LDA_RAD: A Spam review detection method based on topic model and reviewer anomaly degree

Zhou Ya1, Zhou Qingqing1, Wu Yuhan1 and Zhang Shuai1
1Guangxi Key Laboratory of Trusted Software, Guilin University of electronic technology, Guilin 541004, China
E-mail: zsi386@163.com. National Natural Science Foundation of China: (61662015)

Abstract. Online shopping has become a popular activity among consumers. Reviews have become important standards for consumers to buy things. To detect the spam reviews that do not correctly response the right information of goods, a method based on topic model and reviewer anomaly degree is raised. This method divides spam reviews into content-type and deceptive spam review respectively. Firstly, the experimental data set is modelled by LDA topic, which detects the content-type spam reviews with different themes. Then the deceptive spam reviews are detected by the reviewer’s abnormality degree index. It assigns a score to each review according to extracted features and related weights. A score is finally combined with an adaptive weight calculation based on the abnormal period and the reviewer's similarity to obtain the review score. The review with a high score is considered to be a spam review, and the review with a low score is a true review. Experiments show that this method has a certain improvement on the recognition rate of spam reviews.

1. Introduction
Most e-commerce platforms allow users to review on the products they purchase, to share the consumer's experience in the consumer process and to evaluate the quality of the product. These reviews have important guiding effects on the user's consumption process. Relevant statistics show that [1], about 81% of online shoppers in the United States will observe their product reviews before purchasing a certain product. The research issue of junk review was first proposed by Jindal and Liu [2, 3] in 2008. Some spam detection studies based on emotional polarity [4,5,15]. Ott et al. [7,8], Shojaee et al. [9] and Li [10] et al., use the data set constructed by Amazon Mechanical Turk, utilized word bag features, part-of-speech features and text features, etc. and applied support vector machine classifier to obtain 84%~89.6% accuracy in spam review detection. Baberjee [11] et al. conducted an exploratory study on the writing process of spam reviews. Chang [12] used text mining technology to extract three features of the review content, Liu et al. [13] proposed a spam review detection method based on user behavior and review text. Li [14] proposed a spam review detection method based on PU learning.

2. LDA-RAD method
There are four main stages of our spam detection based on LDA model and reviewer abnormality degree. Figure1 is the flowchart.
Firstly, processing review samples and corpus data and use uses the processed corpus data to train the LDA model. Next, examine if the length of processed corpus data is less the 7, if so, we directly extract the abnormality degree feature, otherwise, extracting the theme and then extracting the feature need to be implemented. Then, combine the calculated feature weight and reviews’ adaptive weight, the degree of review abnormality is obtained, through setting the threshold, fake reviews can be examined. Finally, the experiments can be evaluated by combing the LDA and the method above.

2.1. LDA-based content type spam review detection

The LDA topic model is a method of modelling the topic information of text data and the advantage is that the text can be distributed in an unsupervised manner. This article collected nearly 100,000 reviews from four fields in Jingdong, Dianping and Taobao to train LDA topic models.

Since the LDA topic model does not treat the review as a single topic set, but treats it as a combination of multiple topics, the output will be a topic string. The output of the LDA model is set to top 6 topic with the maximum probability. Since the experimental data set of this paper is the dietary review data, the six themes output by the model need to be filter. If the diet topic is included, it is judged that the review is a dietary field review, otherwise it is judged as other field category review, so the Content-based spam review can be obtained.

2.2. Reviewer abnormality degree extraction

In this section, the identification of spam review is performed from the reviewer's abnormality, and the seven-dimensional feature is extracted to calculate the reviewer's abnormality degree.

Maximum number of abnormal comments per day (F1). Since the consumption times for a reviewer are limited, the larger the maximum number of reviews per day, the more likely it is to do some spam review activities. Define the max-comments as $f$, totalmin and totalmax as the max- and min-value off. The formula is shown in (1).

$$f = \max \{ daily \_review \_j \}, d \in D$$
$$total_{min} = \min \{ f_i \}, i = 1, 2, ..., n$$
$$total_{max} = \max \{ f_i \}, i = 1, 2, ..., n$$
The \( f_i \) in the formula is the max-reviews for the \( i \)-th reviewer, and \( \text{daily\_review}_d \) is the number of reviews with date \( d \) in the reviewer's historical review. \( D \) is the set of all review times for the \( i \)-th reviewer, and \( n \) is the number of all reviewers. After normalization processing, F1 feature formula is shown in (2).

\[
F1 = \frac{f_i - \text{total}_i}{\text{total}_i \text{max} - \text{total}_i \text{min}}
\]

The maximum abnormal offset distance (F2). F2 is based on the reviewer's historical review data, which is the maximum of the ratio of the distance between the neighbouring merchants of the reviewer reviewed and the neighbouring review time interval. Because everyone has limited energy, normal reviewers will not move too far in a short period of time, so the larger the indicator, the more suspicious the user is. Define the \( i \)-th review’s maximum offset distance as \( l_i \), the formula is shown in (3).

\[
l_i = \max \left\{ \frac{\text{Dis}(r_j, r_{j+1})}{|r_j.t+j+1 - r_{j+1}.t|}, j = 1, 2, ..., m \right\}
\]

Where \( r_j \) represents the address of the \( j \)-th review merchant of the reviewer's historical review data, \( m \) is the number of reviews of the reviewer's history, and \( \text{Dis}(r_j, r_{j+1}) \) is the distance between the \( j \)-th review merchant and the \( (j+1) \)-th merchants the reviewer reviewed. This article uses the Baidu map API to get the distance between the two merchants. \( r_j.t \) is the time when the \( j \)-th review is published. The denominator \( |r_j.t+j+1 - r_{j+1}.t| \) is to prevent the neighboring reviews from being published on the same day, resulting in 0 exception.

Define \( \text{total}_i \text{min} \) as the minimum of the maximum offset distances of all reviewers. \( \text{Total}_i \text{max} \) is the maximum of the maximum offset distances of all reviewers. After normalization of \( l_i \), the F2 feature formula is shown as (4).

\[
F2 = \frac{l_i - \text{total}_i \text{min}}{\text{total}_i \text{max} - \text{total}_i \text{min}}
\]

It's perhaps that there have some abnormality (F3). That is, the deviation rate of whether the review has the behavior of the image upload and the historical behavior. The larger the indicator, the more suspicious the current behavior of the reviewer does not conform to historical habits. The formula to define this feature is shown in (5).

\[
F3 = 1 - \frac{\text{current}_i \text{num}}{n}
\]

Where \( n \) is the current reviewer’s all historical data, \( \text{current}_i \text{num} \) is the number of same image-within state of historical reviews and the current review (the data set has reviews with image-within information). If the current state does not conform to the habit, the F3 is larger.

Sentiment-score abnormality (F4). Sentiment-score bias is the degree of agreement between the sentiment polarity of the review and the overall score and the variance of the historical review. The definition \( \text{senti}_i \text{_hab}_i \) is the \( i \)-th reviewer's sentiment-scoring habit deviation, and its formula is shown as (6).

\[
\text{senti}_i \text{_hab}_i = \frac{1}{n} \sum_{j=1}^{n} (\delta \text{current}_i \text{_review} - \delta_j)
\]

Among them, \( \delta \) is the degree of agreement of the sentiment-score of the review, \( \delta_j \) is the \( j \)-th historical review of current review user, \( f_\text{total} \) is the full score, and \( \text{senti}_i \text{_score} \) is the Sentiment polarity of the review text. The formula for calculating \( \delta \) is shown in (7).

\[
\delta = \frac{\text{senti}_i \text{_score}}{f_\text{max} - \frac{1}{2} f_\text{total}}
\]

Define \( \text{total}_i \text{min} \) as the minimum of the maximum number of reviews for all reviewers, and \( \text{total}_i \text{max} \) normalizes \( \text{senti}_i \text{_hab}_i \) for the maximum of the maximum number of reviews for all reviewers. After normalization, the F4 feature formula is shown in (8).
The review length abnormality (F5). The length deviation of the review is the variance of the length of current review and the length of the historical review. The normal user can also give spam review due to the interest. If the variance of the current review length and the length of the historical review is too large, the user is considered to have written a spam review. Possibly, this review can be considered a potential spam review. The \( \text{review}_\text{length}_i \) is defined as the review length deviation of the \( i \)-th reviewer, and \( \gamma_j \) is the length of the \( j \)-th historical review, and the formula is as shown in (9).

\[
\text{review}_\text{length} = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (F1 - \gamma_j)^2}
\]

(9)

The \( \text{review}_\text{length}_{\text{min}} \) is defined as the minimum value of the maximum number of review of all the reviewers per day, and the \( \text{review}_\text{length}_{\text{max}} \) is the maximum value of the maximum number of reviews of all the reviewers per day. The \( \text{review}_\text{length}_i \) is normalized to obtain the F5 feature formula shown in (10).

\[
F5 = \frac{\text{review}_\text{length}_i - \text{review}_\text{length}_{\text{min}}}{\text{review}_\text{length}_{\text{max}} - \text{review}_\text{length}_{\text{min}}}
\]

(10)

Extreme rating behavior (F6). That is, the rating of the product in the reviewer is rated as the highest and lowest score. An extreme rating of the product indicates that the review has the potential to deliberately praise and discredit the product. The greater the proportion of this behavior, the more suspicious the reviewer is. The feature formula is shown in (11), where \( \text{extreme}_\text{num} \) is the number of reviews for extreme ratings.

\[
F6 = \frac{\text{extreme}_\text{num}}{n}
\]

(11)

Maximum review interval (F7). That is, the time interval between the reviewer’s adjacent historical reviews. If it is a normal user, its behavior of reviewing is spontaneous and the number of days is not too much. Spam reviewers use this account to only post reviews when they receive a mission to do spam-reviewing, so the time interval between adjacent reviews will be relatively large. The feature formula is shown in (12).

\[
F7 = \max |t_r - t_r - 1, A|
\]

(12)

2.3. Abnormal feature weight calculation

Here, the Fisher criterion is used to calculate the contribution degree of each feature to the spam review detection, and the Fisher value of the single feature is used as the weight of the feature. Let the data in the data set \( Y = \{y_1, y_2, ..., y_n\} \) be divided into two categories, \( n_1 \) data belong to category \( \gamma_1 \), \( n_2 \) data belong to classification \( \gamma_2 \), and \( n = n_1 + n_2 \), each data \( y_i \) contains \( m \)-dimensional features, \( S_{w} \) and \( S_{b} \) are the intra-class variances and inter-class variances of the \( k \)-th (\( k \in \{1, 2, ..., m\} \)) dimension features on the dataset, the formula is (13) and (14).

\[
S_{w} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (y_{ij} - \mu_{i}^j)^2
\]

(13)

\[
S_{b} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (\mu_{i}^j - \mu_{i}^j)^2
\]

(14)

Where \( y_{ij} \) is the value of data \( y_i \) in the \( k \)-th dimension, \( \mu_{i}^j \) is the mean of the \( i \)-th class of data in the \( k \)-th dimension, and \( \mu_{i}^j \) is the mean of all samples in the \( k \)-th dimension.

Then the Fisher criterion for a single feature represents Equation (15).

\[
F(k) = \frac{S_{w}^{-1}}{S_{w}^{-1} + S_{b}^{-1}}
\]

(15)

\( F(k) \) is expressed as the Fisher value of the \( k \)-th dimension. If the \( F(k) \) value is larger, it indicates that the dimension feature has a greater effect on the classification of the category, and can make a greater contribution to the classification work.
2.4. Adaptive weight calculation using abnormal period & reviewer similarity

This section will assign adaptive weights to the reviews based on the abnormal period and the reviewer's similarity.

2.4.1. Calculation of abnormality degree. For the sake of profit, merchants will buy services from someone to do spam-reviewing. And the server providers will issue tasks to spam-reviewers to generate large spam-reviews. By using EMA curves, the abnormal growth value in the time series can be obtained. EMA (Exponential Moving Average), the index average index, also known as the EXPMA indicator, is also a trend-oriented indicator. Assuming that the data set X={x1, x2,..., xn}, where n is the number of data, and the value of each point of the EMA curve of data set X is Yi, (16) is the formula.

\[ Y_i = \left\{ \begin{array}{ll}
  x_i & , i = 1 \\
  \alpha x_i + (1 - \alpha) Y_{i-1} & , i \geq 2 
\end{array} \right. \]  

(16)

It can be concluded that the specific formula of Yi is as shown in (17).

\[ Y_i = \alpha x_i + (1 - \alpha) x_{i-1} + (1 - \alpha)^2 x_{i-2} + ... \]  

(17)

Here \( \alpha = \frac{2}{n+1} \), it can be seen from the above equation that the most significant influence on Yi is the data closest to it, so the most recent three data are considered here.

2.4.2. Abnormal quantity growth (AQG). The outlier is the difference between the number of real reviews increased and the theoretical expected increase in each period of time. The larger the difference, the more abnormal the reviewer in this interval. In this paper, the difference between the existing curve value and the EMA curve value (Yi) is used as the calculation method, and the formula for calculating the deviation value of the i-th month is as shown in (18).

\[ AQG_i = |x_i - Y_i| \]  

(18)

In order to be easily understanding, here is a calculation of the flow of people for all the reviews of a store for three years, setting a month as a recording time interval, and the fluctuation of flow of people in the store as shown in Fig.2 (a).

2.4.3. Average score increase abnormally (ASA). The outlier is the difference between the average product score and the theoretical expected average score for each period of time. The larger the difference, the higher the average score is, and the more abnormal the reviewer is during this period, the difference between the existing curve value and the EMA curve value is used to calculate it. The calculation formula for the deviation value of the i-th month is as shown in (19).

\[ ASA_i = |x_i - Y_i| \]  

(19)

Based on the existing data, the average score curve is plotted. The time interval is also one month. The average score curve is shown in Fig.2 (b).

Reviewer similarity calculation. In the manufacture of fake reviews, the publishers of reviews are usually divided into two categories. One is to control multiple accounts by one person to make spam reviews. The other is a group of people collaboratively to publish spam reviews. Both types of reviewers
will inevitably produce the same batch of fake reviewers who will post spam reviews in two identical
merchants, which will result in the same rating merchants between different accounts.

In this paper, the Jaccard algorithm is used to calculate the similarity between reviewers, and the i-th and j-th reviewers are calculated similarly. Here, two \( u_i \) and \( u_j \) reviewers are set to review the merchants as \( p_i \) and \( p_j \), respectively. Then the calculation of the similarity between the two is the ratio of the number of intersections of the two sets to the number of unions of the two sets, and the formula is as shown in (20).

\[
\text{Jaccard} \left( u_i, u_j \right) = \frac{p_i \cap p_j}{p_i \cup p_j}
\]  

(20)

Since a reviewer may be similar to several different reviewers, \( \text{Sim} \_\text{score}_i \) of this paper takes the maximum value of Jaccard similarity, as shown in formula (21), where \( n \) is the number of reviewers.

\[
\text{Sim} \_\text{score}_i = \max \left\{ \text{Jaccard} \left( u_i, u_j \right) \right\}, j = 1, 2, ..., n
\]  

(21)

2.4.4. **Review adaptive weight.** This section assigns adaptive weights to reviews based on anomalous periods and reviewer similarities. The main idea is that one reviewer and another reviewer all have reviews in different stores. The more shops they have reviews together, the more suspicious the reviewers. If two reviewers who exist in the abnormal period simultaneously have a high degree of similarity, the degree of suspiciousness is greatly increased. In this paper, the form of product is used to evaluate the assignment of adaptive weights. The \( i \)-th reviewer dynamic weight \( \text{Dynamic} \_\text{weight}_i \) defines the product of the deviation and reviewer's similarity between the AQG and ASA periods, as shown in (22).

\[
\text{Dynamic} \_\text{weight}_i = \text{AQG} \times \text{ASA} \times \text{Sim} \_\text{score}_i
\]  

(22)

After normalization, we get \( \text{D}_\text{weight}_i \), as shown in (23).

\[
\text{D}_\text{weight}_i = \frac{\text{Dynamic} \_\text{weight}_i - \text{total} \_\text{dw}_\text{ana}}{\text{total} \_\text{dw}_{\text{ana}} - \text{total} \_\text{dw}_{\text{ana}}}
\]  

(23)

2.5. **Deceptive spam review detection based on reviewer anomaly degree**

This section describes the calculation process of the reviewer's abnormality degree in detail. Firstly, the result of multiplying the seven-dimensional feature and its corresponding feature weight is added to obtain the feature combination score of the reviewer; then the review adaptive weight is multiplied by the feature combination score of the reviewer to finally obtain the reviewer's abnormality degree. The score is used to represent the abnormality degree of the review. (24) shows the formula.

\[
\text{Outlier} \_\text{score}_i = \text{D}_\text{weight}_i \sum_{k=1}^{7} \left( F_i \times F(k) \right)
\]  

(24)

After all the reviews are evaluated for the degree of abnormality, the final result is represented by a value for each review, which is ranked from large to small according to the value. The higher of the ranking, the higher the score of the abnormality degree and more likely it is to be a spam review. Here, a threshold \( \theta \) is set, and the top \( n*\theta \) review is regarded as a spam review.

3. **Experimental results and analysis**

3.1. **Introduction to experimental data sets**

All experimental data in this article are from the user reviews of the public review network (http://www.dianping.com). The test data is shown in Table 1.

| The data type       | Number of data |
|---------------------|----------------|
| Non-spam review     | 3009           |
| Spam review         | 1113           |
| Historical review   | 301005         |
| Total number of comments | 305127       |
3.2. Determine the value of the threshold variable

The spam review determined by the LDA_RAD method are derived from the degree of abnormality, which seeks the best detection effect by adjusting the value of the threshold $\theta$. Since the experimental data between 0.1 and 0.2 can already be determined, the experimental data from 0 to 1 is not fully displayed. The evaluation indicators of the experimental results in this paper are precision, recall and $F$ values. The experimental display data was tested with an observation value of 0.01 from 0.1 to 0.2, and Fig. 3 shows the experimental performance effect, as shown below.

![Fig. 3. Comparison of experimental results under different thresholds](image)

As can be seen from the figure, the smaller the threshold, the higher the accuracy rate and the lower the recall rate. This indicates that most of the top reviews are spam reviews, and there are very few spam reviews. As the value of $\theta$ increases, the Precision value begins to decrease at a very high speed after 0.15, the Recall value tends to be gentle, and the $F$ value begins to decrease. Considering comprehensively, the $F$ value is used as the decisive reference standard, and the threshold value that maximizes the $F$ value is selected, so $\theta$ is selected to be 0.15.

3.3. Spam review detection comparison test.

This paper based on topic model and reviewer anomaly spam detection method (LDA_RAD) and spam review detection method based on reviewer anomaly (RAD), topic-based model and Fisher criterion-based spam detection method (LDA_Fisher) and a comparative experiment based on the custom bias and the XGBoost algorithm (HDXG) spam detection method. The thresholds of RAD, LDA_Fisher and LDA_RAD are both 0.15.

![Fig. 4. Comparison diagram of experimental results under different methods](image)

It can be seen from Fig. 4 that the LDA_Fisher method achieves an $F$ value of 0.87, which indicates that the seven-dimensional abnormality degree feature proposed in this paper can reveal the characteristics of spam reviews and has a good characterization of the feature of spam reviews. The RAD method achieved an $F$ value of 0.9. Comparing the $F$ value of 0.93 obtained by LDA_RAD, it indicates that the LDA topic model does have the ability to identify spam reviews. Some content-type spam reviews in the data set cannot be detected by the reviewer's abnormality score. In contrast, the experimental effect of the LDA_RAD method has a small increase in HDXG. The reason is analyzed: the LDA_RAD method provides a more in-depth analysis of the fake reviewer's behavior, and uses the
scientific period of the HDXG method more scientifically. The HDXG method only divides it by 0,1, and ignores its outbreak. The degree is different. The LDA_RAD method is calculated by the deviation value, and the difference can be provided in more detail. Therefore, the LDA_RAD method can better reflect the potential behavior of fake reviewers, and thus obtains better experimental results.

4. Conclusions
This paper proposes a spam review recognition method (LDA_RAD) based on topic model and reviewer abnormality degree. This method detects spam reviews from two aspects: content spam reviews and deceptive spam reviews. The experimental results show that the LDA model has the ability to detect content spam reviews, and the anomaly degree of reviewer plays a great role in the detection of deceptive spam reviews. Future work will focus on increasing the efficiency with large scale data and optimize the model dynamically.

References
[1] Kelsey group. Online consumer-generated reviews have significant impact on offline purchase behavior[R/OL]. Kelsey group, 2007.
[2] Jindal N, Liu B. Opinion spam and analysis[C]//Proceedings of the 2008 international conference on web search and data mining. California, USA, 2008: 219-230.
[3] Liu B. Sentiment Analysis and Opinion Mining[C]// Synthesis Lectures on Human Language Technologies 5.1 (2012): 1-167. Morgan & Claypool, 2012.
[4] Runa D, Zhang X, Zhai Y. Try to Find Fake Reviews with Semantic and Relational Discovery[C]//2017 13th International Conference on Semantics, Knowledge and Grids (SKG). Beijing, China: IEEE, 2017: 234-239.
[5] Wang C C, Day M Y, Chen C C, et al. Temporal and Sentimental Analysis of A Real Case of Fake Reviews in Taiwan[C]//Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. Sydney, Australia: ACM, 2017: 729-736.
[6] Ansari G, Ahmad T, Doja M N. Review ranking method for spam recognition[C]// Ninth International Conference on Contemporary Computing. Noida, India: IEEE, 2017:1-5.
[7] Ott M, Choi Yejin, Cardie C, et al. Finding deceptive opinion spam by anystretch of the imagination[C]// Proc of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon: 2011: 309-319.
[8] Ott, M., Cardie, C. & Hancock, J. (2013). Negative Deceptive Opinion Spam. Proceedings of NAACLHLT. pp. 497–501.d, USA, June 23-25 2014. ACL
[9] Shojaee S, Murad M A A, Azman A B, et al. Detecting deceptive reviews using lexical and syntactic features[C]// International Conference on Intelligent Systems Design & Applications, 2014.
[10] Li J, Ott M, Cardie C, Hovy E (2014) Towards a general rule for identifying deceptive opinion spam. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1566–1576, Baltimore, Maryland
[11] Banerjee S, Chua A Y K. Understanding the process of writing fake online reviews[C]//Ninith International Conference On Digital Information Management (ICDIM 2014). Phitsanulok, Thailand: IEEE, 2014: 68-73.
[12] Liu L, Zhao X, Wang H, et al. Research on Identification Method of Anonymous Fake Reviews in E-commerce[J]. 2016, 14(4):1510.
[13] Li H, Chen Z, Liu B, et al. Spotting Fake Reviews via Collective Positive-Unlabeled Learning[C]// 2014 IEEE International Conference on Data Mining (ICDM). Shenzhen, China: IEEE Computer Society, 2014: 899-904
[14] Chonglin Gu, Zhixiang He, Shi Chen, et al. Detecting spammers using review graph[J]. IJHPCN 2017, 10(4/5): 269-278.