The Journal of Supercomputing (2023) 79:419–434
https://doi.org/10.1007/s11227-022-04685-z

Deep feature fusion for cold-start spam review detection

Lingyun Xiang[1,2,3] · Huiqing You[2] · Guoqing Guo[2] · Qian Li[4]

Accepted: 21 June 2022 / Published online: 11 July 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

The cold-start problem in spam review detection is a significant challenge referring to identifying the authenticity of the first review posted by new users. For generating more sensitive features to identify new reviews, existing methods mainly leverage text-similarity of review to find relevant features to approximate the incomplete behavior features of new reviews. However, they over-rely on the text information of new reviews while ignoring the mutual behavioral information in the review system, leading to a decrease in the sensitivity of features. To address the issue, we propose a deep feature fusion method, which balances the importance of text information and behavior information to enhance features' sensitivity. Specifically, we construct a heterogeneous graph, where products and users serve as vertices connected by edges representing reviews. Then, we perform graph convolution calculation on this graph in the first feature fusion stage. We utilize the mutual behavioral information in the review system to compensate for the incomplete behavior feature of new reviews. Furthermore, we design a co-attention network, which can give features different weights in the global feature fusion stage, to gain features with high sensitivity of identifying new reviews. Extensive experiments on Yelp-hotel and Yelp-restaurant datasets demonstrate that our proposed approach yields better classification performance over existing methods.

Keywords Co-attention network · Cold-start · Graph convolution network · Spam review detection

✉ Huiqing You
youhuiqing@stu.csust.edu.cn

[1] Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha 410114, China

[2] School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China

[3] Hunan Provincial Key Laboratory of Smart Roadway and Cooperative Vehicle-Infrastructure Systems, Changsha University of Science and Technology, Changsha 410114, China

[4] North China Institute of Computing Technology, Beijing 100083, China

1 Introduction

In the era of rapid development of online consumption, for businesses, the quality of online reviews is closely linked to their profit level, especially spam reviews [1]. So it is of vital significance to detect spam reviews. For example, spam reviews will mislead consumers into inauthenticity first impression of goods, resulting in a lower profit of merchants and bad consumption experience of consumers [2, 3]. Existing spam review detection methods mainly focus on extracting features to identify the authenticity of reviews [4]. These features can be divided into linguistic features and behavior features. It is difficult for humans to distinguish the authenticity of reviews just by reading their text [5], so using linguistic features separately has also proved ineffective in detecting spam reviews [6, 7], while the extraction of behavior features usually requires a large number of samples, which cost high compute resource [8]. When facing new users, because they only post a new review, it is hard to extract their behavior features, and linguistic features of the new review are limited. The above reasons result in extracting sensitive features to identify new reviews hardly, which is the cold-start problem in the field of spam review detection [9]. So it is a significant challenge that identifies new reviews with limited information.

Recently, to solve the cold-start problem, many researchers have studied it. [7, 10] adopted the knowledge graph embedding method to model the relationship among the three components, namely review, user, and product. In contrast, [11] used heterogeneous information networks to aggregate the linguistic information among the three components. For the sake of learning their representation, respectively, Wang and You et al. applied TransE [12] embedding model and attempted to jointly learn significant features of each of the three components. Although the TransE model is simple and effective in capturing multiple relationships, its well-known limitation is that it only works for 1-to-1 relationships, not 1-to-N or N-to-1 relationships [13]. Shehnepoor et al. adopted convolutional neural network (CNN) pretraining to obtain the word embedding, then used graph learning to gain the representation of each component, which solved the problem that the TransE model could not capture 1-to-N or N-to-1 relationships. However, it still neglected the original and incomplete behavior features of new users. Furthermore, for cold-start users, the linguistic feature of the review and user components is consistent, leading to problems in information aggregating. So this method is still inadequate in utilizing the mutual behavioral information in the review system and over-rely on text information. The mutual behavioral information contains the extra important information for the new reviews in the cold-start environment, so making full use of mutual behavioral information is helpful to solve the cold-start problem.

In recent years, GCN has been widely studied in association information in a graph. The core idea of GCN is to extend the convolution operation and generate a new representation of the nodes in the graph through the mapping function, which can aggregate the features of the node itself and the features of the neighbor nodes. So it can effectively process the graph data [14]. Xu [15] et al. proposed a GCN with the role-constrained conditional random field, which is used to learn the feature representation of applicants in financial loans, to detect loan

fraud by utilizing user roles and multiple types of social association information among users. Kudo [16] et al. proposed a GCN framework with augmented balance theory for spammer detection on social platforms. Based on GCN, Zhang [17] et al. proposed a user representation learning method for the detection of fraudsters in recommendation systems. The above research works show that GCN can effectively learn the mutual behavioral information in graph data nodes. However, it is not enough that only leveraging the mutual behavioral information, a more sensitive feature representation learning method is needed to take full advantage of text information and mutual behavior information in the review system.

In order to solve the problem mentioned above, this paper proposes a deep feature fusion method, which first performs behavior feature fusion to obtain behavior association features (BAFs) by leveraging graph convolutional network (GCN). It makes full use of the mutual behavioral information in the review system to learn the BAFs. Subsequently, through the co-attention network, we combine linguistic features and BAFs for global feature fusion, which makes up for the deficiency of the sensitive features of new reviews in the cold-start environment to improve the sensitiveness of new reviews detection.

Although new users only comment on one product in the cold-start environment, other users comment on this product in the meantime. Moreover, other users also comment on many products, so the relationship in them can be effectively modeled by constructing a heterogeneous graph. We leverage users' activity posting reviews on products, which associates users with products, to build a heterogeneous graph of users and various products. This graph includes direct and indirect behavior association information in the review system. This paper adopts GCN to utilize the mutual behavioral information in the review system for behavior feature fusion. Therefore, each review can learn adequate behavior association information from the users and products associated with it, which solves the insufficient use of mutual behavioral information in the review system.

Due to the first step of feature fusion only focusing on behavior information and neglecting the text information of new reviews, we should leverage text information and balance their importance at the global feature fusion stage. Obtaining effective representation of new review text also has an indispensable impact on this stage. Therefore, we leverage BERT to make more effective use of the review text information at the sentence level, combined with the word representation of the context. Moreover, learn the common (average) representation of words through the fine-tuning BERT model that is based on a large number of corpus training so that each comment text can obtain better self-representation [18].

In order to more effectively balance linguistic features and BAFs to gain the final classification feature, we use a co-attention network to give linguistic features and BAFs different weights in the global feature fusion stage. So we can avoid the adverse effect of ignoring the different importance of the linguistic feature and BAFs at different hierarchies on the final classification. The final classification features obtained by the co-attention network are sensitive to identifying new reviews.

The research on cold-start spam review detection almost performs experiments on two Yelp datasets; for better comparison, we also run all the experiments on these two public datasets. Extensive experiments show that this method has good detection performance under a cold-start environment.

Our contributions of this work can be summarized as follows:

- We propose a novel deep feature fusion method. In the behavior feature fusion stage, we leverage GCN to make full use of the behavior features of users and products in a cold-start environment and learn BAFs to compensate for the incomplete behavior features of new reviews. In addition, global feature fusion solves the deficiency-sensitive features of the new review by fusing the linguistic feature and BAFs through the Co-attention network.
- To our best knowledge, this is the first work that leverages GCN to perform behavior feature fusion to learn BAFs representation of new reviews. Comparative experiments prove that the BAFs learned through behavior feature fusion can effectively improve cold-start spam review detection.
- The results of contrast experiments give reasonable confidence that a co-attention network can improve the effectiveness of global feature fusion, and the review text can obtain better self-representation through BERT.

The rest of this paper is structured as follows. In Sect. 2, we present the details of the proposed method. Then, we show experiments and analysis to evaluate the proposed method in Sect. 3. Finally, we conclude this paper with an outlook to the future in Sect. 4.

2 Proposed cold-start spam review detection method

This research proposes a spam review detection method for cold-start problems via deep feature fusion. The framework of the proposed method is shown in Fig. 1. It can be divided into two feature fusion stages: behavior feature fusion and global feature fusion.

We adopt two steps to achieve behavior feature fusion in the first stage to generate BAFs for reviews. Firstly, we model a review system as a heterogeneous information graph. Each node is a user or product, and the edge indicates that the user has commented on the product. Behavior features of users and products are taken as values of user nodes and product nodes, respectively. The heterogeneous information graph constructed by this method can store the behavior association information among users, products, and reviews. Subsequently, the GCN is used to learn user-based BAF and product-based BAF, and we combine them as BAFs to take full advantage of behavior association information in the review system.

The global feature fusion stage can also be divided into two steps: extracting linguistic features and leveraging a co-attention network to fuse full features for final classification. When extracting linguistic features, this paper leverages BERT to learn global semantic information features from the text content of reviews. The first review can also utilize global information to obtain better self-representation.

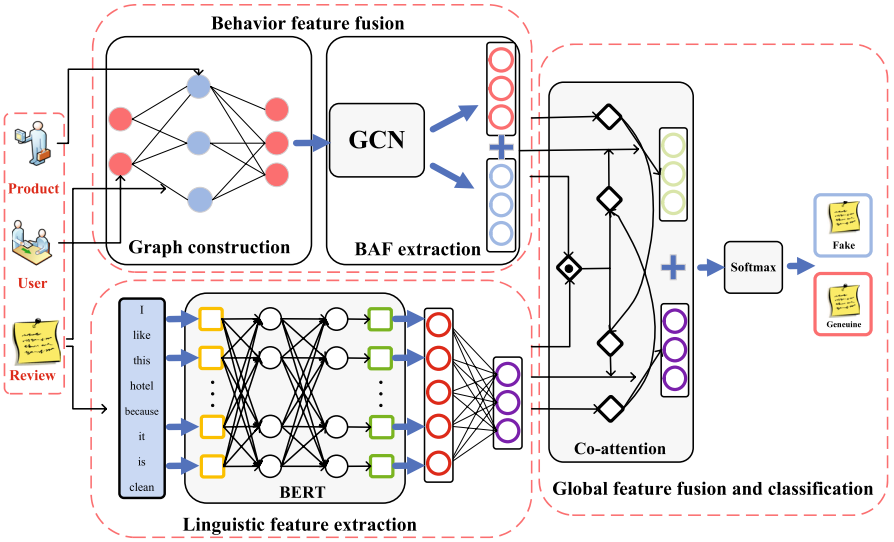

Fig. 1 The framework of the proposed cold-start spam review detection method. In the behavior feature fusion process, we leverage GCN to obtain BAFs. Meanwhile, we put the text of the review into BERT to gain the linguistic features. In the Global feature fusion and classification process, we use a co-attention network to learn the different weights for BAFs and linguistic features to fuse them, and we get the final classification results with a softmax layer

Utilizing fusing the linguistic feature of reviews with the learned BAFs through a co-attention network, which can alleviate the destructive impact of ignoring the importance of different features, we can generate sensitive features representation of new reviews. Finally, this feature of new reviews is fed into a softmax classifier to identify whether new reviews are genuine reviews or not.

2.1 Heterogeneous graph construction

Unlike the existing modeling review system, to better extract and utilize the behavior association between products and users related to new reviews, this paper constructs a heterogeneous graph with users and products serving as nodes. The graph includes two types of relationships: 1.review-based relationship (user, *review*, product), 2.product-based relationship (product, *be reviewed*, user). A user can review multiple products, and a product can also be reviewed by multiple users. Through these two types of relationships, we can better connect the old users and products with the new review in the cold start environment.

When constructing the graph, if the user has reviewed the product, an edge is built from the user to the product. Meanwhile, in this circumstance, the product has been reviewed by the user, and another edge from the product to the user is built. Features of user nodes use the behavior features BF_u, features of products node use the behavior features BF_p.

$$BF_u = \{uMNR, uPR, uNR, uERD, uavgRD, uBST\} \tag{1}$$

$$BF_p = \{pMNR, pPR, pNR, pavgRD, pERD\} \tag{2}$$

where BF_u and BF_p are extracted by the existing method [19], and the meanings of all eigenvalues can be found in Table 1.

2.2 BAFs extraction

After constructing a heterogeneous graph, to avoid the negative impact of the decline in feature sensitivity caused by excessive dependence on ext information, we conduct graph convolution calculation on this graph to conduct behavior feature fusion. Due to the graph includes two types of relationships, there are two types of (source node, target node): 1. (the user node, the product node), 2. (the product node, the user node). After graph convolution calculation, the target node will learn the new representation. Therefore, the user node and the product node will capture the deep information from the products-based BAFs and the user-based BAF, respectively. The user-based BAFs and product-based BAFs are obtained using the behavioral association information in this graph, which compensates for the incomplete behavior features of new reviews.

The behavior feature fusion process under cold-start environment, i.e., the process of behavior association information aggregation of the new review, is shown in Fig. 2. After inputting heterogeneous graph and corresponding behavior feature matrix into GCN, the aggregation process of behavior association information corresponding to the new review is shown in the left part of Fig. 2. Through graph convolution operation, both p1 node and u1 node can aggregate behavior information of

Table 1 Behavior features of users and products

Features	Meaning
pMNR	Maximum number of reviews that a product received within a day [6]
pPR	The ratio of positive reviews (4–5 star) in all of the product's reviews [6]
pNR	Ratio of negative reviews (1–2 star) in all of the product's reviews [6]
pavgRD	Average deviation rate [6]
pERD	Distribution entropy of the average evaluation score obtained [20]
uMNR	Maximum number of reviews that a user posted within a day [6]
uPR	The ratio of positive reviews (4–5 star) in all reviews posted by this user [6]
uNR	The ratio of negative reviews (1–2 star) in all reviews posted by this user [6]
uERD	Distribution entropy of user evaluation scores [20]
uavgRD	Average deviation rate [6]
uBST	Burstiness, $x_{BST}(i) = \begin{cases} 0, \text{ if } L(i) - F(i) > \tau \\ 1 - \frac{L(i)-F(i)}{\tau}, \text{ otherwise} \end{cases}$, where L(i) − F(i) describes days between last and first review, and $\tau = 29$ (day) [6].

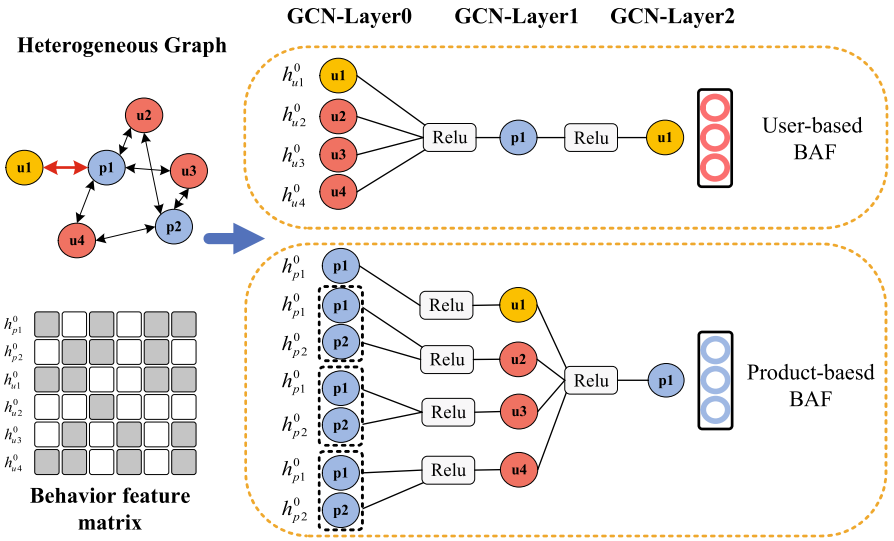

Fig. 2 The new review behavioral association information aggregation process. The edge marked in red is a review posted by the new user, u1 is a node of the new user, p1 is a node of a product reviewed by the new user, u2, u3, u4 are nodes of users who have commented on p1 and p2. The behavior feature matrix at the bottom left is a matrix composed of values of each node in the heterogeneous graph

neighbor nodes, including itself, and update their node features to obtain product-based BAF and user-based BAF corresponding to the new review, respectively. The mathematical definition of graph convolution in a heterogeneous graph is as follow:

$$h_{s_{dst}}^{(l+1)} = \underset{r \in \mathcal{R}, r_{dst}=dst}{AGG} (f_s(g_s, h_{s_{src}}^l, h_{s_{dst}}^l)) \tag{3}$$

where f_s represents convolution module corresponding to each relationship s, AGG is an aggregation function, and $h_{s_{src}}^l$ is feature of source node of the relation s, $h_{s_{dst}}^l$ is feature of target node of the relation s. During initialization, if the node type is the user, its eigenvalues h are behavior features of user BF_u corresponding to the node. If the node type is a product, its eigenvalues h are behavior features of product BF_p corresponding to the node. The aggregation function used in this paper is sum, and the convolution module uses the graph convolution method proposed by Kipf et al. [21]; we can determine it is by:

$$h_i^{(l+1)} = \sigma\left(b^l + \sum_{j \in N(i)} \frac{1}{c_{ji}} h_j^l W^l \right) \tag{4}$$

where $N(i)$ is neighbor nodes set of node i, $c_j i$ is the product of the square root of the node degree, i.e., $c_{ji} = \sqrt{|N(j)|}\sqrt{|N(i)|}$, h_j^l represents the feature of node j, W^l represents learnable weights, b^l represents bias, σ is activation function, we used Relu in this paper.

After the convolution operation on the heterogeneous graph, each edge will learn its source node BAFs h_{src} and target node BAFs h_{dst}, these two hidden features fully utilize the behavior association information under the cold-start environment. h_{src} and h_{dst} represent user-based BAFs or product-based BAFs based on different relationships. For example, in a (user, review, product) relationship, h_{src} is a user-based BAFs, and h_{dst} is a product-based BAFs. At the end, we combine h_{src} and h_{dst} as h_i. Y is the BAFs map containing all h_i of reviews.

$$h_i = h_{src} \oplus h_{dst} \tag{5}$$

2.3 Linguistic feature extraction

The acquisition of linguistic features of reviews depends on the text content of the review itself. Under the cold-start environment, only text information of the new review is complete, so extracting more useful linguistic features is also the key to improving the effectiveness of cold-start spam review detection. Based on the principle of BERT extracting linguist features described in [22], this paper improves the linguistic feature extraction method in [23], using fake review text and genuine review text to train a BERT-based linguistic feature extraction model. Specifically, we construct the sentence-pair input:[*CLS*] *sentenceA* [*SEP*] *sentenceB* [*SEP*], where [*CLS*] and [*SEP*] are special embeddings for classification and separating sentences. Moreover, in the fine-tuning BERT model training process, we do not fix the collocation; in other words, we only ensure that the proportion of genuine reviews and fake reviews is 50 %, but the order is random.

Then, we use a pre-trained fine-tuning BERT model to vectorize each review text, and on the structure described in [23], a fully connected layer with an output dimension of 32 is added. Moreover, the softmax activation function is used for processing to realize the two-classification of text content. The process is described as follow:

$$class_X = softmax\left(W_X \cdot X(i) + b_X\right) \tag{6}$$

where $X(i)$ is the value obtained by the review text i through BERT, W_X is the learnable weight matrix, b_X represents bias. After training the BERT-based linguistic feature extraction model, for each review i, $X(i)$ is used as the linguistic feature of review for subsequent global feature fusion.

2.4 Global feature fusion and classification

Due to the different importance of linguistic features and BAFs in obtaining the final classification features, we balance the importance between them and prevent the excessive impact of one of them. At the global feature fusion and classification stage, this paper designs a co-attention network to perform global feature fusion. In this way, we can get the final features with high sensitivity that identify new reviews.

According to the Co-Attention network proposed in [24], we take linguistic feature and BAFs as input and generate linguistic feature and BAFs attention at the same time. The Co-attention network focuses on both linguistic features and BAFs, connecting linguistic features and BAFs by calculating the similarity of linguistic features and BAFs between all pairs of linguistic feature-location and BAFs-location.

Specifically, given a linguistic feature map $X \in R^{d\times N}$, and the BAFs map $Y \in R^{d\times T}$, the affinity matrix $C \in R^{T\times N}$ is calculated by

$$C = \tanh(Y^T W_b X) \tag{7}$$

where $W_b \in R^{d\times d}$contains the weights. After computing this affinity matrix, we consider this affinity matrix as a feature and learn to predict linguistic feature and BAFs attention maps via the following:

$$\begin{cases} H^x = \tanh(W_x X + (W_y Y)C) \\ H^y = \tanh(W_y Y + (W_x X)C^T) \end{cases} \tag{8}$$

$$\begin{cases} a^x = softmax(w_{hx}^T H^x) \\ a^x = softmax(w_{hy}^T H^y) \end{cases} \tag{9}$$

where W_x,$W_y \in R^{k\times d}$,w_{hx},$w_{hy} \in R^k$ are the weight parameters.$a_X \in R^N$ and $a_y \in R^T$ are the attention weight matrix of linguistic feature and BAFs, respectively. The affinity matrix C transforms linguist feature attention space to BAFs attention space (vice versa for C^T). Based on the above attention weights, the linguistic feature and BAFs attention vectors are calculated as the weighted sum of the linguistic feature and BAFs, i.e.,

$$\hat{x} = \sum_{n=1}^{N} a_n^x x_n, \quad \hat{y} = \sum_{t=1}^{T} a_t^y y_t \tag{10}$$

In the field of cold-start spam review detection, there is a situation where the text of spam review is similar to that of genuine review, and softmax is better than SVM in distinguishing samples which has similar representations but with different labels [11]. Therefore, the softmax activation function is added in the fully connected layer for final classification.

$$r = softmax\left(W_F\left(\hat{X} \oplus \hat{Y}\right) + b_F\right) \tag{11}$$

where r is the final classification result, W_F is learnable weight matrix, b_F is bias, $\hat{X}$ and $\hat{Y}$ are the total collection of $\hat{x}$ and $\hat{y}$, respectively.

Table 2 Two Dataset Statistics

Domain	Yelp-Hotel	Yelp-Restaurant
#reviews	688328	788471
#reviewers	5132	35593
Date range	2004.10.23 2012.09.26	2004.10.12 2012.10.02
%before 2012.01.01	99.01%	97.04%

3 Experiments and analysis

3.1 Experimental settings

(1) *Dataset*: We conduct experiments on the following two subset of Yelp dataset, the statistics of these two dataset are listed in Table 2:

- **Yelp-hotel** [20, 25]: This review dataset contains 688328 reviews on hotels, the time when reviews were posted, the rating of reviews, and the label of reviews.
- **Yelp-restaurant** [20. 25]: This review dataset is similar to the Yelp-hotel dataset, but it collects reviews on the restaurant, including 788471 reviews.

In order to solve the cold-start problem, this paper refers to Wang et al. [7], using the first labeled review posted by the new user after January 1, 2012, as the test set, and the first labeled review posted before January 1, 2012, is used as the train set to train GCN-based behavioral association feature extraction model and Co-attention network. In addition, this paper uses all labeled review data before January 1, 2012, to train the BERT-based linguistic feature extraction model.

(2) *Comparison methods*: We compare our method with baseline methods as follow:

- **LF** [26]: The SVM classification results only across bigrams linguistic feature.
- **Supervised-CNN** [7]: This method only uses supervised-CNN to detect spam reviews.
- **LF+BF** [26]: Combined linguistic feature and behavior features for detecting spam reviews.
- **BFEditSim+LF** [7]: the SVM classification results by the intuitive method that finding the most similar existing review by edit distance ratio and take the found reviewers' behavioral features as approximation.
- **BFW2Vsim+W2V** [7]: This method obtains SVM results by averaging pre-trained word embeddings (using Word2Vec) to find the most similar existing reviews.
- **RE+RRE+PRE** [7]: This method uses three new features, which are the learnt review embeddings (RE), the learnt review's rating embeddings (RRE), the learnt product's average rating embeddings (PRE), to perform spam review detection.

(3) *Parameter settings*: In this paper, the output dimension of the GCN-based BAFs extraction model is set to 15, the optimizer uses Adam, the default learning rate is 0.001, the epoch is set to 1000, the loss function uses focal-loss, and the parameter is set to $\alpha = 0.25$. In the training process of the pre-trained BERT-based linguistic feature extraction model, this paper sets the linguistic feature-length to 32, the learning rate is set to 0.00001, and the epoch is set to 1000. The model uses cross-entropy loss and sets the weight ratio of genuine review and spam review to 1:10, which is used to alleviate the imbalance of genuine review and spam review and save the model with the highest F1 during the training process as the final linguistic feature extraction model.

(4) *Metrics*: This paper adopts the same evaluation metrics as [6][8], namely precision (P), recall (R), F1-Score (F1), and accuracy (Acc), to better compare with the existing baseline method.

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{14}$$

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{15}$$

where TP is the number of spam reviews correctly detected as fake reviews, FN is the number of spam reviews incorrectly detected as genuine reviews, FP is the number of genuine reviews incorrectly detected as fake reviews, and TN is the number of genuine reviews correctly detected as genuine reviews.

Table 3 Cold-start spam review detection methods comparison

	Feature	Hotel				Restaurant			
		P	R	F1	Acc	P	R	F1	Acc
1	LF	54.5	71.7	61.7	55.9	53.8	80.8	64.6	55.8
2	Supervised-CNN	61.2	51.7	56.1	59.5	56.9	58.8	57.8	57.1
3	LF+BF	63.4	52.6	57.5	61.1	58.1	61.2	59.6	58.5
4	BFEditSim+LF	55.3	69.7	61.6	56.6	53.9	82.2	65.1	56.0
5	BFW2Vsim+W2V	58.4	65.9	51.9	59.5	56.3	73.4	63.7	58.2
6	RE+RRE+PRE	63.6	71.2	67.2	65.3	59.0	78.8	67.5	62.0
7	LF+BAFs(ours)	**78.8**	**83.6**	**81.1**	**69.9**	**76.9**	**87.0**	**81.4**	**69.9**

Bold values represent the best experimental results in the compared experiment

3.2 Comparison with baseline

In order to prove the effectiveness of our deep feature fusion method, the proposed model is compared with the other six baseline methods. Because the proposed method uses the same dataset and data partitioning method as Wang et al. [7], this paper directly uses the actual results as a comparison. Table 3 shows spam review detection results of the same dataset using different cold-start spam review detection methods.

The proposed method is superior to the comparison method in all evaluation metrics; in particular, it has the most noticeable improvement over other methods on precision, which shows that this method can more accurately identify spam reviews in cold-start scenarios.

In addition, through the analysis of Table 3, the recognition accuracy of the LF based on binary grammar features is the lowest among all the comparison methods. The F1 of the method based on Supervised-CNN is the lowest compared with other methods, the adequate information extracted by simple linguistic feature in cold-start spam review detection is limited, and its performance is not good. Combining behavior features can improve the detection effect in a cold-start environment to some extent. From the results of Model 3, it can be seen that the combination of behavior features and linguistic features increases the adequate information for the first review under a cold-start environment and improves the detection accuracy of fake reviews. However, R and F1 of model 3 were reduced, indicating that this method would lead to more fake reviews identified as genuine reviews under a cold-start environment. The reason is that the original behavior features of new users are incomplete, the direct use of this feature leads to information redundancy, and there is a camouflage problem [8]. Model 4 and Model 5 conduct spam reviews detection by feature replacement from the user's perspective and text similarity, respectively. The experimental results show that replacing the behavior features of reviews to be detected directly with similar review behavior features under a cold-start environment performs the poor effect, which may be because the behavior association information between reviews, users, and products are neglected when replacing the features. Furthermore, models 6,7 construct the behavior features of cold-start reviews by extracting correlation information from existing reviews and combining them with the original behavior features. Compared with other methods, the detection effect is greatly improved.

Because the features based on graph convolution learning utilize the mutual behavioral information in the review system, the problem of missing sensitive features of new reviews is improved by combining practical linguistic features with a co-attention network. Compared with other baseline methods, the method proposed in this paper is superior to other comparison methods in all evaluation metrics.

3.3 Linguistic feature extraction method and global feature fusion study

The review text is the complete original information of the spam review in the cold-start environment. Extracting more useful linguistic features is indispensable

Table 4 Comparison of linguistic feature extraction methods

Method	Hotel				Restaurant			
	P	R	F1	Acc	P	R	F1	Acc
LF(BERT)+BAFs	78.78	83.62	81.13	69.92	76.86	86.99	81.46	69.86
LF(textCNN)+BAFs	78.07	83.59	80.73	69.42	76.60	86.57	81.42	69.69

Table 5 Ablation experiment result

Method	Hotel				Restaurant			
	P	R	F1	Acc	P	R	F1	Acc
LF+BAFs[1]	77.46	82.66	80.44	69.08	76.53	86.12	81.06	69.32
LF+BAFs(ours)	78.78	83.62	81.13	69.92	76.86	86.99	81.46	69.86

[1] We directly splice the linguistic feature and BAFs in this method

in improving final classification results. The existing linguistic feature extraction methods are BERT and textCNN. Both methods consider surrounding information of the word to realize characterization of the word and obtain the word embedding. However, the two methods use surrounding information in different ways. Model architecture and training methods are different, resulting in different sentence representation effects between the two methods.

In order to study the influence of the linguistic feature extracted by BERT and textCNN on the final classifier, we use the same dataset to train two linguistic feature extraction models based on BERT and textCNN according to the paper [18, 19, 23], respectively. Then, the linguistic feature extracted by the two models is fused with BAFs to construct the final classifier. The experimental results of the classification are shown in Table 4.

Through the analysis of Table 4, the pre-training process of BERT uses multi-task training, including two tasks: mask language model and next sentence prediction. Through the task of next sentence prediction, BERT can use the information of sentence granularity to achieve a better representation of sentence information. Therefore, the classification effect of the final classifier constructed with the linguistic feature extracted by BERT is better than with the linguistic feature extracted by textCNN.Therefore, in this scenario, BERT can extract more useful linguistic features than textCNN.

To investigate the influence of the global feature fusion method on the final classifier, after obtaining linguistic feature extracted by BERT and BAFs extracted by GCN, we directly splice the linguistic feature and BAFs and then input them into the classifier as the final features for classification. Subsequently, we give linguistic features different weights from BAFs according to the co-attention network proposed in [24] to obtain the final features. The classification results of these two global feature fusion methods are shown in Table 5.

The co-attention network has improved the model effect. Because direct splicing of two features neglects the critical difference between the two features, through the co-attention network, we can generate both linguistic feature attention and BAFs attention at the same time and calculate the similarity between linguistic feature and BAFs in all pairs of linguistic feature-location and BAFs-location to connect the linguistic feature and BAFs, to avoid the separation of those two features. Therefore, by adopting a co-attention network, we can better fuse linguistic features and BAFs.

4 Conclusion

Aiming at the insufficient sensitive features of the first review issued by new users, this paper proposes a deep feature fusion method framework for spam review detection under a cold-start environment by fusing BAFs and LF. Unlike the previous modeling methods for social review platforms, we take users and products as nodes here. We use reviews as the edges connecting users and products. After graph convolution learning, each review can obtain user-based BAFs and product-based BAFs by fusing behavior features, which effectively use the original behavior features between users and products. That means reviews can collect behavior association information from associated users and products. Subsequently, after obtaining a more effective self-representation of the review text, by fusing BAFs and linguistic features by the co-attention network, we can obtain the final feature for classification to compensate for the lack of sensitive features of the new review. The experimental results show that the method has high detection performance in cold-start spam review detection.

In the future, we will extend from GCN to graph attention network (GAT), giving different importance to each node to study spam review detection under cold-start environments. This method can solve the problem caused by GCN sharing weights [27], and it applies to a cold-start environment, where the importance of new and old users and their comments are inconsistent.

Acknowledgements This project is supported by National Natural Science Foundation of China under Grant 61972057, and 62172059, Hunan Provincial Natural Science Foundation of China under Grant 2022JJ30623, Scientific Research Fund of Hunan Provincial Education Department of China under Grant 21A0211, Hunan Provincial Innovation Foundation For Postgraduate under Grant CX20210812.

Data availability The datasets generated and/or analyzed during the current study are not publicly available due to privacy and confidentiality agreements as well as other restrictions, but are available from the corresponding author on reasonable request.

Declarations

Conflict of interest We have no conflicts of interest.

References

1. Luca M (2016) Reviews, reputation, and revenue: the case of yelp. com. Com (March 15, 2016). Harvard Business School NOM Unit Working Paper (12-016)
2. Ho-Dac NN, Carson SJ, Moore WL (2013) The effects of positive and negative online customer reviews: do brand strength and category maturity matter? J Mark 77(6):37–53
3. Zhu F, Zhang X (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. J Mark 74(2):133–148
4. Hussain N, Mirza HT, Hussain I, Iqbal F, Memon I (2020) Spam review detection using the linguistic and spammer behavioral methods. IEEE Access 8:53801–53816
5. Mohawesh R, Xu S, Tran SN, Ollington R, Springer M, Jararweh Y, Maqsood S (2021) Fake reviews detection: a survey. IEEE Access 9:65771–65802
6. Mukherjee A, Venkataraman V, Liu B, Glance N (2013) What yelp fake review filter might be doing? In: Seventh International AAAI Conference on Weblogs and Social Media
7. Wang X, Liu K, Zhao J (2017) Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 366–376
8. Shehnepoor S, Salehi M, Farahbakhsh R, Crespi N (2017) Netspam: a network-based spam detection framework for reviews in online social media. IEEE Trans Inf Forensics Secur 12(7):1585–1595
9. Dou Y (2019) A review of recent advance in online spam detection
10. You Z, Qian T, Liu B (2018) An attribute enhanced domain adaptive model for cold-start spam review detection. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1884–1895
11. Shehnepoor S, Togneri R, Liu W, Bennamoun M (2021) Dfraud^3: Multi-component fraud detection free of cold-start. IEEE Trans Inf Forensics Secur 16:3456–3468
12. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. Adv Neural Inform Process Syst **26**
13. Huynh V-P, Papotti P (2018) Towards a benchmark for fact checking with knowledge bases. In: Companion Proceedings of the The Web Conference 2018, pp. 1595–1598
14. Gori M, Monfardini G, Scarselli F (2005) A new model for learning in graph domains. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., vol. 2, pp. 729–734 . IEEE
15. Xu B, Shen H, Sun B, An R, Cao Q, Cheng X (2021) Towards consumer loan fraud detection: Graph neural networks with role-constrained conditional random field. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 4537–4545
16. Kudo W, Nishiguchi M, Toriumi F (2020) Gcnext: graph convolutional network with expanded balance theory for fraudulent user detection. Soc Netw Anal Min 10(1):1–12
17. Zhang S, Yin H, Chen T, Hung QVN, Huang Z, Cui L (2020) Gcn-based user representation learning for unifying robust recommendation and fraudster detection. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 689–698
18. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
19. Xiang L, Guo G, Li Q, Zhu C, Chen J, Ma H (2020) Spam detection in reviews using lstm-based multi-entity temporal features. Intell Automat Soft Comput
20. Rayana S, Akoglu L (2015) Collective opinion spam detection: Bridging review networks and metadata. In: Proceedings of the 21th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 985–994
21. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
22. Du C, Sun H, Wang J, Qi Q, Liao J (2020) Adversarial and domain-aware bert for cross-domain sentiment analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4019–4028
23. Zhang X, Lai H, Feng J (2018) Attention-aware deep adversarial hashing for cross-modal retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 591–606
24. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. Adv Neural Inf Process Syst 29:289–297

25. Mukherjee A, Kumar A, Liu B, Wang J, Hsu M, Castellanos M, Ghosh R (2013) Spotting opinion spammers using behavioral footprints. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 632–640
26. Mukherjee A, Venkataraman V, Liu B, Glance N et al (2013) Fake review detection: Classification and analysis of real and pseudo reviews. UIC-CS-03-2013. Technical Report
27. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint arXiv:1710.10903

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.