Based on Logistic Regression the online Reviews False Identification Model Study

Mei-Mei ZHANG\textsuperscript{1,a,*}

\textsuperscript{1}North China Electric Power University, Baoding, China, 071000
\textsuperscript{a,zmm1976\textsuperscript{1}@126.com,}
\textsuperscript{*Corresponding author}

Keywords: False comments, Logistic regression, ANOVA, User perception.

Abstract. In this paper, a questionnaire survey was made on the key factors to identify the false comments from the perspective of consumer perception, including words, number, comment tone, detailed degree, three days before the comments number, first week comments, total comments number, 2 weeks before comments number, additional comments number, photo show number, good review, medium comments or bad comments, sellers reputation, from 0 to 9 o'clock and 21 measurement items, in the case of taobao commodity comments, artificial traversal 60 kinds of commodities in 6885 comments to collect data, it is concluded that additional comments number, slide show number, and medium review number of significant differences when consumers perceive false comments. By constructing the Logistic regression model, detailed degree, three days before the comments number and slide show number are the key influencing factors, overall percentage prediction rate was 85%, the discriminant function effect is better. The results show that the three indexes can be used as a consumer in judging comment whether false goods, shopping is the basis of risk.

Introduction

In online shopping, consumers could not be touch goods and lack of trust to the businessman often turn to commodities comments to get more information to avoid risk, in fact, more and more consumers to online comments as an important reference of the consumer decision-making, in order to encourage traders were published after each transaction feedback, many businesses induce customers give a high praise by cash back, sending coupons, red envelope for all sorts of ways, which results in a deterioration in the authenticity and reliability of online reviews. In addition, business by hiring online water army published false online reviews to exaggerate the decoration of the goods, improve the shop's ranking, sales and praise with the genuine way of shopping, to attract customers, and then mislead consumers to achieve consumption. In August 2016 of 1046 consumers in quantitative research found that 62% of consumers will refer to “the commodity sales records and buyers comments” in online shopping; but 70% of consumers think there are a lot of fraud and false propaganda online shopping. Product reviews are based on the initial purchase amount, which formed the online shopping consumption cycle “Consumption based on commodity comments-Lack of initial purchases and reviews-No comment, no buy”, some businesses in order to break this vicious cycle made false comments, that is "Consumption based on commodity comments-Lack of initial sales volume and comments-Creating false comments by scalping-Consumption based on commodity comments". But with the rapid development
of electronic commerce, false trading of hype credit and scalping has formed a dark gray industrial chain, including recruitment scalping staff, logistics, Selling personal information. "Hype credit" has characteristics of occupation and specialization, the scale of the industrial chain is more and more big, the harm to the healthy development of electronic commerce is becoming more and more prominent. In order to avoid regulation, online water army is constantly improving. In order to avoid the identification of more and more vigilant consumers, they began to use a puppet account, they are more like a normal consumer when buying and commenting, which makes the identification more difficult. The new online water army is to imitate the attitude of the user's remarks, to create a large number of false information, thus affecting consumers.

False online comments not only raises the operating costs and integrity of the cost of electricity providers, but also to mislead consumers and consumers to bring post-purchase risk, making consumers the value of online reference to reduce the electricity business platform, the deterioration of the ecological environment. With the lack of business integrity and consumer online shopping experience growth, wait without buying, impulse buying and return, sellers buy good comments by cash back, which becomes the norm.

This article focuses on the authentication of commodity reviews from the perspective of consumer perception. In particular, this authentication refers to whether there are false comments in the whole product review instead of in one certain commodity review. We extract features as the number of the words in comments, mood, detailed degree, text self-similarity, published time intensity, pre-and-post sales, positive comments and the star rating of business, which are the input parameters of logistic regression model. Then the model is trained and applied to the authentication of false comments. This paper is divided into five parts, the second part mainly introduces the related research, the third part includes data acquisition, feature selection and introduction of relevant algorithm principle, the fourth part is the experimental process and results, the fifth part is the conclusion and summary.

Related Research

There has been a lot of literature reviews proving that online reviews have become the important factor in consumer's observation of business reputation and public praise [1], and the key indicator of whether consumers are willing to buy [2]. Online reviews are related to brand trust and willingness to buy, which can affect the consumption decision of hotel [3], restaurant [4], etc. The negative online reviews are more useful than positive comments for risk-averse travelers. As for positive online reviews, risk-averse travelers think the post with expert reviews, the products of tourism products, and positive online reviews of famous brand are more valuable. Regardless of the level of brand reputation, there exists positive interrelated relationship between the depth and the usefulness of the review. The usefulness of timeliness of comments to commodity review with a high reputation is stronger than that with a low reputation. The quantity and the value of the comments have a significant positive effect on sales volume, while the difference comments has a significant negative effect on it.

Jyoti et al. developed models based on machine learning that can predict the helpfulness of the consumer reviews using several textual features such as polarity, subjectivity, entropy, and reading ease. The model will automatically assign helpfulness values to an initial review as
soon as it is posted on the website so that the review gets a fair chance of being viewed by
other buyers[5]. Serial mediation tests show that the causal link between language typicality /
perceived reviewer expertise explains the language style effect on consumers' pre-purchase
evaluations[6].

Design of the Identification Model of False Commodity Comments

Extraction and Analysis of Perception Factors in False Comments

The false comments identification can be carried out from two angles, a point of view is
platform supervisor side based on large data extraction online water army released false
comments features to establish identification model for identification, another perspective is
based on consumer perception as a starting point, identify the consumer's perception of false
comments and thus establish the identification model, this paper selects the second
perspective to study. Taobao allows customers to appraise the conformity between physical
goods and description, seller's service attitude and delivery speed with a one-to five-star rating.
One star, the lower score, means dissatisfied. Three stars, the intermediate score, means
acceptable. Five stars, the higher score, means satisfied. Besides, Taobao also supports image
uploading and additional evaluation. Online water army gives false comments to commodities
through the way provided by Taobao. Considering that false comments aim to exaggerate the
quality of goods so as to induce the purchase of customers, their internal and external
characteristics are as follows. The comments have too much words and details, which are
always provocative and appear intensively in a certain period of time. The sale of new
products surges after they hit shelves. The text self-similarity is too high. (There is too much
similarity in the words used by the same user.) The comments are too professional to the
layman to understand. One person buys several same products.

There are identical comments and a large number of star buyers. (Online water army uses
the puppet account, which is the star account that will be found and then closed down by
Taobao after being used for a period of time.) The normal customers will not make comments
as above; most of them will just evaluate the products, delivery and service instead of writing
down too much. However, sometimes this has some drawbacks. Some regular customers who
are very faithful to the store usually make comments which researchers tend to regard as
online water army comments, while they want to express their satisfaction to the products.

Above all, as the behavior of online water army is more hidden and similar to the normal
customers, we need to extract the underlying feature of online water army from the
perspective of user perception.

In summary consumer perception features in false comments showed the following
characteristics, Including words number, comment tone, detailed degree, three days before the
comments number, first week comments, total comments number, 2 weeks before comments
number, additional comments number, photo show number, good review, medium comments
or bad comments, sellers reputation, from 0 to 9 o'clock and 21 measurement items, in the
case of taobao commodity comments, artificial traversal 60 kinds of commodities in 6885
comments to collect data.
Analysis of Perceptual Differences in False Comments

Firstly the comparison of the users in the commodity review of false discrimination on cognitive differences analysis, The differences of perceived factors were compared between words number, comment tone, detailed degree, three days before the comments number, first week comments, total comments number, 2 weeks before comments number, additional comments number, photo show number, good review, medium comments or bad comments, sellers reputation, from 0 to 9 o'clock and 21 measurement items. By analysis of variance, there were significant differences in the three factors in “photo show number”, “medium comments” and “additional comments number”, as shown in Table 1.

Table 1. Consumer Perceived Factor in false comment ANOVA table.

<table>
<thead>
<tr>
<th>Photo show number *classification</th>
<th>Sum of squares</th>
<th>df</th>
<th>mean square</th>
<th>F</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>(combination)</td>
<td>4824.011</td>
<td>1</td>
<td>4824.011</td>
<td>7.271</td>
</tr>
<tr>
<td>In the group</td>
<td>38480.839</td>
<td>58</td>
<td>663.463</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>43304.850</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium comments *classification</td>
<td>14.015</td>
<td>1</td>
<td>14.015</td>
<td>4.169</td>
<td>.046</td>
</tr>
<tr>
<td>In the group</td>
<td>194.969</td>
<td>58</td>
<td>3.362</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>208.983</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional comments number *classification</td>
<td>604.353</td>
<td>1</td>
<td>604.353</td>
<td>4.546</td>
<td>.037</td>
</tr>
<tr>
<td>In the group</td>
<td>7710.897</td>
<td>58</td>
<td>132.947</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8315.250</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Construction of Logistic Regression Model for Identification in False Comments

Because online water army and normal users have the essential difference in the goal, causes the commodity commentary content characteristic and the external characteristic difference which published. From the user perceived factor perspective, commodity reviews have two results with perception reality or falsity, so identification problem of false comments can be converted to a two classification problem, this article from the user perceived factor angle to identify product reviews is true or false, so the existence of false comments on the goods can be used to determine the depth of user research methods, please ten in-depth customer to determine whether the product review is true or false, the result is 1 if 8-10 user perceptions are true, false is 0. Such information can be described in mathematical language as a training set:

\[ T = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subseteq (\mathbb{R}^n \times Y) \]
Where $x_i = ([x]_1, \cdots, [x]_n)^T \in \mathbb{R}^n$ is the input, $y_i \in \{0, 1\}$ is the output, $i = 1, \ldots, l$. Specifically, each sample point in the training set (1) represents a commodity with a total of $= 60$ items and is entered as $n = 21$ value attributes in Table 1, where $[x]_j$ means the $j$-th value attribute. $y = 1$ to express perception reality, $y = 0$ to express perception falsity.

In order to analyze the problem, writer constructed one model to practice logistic regression. Using LR forward step wise logistic regression analysis method, These three measures of "detailed degree", "three days before the number of comments", "photo show number" eventually enter the model of the independent variable. This shows that these three factors are the main considerations when users judge whether the commodity exists false comments, the results in Table 2. Classification table shows that in judging the existence of goods false comments correct rate is 82.1%. The correct rate of "1" that is commodity does not exist false comments is 87.5%, he overall percentage prediction accuracy was 85%, results are shown in Table 3.

<table>
<thead>
<tr>
<th>Step 3c</th>
<th>B</th>
<th>S.E,</th>
<th>Wals</th>
<th>df</th>
<th>Sig.</th>
<th>Exp (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detailed degree</td>
<td>-.384</td>
<td>.121</td>
<td>10.153</td>
<td>1</td>
<td>.001</td>
<td>.681</td>
</tr>
<tr>
<td>Three days before comments number</td>
<td>.323</td>
<td>.128</td>
<td>6.372</td>
<td>1</td>
<td>.012</td>
<td>1.381</td>
</tr>
<tr>
<td>Photo show number</td>
<td>-.262</td>
<td>.083</td>
<td>9.820</td>
<td>1</td>
<td>.002</td>
<td>.770</td>
</tr>
<tr>
<td>Constant</td>
<td>5.823</td>
<td>1.559</td>
<td>13.952</td>
<td>1</td>
<td>.000</td>
<td>338.102</td>
</tr>
</tbody>
</table>

a. Variables entered in step 1: photo show number.
b. Variables entered in step 2: detailed degree.
c. Variables entered in step 2: three days before comments number.
Table 3. False comments logistic regression classification table.

<table>
<thead>
<tr>
<th>Step 3</th>
<th>Already observed</th>
<th>Already observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classification</td>
<td>Percentage correction</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Classifica</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>tion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>Total percentage</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Cut value .500

Conclusion

Variance analysis showed that there are significant differences between the three perceived factor factors of "photo show number", "medium comments" and "additional comments number", which are important factors to distinguish true and false, therefore, the identification of genuine commodities should be reflected in these three aspects of the difference.

Users in judging whether the existence of commodity false comments mainly consider these three indicators of "detailed degree", "three before comments number", "photo show number". This behavioral characteristics and commercial purposes is also very consistent with the behavior of online water army and commercial purposes, Online water army will carefully describe and exaggerate all aspects of goods to guide customers to spend. Consumer online shopping decision-making reference to rely mainly on commodity reviews, no comments are greatly reduced consumer desire to buy, the online water army focused on the goods just bought and after receipt to make comments to induce consumers. Photo show number greatly increased the cost of user comments, according to Lotka Distribution only a small part of the user is willing to make a detailed and profound text description and map display, and online water army is the purpose of commercial operation, willing to make rich and detailed with map of the product reviews. Accordingly, when consumers online shopping can avoid such comments of many “photo show number”, very complete “detailed degree”, Very focused “three days before comments number”, so as to get more accurate commodity information, make the right shopping decisions.

Acknowledgment

This research was financially supported by the Central Universities and Hebei province Department of Education fund.

References


