Decision Support Systems 157 (2022) 113765

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A novel approach for fraudulent reviewer detection based on weighted topic modelling and nearest neighbors with asymmetric Kullback–Leibler divergence

Wen Zhang [a,*], Rui Xie [a], Qiang Wang [a], Ye Yang [b], Jian Li [a]

[a] College of Economics and Management, Beijing University of Technology, Beijing 100124, China
[b] School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030, USA

ARTICLE INFO

Keywords:
E-commerce
Fraudulent reviewer detection
Imbalanced data
Weighted LDA
Kullback–Leibler divergence

ABSTRACT

The task of detecting fraudulent reviewers is of great importance to *E*-commerce platforms. Existing research has invested much effort into developing comprehensive features and advanced techniques to detect fraudulent reviewers. However, most of these studies have ignored the data imbalance problem inherent in fraudulent reviewer detection: non-fraudulent reviewers are the majority, while fraudulent reviewers are the minority in real practice. To fill this gap, we propose a novel approach called ImDetector to detect fraudulent reviewers while handling data imbalance based on weighted latent Dirichlet allocation (LDA) and Kullback–Leibler (KL) divergence. Specifically, we develop a weighted LDA model to extract the latent topics of reviewers distributed on the review features. Asymmetric KL divergence is adopted to make the similarity measure between reviewers biased toward the fraudulent minority when using the *K*-nearest-neighbor for classification. By mapping the reviewers to the latent topics of features derived from the weighted LDA model and measuring the similarities between reviewers using asymmetric KL divergence, the data imbalance problem in fraudulent reviewer detection is alleviated. Extensive experiments on the Yelp.com dataset demonstrate that the proposed ImDetector approach is superior to the state-of-the-art techniques used for fraudulent reviewer detection. We also explain the experimental results and present the managerial implications of this paper.

1. Introduction

Online reviews play an important role in shaping consumers' purchase decisions and affecting product sales. On the one hand, online reviews affect customers' online purchasing behaviors and purchase decisions to a great extent. According to a survey by PowerReviews, 95% of consumers read online reviews, and 86% consider online reviews important in purchase decision making [1]. In 2010, 67% of consumers read online reviews on *E*-commerce platforms; in 2020, this percentage increased to 87% [2]. Consumers' purchase intention for an online product with five online reviews was 270% greater than that without online reviews [3]. On the other hand, the revenue of firms is positively correlated with review ratings. Existing studies have shown that a one-star increase in review rating can increase the revenue of a restaurant by 5%–9% [4], and that a restaurant would increase more than 50% of reservations with an extra half-star [5].

The surge of review manipulation on *E*-commerce platforms has resulted in the prevalence of online fraudulent reviews. The Washington Post reported that over 61% of online reviews for electronics on Amazon.com are fake [6]. Although online platforms have invested much in fighting against fraudulent reviews, this problem remains unresolved in online shopping. A recent survey has shown that 80% of consumers have encountered a fraudulent review and that 33% of them have spotted multiple reviews [2]. Fraudulent reviews mislead consumers; in fact, they cause each US consumer to waste $125 on average in online shopping [7]. Although Amazon.com spends more than $500 million and employs more than 8000 employees each year to fight against online review fraud [8], 15%–30% of online reviews are still manipulated by fraudulent reviewers on the platform [4].

The problem of review fraud detection is divided into three categories: fraudulent review detection [9–12], fraudulent reviewer detection [13–15], and fraudulent reviewer group detection [16–18]. The

* Corresponding author.
E-mail addresses: zhangwen@bjut.edu.cn (W. Zhang), xierui@emails.bjut.edu.cn (R. Xie), wangqiang@emails.bjut.edu.cn (Q. Wang), ye.yang@stevens.edu (Y. Yang), lijiansem@bjut.edu.cn (J. Li).

https://doi.org/10.1016/j.dss.2022.113765
Received 18 July 2021; Received in revised form 18 January 2022; Accepted 22 February 2022
Available online 1 March 2022

current study falls into the second category of fraudulent reviewer detection. In the current stage, researchers are striving to develop review features to characterize anomalies in review content and reviewer behavior patterns to detect fraudulent reviewers [15]. For example, Ghose and Ipeirotis make use of the parts of speech in online reviews, such as adjectives, prepositions, and nouns, to detect fraudulent reviewers [19]. Wu uses reviewer behavior patterns, such as review time, review valence, and review volume, to detect fraudulent reviewers [20]. Hooi et al. [20] and Rayana et al. [21] employ heterogeneous network analysis in reviews, reviewers, and products to identify fraudulent reviewers through belief cascade and trust propagation.

Most of the existing studies ignore the problem of data imbalance inherent in fraudulent reviewer detection—non-fraudulent reviewers are the majority, while fraudulent reviewers are the minority—because the number of legitimate reviewers is far greater than that of fraudulent reviewers on *E*-commerce platforms. Kumar et al. [15] find that 20% of reviewers are fraudulent, while 80% are non-fraudulent in the Yelp.com dataset. Nevertheless, methods for fraudulent reviewer detection suffer from the problem of data imbalance. On the one hand, the classifier is prone to classifying unlabeled reviewers as non-fraudulent rather than fraudulent because it has less knowledge about fraudulent reviewers than non-fraudulent reviewers due to the skewed distribution of the two classes [21]. Thus, traditional methods of fraudulent reviewer detection will be overwhelmed by the majority, thus making the model extremely biased toward non-fraudulent reviewers [22]. On the other hand, for online consumers, the cost of misclassification of fraudulent reviewers as non-fraudulent is much larger than that of misclassification of non-fraudulent reviewers as fraudulent. The reason is that in the former case, online consumers are deceived by fraudulent reviewers and thus make wrong purchasing decisions, resulting in a disappointing online shopping experience. In the latter case, online consumers disregard these misclassified non-fraudulent reviewers and believe in other non-fraudulent reviewers. Therefore, it is crucial to deal with the problem of imbalanced data in fraudulent reviewer detection for *E*-commerce to succeed.

Studies have been conducted to address the problem of data imbalance, and they can be roughly divided into three categories: data-processing method, feature selection method, and algorithmic-centered method [23]. Nevertheless, some problems still need to be resolved. For the data-processing method, oversampling techniques result in overfitting due to the overduplication of samples in the minority class, and undersampling techniques result in important information loss due to the elimination of samples in the majority class [23,24]. The feature selection method may suffer from trivial features and cause a significant performance drop [33,34]. For the algorithmic-centered method, post-adaptation produces a large error for the samples in the majority class, in that post-processing will make the trained classifiers biased toward the minority class. Cost-sensitive learning brings about great difficulty in setting misclassification costs between the minority class and the majority class [14].

To fill this research gap, we propose a novel approach called ImDetector to fraudulent reviewer detection with data imbalance based on the weighted latent Dirichlet allocation (LDA) model and *K*-nearest-neighbor (KNN) with asymmetric Kullback–Leibler (KL) divergence. Specifically, the weighted LDA model is proposed to model the latent topics of both fraudulent and non-fraudulent reviewers distributed in the review features. The asymmetric KL divergence is adopted to make the similarity measure between reviewers biased toward the fraudulent minority with the KNN classifier. By mapping the reviewers to the latent topics of features derived from the weighted LDA model and measuring the similarities between reviewers using asymmetric KL divergence, the data imbalance problem in fraudulent reviewer detection is alleviated. Extensive experiments on the Yelp.com dataset demonstrate that the proposed ImDetector approach is superior to state-of-the-art techniques in fraudulent reviewer detection.

The advantage of the proposed ImDetector approach lies in at least three aspects. The first aspect is that the model training of the proposed ImDetector approach is completely dependent on the existing data, rather than on the adoption of sampling techniques to deal with the imbalanced data. Thus, it can avoid possible overfitting and information loss in detecting fraudulent reviewers with imbalanced data. The second is that the proposed ImDetector approach makes use of the latent topics inherent in imbalanced data for reviewer representation, rather than adopting feature selection techniques to construct a partial set of features biased to the minority class. Thus, it can avoid trivial features and improve the robustness of fraudulent reviewer detection with imbalanced data. The third aspect is that the proposed ImDetector approach makes the KNN classifier with an asymmetric KL divergence biased toward the minority class for fraudulent reviewer identification. Therefore, it is neither necessary to post-process the classifier nor to set the costs for the minority and majority classes. That is, it avoids large errors in the majority class due to post-adaptation and the possible inappropriate setting in the costs of cost-sensitive learning. We admit that the proposed ImDetector approach falls into the stream of algorithmic-centered methods for dealing with imbalanced data. However, we argue that it is different from existing methods in that it addresses imbalanced data using a sample representation with the weighted LDA model and a sample classification by KNN with the asymmetric KL divergence rather than classifier adaptation and cost-sensitive learning.

This study contributes three aspects. First, we present a review of existing studies on the classification of imbalanced data and fraudulent reviewer detection. We show that the three source features—review content feature (RCF), reviewer behavior feature (RBF), and review target feature (RTF)—are commonly used for review fraud detection. Second, we propose the ImDetector approach for fraudulent reviewer detection with imbalanced data. We use the weighted LDA model to extract latent topics biased toward the minority reviewers for representation and the KNN classifier with the asymmetric KL divergence biased toward the majority reviewers for classification. Third, we conduct extensive experiments on the Yelp.com dataset to compare the proposed ImDetector approach with state-of-the-art techniques in fraudulent reviewer detection.

The rest of the paper is organized as follows. Section 2 presents the literature review. Section 3 proposes the ImDetector approach for fraudulent reviewer detection. Section 4 conducts the experiments. Section 5 concludes the paper.

2. Literature review

2.1. Fraudulent reviewer detection

With the rapid development of *E*-commerce, online review fraud has become a serious problem [10,25]. Existing studies on fraudulent reviewer detection can be summarized into two aspects. The first aspect is the development of comprehensive features of the review contents, reviewer behaviors, and review targets to characterize the anomalies of reviewers [15]. For example, Kumar et al. [14] derive the distribution of key RBFs to reflect the behavioral differences between non-fraudulent and fraudulent reviewers. They combine this heterogeneous distribution into a unified model to detect fraudulent reviewers. Zhang et al. [11]explore the detection approach for fraudulent reviews from multi-view features. These features can be either lexical features derived from the natural language texts of the reviews posted by the reviewer (i.e., verbal features) or behavioral features (i.e., non-verbal behavioral features) of the reviewer in promulgating online reviews. Rayana et al. [21] utilize clues from review targets, review texts, and review ratings, as well as their relational network, to collectively spot suspicious reviewers and products in a holistic manner under a unified framework.

The second aspect is the development of advanced techniques based on machine learning and network analysis to detect fraudulent reviewers [26]. For example, Wang et al. [18] exploit the topological structure of the underlying reviewer graph to detect fraudulent reviewer

groups. They use bi-connected graphs to model the relationship among reviewers and develop spammer indicators to measure the maliciousness of a spammer group. Kumar et al. [15] propose an unsupervised hierarchical method by combining the univariate and multivariate distributions of reviewer behaviors using a finite mixture model to detect anomalous online reviewers. Wang et al. [17] consider network effects and time effects in co-reviewing behaviors of reviewers using a pairwise Markov random field to characterize group spamming for fraudulent reviewer group detection.

However, most existing studies ignore the problem of data imbalance inherent in fraudulent reviewer detection. That is, in real practice, non-fraudulent reviewers are the majority, while fraudulent reviewers are the minority because the number of non-fraudulent reviewers is much more than the number of fraudulent reviewers. The disregard for data imbalance will make the classifier biased toward the majority class because it has less knowledge about the minority class. Therefore, a large number of false-negative samples (i.e., type I error [27]) will emerge if the trained classifier is employed to detect fraudulent reviewers. Consequently, fraudulent reviewers will be identified as non-fraudulent, thus misleading online consumers in their online purchase decision making.

2.2. Classification of imbalanced data

The problem of classifying imbalanced data has attracted great interest from researchers of data mining and machine learning due to its widespread applications in many fields, such as fraud detection in online banking [28], software defect prediction [29], and bioinformatics [30]. We can roughly divide existing studies to address this problem into three categories: data-processing methods, feature selection methods, and algorithmic-centered methods.

The data-processing methods use sampling techniques to rebalance the imbalanced dataset using either oversampling or undersampling. The basic idea of oversampling is to increase the size of the minority class to rebalance it with the majority class by randomly duplicating the samples in the minority class. For example, Chawla et al. propose the synthetic minority oversampling technique (SMOTE) by generating synthetic samples using minority class samples [31]. Zhang and Li introduce oversampling with random walk for classifying imbalance by generating samples to increase the number of samples in the minority class [32]. However, the oversampling method may enable the classifier to learn more specific information about the minority class, as the samples of the minority class are duplicated. This causes a large amount of overfitting in the learning process due to excessive information about the minority class caused by duplication [23].

The main idea of undersampling is to randomly remove samples from the majority class to rebalance it with the minority class. For example, Galar et al. propose an enhanced ensemble approach with random undersampling and a boosting algorithm to handle the classification with imbalanced data [33]. Yu et al. present ant colony optimization-based undersampling to address the imbalance data distributions in the DNA micro-array dataset [34]. However, undersampling may enable the classifier to disregard the samples removed from the majority class in the learning process, causing a great loss of information from the dataset in the training phase due to the removal of real instances [23].

Feature selection methods use dimensionality reduction techniques to select features in favor of the identification of the minority class in an imbalanced dataset. For example, Maldonado et al. propose a backward elimination approach for successively selecting relevant features for target class identification with a rebalanced loss function [35]. Yin et al. divide the majority class into small pseudo-subclasses and conduct feature selection on the new decomposed data for computing the goodness measurement of features [36].

Generally, the feature selection method for imbalanced data classification uses only a partial set of all the features of the dataset for classification, and this partial set of features is decided mostly by the minority class. Therefore, it may induce a large number of trivial features in the partial set. In the training data, these trivial features occur frequently in the minority class and infrequently in the majority class. However, in test data, these trivial features may not occur in the minority class but rather in the majority class due to the much smaller size of the minority class compared with the majority class. As a result, samples in the majority class can be misclassified as those in the minority class, significantly decreasing performance.

Algorithmic-centered methods include post-adaptation and cost-sensitive learning. The former adapts the parameters of trained classifiers, such as the boundary of neural networks [37] and the hyperplane of the support vector machine (SVMs) [38], to overdrive the classifier toward the minority class through output compensation with post-processing. For example, Du et al. propose a novel imbalance learning method for binary classification called post-boosting of classification boundary for imbalanced data (PBI) to readjust the classification boundary of trained neural networks [37]. Imam et al. introduce z-SVM to improve imbalanced classification by making the decision boundary of SVM skewed toward the minority class [38].

The latter uses class-specific costs associated with the misclassified samples to give more loss to the misclassification of the minority samples than that of the majority samples. For example, Cao et al. propose a wrapper framework to incorporate the evaluation measure into the objective function of cost-sensitive SVM to simultaneously optimize the collocation of feature subsets, intrinsic parameters, and misclassification cost parameters [39]. Khan et al. propose a cost-sensitive deep neural network to learn robust feature representations for both the majority and minority classes while jointly optimizing the class-dependent costs and the neural network parameters [40]. Zhang et al. propose feature weighted confidence to incorporate prior knowledge into SVM for cost-sensitive learning [41]. They use prior features to express prior knowledge and make SVM biased to assign larger costs of the prior features in the slope vector than those of the non-prior features.

However, there are also some drawbacks to algorithmic-centered methods in classifying imbalanced data. The post-adaptation method usually causes a large error for the samples in the majority class due to its post-processing of the parameters of the trained classifiers toward the minority class. For the cost-sensitive learning method, it is difficult to set the misclassification costs between the minority class and the majority class because the costs are introduced to the classifier in an ad hoc manner by human beings rather than learned purely from the training data [42].

3. The proposed approach

3.1. Overview of the ImDetector approach

Fig. 1 illustrates an overview of the proposed ImDetector approach for fraudulent review detection. First, the training reviewers are fed into the weighted LDA model to extract the latent topics of features. Second, both the training reviewers and test reviewers are represented by the extracted latent topics according to the features occurring in them. Third, the KNN classifier with KL divergence is used to find the predefined number of similar reviewers given a test reviewer and label it as either fraudulent or non-fraudulent by major voting. We present the details of the weighted LDA model and the KNN classifier with KL divergence in the following sections.

3.2. The weighted LDA model

Assume that we have a collection of online reviewers, $R = \{r_1, \ldots, r_i, \ldots, r_{|R|}\}$, and each reviewer r_i in R can be represented as its feature set as $r_i = \{f_{i1}, \ldots, f_{ij}, \ldots, f_{i|r_i|}\}$. Here, a feature f_{ij} of the reviewer r_i is defined as one of the predictive features (i.e., RCFs, RBFs, and RTFs) used in the paper (see Table 5 in Section 4.2). On the whole, all the features contained in the reviewers R constitute a feature vocabulary F for reviewer

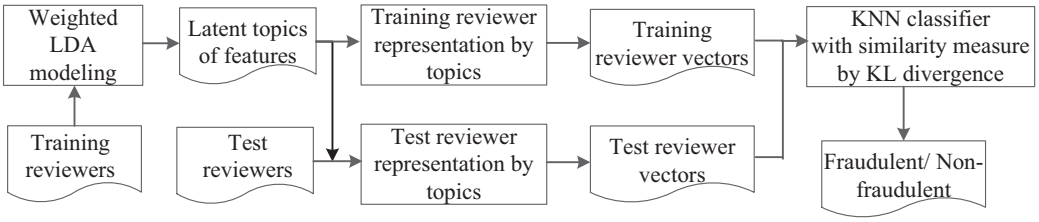

Fig. 1. Overview of the proposed ImDetector approach for fraudulent reviewer detection.

representation. Note that in this study, we do not develop new features and just follow the existing studies [14,15,26] to adopt their features of the reviews, reviewers, and targets being reviewed for fraudulent reviewer detection.

With the co-occurrences of features in the reviewers, we can discover the latent topics of features from the reviewers. Specifically, let Z_{ij} be the latent topic of feature f_j in reviewer r_i. That is, $Z_{ij} = topic_k$ means the topic of the feature f_j in reviewer r_i is $topic_k$. Then, ϕ_k is a multinomial feature–topic distribution that describes the probability of each feature appearing in the latent topic$topic_k$, i.e., ϕ_k=$\{p_1^k,\ldots,p_j^k,\ldots,p_{|F|}^k\}$, where p_j^k is the probability of feature f_j in the latent topic $topic_k$. Note that f_j is the jth feature in the feature vocabulary F, and f_{ij} refers to the specific feature f_j in the reviewer r_i. For the whole collection of reviewers R, we set the number of topics as K, i.e., $1 \le k \le K$.

Furthermore, the parameterθ_i is another multinomial topic–reviewer distribution of topics that describes the probability of each topic $topic_k$ occurring in the reviewer r_i, i.e., $\theta_i = \{p_1^i,\ldots,p_k^i,\ldots,p_K^i\}$, where p_k^i is the probability of topic $topic_k$ in the reviewer r_i. Without a loss of generality, the feature–topic distribution ϕ_k and the topic–reviewer distribution θ_i conform to Dirichlet distributions $D(\alpha)$ and $D(\beta)$, where α and β are two hyperparameters. The larger the value of parameter α, the greater the probability that reviewer r_i is associated with multiple topics $topic_k$. The larger the value of parameter β, the greater the probability that topic $topic_k$ is associated with multiple features f_j. For convenience, the notation is summarized in Table 1, and the graphical model representations of the weighted LDA model are shown in Fig. 2.

As shown in Fig. 2, the difference between the weighted LDA model and the classic LDA model (see Fig. 1 in Blei et al. [43]) lies in the observed variable $c(l) \cdot f_{ij}$, where $c(l)$ is the weight of the reviewer r_i in class l and f_{ij} is an observed feature f_j from the reviewer r_i. That is, in the classic LDA model, the observed variable is a word in a document. However, in the weighted LDA model, the observed variable is the feature f_j in the reviewer r_i with the class label l. In practice, we define the class-dependent function $c(l)$ in Eq. (1) as follows:

Table 1
Notations used in the weighted LDA model.

Symbol	Description
R	Set of all the reviewers.
r_i	ith reviewer in the reviewer set R.
F	Set of all the features.
f_{ij}	jth feature of reviewer$r_i$and $f_{ij} \in F$.
$topic_k$	kth latent topic of features.
z_{ij}	Latent topic associated with the jth feature in reviewer r_i.
ϕ_k	Multinomial distribution of features with respect to topic $topic_k$.
θ_i	Multinomial distribution of topics with respect to reviewer r_i.
n_{ij}^k	Number of feature f_j in reviewer r_i assigned to the topic $topic_k$.
$n_{i,(\cdot)k,-(i,j)}$	Number of features in reviewer r_i assigned to the topic $topic_k$without considering the featuref_{ij}.
$n_{(\cdot),jk,-(i,j)}$	Number of occurrences of the featuref_j assigned to the topic $topic_k$without considering the featuref_{ij}.
α	Hyperparameter for the Dirichlet distribution of features with respect to latent topics.
β	Hyperparameter for the Dirichlet distribution of topics with respect to reviewers.
γ	Ratio of the proportion of the majority class to the proportion of the minority class in the dataset.

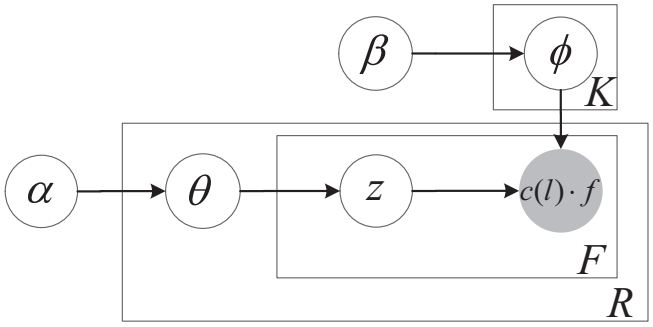

Fig. 2. Graphical model representation of the weighted LDA model.

$$c(l_i) = \begin{cases} 1, if\ r_i\ belongs\ to\ the\ majority\ class. \\ \gamma, if\ r_i\ belongs\ to\ the\ minority\ class. \end{cases} \quad (1)$$

As the ratio γ is larger than one, we give larger weights to the features in the reviewers of the minority class than in those of the majority class. By combining the weights of the reviewers of different classes and the actual occurrences of features in the reviewers, we handle the data imbalance problem by developing a weighting feature generation process (Fig. 3); this is the main idea of the proposed weighted LDA model. We argue that the weighting scheme is devised at the feature level to augment the weight of the features in the minority class rather than at the reviewer level to augment the weight of the reviewers.

Existing studies, such as Moreo et al. [44], Chawla et al. [31], and Pereira et al. [45] propose the random sampling techniques such as oversampling and undersampling to rebalance imbalanced data in text classification. However, we claim that the weighted LDA model proposed in this study is completely different from random sampling techniques for data rebalancing. The methods proposed by existing studies [29,42,43] duplicate the data samples of the minority class or prune the data samples of the majority class randomly to rebalance the data samples in the different classes. If the sampling method is adopted here, a feature f_{ij} of the different copies of the reviewer r_i can be assigned to different topics, and the generative process of the feature $f_{ij}$is

1. For each topic $topic_k$ $k \in \{1,\ldots,K\}$:
2. Draw a feature-topic distribution $\phi_k \mid \beta \sim Dirichlet(\beta)$
3. For each reviewer r_i in:
4. Let the frequency of all the features in r_i be dependent on its class label, i.e., $n_{i,(\cdot)} \leftarrow c(l_i) \cdot n_{i,(\cdot)}$
5. Draw a topic-reviewer distribution for r_i, i.e., $\theta_i \mid \alpha \sim Dirichlet(\alpha)$
6. For each feature f_j in r_i :
7. Draw a topic $topic_k \mid \theta_i \sim Multinomial(\theta_i)$
8. Draw the feature f_j with the feature-topic distribution ϕ_k of $topic_k$ $f_j \mid \phi_k \sim Multinomial(\phi_k)$

Fig. 3. The generative process of the weighted LDA model.

stochastic. That is, the topic of the feature f_{ij} in the reviewer r_i is drawn multiple times and may be assigned to different topics. Nevertheless, in the proposed weighted LDA model, we augment the frequencies of the features of the reviewers in the minority class using the class-dependent function $c(l)$. Thus, the topic of a feature f_j in the reviewer r_i is drawn only once, and after the generative process, the topic of the feature f_{ij} is deterministic.

We argue that the stochastic generative process could produce a large amount of noise to assign the topic of the feature f_j in the reviewer r_i because the assignments at different times could be completely different, thus evenly distributing the feature f_j to all the topics. By contrast, in the generative process of the weighted LDA model, the occurrence of the feature f_j in the reviewer r_i is weighted by $c(l_i)$, and the weight $c(l_i)$ sharpens the feature–topic distribution of the features f_j in the minority class to make them more discriminative in the classification. Essentially, the methods proposed by Moreo et al. [44], Chawla et al. [31], and Pereira et al. [45] also weight the data samples in the dataset using sampling techniques. However, the weighted LDA model directly weights the features rather than the data samples in the generative process. We also compare the methods proposed by Moreo et al. [44], Chawla et al. [31], and Pereira et al. [45] with the weighted LDA model in fraudulent reviewer detection in the experiments.

3.3. Inference of the weighted LDA model

The parameters that need to be inferred from the weighted LDA model include the feature–topic distribution ϕ_k and the topic–reviewer distribution θ_i. According to the generative process of the weighted LDA model, the total probability of the model is described in Eq. (2).

$$P(F,Z,\theta,\phi;\alpha,\beta,l)=\prod_{k=1}^{K}P(\phi_k;\beta)\prod_{i=1}^{|R|}P(\theta_i;\alpha)\prod_{j=1}^{|F|}P(Z_{i,j}|\theta_i)P\left(c(l_i)F_{i,j}\,|\phi_{Z_{i,j}}\right) \tag{2}$$

First, we integrate out the parameters ϕ_k and θ_i in Eq. (3) for further computation.

$$\begin{aligned}P(F,Z;\alpha,\beta,l)&=\int_i\int_k P(F,Z,\theta,\phi;\alpha,\beta,l)\,d\phi d\theta\\&=\int_\phi\prod_{k=1}^{K}P(\phi_k;\beta)\prod_{i=1}^{|R|}\prod_{j=1}^{|F|}P\left(c(l_i)F_{i,j}\,|\phi_{Z_{i,j}}\right)d\phi\int_\theta\prod_{i=1}^{|R|}P(\theta_i;\alpha)\prod_{j=1}^{|F|}P(Z_{i,j}|\theta_i)d\theta\end{aligned} \tag{3}$$

As the parameters ϕ and θ are independent of each other, we can integrate them out separately. Taking the parameter θ as an example, the integration is described in Eq. (4).

$$\int_\theta\prod_{i=1}^{|R|}P(\theta_i;\alpha)\prod_{j=1}^{|F|}P(Z_{i,j}|\theta_i)d\theta=\prod_{i=1}^{|R|}\int_\theta P(\theta_i;\alpha)\prod_{j=1}^{|F|}P(Z_{i,j}|\theta_i)d\theta_i \tag{4}$$

Second, we can derive Eq. (5) from the integration of Eq. (4). Here, Γ is a gamma function, where $\Gamma(a)=(a-1)!$ if a is an integer and $\Gamma(a)=\int_0^\infty x^{a-1}e^{-x}dx$ if a is a non-integer real number. With this definition, we can obtain $\Gamma(a)=a\Gamma(a-1)$. $P(\theta_j;\alpha)$ is a Dirichlet distribution with the hyperparameter a.

$$\int_{\theta_i}P(\theta_i;\alpha)\prod_{j=1}^{|F|}P(Z_{i,j}|\theta_i)d\theta_i=\int_{\theta_i}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_i\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{i,k}^{\alpha_k-1}\prod_{j=1}^{|F|}P(Z_{i,j}|\theta_i)d\theta_i \tag{5}$$

According to the generative process of the weighted LDA model illustrated in Fig. 3, we define $c(l_i)\cdot n_{i,jk}$ as the times that the feature f_{ij} is assigned to $topic_k$, $c(l_i)\cdot n_{i,(\cdot)k}$as the sum of all the features in r_i assigned to $topic_k$, and $c(l_i)\cdot n_{(\cdot),jk}$as the times that the feature f_j is assigned to $topic_k$ among all the reviewers. As $P(Z_{i,j}|\theta_i)$ is a multinomial distribution, it can be rewritten as Eq. (6), where $\theta_{i,k}^{c(l_i)n_{i,(\cdot)k}}$ is the product of the probabilities of the features in r_i assigned to $topic_k$. Note that $c(l_i)\cdot n_{i,(\cdot)k}$ is not necessarily an integer, but this will not affect the computability of $P(Z_{i,j}|\theta_i)$.

$$P(Z_{i,j}|\theta_i)=\prod_{k=1}^{K}\theta_{i,k}^{c(l_i)n_{i,(\cdot)}^{k}} \tag{6}$$

Third, we can further integrate θ_j from Eqs. (5)–(7) with the product of all the features in the feature set F.

$$\int_{\theta_i}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{i,k}^{\alpha_k-1}\prod_{k=1}^{K}\theta_{i,k}^{c(l_i)n_{i,(\cdot)}^{k}}d\theta_i=\int_{\theta_j}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{i,k}^{c(l_i)n_{i,(\cdot)}^{k}+\alpha_k-1}d\theta_i \tag{7}$$

According to the mathematical property of a Dirichlet distribution, we can obtain $\int_{\theta_i}\frac{\Gamma\left(\sum_{k=1}^{K}c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}\prod_{k=1}^{K}\theta_{i,k}^{c(l_i)n_{i,(\cdot)}^{k}+\alpha_k-1}d\theta_i=1$. Thus, we can rewrite Eq. (7) into Eq. (8).

$$\begin{aligned}\int_{\theta_i}P(\theta_i;\alpha)\prod_{j=1}^{|F|}P(Z_{i,j}|\theta_i)d\theta_i&=\int_{\theta_i}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{i,j}^{c(l_i)n_{j,(\cdot)}^{k}+\alpha_k-1}d\theta_i\\&=\frac{\Gamma\left(\sum_{k=1}^{K}c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}\int_{\theta_j}\frac{\Gamma\left(\sum_{k=1}^{K}c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}\prod_{k=1}^{K}\theta_{i,j}^{c(l_i)n_{j,(\cdot)}^{k}+\alpha_k-1}d\theta_j\\&=\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\Gamma\left(\sum_{k=1}^{K}c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}\end{aligned} \tag{8}$$

By analogy, the integral, with respect to ϕ in Eq. (5), can be rewritten as Eq. (9), where $c(l_t)n_{t,j}^{k}$ denotes the weighted occurrences of f_{tj} in the reviewer r_t, and $\phi_{k,j}^{\sum_{t=1}^{|R|}c(l_t)n_{t,jk}}$ is the product of the probabilities of f_j assigned to $topic_k$ among all the reviewers. Using the mathematical property of a Dirichlet distribution [46], we can obtain $\int_{\phi_k}\frac{\prod_{j=1}^{|F|}\Gamma\left(\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k}+\beta_j\right)}{\Gamma\left(\sum_{j=1}^{|F|}\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k}+\beta_j\right)}\prod_{j=1}^{|F|}\phi_{k,j}^{\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k}+\beta_j-1}d\phi_k=1$.

$$\int_{\phi}\prod_{k=1}^{K}P(\phi_k;\beta)\prod_{i=1}^{|R|}\prod_{j=1}^{|F|}P\left(c(l_i)f_{ij}\,|\phi_{z_{i,j}}\right)d\phi=\prod_{k=1}^{K}\int_{\phi_k}P(\phi_k;\beta)\prod_{i=1}^{|R|}\prod_{j=1}^{|F|}P\left(c(l_i)f_{ij}\,|\phi_{z_{i,j}}\right)d\phi_k$$

$$=\prod_{k=1}^{K}\int_{\phi_k}\frac{\Gamma\left(\sum_{j=1}^{|F|}\beta_j\right)}{\prod_{j=1}^{|F|}\Gamma(\beta_j)}\prod_{j=1}^{|F|}\phi_{k,j}^{\beta_j-1}\prod_{j=1}^{|F|}\phi_{k,j}^{\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k}}d\phi_k=\prod_{k=1}^{K}\int_{\phi_k}\frac{\Gamma\left(\sum_{j=1}^{|F|}\beta_j\right)}{\prod_{j=1}^{|F|}\Gamma(\beta_j)}\prod_{j=1}^{|F|}\phi_{k,j}^{\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k}+\beta_j-1}d\phi_k$$

$$=\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{j=1}^{|F|}\beta_j\right)}{\prod_{j=1}^{|F|}\Gamma(\beta_j)}\frac{\prod_{j=1}^{|F|}\Gamma\left(\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k}+\beta_j\right)}{\Gamma\left(\sum_{j=1}^{|F|}\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k}+\beta_j\right)} \tag{9}$$

With Eqs. (8) and (9), we can derive Eq. (10) by substituting them into Eq. (3). That is, given the hyperparameters α and β and the class labels l for all the reviewers, we can obtain the total probability of all the features in the feature set F of the reviewer set R assigned to their corresponding topics in the generative process (Fig. 3).

$$P(F,Z;\alpha,\beta,l)=\prod_{i=1}^{|R|}\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\Gamma\left(\sum_{k=1}^{K}c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(c(l_i)n_{i,(\cdot)}^{k}+\alpha_k\right)}\times\prod_{k=1}^{K}\frac{\Gamma\left(\sum_{j=1}^{|F|}\beta_j\right)}{\prod_{j=1}^{|F|}\Gamma(\beta_j)}\frac{\prod_{j=1}^{|F|}\Gamma\left(\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k}+\beta_j\right)}{\Gamma\left(\sum_{j=1}^{|F|}\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k}+\beta_j\right)} \tag{10}$$

Following the idea of the classical LDA model proposed by Blei et al. [43], we also adopt Gibbs sampling to estimate the parameters ϕ and θ by observing the latent topic z_{ij} associated with each feature f_{ij} in the reviewer r_i in the generative process. Specifically, let $P(Z_{i,j}=topic_k|Z_{-(i,j)},F;\alpha,\beta,l)$ in Eq. (11) be the marginal probability of the feature f_{ij} in the reviewer r_i assigned to $topic_k$ on the condition that the feature f_{ij} under investigation is not considered.

$$P\left(Z_{i,j}=topic_k|Z_{-(i,j)},F;\alpha,\beta,l\right)=\frac{P\left(Z_{j,i}=topic_k,Z_{-(i,j)},F;\alpha,\beta,l\right)}{P\left(Z_{-(i,j)},F;\alpha,\beta,l\right)} \tag{11}$$

For all the possible $topic_k$ ($1\le k\le K$) in Eq. (11), the denominator $P(Z_{-(i,j)},F;\alpha,\beta,l)$ is a constant, thus making the marginal probability $P(Z_{i,j}=topic_k|Z_{-(i,j)},F;\alpha,\beta,l)\propto P(Z_{i,j}=topic_k,Z_{-(i,j)},F;\alpha,\beta,l)$. Further, $P(Z_{i,j}=topic_k,Z_{-(i,j)},F;\alpha,\beta,l)$ can be computed as Eq. (12).

Finally, with Eq. (12), it is not difficult to derive the estimates of the feature–topic distribution ϕ_k and the topic–reviewer distribution θ_i, as shown in Eqs. (13) and (14), respectively.

$$\widehat{\phi}_k^{j}=\frac{\sum_{i=1}^{|R|}c(l_i)n_{i,j}^{k}+\beta_j}{\sum_{s=1}^{|F|}\sum_{i=1}^{|R|}c(l_i)n_{i,s}^{k}+\beta_s} \tag{13}$$

$$\widehat{\theta}_i^{k}=\frac{c(l_i)n_{i,(\cdot)}^{k}+\alpha_k}{\sum_{v=1}^{K}c(l_i)n_{i,(\cdot)}^{v}+\alpha_v}=\frac{c(l_i)n_{i,(\cdot)}^{k}+\alpha_k}{c(l_i)\sum_{v=1}^{K}n_{i,(\cdot)}^{v}+\alpha_v} \tag{14}$$

As shown in Eq. (13), for a feature f_j, if it appears in the fraudulent reviewer r_i (i.e., the minority class), then it will be given larger weight as $c(l_i)$ in $topic_k$ than the features in the non-fraudulent reviewers (i.e., the majority class). However, as shown in Eq. (14), the probability estimate of a latent $topic_k$ in the reviewer r_i as $\widehat{\theta}_i^{k}$ is less affected by the weight $c(l_i)$ than the probability estimate of the feature f_j in the latent $topic_k$ as $\widehat{\phi}_k^{j}$. In addition, if we set all the weights of the classes $c(l_i)$ as equal to one, then the weighted LDA model is degenerated into the classic LDA model. That

$$P\left(Z_{i,j}=topic_k,Z_{-(i,j)},F;\alpha,\beta,l\right)=\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)^{|R|}\prod_{t\neq i}\frac{\Gamma\left(c(l_t)n_{t,(\cdot)}^{k}+\alpha_k\right)}{\Gamma\left(\sum_{v=1}^{K}c(l_t)n_{t,(\cdot)}^{v}+\alpha_v\right)}\left(\frac{\Gamma\left(\sum_{s=1}^{|F|}\beta_s\right)}{\prod_{s=1}^{|F|}\Gamma(\beta_s)}\right)^{K}$$

$$\times\prod_{s\neq j}\Gamma\left(\sum_{t=1}^{|R|}c(l_t)n_{t,s}^{k}+\beta_s\right)\frac{\Gamma\left(c(l_t)n_{t,(\cdot)}^{k}+\alpha_k\right)}{\Gamma\left(\sum_{v=1}^{K}c(l_t)n_{t,(\cdot)}^{v}+\alpha_k\right)}\frac{\Gamma\left(\sum_{t=1}^{|R|}c(l_t)n_{t,s}^{k}+\beta_s\right)}{\Gamma\left(\sum_{s=1}^{|F|}\sum_{t=1}^{|R|}c(l_t)n_{t,s}^{k}+\beta_s\right)} \tag{12}$$

$$\propto\frac{c(l_i)n_{i,(\cdot)}^{k,-(i,j)}+\alpha_k}{\sum_{v=1}^{K}c(l_i)n_{i,(\cdot)}^{v,-(j,i)}+\alpha_s}\frac{\sum_{t=1}^{|R|}c(l_t)n_{t,j}^{k,-(i,j)}+\beta_j}{\sum_{s=1}^{|F|}\sum_{t=1}^{|R|}c(l_t)n_{t,s}^{k,-(i,s)}+\beta_s}$$

Input: *TrainSet* - A set of reviewers labeled as fraudulent or non-fraudulent.
TestSet - A set of reviewers needed to be labeled.
K - The number of latent topics in the weighted LDA model.
h -The odd number of neighbors in K-nearest neighbor.
Output: The labels for test reviewers in *TestSet*.

1. Use the training reviewers *TrainSet* to produce the feature-topic distribution ϕ_k^j and topic-reviewer distribution θ_i^k with the weighted LDA model described in Section 3.2.
2. For each reviewer r_i in *TrainSet*, it is represented as (θ_i, l_i) where θ_i is the vector of topic probability distribution in r_i and l_i is the class label of r_i.
3. For each reviewer $r_{i'}$ in *TestSet*
4. Initialize the topic distribution vector $vec[i']$ for $r_{i'}$ as a zero vector.
5. For each feature $f_{i'j}$ in $r_{i'}$
6. $vec[i'][k] = vec[i'][k] + \phi_k^j$;
7. $\theta_{i'} = vec[i']$;
8. Find the h nearest neighbors of $r_{i'}$ by using the KL divergence between θ_i and $\theta_{i'}$ as shown in Equation (16) and the similarity measure shown in Equation (17).
9. For all the h nearest neighbors $r_{h'}$ of $r_{i'}$
10. If $l_{h'}$ is fraudulent, $N[fraud]++$;
11. Else if $l_{h'}$ is non-fraudulent, $N[nonf]++$;
12. If $N[fraud] > N[nonf]$, $l_{r_{i'}} = fraud$;
13. Else if $N[fraud] < N[nonf]$, $l_{r_{i'}} = nonf$;

Fig. 4. The proposed ImDetector approach for detecting fraudulent reviewers.

is, the weight $c(l_i)$ adjusts the probability of the features in the minority class to a large extent while making a slight change to the topic distribution in the reviewers. In this way, the minority-specific features with a small number in the dataset will not be annihilated by the majority-specific features with a large number in the dataset. Consequently, minority-specific topics containing minority-specific features will be automatically produced from the generative process of the weighted LDA model.

3.4. Fraudulent reviewer detection

By using the weighted LDA model on the reviewers in the training dataset, we can represent each reviewer r_i using the topic–reviewer distribution vector $\theta_i = \{p_1^i, \ldots, p_k^i, \ldots, p_K^i\}$, where p_k^i denotes the probability of $topic_k$ in the reviewer r_i and $\sum_{k=1}^{K} p_k^i = 1$. The reviewer $r_{i'}$ is represented by the features it has as $r_{i'} = \{f_{i'1}, \ldots, f_{i'j}, \ldots, f_{i'|r_{i'}|}\}$ in the test dataset. With this representation, the test reviewer $r_{i'}$ can be further represented by its topic distribution as $\theta_{i'} = \{p_1^{i'}, \ldots, p_k^{i'}, \ldots, p_K^{i'}\}$, where $p_k^{i'}$ is defined in Eq. (15) and $\sum_{k=1}^{K} p_k^{i'} = 1$. Thus, the similarity $sim(r_i, r_{i'})$ between the training reviewer r_i and the test reviewer $r_{i'}$ can be measured using the asymmetric KL divergence [47], as shown in Eqs. (16) and (17).

$$p_k^{i'} = \sum_{j=1, f_j \in r_{i'}}^{|F|} \phi_k^j \tag{15}$$

$$D_{KL}(\theta_i \| \theta_{i'}) = \sum_{k=1}^{K} p_k^{i'} log\left(\frac{p_k^{i'}}{p_k^i}\right) \tag{16}$$

$$sim(r_i, r_{i'}) = 1 - D_{KL}(\theta_i \| \theta_{i'}) \tag{17}$$

The asymmetric KL divergence $D_{KL}(\theta_i \| \theta_{i'})$ measures the logarithmic distance of the corresponding latent topics in the training reviewer r_i and that in the test reviewer $r_{i'}$ to derive its similarity. The more approximate $p_k^{i'}$ is to p_k^i (i.e., each latent topic $topic_k$ has an approximate probability in the training reviewer r_i to that in the test reviewer $r_{i'}$), the smaller the KL divergence $D_{KL}(\theta_i \| \theta_{i'})$ and the larger the similarity $sim(r_i, r_{i'})$ between the two reviewers. Unlike the traditional methods, such as those of Davis et al. [48], Wang et al. [49], and Kapoor et al. [50], for using KL divergence to measure the difference between the actual distribution and the theoretical distribution of data samples, we use the asymmetric property of KL divergence to make the similarity measure between reviewers biased toward the fraudulent ones. Note that in Eq. (16), the KL divergence between the training reviewer r_i and the test reviewer $r_{i'}$ (i.e., $D_{KL}(\theta_{i'} \| \theta_i)$) is not equal to the KL divergence between the test reviewer $r_{i'}$ and the training reviewer r_i (i.e., $D_{KL}(\theta_i \| \theta_{i'})$).

As shown in Eq. (16), the probability of $topic_k$ in the test reviewer $r_{i'}$ as $p_k^{i'}$ is more important than that in the training reviewer r_i as p_k^i in measuring the distance $D_{KL}(\theta_i \| \theta_{i'})$. For example, we assume that we have a test reviewer represented as $r_{i'} = (0.5, 0.2, 0.2, 0.1)$ ($k = 4$) and two training reviewers as $r_i = (0.5, 0.1, 0.1, 0.3)$ and $r_j = (0.1, 0.2, 0.2, 0.5)$. It

Table 2
Distribution of the number of reviewers among 5044 restaurants in the Yelp dataset.

Range of reviewers #	# of restaurants
1 ≤ # of reviewers ≤ 50	2995
50 ≤ # of reviewers ≤ 100	717
100 ≤ # of reviewers ≤ 200	584
201 ≤ # of reviewers ≤ 500	485
501 ≤ # of reviewers	263

Table 3
Distribution of the number of restaurants reviewed by 262,277 reviewers in the Yelp dataset.

Range of restaurants #	# of reviewers
1 ≤ # of restaurants ≤ 2	210,258
3 ≤ # of restaurants ≤ 4	26,188
5 ≤ # of restaurants ≤ 6	9498
7 ≤ # of restaurants ≤ 10	7997
11 ≤ # of restaurants	6336

can be found that with the KL divergence defined in Eq. (16), the test reviewer$r_{i'}$has a smaller distance with the training reviewerr_i (0.1674) than that with the training reviewerr_j(0. 6438). The reason lies in the fact that on the first dominant topic(i.e.,$topic_1$)of the test reviewer$r_{i'}$, the training reviewerr_i has the same probability as that of test reviewer$r_{i'}$(i.e., ${p_1}^{i'} = {p_1}^{i}$). Although the test reviewer$r_{i'}$and the training reviewer$r_j$have the same probabilities on both $topic_2$ and $topic_3$ (i.e.,${p_2}^{i'} = {p_2}^{j}$ and${p_3}^{i'} = {p_3}^{j}$), these two topics are not the first dominant topic of the test reviewer $r_{i'}$.

The basic idea behind Eq. (16) is that, as shown in Fig. 5 (see Section 4.4), most fraudulent reviewers are dominated by less than five latent topics, and non-fraudulent reviewers are diversely distributed in the remaining 15 latent topics. Therefore, if the dominant topics of fraudulent reviewers are given more importance, the asymmetric KL divergence will make the similarity measure between reviewers in KNN [46] biased toward the fraudulent reviewers, who are the minority in imbalanced classification. With the asymmetric KL divergence for similarity measure between reviewers, we adopt KNN to classify the unlabeled reviewers in the test dataset to detect fraudulent reviewers. Given an unlabeled test reviewer $r_{i'}$, we first find its h nearest neighbors in the similarity defined in Eq. (17). To simplify the labeling process, we usually set the parameter h as an odd number [46]. Then, we vote the label for the unlabeled test reviewer $r_{i'}$ using the labels of its h nearest neighbors. If the number of fraudulent reviewers is larger than that of non-fraudulent reviewers among all the h nearest neighbors, we assign the label of the unlabeled test reviewer $r_{i'}$ as fraudulent. Otherwise, its label will be assigned as non-fraudulent. Note that as the parameter h in KNN is an odd number, it is impossible that the number of fraudulent reviewers is equal to that of non-fraudulent reviewers in the nearest neighbors. The complete algorithm for fraudulent reviewer detection is illustrated in Fig. 4.

Table 4
Statistics of the 10 subsets of restaurants. FD means "fraudulent."

Subset no.	# of restaurants	# of reviewers	# of FD reviewers	# of reviews	Time span
1	504	26,655	6162	65,305	07–2010 to 05–2014
2	505	27,134	6438	57,931	05–2011 to 12–2014
3	505	25,341	6183	62,836	02–2011 to 06–2014
4	505	26,966	6396	67,550	07–2010 to 12–2013
5	504	26,219	6224	64,630	01–2011 to 01–2014
6	504	25,944	6149	55,119	09–2010 to 06–2014
7	505	26,107	6187	60,568	01–2011 to 12–2014
8	504	25,738	6118	55,723	11–2010 to 08–2014
9	504	26,117	6193	55,608	06–2011 to 12–2014
10	504	26,056	6178	63,328	12–2010 to 05–2014

3.5. Complexity of the proposed ImDetector approach

The computation complexity of ImDetector comprises two parts: weighted LDA modelling and the KNN classifier. The computation complexity of the weighted LDA modelling is the same as that of the LDA modelling, $O(|R||F|KT)$ [51], where $|R|$ is the number of reviewers in the reviewer set R, $|F|$ is the number of features in the feature vocabulary F, K is the number of topics, and T is the number of iterations to make the weighted LDA model convergent and stable. The computation complexity of the KNN classifier is $O(|R||F|h)$, where h is the predefined number of neighbors [52]. Therefore, we can speculate from Fig. 4 that the computation complexity of the ImDetector is $O(|R||F|KT)+O(|R||F|h)$. In the weighted LDA model, the number of reviewers $|R|$ (Table 4) and the number of features $|F|$ (Table 5) are much larger than the number of topics K and the number of iterations T. In the same way, the number of reviewers $|R|$ (Table 4) and the number of features $|F|$ (Table 5) are much larger than the number of nearest neighbors h in the KNN classification. That is, the computation complexity of ImDetector mostly depends on the multiplication of the number of reviewers $|R|$ and the number of features $|F|$, i.e., $|R||F|$.

4. Experiments

4.1. The dataset

The dataset used in the experiment is the restaurant review data collected from the Yelp.com website,[1] which is available online from Rayana and Akoglu [26]. This dataset contains 608,598 reviews of 5044 restaurants posted by 262,277 reviewers. Among the 262,277 reviewers, 62,228 are labeled fraudulent, and the remaining 200,049 reviewers are labeled non-fraudulent. The dataset also contains the complete information of each review, including its star rating, textual content, review date, and target restaurant.

Table 2 shows the distribution of the 5044 restaurants among the 608,598 reviewers. Most restaurants have fewer than 201 reviewers, and a relatively small number of restaurants have more than 500 reviewers. Table 3 shows the distribution of the 608,598 reviewers among the 5044 restaurants. Most reviewers have reviewed fewer than three restaurants, and a relatively small number of reviewers have reviewed more than 10 restaurants. Although most restaurants are reviewed by more than 50 reviewers on average, most reviewers have reviewed only two restaurants on average, with nearly 2–3 reviews. On the whole, the interests of all of the reviewers in the restaurant are dispersive and different from each other. However, the intertest of an individual reviewer in restaurants is concentrated on a small number of restaurants.

We adopt the randomized snowball sampling [53] to partition the dataset into 10 subsets by restaurant. First, we initialize a subset, sample a restaurant randomly from the Yelp.com dataset, and then locate its reviewers. Second, we use the located reviewers to find more restaurants to be added to the subset. Third, we repeat this process using the newly added restaurants to locate more reviewers. Finally, when the number of

[1] Restaurants, Dentists, Bars, Beauty Salons, Doctors-Yelp. Online: https://www.yelp.com

Table 5
Predictive features as the review content features, reviewer behavior features, and review target features, with their bins as the dummy features.

Feature category	Feature name (abbreviation)	Description	Data type	Range	# of bins
Review content features (RCFs) (8)	Average sentence length (ASL)	Average number of words per sentence in the review	Continuous	[0, 199]	15
	Similarity (SI)	Average cosine similarity of the reviews posted by a reviewer	Continuous	[0, 99]	15
	Lexical diversity (LD)	Ratio of unique words in the review to all words	Continuous	[0, 79.17]	15
	Affective diversity (AD)	Ratio of unique affective words in the review to all affective words	Continuous	[0,100]	14
	Cognitive diversity (CD)	Ratio of unique cognitive words to all cognitive words	Continuous	[0, 58.34]	14
	Perceptual diversity (PD)	The ratio of unique perceptual words to all perceptual words	Continuous	[0, 100]	15
	Pronoun counts (PC)	Total number of pronouns in the review	Continuous	[0, 58.25]	12
	Self-reference diversity (SD)	Ratio of first-personal pronouns to all words in the review	Continuous	[0, 1.01]	14
Reviewer behavior features (RBFs) (7)	Rating score (RS)	Mean rating score posted by a reviewer	Continuous	[1, 5]	8
	Rating deviation (RD)	Difference in the mean of the reviewers' ratings and target restaurant ratings	Continuous	[0, 3.44]	15
	Rating entropy (RE)	Review rating randomness of a reviewer	Continuous	[0.45, 5.25]	5
	Review length (RL)	Average number of words of all reviews posted by a reviewer	Continuous	[0, 480]	14
	Review gap (RG)	Average time interval of all successive reviews posted by the same reviewer	Continuous	[0, 3046]	13
	Review count (RC)	Review volume posted by the same reviewer	Continuous	[2, 205]	15
	Review time difference (RTD)	Mean time difference between a reviewer's review time and the target restaurant's first review time	Continuous	[0, 3602]	13
Review target features (RTFs) (8)	Brand chain (BC)	Whether the restaurant is a brand chain operation or an independent operation	Nominal	{0,1}	–
	# of nearby competitors (NNC)	Number of competitors within four blocks of the target restaurant	Continuous	[4, 27]	13
	Target rating gap (TRG)	Mean difference in ratings between nearby competitors and the target restaurant	Continuous	[−1.22, 1.40]	15
	Volume gap (VG)	Difference in review volume between nearby competitors and the target restaurant	Continuous	[−73.93, 60.95]	15
	Mean rating (MR)	Mean rating of a restaurant posted by all reviewers	Continuous	[1.88, 5]	10
	Review volume (RV)	Review volume of the target restaurant	Continuous	[3. 7005]	15
	Target review duration (TRD)	Mean time difference between all reviews and the first review of the target restaurant	Continuous	[7, 3683]	15
	Review filtered rate (RFR)	Ratio of filtered reviews to all reviews of the target restaurant	Continuous	[0.13, 1]	15

restaurants reaches 505 or 504 (i.e., partition with equal size), we stop the snowball sampling for the current subset and start the snowball sampling for the next subset. We admit that the randomized snowball sampling cannot completely separate the reviewers of a subset from those of other subsets without overlapping, which causes a loss of information when extracting RCFs and RBFs (Table 5). However, as most reviewers have reviewed nearly two restaurants, we minimize the overlapping of reviewers among different subsets to alleviate the effect of reviewer overlapping on the experimental results.

As a result, each subset contains approximately 500 restaurants. Table 4 shows the statistics for each subset of the sampled restaurants. For all 10 subsets, the numbers of restaurants, reviewers, reviews, and the review time spans are almost in the same order of magnitude in the crosswise comparison. That is, each subset has nearly 500 restaurants, 26,000 reviewers, and 60,000 reviews. Nearly 6000 of the reviewers in each subset are fraudulent, and the remaining are non-fraudulent. With this observation, we hypothesize that the difference in the values of the features in Table 5 (see Section 4.2) is insignificant. We test this hypothesis using the Wilcoxon signed-rank test [54] on all the subset pairs of the feature values, with $p > 0.1$.

4.2. Predictive features

The predictive features for fraudulent reviewer detection come from three aspects: RCFs, RBFs, and RTFs (Table 5). According to Zhang et al. [11], the RCFs have the following eight features: average sentence length (ASL), similarity (SI), lexical diversity (LD), affective diversity (AD), cognitive diversity (CD), perceptual diversity (PD), pronoun counts (PC), and self-reference diversity (SD). Following Kumar et al. [14,15], we use seven RBFs to characterize the anomalous behaviors of the reviewers: rating score (RS), rating deviation (RD), rating entropy (RE), review length (RL), review gap (RG), review count (RC), and review time difference (RTD). Following Rayana and Akoglu [26] and Leman et al. [55], we use the following eight RTFs: brand chain (BC), number of nearby competitors (NNC), target rating gap (TRG), volume gap (VG), mean rating (MR), review volume (RV), target review duration (TRD), and review filtered rate (RFR).

As shown in Table 5, 22 of the 23 features are continuous variables that cannot be directly processed by the weighted LDA model. For this reason, we discretize the continuous features into categorical variables and transform them into Boolean dummy features using the classification and regression tree algorithm [56]. For a continuous feature containing m feature values, first, we sort the feature values in ascending order and use the average of two adjacent values as the split points. Second, we compute the Gini index [57] of each split point as the binary classification point, one by one. Third, we use the split point with the smallest Gini index as the optimal split point for this continuous feature and partition the feature values into two parts. Finally, we repeat the above steps to split the continuous feature values until the minimum number of leaf nodes is less than 5% of the total sample size [58]. Through this method, we discretize the 22 continuous features into 293 discrete variables, as shown in Table 5, where "# of bins" indicates the number of dummy variables derived from each of the corresponding continuous features. As a result, we obtain a total of 295 dummy variables by combining the 293 dummy variables and the one Boolean variable as the BC in RTFs. In this process, each reviewer r_i in R is represented by the 295 dummy variables. Thereafter, we refer to the 295 dummy variables as dummy features to be consistent with the description of the weighted LDA model in Section 3. Thus, the feature–topic distribution ϕ_k and the topic–reviewer distribution θ_i are derived using the weighted LDA model in the 295 dummy features.

4.3. Experimental setup

For the proposed ImDetector approach described in Fig. 3, two parameters are set in model training: the number of latent topics k used in the weighted LDA model and the number of nearest neighbors h in the KNN classifier to detect fraudulent reviewers. For the first parameter k, we use the Gibbs sampling tool called GibbsLDA++ [59] to implement the weighted LDA model and tune the number of latent topics k in the weighted LDA model to 5, 10, 15, 20, and 25 to optimize the performance by trial and error. To the best of our knowledge, there is no known method for estimating the a priori number of topics k in LDA modelling. Therefore, we use the default values in GibbsLDA++ as α=50/k and β=0.01 to set the hyperparameters. For the second parameter h, we tune the number of nearest neighbors to 51, 101, 151, 201, and 251 for the performance comparison by trial and error. Note that the two parameters k and h are set for all 10 subsets shown in Table 4.

We conduct two types of studies to investigate the effectiveness of the proposed ImDetector approach in fraudulent reviewer detection. The first study compares the proposed ImDetector approach with the baseline methods for fraudulent reviewer detection. Akoglu et al.'s [55] FraudEagle approach, Rayana and Akoglu's [26] SpEagle approach, the hierarchical supervised learning (HSL) approach [14], and Kumar et al.'s [15] STK approach are used as the baseline for the performance comparison with the proposed ImDetector approach for fraudulent reviewer detection.

We follow Akoglu et al. [55] in setting the three hyperparameters for the FraudEagle approach in the experiment: prior potential to 0.5, threshold of fraud score to 0.89, and number of top malicious users to 100. We follow Rayana and Akoglu [26] in setting the prior for fraudulent and non-fraudulent users to 0.5 and the prior for non-fraudulent users writing positive reviews for a poor-quality product to 0.1. We follow Kumar et al. [14] in using the log-log transformation to smooth the values of the predictive features (Table 5) and logistic regression for supervised learning. However, we do not follow Kumar et al. [14] in modelling the users' set of rating scores as a categorical distribution (i.e., Dirichlet distribution [60]) because we argue that users' rating of 1–5 is essentially ordinal numeric rather than a simple category. We follow Kumar et al. [15] in adopting the log-log transformation, the log-norm transformation, and the beta transformation to smooth the values of the predictive features with the best distribution fitting. The joint distribution between the ratings and the restaurants is modeled by a Dirichlet distribution initialized by the moment-matching estimation method [23].

The second study compares the proposed ImDetector approach with the baseline methods for imbalanced classification for fraudulent reviewer detection. The baseline methods, including two oversampling techniques (SMOTE [31] and latent Dirichlet oversampling (LDO) [44]) and one undersampling technique (Tomek link [45]) are used to handle imbalanced data for fraudulent reviewer detection. The SMOTE technique [31] selects a sample from the minority class and obtains its nearest neighbors. It takes the difference between the sample and one of its nearest neighbors using random selection. Then, it generates a synthetic sample by adding the difference multiplied by a random number between 0 and 1 to the sample. In this way, the SMOTE technique oversamples the predefined number of synthetic samples to increase the size of the minority class. We follow Chawla et al. [31] in setting the number of nearest neighbors of SMOTE to 5 to produce the synthetic samples and oversample the fraudulent reviewers to the same size as the non-fraudulent reviewers.

The LDO technique [44] uses the LDA model to obtain the topic–reviewer distribution θ_i and the feature–topic distribution ϕ_k from the training data. For each sample in the majority class, LDO generates one feature vector using the topic–reviewer distribution of the sample and the word–topic distribution. All samples are represented by the concatenation of their original feature vectors and the generated feature vectors. As a result, each sample in the minority class has a predefined number of synthetic samples by concatenating its original feature vector and each of its generated feature vectors. Then, each sample in the majority class has only one synthetic sample by concatenating its original feature vector and the generated feature vector. We follow Moreo et al. [43] in adopting a perplexity measure to estimate the LDA model for LDO among the reviewers and oversample the fraudulent reviewers to the same size as the non-fraudulent reviewers.

As a classical undersampling method, the Tomek link [45] has the following procedure: Given two samples r_i and r_j, which belong to the two different classes of reviewers, and $dis(r_i,r_j)$, which is the distance between r_i and r_j, then a pair (r_i,r_j) is called a Tomek link if there is no sample r_l, such that $dis(r_i,r_l) < dis(r_i,r_j)$ or $dis(r_j,r_l) < dis(r_i,r_j)$. If two samples form a Tomek link, then either one is noise, or both samples are borderline cases. As this is an undersampling method, if two samples form a Tomek link, the non-fraudulent sample should be removed. With a Tomek link, the non-fraudulent reviewers are undersampled to the size of the fraudulent reviewers. For the above baseline methods, we follow Moreo et al. [44] in using SVM with a linear kernel for reviewer fraud detection after class rebalancing using data preprocessing. In our previous studies [41,61], we also find that SVM with a linear kernel performs better than other methods in text classification.

We use the four metrics of precision, recall, F1-score, and area under the curve score (AUC) [62], as the performance measures, as described in Eqs. (20), (21), (22), and (23), respectively. Here, TP means true positive (i.e., fraudulent classified as fraudulent), NP means false positive (i.e., non-fraudulent classified as fraudulent), and FN means false negative (i.e., fraudulent classified as non-fraudulent). In Eq. (23), M is the number of positive samples, N is the number of negative samples, $rank_i$ is the rank of sample i of all the test samples sorted by its probabilities of being fraudulent in ascending order, and $\sum_{i \in positiveClass} rank_i$ refers to the sum of the ranks of all the true positive samples in the prediction list. The AUC score measures the area under the receiver operating characteristic curve and indicates the predictive ability of a binary classifier when its discrimination threshold is varied [63]. The larger the AUC score, the better the classifier performs.

Following Twyman et al. [64] and Kumar et al. [14,15], we evaluate the performance of the proposed model by conducting a five-fold cross-validation on each subset. We divide each subset (Table 4) into five folds and use one fold for model testing and the remaining four folds for training the model. The outcomes of fraudulent reviewer detection on the five folds are averaged for model performance evaluation on the

Table 6
Performance of the proposed ImDetector approach for fraudulent reviewer detection under different numbers of latent topics in the weighted LDA model (h = 151).

# of latent topics	Precision	Recall	F1 score	AUC score	Time
k = 10	0.832	0.671	0.743	0.737	89
k = 15	0.851	0.800	0.825	0.864	121
k = 20	0.854	0.830	0.842	0.887	173
k = 25	0.842	0.785	0.813	0.842	235
k = 30	0.838	0.720	0.775	0.789	312

Table 7
Performance of the proposed ImDetector approach for fraudulent reviewer detection under different numbers of nearest neighbors in the KNN classifier (k= 20).

# of neighbors	Precision	Recall	F1 score	AUC score	Time
h=51	0.812	0.696	0.750	0.674	146
h=101	0.828	0.742	0.783	0.752	158
h=151	0.854	0.830	0.842	0.887	173
h=201	0.825	0.594	0.691	0.707	189
h=251	0.811	0.581	0.677	0.621	208

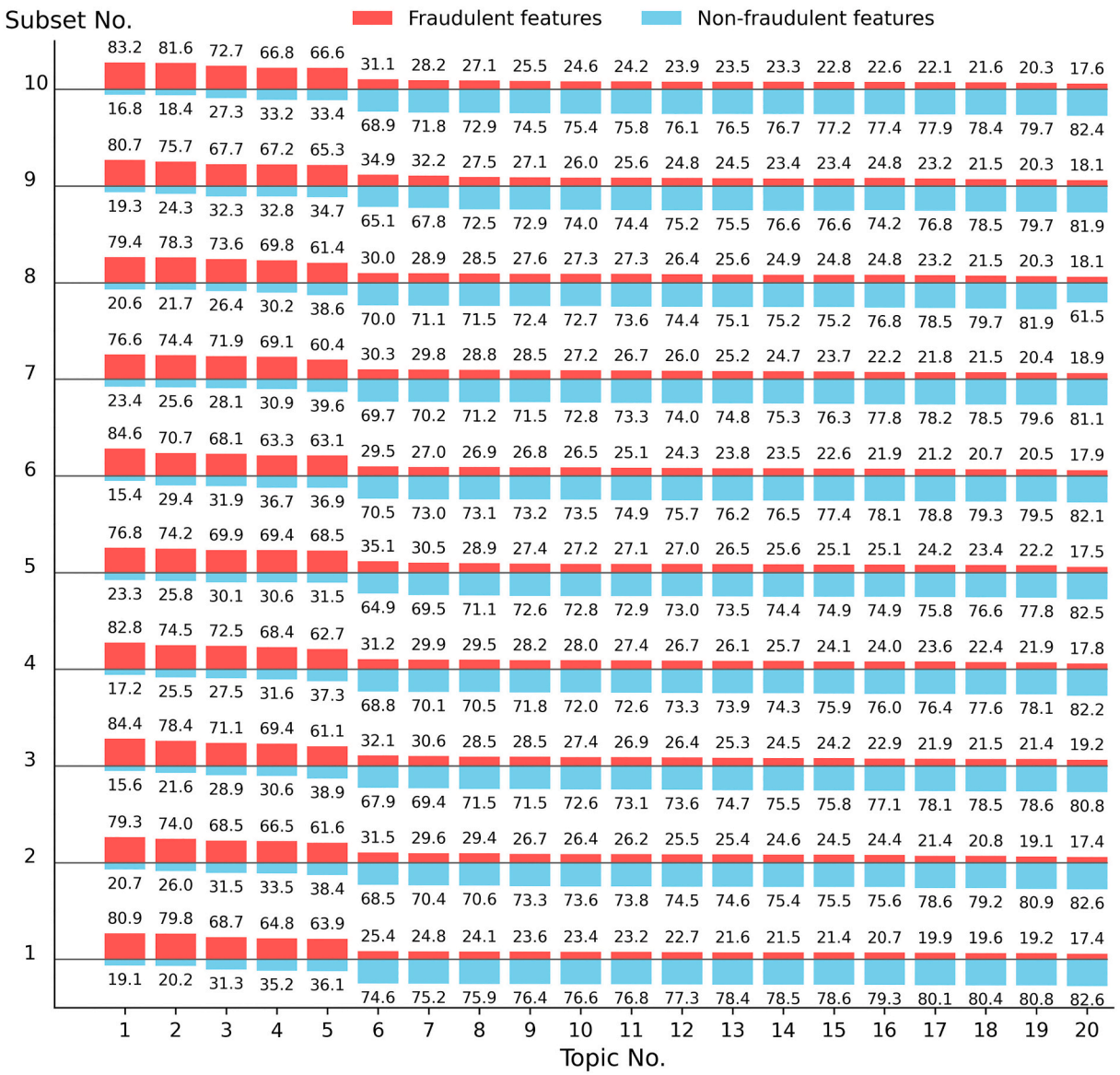

Fig. 5. Probability distribution of fraudulent features versus non-fraudulent features in the latent topics of each subset of reviewers.

subset. Note that the final performance of our method is gauged by averaging the performances derived from the 10 subsets of the reviewers described in Table 4.

$$precision = \frac{TP}{TP + FP} \tag{20}$$

$$recall = \frac{TP}{TP + FN} \tag{21}$$

$$F1 = \frac{2*Precision*Recall}{Precision + Recall} \tag{22}$$

$$AUC = \frac{\sum_{i \in positiveClass} rank_i - \frac{M(M+1)}{2}}{M \times N} \tag{23}$$

Table 8
Performance averaged on the 10 subsets of the proposed ImDetector approach and the baseline methods for fraudulent reviewer detection.

Algorithm	Precision	Recall	F1 score	AUC score	Time cost
FraudEagle	0.62	0.59	0.60	0.61	247
SpEagle	0.67	0.63	0.65	0.69	253
HSL	0.83	0.71	0.76	0.82	112
STK	0.68	0.65	0.66	0.70	194
Weighted-LDA-LR	0.75	0.70	0.72	0.71	138
ImDetector	0.85	0.83	0.84	0.88	173

Table 9
Performance in precision, recall, F1 score, AUC score, and time of the proposed ImDetector approach compared with the baseline methods for fraudulent reviewer detection.

Method pair	Precision	Recall	F1 score	AUC score	Time cost
ImDetector vs. FraudEagle	>>	>>	>>	>>	>>
ImDetector vs. SpEagle	>>	>>	>>	>>	>>
ImDetector vs. HSL	~	>>	>>	>>	<<
ImDetector vs. STK	>>	>>	>>	>>	<
ImDetector vs. Weighted-LDA-LR	>>	>>	>>	>>	<<

4.4. Results

Table 6 shows the performance of the proposed ImDetector approach for fraudulent reviewer detection under different numbers of latent topics k in the weighted LDA model. Here, we set the number of the nearest neighbors h to 151. As shown in Table 6, the proposed ImDetector approach performs best when the number of latent topics k in the weighted LDA model is 20, with precision of 0.854, recall of 0.830, F1 score of 0.842, and AUC score of 0.887. The performance of the proposed ImDetector approach increases when the number of latent topics k increases from 10 to 20. However, the performance decreases when the number of latent topics k is larger than 20.

This outcome can be explained as follows: when the number of latent topics k is smaller than 20, the information inherent in the latent topics overlaps and mixes with each other. This can result in poor representation of reviewers by the derived latent topics from the weighted LDA model and make it difficult to differentiate fraudulent from non-fraudulent reviewers using the KNN classifier. Nevertheless, when the number of latent topics k is larger than 20, we find that there are many trivial topics produced by the weighted LDA model. That is, the probability p_k^i of $topic_k$ in all the reviewers ($1 \leq i \leq |R|$) is small. These trivial latent topics dominate the similarities of reviewers when using the asymmetric KL divergence for the similarity measure in Eq. (17). This makes all the reviewers similar and induces a great amount of noise to the neighbors of a test reviewer when using the KNN classifier to vote its label.

Table 7 shows the performance of the proposed ImDetector approach for fraudulent reviewer detection under different numbers of nearest neighbors in the KNN classifier. We set the number of latent topics in the weighted LDA model to 20. As shown in Table 7, the proposed ImDetector approach performs best when the number of nearest neighbors in the KNN classifier h is set to 151. When the parameter h is smaller than 151, the performance is improved when the number of nearest neighbors increases in the KNN classifier. However, when the parameter h is larger than 151, the performance degenerates if the number of nearest neighbors is further increased. The reason is that, when the parameter h is smaller than 151, the number of neighbors is not sufficient to feed to the KNN classifier for fraudulent reviewer detection. Nevertheless, when the parameter h is larger than 151, through manual checking, we find too many noisy neighbors that mislead the KNN classifier in labeling test reviewers.

Fig. 5 shows a snapshot of the feature–topic distributionsϕ_k ($1 \leq k \leq 20$) on the 20 latent topics derived from the weighted LDA model. We cannot list all the distributions of the five-fold cross-validation on all 10 subsets because of space limitations. As shown in Fig. 5, there are 295 dummy features in total after discretization. These dummy features can be roughly categorized into two types: 1) the fraudulent feature, in which the percentage of occurrences of the feature in the fraudulent reviewers is larger than that in the non-fraudulent reviewers, and 2) the non-fraudulent feature, which is the opposite case. For example, the dummy feature "AD_12" (i.e., the 12th bin of the values under the AD feature) occurs at 83.75% among fraudulent reviewers, and the dummy feature "BC_0" (i.e., the zeroth bin of the values under the BC feature) occurs at 77.05% among non-fraudulent reviewers. Therefore, we can regard the dummy feature "AD_12" as a fraudulent feature and the dummy feature "BC_0" as a non-fraudulent feature.

As shown in Fig. 5, when the number of latent topics is set to 20, the most fraudulent features are distributed to topics 1–5, while the most non-fraudulent features are distributed to topics 6–20. If a test reviewer has more non-fraudulent features than fraudulent features, then in its topic–reviewer distribution, the probability of topics 6–20 will be larger than that of topics 1–5. In this way, the non-fraudulent reviewers are aggregated to topics 6–20, while the fraudulent reviewers are aggregated to topics 1–5. The imbalanced data, which are characterized by fraudulent reviewers as the minority and non-fraudulent reviewers as the majority, will cause the number of non-fraudulent topics to be larger than that of fraudulent topics. Therefore, there should be some fraudulent topics that dominate their representation among fraudulent reviewers. That is, the probabilityp_k^i of some fraudulent topics will be greater than that of non-fraudulent topics among the fraudulent reviewers because of their small number. Furthermore, with the asymmetric KL divergence in Eq. (16), the similarity measure between reviewers will be biased toward the fraudulent reviewers.

4.5. Comparison with baseline methods in fraudulent reviewer detection

Table 8 shows the performances averaged on the 10 subsets of the proposed ImDetector approach and the baseline methods for fraudulent reviewer detection. To better illustrate the effectiveness of each method, the classic non-parameter Mann–Whitney U test [54] is employed. Table 9 shows the results of the Mann–Whitney U test of the performances of ImDetector and the baseline methods in precision, recall, F1 score, AUC score, and time cost. The following codification of the P-value in ranges is used: ">>" ("<<") means that the P-value is less than or equal to 0.01, indicating strong evidence that the proposed ImDetector approach performs better (worse) than the compared baseline method; "<" (">") means that the P-value is greater than 0.01 and less than or equal to 0.05, indicating weak evidence that the proposed ImDetector approach performs better (worse) than the compared baseline method; and "∼" means that the P-value is greater than 0.05, indicating that the compared methods do not have significant differences in performance.

As shown in Tables 8 and 9, the proposed ImDetector approach performs the best among all the compared methods in fraudulent reviewer detection, except in time cost. The better performance of the proposed ImDetector approach can be explained in two ways. First, it has the capacity to use the proposed weighted LDA model to deal with the imbalanced distribution of fraudulent and non-fraudulent reviewers in the dataset. Second, we use multiple sources of features—RCFs, RBFs, and RTFs (Table 5)—to characterize a reviewer. The poor performance of the FraudEagle approach illustrates its infeasibility to use the heterogeneous network to model the relationship among reviewers, reviews, and products for fraudulent reviewer detection without an in depth investigation of the features of the nodes in the network. The STK approach and the SpEagle approach have similar performances because both are unsupervised techniques and make use of the behavioral features of reviewers for fraudulent reviewer detection.

The HSL approach performs second best because it uses standard statistical distributions, such as exponential and normal distributions, to conduct feature transformation to smooth the skewness in the original feature values, as well as supervised learning as logistic regression to identify fraudulent reviewers. Feature transformation is beneficial in handling data imbalance, and logistical regression can capture the probability differences in the transformed feature values between fraudulent and non-fraudulent reviewers. In this study, we also attempt to conduct logistic regression after reviewer representation by the latent topics (i.e., the weighted LDA–LR method). However, it performs worse than the KNN classifier with the asymmetric KL divergence. With the above analysis, we can infer that the proposed ImDetector approach significantly outperforms the baseline methods because of its capacity to handle data imbalance and integrate multiple sources of features.

Table 10
Performance averaged on the 10 subsets of the proposed ImDetector approach and the baseline methods for the classification of imbalanced data for fraudulent reviewer detection.

Algorithm	Precision	Recall	F1 score	AUC score	Time cost
SMOTE	0.81	0.79	0.80	0.82	391
LDO	0.78	0.76	0.77	0.79	352
Tomek link	0.76	0.74	0.75	0.73	316
ImDetector	0.85	0.83	0.84	0.88	173

Table 11
Performance in precision, recall, F1 score, AUC score, and time of the proposed ImDetector approach compared with the baseline methods for the classification of imbalanced data for fraudulent reviewer detection.

Method pair	Precision	Recall	F1 score	AUC score	Time cost
ImDetector vs. SMOTE	>>	>>	>>	>>	>>
ImDetector vs. LDO	>>	>>	>>	>>	>>
ImDetector vs. Tomek link	>	>>	>>	>>	<<

In terms of computation complexity, as shown in Table 8, the HSL approach needs the smallest time cost for computation because, when using logistic regression as the classifier, it only needs to solve a least square optimization with a computation complexity of $O(|F|^2)$ [65]. The proposed ImDetector approach requires more time than the HSL approach because the number of reviewers is much larger than that of the features. The FraudEagle approach [55] adopts loopy belief propagation [66] to infer the probability of being fraudulent and non-fraudulent of reviewers, reviews, and restaurants for fraudulent reviewer identification. Therefore, its computation complexity is $O(|F||R||P|)$, where $|P|$ is the number of restaurants in the subset. The SpEagle approach is an adaptation of the FraudEagle approach by using not only relational data but also metadata with loopy belief propagation [66] for fraudulent reviewer identification. Thus, its computation complexity is approximately the same as that of the FraudEagle approach. The STK approach [14] uses an expectation maximization (EM) algorithm with a mixed model of probability distributions of the features for fraudulent reviewer identification. Therefore, its computation complexity is $O(|F||R||T)$, where T is the number of iterations needed to make the objective function of the EM algorithm convergent. Considering that $|R|$ and $|F|$ are usually much larger than T, we can speculate that the computation complexity of the STK algorithm mostly depends on $|F||R|$. This is also consistent with the observation in Table 8 that the time cost of the STK approach is comparable with that of the proposed ImDetector approach.

4.6. Comparison with the baseline methods for imbalanced data classification

Table 10 shows the performance averaged on the 10 subsets of the proposed ImDetector approach and the baseline methods for imbalanced classification for fraudulent reviewer detection. We use the classic non-parameter Mann–Whitney U test [54] to illustrate the effectiveness of the compared methods. Table 11 presents the results of the Mann–Whitney U test of the performance of ImDetector and the baseline methods in precision, recall, F1 score, AUC score, and time cost. The same codification as that used in Table 9 is also adopted in Table 11.

As shown in Tables 10 and 11, the proposed ImDetector approach outperforms all the baseline methods for imbalanced data classification for fraudulent reviewer detection. The proposed ImDetector approach uses the topic–reviewer distribution produced by the weighted LDA model for reviewer representation. Although the topics dominated by fraudulent features have a smaller number than those dominated by non-fraudulent features (Fig. 5), these topics are capable enough to characterize the feature distribution of fraudulent reviewers. Therefore, when the topics dominated by the fraudulent and non-fraudulent features produced by the weighted LDA are used to represent the reviewers, the distributions of the topics to the fraudulent and non-fraudulent reviewers differ from each other. Moreover, the proposed ImDetector approach uses the asymmetric KL divergence in Eqs. (16) and (17) to measure the similarity between the test reviewer $r_{i'}$ and the training reviewer r_i, i.e., $D_{KL}(\theta_i||\theta_{i'})$. As the distributions of topics to the fraudulent and non-fraudulent reviewers are different from each other, and the asymmetric KL divergence is biased toward the fraudulent reviewers, then $D_{KL}(\theta_i||\theta_{i'})$ will be large, while the similarity $sim(r_i, r_{i'})$ between the test reviewer $r_{i'}$ and the training reviewer r_i will be small if the two reviewers r_i and $r_{i'}$ belong to different classes. Furthermore, the nearest neighbors of the fraudulent reviewers in the test set will be fraudulent, and by analogy, the nearest neighbors of the non-fraudulent reviewers in the test set will be non-fraudulent if the KNN with the similarity measure as the asymmetric KL divergence is used for classification.

The SMOTE method performs second best among all the compared methods. The distribution of the fraudulent and non-fraudulent reviewers in the Yelp.com dataset is similar to that in Chawla et al. [31], who also use the SMOTE method for imbalanced data classification. As shown in Table 4, for most of the 10 subsets, the number of fraudulent reviewers is nearly 6000, and the number of non-fraudulent reviewers is almost 20,000. Therefore, the large number of fraudulent reviewers in the Yelp.com dataset makes SMOTE able to generate synthetic samples in a large space rather than focusing on a specific region, resulting in overfitting. Generally, the SMOTE method generates synthetic samples through the linear combination of neighborhood samples. However, the LDO method generates synthetic samples by expanding the feature space of the samples, and this could induce a large amount of noise in the synthetic samples because there is much more freedom for feature space expansion than in the linear combination of existing samples. The worst performance of the Tomek link method shows that the undersampling technique is not suitable for fraudulent reviewer detection in imbalanced data because it induces a large error when identifying non-fraudulent reviewers. As SVM has a computation complexity of $O(|R|^3)$ [67], all three methods have a larger time cost than the proposed ImDetector approach. This is consistent with the outcome of the experiment shown in Table 10.

Nevertheless, in most cases, the methods for imbalanced data classification perform better than the baseline methods in fraudulent reviewer detection, except the HSL approach (Tables 8 and 10). Note that the HSL approach also uses feature transformation to smooth the skewed distribution of the original features through exponential and normal distributions. Considering the problem of imbalanced data is beneficial to fraudulent reviewer detection. This is consistent with our motivation to propose the ImDetector approach for fraudulent reviewer detection in the Yelp.com dataset to handle the imbalanced data problem.

5. Concluding remarks

5.1. Implications

With the success of *E*-commerce, business stakeholders have recognized the importance of online reviews in winning market competition, and online consumers have accepted that online reviews are valuable and indispensable in making online purchase decisions. In this situation, online review manipulation can mislead online consumers' decision making, thus benefitting review manipulators with deceptive reviews [9]. Therefore, a series of studies has been conducted on fraudulent review detection [9–11], fraudulent reviewer detection [13–15] and fraudulent reviewer group detection [16–18] to identify fraudulent reviews and reviewers. This study deals with the problem of fraudulent reviewer detection to detect those who post fraudulent online reviews.

The managerial implications of this study are summarized in two aspects. First, the performance of fraudulent reviewer detection can be improved if the problem of imbalanced data is considered. Existing studies have invested much effort in developing comprehensive features, such as RCFs, RBFs [15], and RTFs [26,55], as well as novel techniques, such as mixture model learning [15], heterogeneous network analysis [26], and loopy belief propagation [55] for fraudulent reviewer detection. Unlike existing studies, this study focuses on the problem of imbalanced data in fraudulent reviewer detection.

We are motivated by the observation that non-fraudulent reviewers are the majority and that fraudulent reviewers are the minority in real

practice. This causes the classifier in fraudulent reviewer detection to be biased toward the non-fraudulent reviewers because it has less knowledge of the fraudulent reviewers due to the unbalanced data distribution of the two classes [21]. As shown in Tables 8 and 10, the classification of imbalanced data performs better than the existing methods, which do not consider the problem of imbalanced data in fraudulent reviewer detection. This is consistent with our observation from the Yelp.com dataset, in which almost 20% of the reviewers are fraudulent and 80% are non-fraudulent. With the prevalence of online review manipulation, the detection of fraudulent online reviewers is crucial for the success of E-commerce. Therefore, the problem of imbalanced data should be taken into account when data-driven methods, such as statistical machine learning and deep learning, are adopted to identify fraudulent reviewers accurately.

Second, fraudulent reviewer detection can be further improved using the proposed ImDetector approach. Existing studies on imbalanced classification can be categorized into three streams: data-processing method, feature selection method, and algorithmic-centered method [23]. The proposed ImDetector approach falls under the algorithmic-centered method. However, it is different from the traditional algorithmic-centered method in that it handles imbalanced data by improving the sample representation with the proposed weighted LDA and the sample classification using KNN with the asymmetric KL divergence rather than classifier adaptation or cost-sensitive learning. We can regard the proposed ImDetector approach as a novel algorithmic-centered method to address the problem of imbalanced data by combining weighted LDA modelling and the KNN classifier with asymmetric KL divergence.

The weighted LDA model gives larger weights to the features of fraudulent reviewers and smaller weights to the features of non-fraudulent reviewers to balance the feature–topic distribution. In the KNN classifier with asymmetric KL divergence, the number of non-fraudulent topics is much larger than that of fraudulent topics in the dataset. Therefore, a small number of fraudulent topics will be aggregated to the fraudulent reviewers, making the probabilities of some fraudulent topics much larger than those of the non-fraudulent topics among the fraudulent reviewers. Furthermore, if the dominant topics of fraudulent reviewers are given more importance, it will make the asymmetric KL divergence biased toward the fraudulent reviewers in imbalanced classification with the KNN classifier. Thus, the proposed ImDetector approach can handle the problem of data imbalance between fraudulent and non-fraudulent reviewers on the one hand and augment the importance of the latent topics from the fraudulent features to detect fraudulent reviewers on the other. The experimental results shown in Table 10 validate the superiority of the proposed ImDetector approach to traditional classification methods for imbalanced data for fraudulent reviewer detection.

5.2. Limitations and future work

Although the proposed ImDetector approach shows some promising aspects of fraudulent reviewer detection, we admit that it still has some limitations. First, the RCFs are derived only by the statistical analysis of the words and their parts of speech without an in-depth analysis of the textual contents of the reviews. In fact, the grammatical rules, sentimental polarity, and spatial–temporal words of the fraudulent reviews are significantly different from those of the fraudulent reviews, as reported by previous studies [68]. Second, we only consider the behaviors of individual reviewers without considering the behaviors of reviewer groups. We can cluster the reviewers by restaurant to discover the reviewer groups, and the anomalous behaviors of reviewer groups can be derived and utilized for fraudulent reviewer detection. Third, we assume the independent effects of the RCFs, RBFs, and RTFs on the reviewers while ignoring the network effect of belief propagation on the reviewers, reviews, and review targets.

In the future, we will extend the proposed ImDetector approach in three ways. First, we will use advanced tools in psycholinguistics, such as WordNet [62] and LIWC [69], to extract the linguistic and psychological features of reviews and refine the RCFs for fraudulent reviewer detection. We expect this to improve the interpretability and performance of fraud detection [70]. Second, we will detect fraudulent reviewer groups by considering reviewer group behaviors: we will find fraudulent online teams that work collaboratively to manipulate online reviews [71]. Third, we will develop a heterogeneous network to model the relationships among reviewers, reviews, and target products to improve both fraudulent reviewer detection and fraudulent review detection in a unified framework [18,72]. The proposed ImDetector can be used to identify fraudulent reviewers and thus fraudulent reviews. It can also identify fraudulent reviews automatically using natural language processing and deep learning [9,10,68]. Therefore, with the help of fraudulent reviews, we can further identify fraudulent reviewers.

Declaration of Competing Interest

None.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China under Grant Nos. 72174018 and 71932002; the Beijing Natural Science Foundation under Grant No. 9222001; the Philosophy and Sociology Science Fund from Beijing Municipal Education Commission (SZ202110005001).

References

[1] G. Askalidis, E.C. Malthouse, The value of online customer reviews, in: RecSys 2016 - Proc. 10th ACM Conf. Recomm. Syst, 2016, pp. 155–158.
[2] Rosie Murphy, Local Consumer Review Survey 2020. https://www.brightlocal.com/research/local-consumer-review-survey/, 2020.
[3] T. Collinger, How Online Reviews Influence Sales. https://spiegel.medill.northwestern.edu/how-online-reviews-influence-sales/, 2017.
[4] M. Luca, G. Zervas, Fake it till you make it: reputation, competition, and yelp review fraud, Manag. Sci. 62 (2016) 3412–3427.
[5] M. Anderson, J. Magruder, Learning from the crowd: regression discontinuity estimates of the effects of an online review database, Econ. J. 122 (2012) 957–989.
[6] G. Sterling, Study Finds 61 Percent of Electronics Reviews on Amazon Are 'Fake'. https://martech.org/study-finds-61-percent-of-electronics-reviews-on-amazon-are-fake/, 2018.
[7] Trustpilot, The Critical Role of Reviews in Internet Trust. https://business.trustpilot.com/guides-reports/build-trusted-brand/the-critical-role-of-reviews-in-internet-trust, 2020.
[8] J. Boyce, It is Time to Kick Counterfeit Goods Off Amazon, Facebook Marketplace and other e-commerce Sites. https://www.marketwatch.com/story/it-is-time-to-kick-counterfeit-goods-off-amazon-facebook-marketplace-and-other-e-commerce-sites-11617975097, 2021.
[9] W. Zhang, Q. Wang, X. Li, T. Yoshida, J. Li, DCWord: a novel deep learning approach to deceptive review identification by word vectors, J. Syst. Sci. Syst. Eng. 28 (2019) 731–746.
[10] W. Zhang, Y. Du, T. Yoshida, Q. Wang, DRI-RCNN: an approach to deceptive review identification using recurrent convolutional neural network, Inf. Process. Manag. 54 (2018) 576–592.
[11] D. Zhang, L. Zhou, J.L. Kehoe, I.Y. Kilic, What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews, J. Manag. Inf. Syst. 33 (2016) 456–481.
[12] G. Shan, L. Zhou, D. Zhang, From conflicts and confusion to doubts: examining review inconsistency for fake review detection, Decis. Support. Syst. 144 (2021).
[13] B. Manaskasemsak, J. Tantisuwankul, A. Rungsawang, Fake review and reviewer detection through behavioral graph partitioning integrating deep neural network, Neural Comput. Applic. (2021) 1–14.
[14] N. Kumar, D. Venugopal, L. Qiu, S. Kumar, Detecting review manipulation on online platforms with hierarchical supervised learning, J. Manag. Inf. Syst. 35 (2018) 350–380.
[15] N. Kumar, D. Venugopal, L. Qiu, S. Kumar, Detecting anomalous online reviewers: an unsupervised approach using mixture models, J. Manag. Inf. Syst. 36 (2019) 1313–1346.
[16] F. Zhang, X. Hao, J. Chao, S. Yuan, Label propagation-based approach for detecting review spammer groups on e-commerce websites, Knowl.-Based Syst. 193 (2020).
[17] Z. Wang, R. Hu, Q. Chen, P. Gao, X. Xu, ColluEagle: collusive review spammer detection using Markov random fields, Data Min. Knowl. Disc. 34 (2020) 1621–1641.
[18] Z. Wang, S. Gu, X. Zhao, X. Xu, Graph-based review spammer group detection, Knowl. Inf. Syst. 55 (2018) 571–597.

[19] A. Ghose, P.G. Ipeirotis, Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics, IEEE Trans. Knowl. Data Eng. 23 (2011) 1498–1512.
[20] P.F. Wu, Motivation crowding in online product reviewing: a qualitative study of amazon reviewers, Inf. Manag. 56 (2019), 103163.
[21] H.A. Khorshidi, U. Aickelin, Constructing classifiers for imbalanced data using diversity optimisation, Inf. Sci. 565 (2021) 1–16.
[22] W. Wang, D. Sun, The improved AdaBoost algorithms for imbalanced data classification, Inf. Sci. 563 (2021) 358–374.
[23] H. Kaur, H.S. Pannu, A.K. Malhi, A systematic review on imbalanced data challenges in machine learning: applications and solutions, ACM Comput. Surv. 52 (2019) 1–36.
[24] R. Longadge, S. Dongre, Class Imbalance Problem in Data Mining Review, ArXiv Preprint ArXiv. 1305.1707, 2013.
[25] Y. Wu, E.W.T. Ngai, P. Wu, C. Wu, Fake online reviews: literature review, synthesis, and directions for future research, Decis. Support. Syst. 132 (2020), 113280.
[26] S. Rayana, L. Akoglu, Collective opinion spam detection: Bridging review networks and metadata, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min, 2015, pp. 985–994.
[27] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide to Support Vector Classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2008.
[28] W. Wei, J. Li, L. Cao, Y. Ou, J. Chen, Effective detection of sophisticated online banking fraud on extremely imbalanced data, World Wide Web 16 (2013) 449–475.
[29] T. Menzies, Z. Milton, B. Turhan, B. Cukic, Y. Jiang, A. Bener, Defect prediction from static code features: current results, limitations, new approaches, Autom. Softw. Eng. 17 (2010) 375–407.
[30] S. Wan, Y. Duan, Q. Zou, HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source, Proteomics 17 (2017).
[31] Nitesh V. Chawla, et al., SMOTE, J. Artif. Intell. Res. 16 (2002) 321–357.
[32] H. Zhang, M. Li, RWO-sampling: a random walk over-sampling approach to imbalanced data classification, Inform. Fusion 20 (2014) 99–116.
[33] M. Galar, A. Fernández, E. Barrenechea, F. Herrera, EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling, Pattern Recogn. 46 (2013) 3460–3471.
[34] H. Yu, J. Ni, J. Zhao, ACOSampling: an ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data, Neurocomputing 101 (2013) 309–318.
[35] S. Maldonado, R. Weber, F. Famili, Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines, Inf. Sci. 286 (2014) 228–246.
[36] L. Yin, Y. Ge, K. Xiao, X. Wang, X. Quan, Feature selection for high-dimensional imbalanced data, Neurocomputing 105 (2013) 3–11.
[37] J. Du, C.M. Vong, C.M. Pun, P.K. Wong, W.F. Ip, Post-boosting of classification boundary for imbalanced data using geometric mean, Neural Netw. 96 (2017) 101–114.
[38] T. Imam, K.M. Ting, J. Kamruzzaman, Z-SVM: an SVM for improved classification of imbalanced data, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2006, pp. 264–273.
[39] P. Cao, D. Zhao, O. Zaiane, An optimized cost-sensitive SVM for imbalanced data learning, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2013, pp. 280–292.
[40] S.H. Khan, M. Hayat, M. Bennamoun, F.A. Sohel, R. Togneri, Cost-sensitive learning of deep feature representations from imbalanced data, IEEE Trans. Neural Networks Learn. Syst. 29 (2018) 3573–3587.
[41] W. Zhang, L. Yu, T. Yoshida, Q. Wang, Feature weighted confidence to incorporate prior knowledge into support vector machines for classification, Knowl. Inf. Syst. 58 (2019) 371–397.
[42] B. Krawczyk, M. Woźniak, G. Schaefer, Cost-sensitive decision tree ensembles for effective imbalanced classification, Appl. Soft Comput. J. 14 (2014) 554–562.
[43] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[44] A. Moreo, A. Esuli, F. Sebastiani, Distributional random oversampling for imbalanced text classification, in: SIGIR 2016 - Proc. 39th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr, 2016, pp. 805–808.
[45] R.M. Pereira, Y.M.G. Costa, C.N. Silla, MLTL: a multi-label approach for the Tomek link undersampling algorithm: MLTL: the multi-label Tomek link, Neurocomputing 383 (2020) 95–105.
[46] J.C. Campbell, A. Hindle, E. Stroulia, Latent Dirichlet allocation: extracting topics from software engineering data, in: Art Sci. Anal. Softw. Data, 2015, pp. 139–159.
[47] E. Grivel, R. Diversi, F. Merchan, Kullback-Leibler and Rényi divergence rate for Gaussian stationary ARMA processes comparison, Digit. Signal Proc. 116 (2021).
[48] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: ACM Int. Conf. Proceeding Ser, 2007, pp. 209–216.
[49] H. Wang, L. Feng, X. Meng, Z. Chen, L. Yu, H. Zhang, Multi-view metric learning based on KL-divergence for similarity measurement, Neurocomputing 238 (2017) 269–276.
[50] R. Kapoor, R. Gupta, L.H. Son, S. Jha, R. Kumar, Boosting performance of power quality event identification with KL divergence measure and standard deviation, Measurement 126 (2018) 134–142.
[51] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, M. Welling, Fast collapsed gibbs sampling for latent dirichlet allocation, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min, 2008, pp. 569–577.
[52] C.M. Bishop, Pattern Recognition and Machine Learning Solutions to Exercis, 2006.
[53] D.A. Rolls, G. Robins, Minimum distance estimators of population size from snowball samples using conditional estimation and scaling of exponential random graph models, Comput. Stat. Data Anal. 116 (2017) 32–48.
[54] S. Siegel, Non-parametric Statistics for the Behavioral Sciences, 1956.
[55] L. Akoglu, R. Chandy, C. Faloutsos, Opinion fraud detection in online reviews by network effects, in: Proc. 7th Int. Conf. Weblogs Soc. Media, ICWSM 2013, 2013, pp. 2–11.
[56] A. De Caigny, K. Coussement, K.W. De Bock, A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, Eur. J. Oper. Res. 269 (2018) 760–772.
[57] J. Ross Quinlan, R.L. Rivest, Inferring decision trees using the minimum description lenght principle, Inf. Comput. 80 (1989) 227–248.
[58] L. Rutkowski, M. Jaworski, L. Pietruczuk, P. Duda, The CART decision tree for mining data streams, Inf. Sci. 266 (2014) 1–15.
[59] P. Xuan-Hieu, N. Cam-Tu, Gibbslda++: A c/c++ Implementation of Latent Dirichlet Allocation. http://gibbslda.sourceforge.net/, 2007.
[60] T.P. Minka, Estimating a Dirichlet Distribution, 2020.
[61] W. Zhang, T. Yoshida, X. Tang, Text classification based on multi-word with support vector machine, Knowl. Based Syst. 21 (2008) 879–886.
[62] W. Zhang, T. Yoshida, X. Tang, A comparative study of TF*IDF, LSI and multi-words for text classification, Expert Syst. Appl. 38 (2011) 2758–2765.
[63] S. Rosset, Model selection via the AUC, in: Proceedings, Twenty-First Int. Conf. Mach. Learn. ICML 2004, 2004, pp. 703–710.
[64] N.W. Twyman, J.G. Proudfoot, R.M. Schuetzler, A.C. Elkins, D.C. Derrick, Robustness of multiple indicators in automated screening systems for deception detection, J. Manag. Inf. Syst. 32 (2015) 215–245.
[65] L. Li, A new complexity bound for the least-squares problem, Comput. Math. Appl. 31 (1996) 15–16.
[66] J. Yedidia, W. Freeman, Understanding belief propagation and its generalizations, in: Exploring Artificial Intelligence in the New Millennium, 2001, pp. 239–269. http://dce.hut.edu.vn/personal/vinhlt/Papers/MustReadPapers/UnderstandingBP.pdf.
[67] T. Joachims, Svmlight: support vector machine, in: SVM-light Support Vector Machine Http://Svmlight.Joachims.Org/, University of Dortmund, 1999.
[68] W. Zhang, C. Bu, T. Yoshida, S. Zhang, CoSpa: a co-training approach for spam review identification with support vector machine, Information (Switzerland) 7 (2016).
[69] C.K. Chung, J.W. Pennebaker, Linguistic inquiry and word count (LIWC): pronounced "Luke,". and other useful facts, in: Applied Natural Language Processing: Identification, Investigation and Resolution, 2011, pp. 206–229.
[70] M. Siering, J.A. Koch, A.V. Deokar, Detecting fraudulent behavior on crowdfunding platforms: the role of linguistic and content-based cues in static and dynamic contexts, J. Manag. Inf. Syst. 33 (2016) 421–455.
[71] Z. Wang, S. Gu, X. Xu, GSLDA: LDA-based group spamming detection in product reviews, Appl. Intell. 48 (2018) 3094–3107.
[72] G. Wang, S. Xie, B. Liu, P.S. Yu, Identify online store review spammers via social review graph, in: ACM Trans. Intell. Syst. Technol, 2012, pp. 1–21.

Wen Zhang is a professor of College of Economics and Management, Beijing University of Technology. He received his PhD degree in knowledge science from the Japan Advanced Institute of Science and Technology in 2009. His recent research interests include E-commerce big data analysis, data mining, and information systems.

Rui Xie is a PhD candidate of College of Economics and Management, Beijing University of Technology. His research interest includes data mining and machine learning.

Qiang Wang is a PhD candidate of College of Economics and Management, Beijing University of Technology. He received his BS degree in marketing from Qufu Normal University in 2016. His research interest includes E-commerce big data analysis, data mining, and machine learning.

Ye Yang is an associate professor of School of Systems and Enterprises, Stevens Institute of Technology. Her recent research interests include recommendation systems, big data analytics, machine learning, and crowdsourcing data analysis.

Jian Li is a professor of College of Economics and Management, Beijing University of Technology. He received his PhD degree from Chinese Academy of Sciences in 2007. His recent research interests include big data analytics, data-driven decision making, and digital economics.