ISSN: 2664-9640

Text Sentimental Analysis of Online Reviews Based on Lexicon and Machine Learning

Huiling Cheng

School of management, Shanghai University, Shanghai 201800, China

Abstract

With the development of electronic commerce platform, text information became an important data resource for product research. This paper is mainly focused on natural language processing, which is the little branches of artificial intelligence. This research is mainly composed of two part, one is to clarify the importance of Naïve Bayes model for automatic text classification and the other is to simplify the significance of words frequency and words cloud in product promotion based on the statistics. Xiaomi 10 is chosen for this experiment and have collected number of 2635 comments including 1976 positive comments and 659 negative comments, and the experiment results indicated that the correction can come up to 86.95% the accuracy is 79.08% and the rate of recall is 79.08%. These indicators have proved that the Naïve Bayes Model is an excellent model for text classification. We also found that the more numbers of raw data we have, the more accurate the classifier is when conducting the experiment. In the follow-up works, we have respectively collected the 1869 positive comments and 766 negative comments, then we counted the frequency of words and draw the words cloud. This procedure has proved the role of words frequency and provided the conferences for enterprises in product promotion.

Keywords

Text sentiment analysis; Naïve Bayes model; word frequency； visualization.

1. Introduction

Companies need to do research when making product improvements, to consider what kind of products users like, and how to improve them in order to be right and make them happy (Arghashi and Yuksel, 2022; Zemack-Rugar et al., 2016). To know this, we must use scientific methods to know precisely what our users like, what their evaluation of existing products is, good or bad reviews, and why it is good, where it is good, and why it is bad, where it is bad, so that we can have the right target to improve product reputation and enhance product sales (Arora and Henderson, 2007; Chopdar and Balakrishnan, 2020). With the progress of the times, in the past period, companies studying users usually choose to collect basic information of users, including their age, gender, occupation, education, marital status to make user classification or user portrait, and think about what kind of products this kind of customers would like from the perspective of personal experience, so as to design and recommend products to users (Arghashi and Yuksel, 2022; Ieva et al., 2022; Zemack-Rugar et al., 2016). However, in the Internet era, there are many data processing technologies supporting our mobile e-commerce platforms, mobile social platforms, which allow users to publish and share their product experiences in a timely manner, including their attitudes and emotions towards products and services, and this new data perspective, properly utilized, can allow us to understand our users faster, more comprehensively and more accurately, and we have many new ways to study users (Kwon et al., 2016; Pu et al., 2020). Nowadays, product review data on major e-commerce platforms is also a hot spot for research and has been widely used in enterprise product improvement (Ho et al., 2017; Yang et al., 2019).

ISSN: 2664-9640

2. Literature review

Online reviews are considered to be an important source of product uncertainty reduction for online shopping consumers, and research in recent years has focused on the study of incentives for online reviews and the impact of false reviews in online reviews on consumer decision making. Ke et al. (2020) investigated whether friend contribution prompts motivate consumers to post more higher quality reviews, and the study found through an econometric research approach that friend contribution prompts have a positive positive effect on consumers posting product reviews, with users three times more likely to provide a review after a recent friend's review than after a recent stranger's review, and reviews written after a friend's review tend to be of higher quality, longer, and more original. Huang et al. (2017) investigated how social network integration affects the quantity and quality of online comments posted by users, and found that after Facebook integration, the number of comments increased, as did the expression of positive emotions and the expression of cognitive language, negative comments, and disagreement in online commenting texts. It has increased the quantity of comments while decreasing their quality. Mudambi and Schuff (2010) investigated what kind of review is a helpful review, i.e., how the polarity of the review, the depth of the review, and the product type affect the user's perceived helpfulness, and found that the product type moderates the effect of review polarity on perceived helpfulness, and that for experience-based items, reviews with extreme ratings are less helpful than reviews with moderate ratings. Regarding the study of false reviews in online reviews. Ananthakrishnan et al. (2020) examined the question of whether false reviews should be displayed on online review portals and how consumers respond to false reviews, finding that consumers' trust values are higher when review sites display false reviews and finding that consumers do not effectively deal with false review information. The impact of false reviews on online visibility was studied in Lappas et al. (2016) , which found that even a limited number of false reviews can have a significant impact. The dark side of online reviews was studied in Liu and Karahanna (2017) , which examined the swaying effect of online product reviews on the construction of attribute preferences, summarizing three characteristics of online reviews, the amount of information about attribute-level performance, the degree of information controversy, and the relationship between overall numerical ratings in reviews and attribute-level performance information, which found that three review characteristics affect attribute preferences and that their effects are strong enough that these online review features affect attribute preferences more than the relevance of attributes to consumer decision context.

3. Naïve Bayes Model

3.1. Algorithm introduction

The Naïve Bayes algorithm is one of the top ten classical algorithms for machine learning, and is a supervised learning classification algorithm (Ryoo et al., 2021; Timoshenko and Hauser, 2019). The Naïve Bayes algorithm is a classification method based on Bayes' theorem and the assumption of conditional independence of features, which determines the category of the classified item based on the probability value of the classified item belonging to each category, and the one with the highest probability is the classified category (Fang et al., 2013; Ho et al., 2017). The plain in the algorithm refers to the assumption that the features are independent of each other, that is, the probability of occurrence between feature a and feature b is not affected by each other, and the occurrence of feature a is independent of feature b, and each feature is equally important. The principle is mainly as follows, an existing item X to be classified, consisting of multiple attribute values a_i, is expressed as $X=\{a_1,a_2,a_3,\ldots,a_n\}$; There is a category Y, consisting of several different categories, is expressed as $Y=\{b_1,b_2,b_3,\ldots,b_m\}$. According to

Bayes' theorem, if we want to know whether the classified item X belongs to b_1,b_2, or b_m, we need to compare the size of the values of$P(b_1|X)$、$P(b_2|X)$、…、$P(b_m|X)$，and the one with the largest probability value is considered to belong to the category. The prerequisite of this assumption is to simplify the calculation of the value of $P(b_m|X)$, The assumption is that each attribute value a_1,a_2,a_3,…,a_n of X is independent with each other, and the denominator of each probability formula is the same, the larger the numerator is, the larger the Bayesian formula can be reduced to calculate the numerator, and the calculation of the numerator can be reduced to the product of the conditional probabilities of each attribute value due to the assumption of independence, the reduction process is as follows.：

(1) Bayes algorithm

$$P(bm|X) = \frac{P(X|bm)P(bm)}{P(X)}$$

(2) Calculate molecules

$$P(X|b_m)P(b_m)$$

(3) The " Naïve " assumption is reduced to

$$P(a_1|b_m)\ P(a_2|b_m)\ P(a_3|b_m)\ldots P(a_n|b_m)p(b_m)$$

The formula $P(X|b_m)$ is called the conditional probability, $P(b_m|X)$ is the posterior probability, and $P(b_m)$ is the prior probability. A priori probability is the probability of something happening through experience. The posterior probability is the probability of inferring the cause after the occurrence of the result. The advantage of this algorithm is that the principle is simple and the computational process is simple but the classification effect is indeed high, so the Naïve Bayes algorithm has been widely used in many fields (Arora and Henderson, 2007; Fang et al., 2013; Kaneko et al., 2019). However, the algorithm assumes that the feature variables are independent of each other and that they are normally distributed when they are continuous variables, which can lead to a certain degree of compromise in the accuracy of the algorithm. The premise of this assumption is also to simplify the computational process: the calculation of the probability of a data belonging to a certain category is reduced to the product of the conditional probabilities of the occurrence of all feature values, a process that greatly reduces the computational effort, but also lacks in accuracy. The Naïve Bayes algorithm usually has the following models, usually Gaussian model when the feature variables are continuous type variables, and polynomial model or Bernoulli model when the feature variables are discrete type variables. First, Gaussian model. This model assumes that the continuous type feature variables obey normal distribution, and then the conditional probability value can be calculated according to the following formula:

$$P(x_i|y_k) = \frac{1}{\sqrt{2\pi\sigma_{yk,i}^2}} e^{-\frac{(x_i-u_{yk,i})^2}{2\sigma_{yk,i}^2}}$$

Where $u_{yk,i}$ denotes the mean of the training sample set and $\sigma_{yk,i}^2$ denotes the standard deviation of the training sample set, thus turning the continuous feature variables into nominal discrete variables. Second, Polynomial model. This model is commonly used in text classification, where the feature is a word and the value is the number of occurrences of the word. The most classic case is spam filtering, and the principle is similar to the sentiment classification done in the article, where a certain number of sample emails are used to train the plain Bayes classifier, and then a test set is used to verify the classification accuracy of the classifier. Third, Bernoulli model. Bernoulli model is similar to polynomial model, both are used to deal with discrete data, but Bernoulli model treats repeated words as appearing only once, without considering the number of times the feature variable appears, this method is more

concise and clear, but it does not count the number of times the word appears, and the final effect is also worse, more suitable for dealing with dichotomous (Boolean) model.

3.2. Algorithm evaluation

For a binary classification problem, the predicted results of the machine classifier and the actual ones still have deviations (Timoshenko and Hauser, 2019; Wang et al., 2018). The performance of the classifier can be evaluated by calculating several metrics as follows, and the significance of these evaluation metrics can be more clearly understood by the Confusion Matrix, as shown in Table1.

Table 1. Confusion matrix

	Predict Positive	Predict Negative
True Positive	TP (True Positive)	FN (False Negative)
True Nagative	FP (False Positive)	TN (True Negative)

After understanding the above concepts, the following evaluation metrics will be used to determine how effective a classification algorithm is.

Accuracy rates: The ratio of the number of test samples correctly classified by the classifier to the total number of test samples in a given sample of the test data set.

$$Accuracy\ rates = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Precision rates: The percentage of test sets in which the classifier predicts positive cases that are actually positive. It calculates the percentage of all predicted positive cases that are actually positive. This is because classifiers can misclassify.

$$Precision\ rates = \frac{\mathrm{NP}}{NP + FP}$$

Recall rates: It calculates the percentage of actual positive cases that are correctly predicted as positive, i.e., it calculates all actual positive cases that are accurately predicted as positive.

$$Recall\ rates = \frac{\mathrm{NP}}{\mathrm{NP} + \mathrm{FN}}$$

4. Method

4.1. Data collection

The experimental data object of the article is the review data set of Xiaomi 10 in the flagship store of Xiaomi JD in JD Mall. 2,635 reviews of Xiaomi 10 Titanium Silver Black in the flagship store of Xiaomi in JD Mall were collected, and the data were collected in the order of the latest reviews, which can effectively avoid the impact of the default sorting on the data set by the front of the water army users.

4.2. Data pre-processing

Firstly, the collected data are determined as positive and negative polarity according to the star rating, and those with star rating greater than or equal to 4 are set as positive sentiment polarity, i.e. positive emotion, marked as 1. Conversely, the comment data with star rating less than 4 are set as negative sentiment polarity, marked as 0. The features are separated from the labels, and the X value is set as comment, i.e. comment data, and the Y value is set as label, i.e. positive and negative sentiment polarity. Word Segmentation (Word Segmentation) is a basic step of natural language processing, compared to English word segmentation Chinese sentences are

more difficult (Ryoo et al., 2021; Wang et al., 2020). English sentences are separated by spaces between each word, so you can traverse the document and divide the words into spaces to achieve the purpose of word segmentation, while Chinese word segmentation does not have obvious word segmentation signs, and the Chinese text needs to be separated by complex algorithms to achieve. At present, there are jieba, pangu, yaha and Tsinghua THULAC Chinese word sorting software (Wang et al., 2020). The article uses the jieba lexical database for Chinese lexical classification. jieba lexical classification algorithm is based on the lexical matching lexical pattern, which comes with a dict.txt lexical text containing more than 20,000 words, covering almost all Chinese words. Jieba lexical classification contains three types of lexical patterns. First, full pattern. This means that all possible words are separated, so that the contextual relationship of the words in the sentence is ignored. Second, exact mode. This means that each word is separated precisely. Third, search engine mode. Refers to the precise word separation based on the long words will be cut apart. The article adopts the precise mode of jieba word separation to carry out the word separation process of comment data. And in the default mode, 1/4 proportional data set is randomly divided as the test data set and 3/4 data set as the test data set.

4.3. Feature selection and text vectorization

The article uses the HIT deactivation word list for deactivating words. After the word separation and deactivation, the process of text vectorization is performed. At present, there are three models of text representation methods as follows. First, Boolean model, that is, the appearance of feature words in the text is recorded as 1, not appearing is recorded as 0, after all features are determined, a data set is represented by a two-dimensional data table. Second, VSM vector space model, this model representation method not only records whether the feature words appear but also records the number of times the feature words appear, and the analysis results will be more accurate and appropriate. Third, Word2Vec text representation model, using the given document for feature training, and the document will be represented as a fixed-length vector of feature words, can avoid the problem of dimensional disaster. After feature selection often some words can not be used as features, some feature words need to be filtered out to improve the accuracy of the classifier, the article under the filtering of the deactivated word list and three layers of filtering, through the deactivated word list for the first layer of filtering, frequency statistics methods for the second layer of filtering, and the establishment of regular expressions for the third layer of filtering, this method is simpler and more convenient. We can see that the number of feature columns has changed from 4568 to 1139, and the number of features has been greatly reduced.

5. Results

5.1. Naïve Bayes algorithm evaluation results

In the process of experimentally verifying the Naïve Bayes algorithm we need to preprocess each piece of data and then perform Naïve Bayes classification, but such a process would be very tedious and require re-running many of the above functions each time we change the test data, which is very inefficient. The article directly calls Scikit-learn, a third-party Naïve Bayes classifier library for python, to do the empirical research process. The principle of the toolkit is because the library contains pipeline functions, which can help us connect these sequential jobs and hide the functional sequential associations among them, and complete all the work of sequential definition in one call from outside. According to the confusion matrix of the experimental results, TP=155, FN=41, FP=41, TN=422, the accuracy is 86.95%, the precision is 79.08% and the recall is 79.08%, which proves the important role of the Naïve Bayes algorithm in automatic text classification. In the preliminary stage of the experiment, 1881 data were collected to do the empirical research process of the Naïve Bayes algorithm, including 1410

training data sets and 471 test data sets. The confusion matrix values derived from the classifier after learning from the training data sets were TP=100, FN=30, FP=41, TN=300, of which 400 test data sets were correctly predicted, and the prediction error There were 71 data, and the experiment reached 84.93% accuracy, 70.92% precision and 76.92% recall. In the subsequent experiments, 754 original datasets were added, and the experimental results showed a significant improvement in the performance of the Naïve Bayes classifier. Accuracy improved by 2.63% from 84.93% to 87.56%, precision improved by 8.16% from 70.92% to 79.08%, while recall improved by 2.16% from 76.92% to 79.08%. The experiment illustrates that the more valid data collected, the higher the classification accuracy of the experiment and the better the performance of the classifier. It is also demonstrated here that the quality of the training dataset has a great influence on the performance of the classifier, and the higher the quality of the training dataset that the classifier learns, the better the performance of the classifier is, now the data enhancement effect. The quality of the training dataset is determined by the amount of data, the usefulness of the data, and so on. A comparison of the results of the two empirical studies is shown in Table2.

Table2. Comparison of pre and post experimental results

	Total	Trainning	Testing	Accuracy	Precision	Recall
Pre	1881	1410	471	84.92%	70.92%	76.92%
Post	2635	1976	659	87.56%	79.08%	79.08%

5.2. Word frequency statistics results

After the previous basic data processing is completed, the process of counting the number of valid words begins. It is generally considered that the higher frequency of words appearing in the entire positive review data set we consider that the words can reflect some aspects of the product's evaluation. The word frequency statistics table for positive data is shown in Table3, and the word frequency statistics table for negative data is shown in Table 4.

Table3. Word frequency statistics for the top 30 positive comments

'not bad'	609	'features'	261
'pixel'	181	'good looking'	170
'like'	169	'very quick'	166
'smooth'	160	'curved surface'	141
'hand feel'	140	'clarity'	129
'special'	126	'speakers'	118
'charge'	113	'one hundred million'	113
'very nice'	104	'satisfied'	101
'confortable'	99	'beautiful'	95
'perpect'	89	'high'	88
'great'	86	'good'	85
'865'	79	'battery'	78
'experience'	76	'tone quality'	73
'cost-effectiveness'	73	'JingDong'	73
'system'	67	'1'	63

ISSN: 2664-9640

Table 4. Word frequency statistics for the top 30 negative comments

'*'	147	'disconnection'	122
'standby time'	114	'feature'	107
'Jing Dong'	88	'not bad'	82
'not fluent'	78	'experience'	75
'systems'	73	'signal'	71
'customer service'	71	'camera'	66
'disappointment'	62	'garbage'	60
'Mi'	60	'bad'	58
'charge'	56	'situation'	55
'reboot'	55	'very'	52
'fever'	52	'headphones'	52
'play'	52	'menbrane'	51
'not good'	50	'unlock'	49
'little bit'	49	'film'	49
'voice'	49	'first time'	48

5.3. Mapping the word cloud

After the completion of the word frequency counting process in the previous step, the word cloud plotting process is carried out, and the word cloud plot graphically shows the reasons why the comments are positive or negative sentiments. The scale value of the word cloud is set in this step, which greatly improves the clarity of the drawn word cloud. The word cloud diagram for the positive comment data is shown in Figure1, and the word cloud diagram for the negative comment data is shown in Figure2.

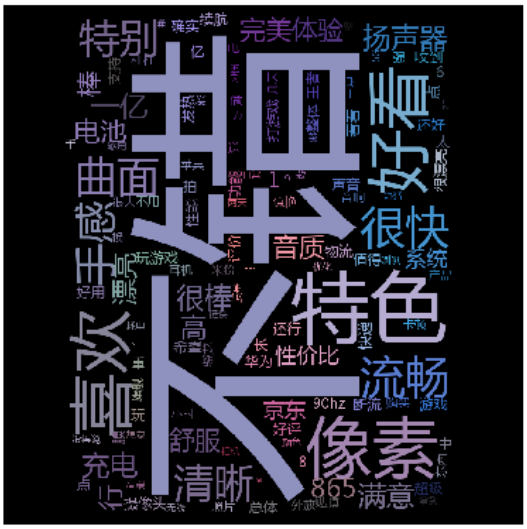

Figure1. Word cloud map of positive review dataset

ISSN: 2664-9640

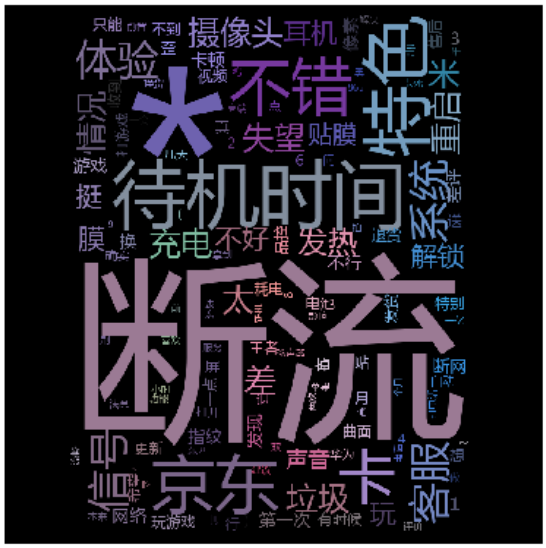

Figure2. Word cloud map of negative review dataset

5.4. Results analysis

Combined with the analysis of word frequency statistics and word cloud chart results in the first two sections, we can have a deeper understanding and appreciation of the strengths and weaknesses of Xiaomi 10 cell phone and the reasons for consumer sentiment. In the word frequency statistics table of 1869 positive reviews, we can see that the overall vocabulary includes noun-level words such as pixel, feel, charging, speaker, system, battery, etc., as well as adjective-level words such as nice, like, smooth, clear, etc., and some emotional words including like, satisfied, great, etc. Combining these words can match some entity and aspect-level combinations, such as smooth system, nice touch, etc. The top three words in the statistical frequency are "not bad", "feel" and "quite", which means that the general evaluation of Xiaomi 10 by consumers with positive reviews is that it feels quite good (609), and there are many words including very fast (166), great (104), good (86), like (169) and satisfied (169). The positive words such as like (169), satisfied (101), perfect (89), etc., as well as the adjectives that follow in the top ranking include smooth (160), clear (129), beautiful (95), feel (140), comfortable (99), etc., reflecting consumers' positive evaluation of Xiaomi 10, the advantages include smooth and clear and good-looking, etc. In the data set of 766 negative comments, the noun aspect-level words include standby time, JD, system, signal, headphones, unlock, etc., while the adjective descriptive words include garbage, poor, bad, heat, etc. The matching description is for bad signal, poor unlock, etc., but this is not an accurate match, but only a general idea of the reasons for the poor reviews. The top three negative word frequency statistics are *, disconnection and standby time, where * is the meaning of bad, that is, the symbol of uncivilized language on the network, the remaining two are noun words, because the word frequency statistics do not take into account the context of the connection and therefore can't determine the evaluation of JD and the system, the ranking of the latter is the system, camera, is also a characteristic noun, including the back of the film (51), unlocking (49). The negative terms such as not smooth (78), disconnection (122), restart (55), heat (52) and other causes can be seen in the Xiaomi 10. The phenomenon of system lag, interruption of flow and reboot, which is also the direction of Xiaomi 10 products need to improve.

6. Conclusion

The main work of the article includes the empirical study of the Naïve Bayes classification algorithm in automatic text classification and the analysis of text sentiment reasons based on statistical principles, which provides some improvement suggestions for enterprise product improvement. During the analysis of the output results, the classification efficiency of the Naïve Bayes algorithm and its many limitations are found. Based on the word frequency and word cloud results of the positive and negative polarity review dataset, the advantages of Xiaomi 10 cell phone in terms of system fluency and beautiful appearance as well as the disadvantages of disconnection and restart lag are more clearly recognized.

On the basis of the empirical study of the article, there are still many aspects of improvement research, and there are problems in the empirical study of the Naïve Bayes algorithm. First, the prerequisite of "plain" means that the features are independent of each other and do not affect each other, but this prerequisite cannot exist in the actual problem solving. There is still much room for improvement. Second, imbalance of data, the article collected a total of 2635 data, of which 1869 comments are characterized as positive sentiment and 766 negative comments, the imbalance of the training data set will affect the performance of the classifier. Third, polarity and star rating mismatch of comments. In the collected data, the user evaluation may give 2 stars, but it is said that the Xiaomi phone is good, such experimental training data will reduce the accuracy of the classifier, and in the case of a large amount of data manually to correct the star rating and the actual comment mismatch to spend more manpower. Problems in the process of text sentiment analysis. First, the number of data sets is not enough, the article only collected 2635 valid data, the amount of data is small, can't fully cover all consumer reviews, the word frequency statistics results and ranking results have a greater impact. Second, the article uses the word frequency statistics to analyze the sentiment reason itself is a coarse-grained rough analysis, and not accurate, for example, Xiaomi cell phone charges fast, which is a description of the physical Xiaomi cell phone in terms of charging fast, but the word frequency statistics in the article will split the charging and fast, which can't reflect the complete aspect-level emotion, and needs a more fine-grained emotional aspect-level research and analysis.

References

[1] Ananthakrishnan, U.M., Li, B., Smith, M.D., 2020. A tangled web: Should online review portals display fraudulent reviews? Inf. Syst. Res. 31, 950–971. https://doi.org/10.1287/ISRE.2020.0925

[2] Arghashi, V., Yuksel, C.A., 2022. Interactivity, Inspiration, and Perceived Usefulness! How retailers' AR-apps improve consumer engagement through flow. J. Retail. Consum. Serv. 64, 102756. https://doi.org/10.1016/j.jretconser.2021.102756

[3] Arora, N., Henderson, T., 2007. Embedded premium promotion: Why it works and how to make it more effective. Mark. Sci. 26, 514–531. https://doi.org/10.1287/mksc.1060.0247

[4] Chopdar, P.K., Balakrishnan, J., 2020. Consumers response towards mobile commerce applications: S-O-R approach. Int. J. Inf. Manage. 53, 102106. https://doi.org/10.1016/j.ijinfomgt.2020.102106

[5] Fang, X., Hu, P.J.H., Li, Z.L., Tsai, W., 2013. Predicting adoption probabilities in social networks. Inf. Syst. Res. 24, 128–145. https://doi.org/10.1287/isre.1120.0461

[6] Ho, Y.C.C., Wu, J., Tan, Y., 2017. Disconfirmation effect on online rating behavior: A structural model. Inf. Syst. Res. 28, 626–642. https://doi.org/10.1287/isre.2017.0694

[7] Huang, N., Hong, Y., Burtch, G., 2017. Social network integration and user content generation. MIS Q. Manag. Inf. Syst. 41, 1035–1058.

[8] Ieva, M., Ziliani, C., Gázquez-Abad, J.C., D'Attoma, I., 2022. I read, therefore I buy? Analyzing the impact of flyer distribution and readership on purchase behaviour. J. Retail. Consum. Serv. 64. https://doi.org/10.1016/j.jretconser.2021.102725

[9] Kaneko, Y., Yada, K., Ihara, W., Odagiri, R., 2019. How game users consume virtual currency: The relationship between consumed quantity, inventory, and elapsed time since last consumption in the mobile game world. IEEE Int. Conf. Data Min. Work. ICDMW 2018-Novem, 848–855. https://doi.org/10.1109/ICDMW.2018.00125

[10] Ke, Z., Liu, D., Brass, D.J., 2020. Do online friends bring out the best in us? The effect of friend contributions on online review provision. Inf. Syst. Res. 31, 1322–1336. https://doi.org/10.1287/isre.2020.0947

[11] Kwon, H.E., So, H., Han, S.P., Oh, W., 2016. Excessive dependence on mobile social apps: A rational addiction perspective. Inf. Syst. Res. 27, 919–939. https://doi.org/10.1287/isre.2016.0658

[12] Lappas, T., Sabnis, G., Valkanas, G., 2016. The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. Inf. Syst. Res. 27, 940–961. https://doi.org/10.1287/isre.2016.0674

[13] Liu, Q. Ben, Karahanna, E., 2017. The dark side of reviews: The swaying effects of online product reviews on attribute preference construction. MIS Q. Manag. Inf. Syst. 41, 427–448. https://doi.org/10.25300/MISQ/2017/41.2.05

[14] Mudambi, S.M., Schuff, D., 2010. What makes a helpful online review? A study of customer reviews on amazon.com. MIS Q. Manag. Inf. Syst. 34, 185–200. https://doi.org/10.2307/20721420

[15] Pu, J., Chen, Y., Qiu, L., Cheng, H.K., 2020. Does identity disclosure help or hurt user content generation? Social presence, inhibition, and displacement effects. Inf. Syst. Res. 31, 297–322. https://doi.org/10.1287/ISRE.2019.0885

[16] Ryoo, J.H., Wang, X., Lu, S., 2021. Do Spoilers Really Spoil? Using Topic Modeling to Measure the Effect of Spoiler Reviews on Box Office Revenue. J. Mark. 85, 70–88. https:// doi.org/10.1177/0022242920937703

[17] Timoshenko, A., Hauser, J.R., 2019. Identifying customer needs from user-generated content. Mark. Sci. 38, 1–20. https://doi.org/10.1287/mksc.2018.1123

[18] Wang, Q., Li, B., Singh, P.V., 2018. Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis. Inf. Syst. Res. 29, 273–291. https:// doi.org/10.1287/isre.2017.0735

[19] Wang, X., Wang, W., Chai, Y., Wang, Y., Zhang, N., 2020. E-book adoption behaviors through an online sharing platform: A multi-relational network perspective. Inf. Technol. People 33, 1011–1035. https://doi.org/10.1108/ITP-10-2018-0482

[20] Yang, M., Ren, Y., Adomavicius, G., 2019. Understanding user-generated content and customer engagement on Facebook business pages. Inf. Syst. Res. 30, 839–855. https:// doi.org/10.1287/isre.2019.0834

[21] Zemack-Rugar, Y., Rabino, R., Cavanaugh, L.A., Fitzsimons, G.J., 2016. When donating is liberating: The role of product and consumer characteristics in the appeal of cause-related products. J. Consum. Psychol. 26, 213–230. https://doi.org/10.1016/j.jcps.2015.06.001