limited set of words [15]. For each topic of the fixed set of topics, the LDA assigns a probability between 0 and 1 to each document, indicating how likely it is that this particular document belongs to a certain topic.

The LDA approach has several advantages over alternative identification approaches. First, this approach can handle large amounts of documents in very short time. Prior literature, e.g., in the field of marketing, has traditionally relied on manual coders to identify topics in online reviews [16]. This approach is very time consuming, costly, and difficult to replicate. Our approach circumvents these limitations. Second, it seems plausible that our underlying data suits the LDA assumption that there is a fixed set of topics underlying the documents. Accordingly, recent studies have highlighted the suitability of LDA to analyze online reviews [15].

Before running the LDA using the web service minemytext.com, we applied standard measures of data pre-processing as suggested in the literature [15]. In particular, we applied stemming, 316 standard stop words (to remove uninformative

words), and we set the n-gram to 1. As the LDA relies on a fixed number of topics that has to be determined by the researcher before running the analysis, the results obtained can depend on the number of topics chosen. As suggested in the literature [15], we tried several different specifications with different numbers of topics (between 15 and 80 topics) and evaluated the quality of the results by reading samples of the reviews marked as failure reviews and comparing the mean rating of the failure reviews with the remaining ones. The LDA that yielded the best results after visual inspection comprises 40 topics, which is within the range of 10 and 50 topics usually proposed in the literature [15].

Within the LDA specification of 40 topics, there was exactly one topic capturing the failure reviews, which we identified by reading the most frequently occurring words. These are camera (7.19%), len(se) (3.02%), problem (2.77%), repair (2.3%), replace (1.35%), fix (1.3%), and defect (0.51%). As each review has a probability between 0 and 1 for each topic, we identified failure reviews as those reviews that had the highest probability, amongst all possible topics, for the above mentioned topic (*failure topic*). In total, our LDA analysis identifies 650 failure reviews within the entire data set of 29,332. As expected, the mean rating of the failure reviews (2.13) is substantially lower than the mean rating of the reviews excluding failure reviews (4.24). As depicted in Figure 1, the shape of the non-failure review distribution resembles the classical J-shape of online reviews [17] whereas the distribution of the failure reviews resembles a perfect inverted J-shape. Therefore, failure reviews are associated with substantially lower rating scores than non-failure reviews. However, there are also some, but not many, 4 and 5 star ratings, which might occur if product failures happened after a longer period of consumption that satisfied the customer.

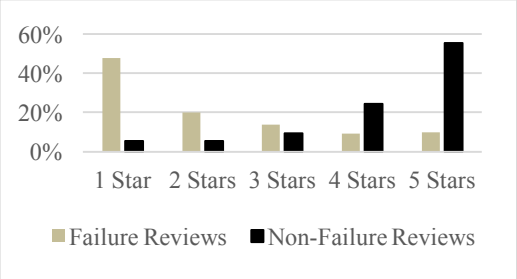

Figure 1. Distribution of Failure and Non-Failure Reviews

3.2 Validation and Estimation Results

In a next step, we validate the results of our topic modeling LDA approach to rule out concerns that our algorithm fails to capture some failure reviews or mistakenly classifies reviews as failure reviews that do not mention any product failure at all. A prime way to do this is by human manual coding and calculate the interrater reliability between the human coding results and the LDA results. We relied on such manual human coding conducted independently by two student assistants. The assistants were asked to read a sample of reviews (without any other information such as the rating or

the brand) and determine review-by-review whether the reviews mentioned product failures or not. Because the entire data contains too many reviews to be handled by humans, we took a random sample of our data comprising 300 reviews (following [18]). Parametric and non-parametric tests of variables such as SALES_RANK, AVG_RATING, and PRICE do not show statistically significant differences between the entire data set and the random sample, thus, we conclude that the random sample is representative of the whole sample. The results of the agreement between the LDA and the human coders are depicted in Table 2.

Table 2. Interrater Agreement between LDA and Human Coders

	%-Agreement	N	Krippendorff's Alpha	Cohen's Kappa
LDA & C 1	97.99%	300	0.740	0.740
LDA & C 2	98.33%	300	0.806	0.806
LDA & C1 & C2	99.00%	300	0.808	0.875

At first glance, one can see that the agreement between the LDA and coder 1 (C1) and coder 2 (C2) is very high in percent (between 97.99% and 99%). This is reconfirmed based on Krippendorff's Alpha and Cohen's Kappa, two conservative standard indices for evaluating interrater agreement [18]. Levels of 0.7 for K's Alpha and C's Kappa are sufficient to conclude interrater agreement and levels above 0.8 indicate large interrater agreement [18]. Based on the results of interrater agreement between the human coders and the LDA, we conclude that our automatic topic modelling approach has reliably captured almost all failure reviews contained in our data set. This also reassuring with regard to our choice of the number of topics (40) and the data pre-processing steps, as these choices have obviously enabled a very robust identification of failure. Finally, we would like to assess the association between the number of failure reviews on a product level and the economic performance of a camera, measured by the sales rank. Like Chevalier and Mayzlin [2], we use the Schnapp - Allwine methodology (LN_SALES_RANK=9.61 – 0.78*ln(SALES_RANK) to translate the sales rank into the natural log of the sales rank for which higher values indicate higher sales.

Table 3. Linear Regression Results

Model	(1)	(2)
Variable	LN_SALES_RANK	LN_SALES_RANK
NUM_FAILS	**-0.0553****	**-0.0601*****
	(0.028)	**(0.0209)**
Control Variables	✓	✓
Brand Fixed Effects	✓	-
Constant	-8.088***	-7.155***
	(0.434)	(0.29)
Observations	460	1,037
R^2	0.398	0.247

Note: Robust standard errors are in parentheses. *p < 0.1; ** p < 0.05; *** p < 0.01.

We conducted a stand linear regression with LN_SALES_RANK as dependent variable controlling the variables shown in Table 1. We computed the variable NUM_FAILS as the sum of failure reviews for each digital camera. We also enriched our data set with information on the camera brand (from synccentric.com) to leverage brand-level fixed effects that control for unobservable time-constant heterogeneity between camera manufacturers (e.g., Fuji, Sony, etc). Because we could not obtain brand information for all cameras, we present also results from regression models without brand-level fixed-effects. The results displayed in Table 3 suggest that there is a significant negative relationship between the number of failure reviews a product has and its sales rank. Based on the coefficient of NUM_FAILS, each failure review is associated with approximately a 5.5% lower (model (1)), respectively 6% lower (model (2)), sales rank.

4 Conclusion

While the positive economic effects of several metrics of online ratings, such as the average rating or the number of ratings, are well-studied, much less attention has been paid to the economic impact of online review texts. This study tries to narrow this gap by proposing and validating a machine learning approach to identify online reviews mentioning product failure and by demonstrating, in turn, that these reviews are negatively associated with the sales of the respective product. Our results show that automated machine learning algorithms based on the LDA approach are capable of reliable identifying failure reviews. Moreover, we can characterize the nature of failure reviews in more detail: (i) they exhibit an inverted J-shape and (ii) they are negatively associated with product sales, after controlling for different metrics of online ratings and digital camera brands. Our results come with implications for practitioners and for future research. Review system designers can use our approach to identify failure reviews and mark them in online review systems, so customers can more easily identify these reviews to make better purchasing decisions. Product designers can filter out failure reviews to learn from the customer feedback and improve their product. Second, researchers can use our identification strategy in order to investigate different consequences of product failures. Additionally, our approach is neither confined to a certain set of products and could also be applied to digital goods such as apps, nor is it limited to identifying failure topics, as it could also be used to identify other topics. Researchers for instance could also investigate failure reviews of apps as antecedents of updates as well as the economic consequences of app updates.

5 Acknowledgements

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Centre “On-The-Fly Computing” (SFB 901). I thank Robin Wulfes for excellent student assistance.

References

1. Anderson, M., Magruder, J.: Learning from the Crowd. Regression Discontinuity Estimates of the Effects of an Online Review Database. The Economic Journal 122, 957–989 (2012)
2. Chevalier, J.A., Mayzlin, D.: The Effect of Word of Mouth on Sales. Online Book Reviews. Journal of Marketing Research 43, 345–354 (2006)
3. Duan, W., Gu, B., Whinston, A.B.: Do online reviews matter? — An empirical investigation of panel data. Decision Support Systems 45, 1007–1016 (2008)
4. Herrmann, P., Kundisch, D., Zimmermann, S., Nault, B.: How do Different Sources of the Variance of Consumer Ratings Matter? Proceedings of the International Conference on Information Systems (ICIS) (2015)
5. Dellarocas, C.: The Digitization of Word of Mouth. Promise and Challenges of Online Feedback Mechanisms. Management Science 49, 1407–1424 (2003)
6. Chen, Y., Xie, J.: Online consumer review. Word-of-mouth as a new element of marketing communication mix. Management Science 54, 477–491 (2008)
7. Blei, D.M.: Probabilistic Topic Models. Commun. ACM 55, 77–84 (2012)
8. Luca, M.: Reviews, Reputation, and Revenue. The Case of Yelp.Com, https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1928601
9. Chintagunta, P.K., Gopinath, S., Venkataraman, S.: The Effects of Online User Reviews on Movie Box Office Performance. Accounting for Sequential Rollout and Aggregation Across Local Markets. Marketing Science 29, 944–957 (2010)
10. Archak, N., Ghose, A., Ipeirotis, P.G.: Deriving the Pricing Power of Product Features by Mining Consumer Reviews. Management Science 57, 1485–1509 (2011)
11. Netzer, O., Feldman, R., Goldenberg, J., Fresko, M.: Mine Your Own Business. Market-Structure Surveillance Through Text Mining. Marketing Science 31, 521–543 (2012)
12. Abrahams, A.S., Fan, W., Wang, G.A., Zhang, Z.J., Jiao, J.: An Integrated Text Analytic Framework for Product Defect Discovery. Prod Oper Manag 24, 975–990 (2015)
13. Winkler, M., Abrahams, A.S., Gruss, R., Ehsani, J.P.: Toy Safety Surveillance From Online Reviews. Decision Support Systems 90, 23–32 (2016)
14. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: Image-Based Recommendations on Styles and Substitutes. In: Proceedings of the 38th ACM SIGIR Conference on Research and Development in Information Retrieval, 43–52 (2015)
15. Debortoli, S., Junglas, I., Müller, O., Vom Brocke, J.: Text Mining For Information Systems Researchers. An Annotated Topic Modeling Tutorial. Communications of the Association for Information Systems Paper 7 (2016)
16. Sridhar, S., Srinivasan, R.: Social Influence Effects in Online Product Ratings. Journal of Marketing 76, 70–88 (2012)
17. Hu, N., Pavlou, P.A., Zhang, J.: On Self-Selection in Online Product Reviews. MIS Quarterly 41, 449–471 (2017)
18. Lombard, M., Snyder-Duch, J., Campanella Bracken, C.: Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. Human Communication Research 28, 587–604 (2002)