Original research article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–16

DOI: 10.1177/20552076241277171
journals.sagepub.com/home/dhj

Sage

How to detect fake online physician reviews: A deep learning approach

Yuehua Zhao*, Tianyi Li*, Qinjian Yuan and Sanhong Deng

Abstract

Objective: The COVID-19 pandemic has spurred an increased interest in online healthcare and a surge in usage of online healthcare platforms, leading to a proliferation of user-generated online physician reviews. Yet, distinguishing between genuine and fake reviews poses a significant challenge. This study aims to address the challenges delineated above by developing a reliable and effective fake review detection model leveraging deep learning approaches based on a fake review dataset tailored to the context of Chinese online medical platforms.

Methods: Inspired by prior research, this paper adopts a crowdsourcing approach to assemble the fake review dataset for Chinese online medical platforms. To develop the fake review detection models, classical machine learning models, along with deep learning models such as Convolutional Neural Network and Bidirectional Encoder Representations from Transformers, were applied.

Results: Our experimental deep learning model exhibited superior performance in identifying fake reviews on online medical platforms, achieving a precision of 98.36% and an F2-Score of 97.97%. Compared to the traditional machine learning models (i.e., logistic regression, support vector machine, random forest, ridge regression), this represents an 8.16% enhancement in precision and a 7.7% increase in F2-Score.

Conclusion: Overall, this study provides a valuable contribution toward the development of an effective fake physician review detection model for online medical platforms.

Keywords

Online physician review, physician rating website, online medical review, fake review detection, online fake review

Submission date: 5 January 2024; Acceptance date: 5 August 2024

Introduction

The advent of Web 2.0 technology has revolutionized traditional industries by providing new platforms and channels for promotion, including the healthcare industry. Over the past decade, there has been a growing demand for Internetization in healthcare, as evidenced by the increasing number of adults in the United States using the Internet to access health information.[1,2] This demand has spurred the creation and development of online healthcare platforms, which combine Internet technologies with the traditional medical industry to facilitate doctor–patient communication and provide medical services. Well-known examples of such platforms include *haodf.com*, *chunyuyisheng.com*, and *zocdoc.com*.[3] These platforms offer patients convenient medical services and physicians the opportunity to expand their reach and earnings. However, the reliability of these platforms for patients is sometimes called into question due to the freedom to post statements online and the use

School of Information Management, Nanjing University, Nanjing, China

*These authors contributed equally to this study.

Corresponding author:
Sanhong Deng, School of Information Management, Nanjing University, No.163, Xianlin Road, Qixia District, Nanjing, China.
Email: sanhong@nju.edu.cn

of reputation management services by physicians to garner more positive reviews and higher ratings.

Similar issues are common in other types of online platforms, such as online hotel and restaurant platforms, which have been extensively researched in recent years.[4,5] In particular, researchers have focused on identifying fake reviews, which mislead readers by offering unobjective and unjust evaluations of target objects.[6] While numerous studies have identified fake reviews on platforms such as Yelp and TripAdvisor, relatively few have focused on identifying fake reviews on online medical platforms. Early research in this area used a dataset of physician reviews constructed by Li et al.[7] but only a few studies have explored the performance of machine learning and deep learning models in detecting fake physician reviews. Moreover, the existing research is limited by the small size of the dataset and the reliance on classical machine learning models such as Support Vector Machine (SVM) for the detection task.[8]

In the wake of the COVID-19 pandemic, the role of online medical services has become even more crucial in addressing the uneven distribution of medical resources. To address the gap in the literature on fake physician review detection, this study proposes to construct a new dataset of fake physician reviews using a crowdsourcing approach and real user review data from a well-known online medical platform. The study will then develop a fake physician review detection model using both classical machine learning methods and deep learning methods.

Literature review

Online health community

The use of the Internet to access health care information has become increasingly popular among patients with diseases.[9] The emergence of Medicine 2.0 or Health 2.0 applications, which use Web 2.0 technologies, has enabled Online Health Communities (OHCs) to provide informational and social support for patients.[10–13] In addition to patients, physicians also participate in OHCs and provide counseling services to patients.[14]

Physicians who participate in OHCs may receive social and financial rewards.[15,16] Studies have been conducted on the factors that influence physicians' rewards from OHCs,[17] as well as on the profitability model of physicians in OHCs from a professional capital perspective.[15] Patients have also been the focus of studies, exploring the positive effects of OHCs on patients, the ways in which patients obtain information they need from OHCs, and how reviews on physicians influence patients' decisions when choosing a physician to consult within the doctor–patient community.[18,19]

Physician rating websites

In addition to providing social support, OHCs can also provide medical information and consultation services, including online consultation services, creating an online patient–physician relationship.[20] Patients are consumers in this relationship and evaluate physicians' services in the same way they evaluate products on e-commerce platforms.[21] Physician Rating Websites (PRWs) have become increasingly popular, with over 40 websites such as Yelp and Angie's List offering patients reviews of healthcare providers.[21] Many patients consult PRWs before choosing a doctor.[22]

Initially, physicians were concerned that PRWs would contain inappropriate and untrue negative reviews that would damage their reputation and work.[21] However, research has shown that online physician ratings are generally around 90/100.[23] Studies on PRWs have focused on the impact of online reviews on patients' choice of physicians,[19] how physicians can increase patient inquiries and profit possibilities through reputation management services or by better building their homepage, and how physicians' reviews differ between regions and departments.[24]

Research has shown that the use of PRWs has increased over the last decade.[25] For PRW users, online reviews strongly influence their decisions,[26] with 65% of German PRW users consulting a doctor based on the ratings provided by these sites.[27] The younger generation is relying more on the Internet when choosing a doctor, with more than a quarter of young parents in the United States reporting that they had selected a pediatrician for their child on the Internet.[28,29] Physicians are using reputation management services to construct and defend their online reputation,[30] with some spending more money to achieve higher ratings[31] or encouraging satisfied patients to write positive reviews.[32]

While reviews on PRWs have a significant impact on patients' choice of care, the authenticity and professional validity of reviews on PRWs need to be verified.[33] Identifying suspicious online physician reviews is meaningful for helping patients make medical choice decisions and for the long-term development of online physician review websites.[34]

Fake review detection

The phenomenon of fake reviews has become increasingly prevalent in online shopping and review websites, leading to significant concerns about the reliability and authenticity of user reviews. Fake reviews can be categorized into three types: untruthful opinions, reviews on brands only, and nonreviews, as proposed by Jindal and Liu.[6] Detecting fake reviews is a challenging task, mainly due to the difficulty in distinguishing between genuine and fake reviews. Therefore, two research focuses have emerged: the construction of the dataset and the development of fake review detection methods.

In terms of dataset construction, Li et al.[7] proposed two primary methods for constructing fake review datasets, namely, manual annotation and crowdsourcing platforms. Research by Ott et al.[35] demonstrated that it is challenging to identify human-written fake reviews manually, leading to lower accuracy in labeling. To address this issue, Ott et al.[35]

Table 1. Datasets for fake review detection.

Domain	Dataset	Volume (Fake%)	Construction Method	Study
Hotel	TripAdvisor	800 (50.00%)	crowdsourcing	[35]
	TripAdvisor	1600 (75.00%)	crowdsourcing	[38]
	TripAdvisor	2848 (13.31%)	manual annotation	[39]
Hotel, Restaurant	YelpChi	67395 (13.23%)	filtering algorithm	[36]
Restaurant	YelpNYC	359052 (10.27%)		
	YelpZip	608598 (13.22%)		
Product	Yelp	18912 (50.00%)	filtering algorithm	[40]
	Amazon	5.8million	similarity check	[6]
	Epinions	6000 (23.30%)	manual annotation	[41]
	Amazon	109,518	/	[42]
	Amazon	6819 (41.30%)	manual annotation	[43]
Hotel, Restaurant, Doctor	TripAdvisor	3032 (37.6%)	crowdsourcing	[7]

created a gold-standard dataset containing 800 reviews, half of which were real and the other half were fake. Subsequently, Li et al.[7] expanded the dataset to include fake reviews from three areas: hotels, restaurants, and doctors, totaling 3032 reviews. Additionally, some studies have used review datasets filtered by review websites.[36,37] Table 1 shows the datasets used in previous studies on fake review detection.

In terms of fake review detection techniques, previous research has focused on the analysis of fake review texts and the identification of fake reviewers' behavior. Li et al.[7] found that true reviews contain more nouns, adjectives, and prepositions, while fake reviews constructed through crowdsourcing contain more verbs, adverbs, and pronouns. Based on text features obtained through syntactic analysis, many researchers have combined SVM models or neural network models for fake review detection tasks, achieving better detection accuracy,[5,7,35,38,44] among which SVMs perform better than other models in detection tasks with small sample sizes.

Semantic similarity computation is a common approach used in the study of fake review text detection based on semantic analysis. Lau et al.[45] concluded that fake reviews have a tendency to copy each other, and fake review detection can be performed by identifying semantic duplicate reviews. Linguistic Inquiry and Word Count (LIWC), a commonly used tool in research on fake review detection based on stylistic features, is capable of extracting multiple text features, including stylistic features, and mapping 4500 keywords into an 80-dimensional vector. This tool was used by Ott et al. and Li et al. in their studies[7,35] and was combined with bag-of-words features to improve the detection effect. The metadata of reviews, including features of attributes other than textual content such as publication time and the number of likes or comments, has been shown to effectively improve the accuracy of fake review detection when combined with textual features.[39,41]

In conclusion, while existing research has made significant contributions to fake review detection in fields such as hotels and restaurants, research targeting the detection of fake reviews on online healthcare platforms is still underdeveloped. In addition, models used to identify fake reviews in domains such as hotels and restaurants often struggle to achieve better performance when identifying fabricated reviews in medical domains.[8,46,47]

This study aims to address the research gap by building a specialized dataset and integrating deep learning models to achieve better results in the detection of fake physician reviews. Most of the existing studies on the detection of fake physician reviews are based on the gold-standard dataset constructed by Li et al.,[7] which is of high quality but contains only 432 physician reviews. Therefore, this study employs the dataset construction method of Li et al.[7] to build a more sizable and more specialized dataset for the online medical context. By doing so, we can mine more representative features of fake physician reviews and achieve better detection results. Second, in

terms of detection methods, previous research has focused on the features of the review text and classical machine learning methods such as SVM and k-nearest neighbor.[8,46,47] However, these methods often struggle to achieve better performance in detecting misinformation compared with deep learning methods.[48–50] Therefore, this study employs both machine learning models such as Logistic Regression (LR), SVM, Random Forest (RF) and Ridge Regression (Ri), and deep learning models such as Bidirectional Encoder Representations from Transformers (BERT), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) to achieve better results in the detection of fake physician reviews.

Methods

To construct a dataset of fake online physician reviews, we employed a crowdsourcing approach where writers were hired to produce fake reviews. In addition, we obtained true online physician reviews by crawling data from a well-known online medical platform (haodf.com) in China using a web crawler tool. We integrated the fake and true reviews to create an experimental dataset for this study. Then, we employed both classical machine learning algorithms (LR, SVM, and RF) and deep learning algorithms (BERT, CNN, and RNN) to develop the fake online physician review detection model. We evaluated the model's accuracy, recall, loss value, and other relevant metrics to compare the performance of the different algorithms. Figure 1 shows the process of the dataset construction and model development.

Dataset construction

The experimental dataset comprised two parts: the fake review dataset and the true review dataset. To construct

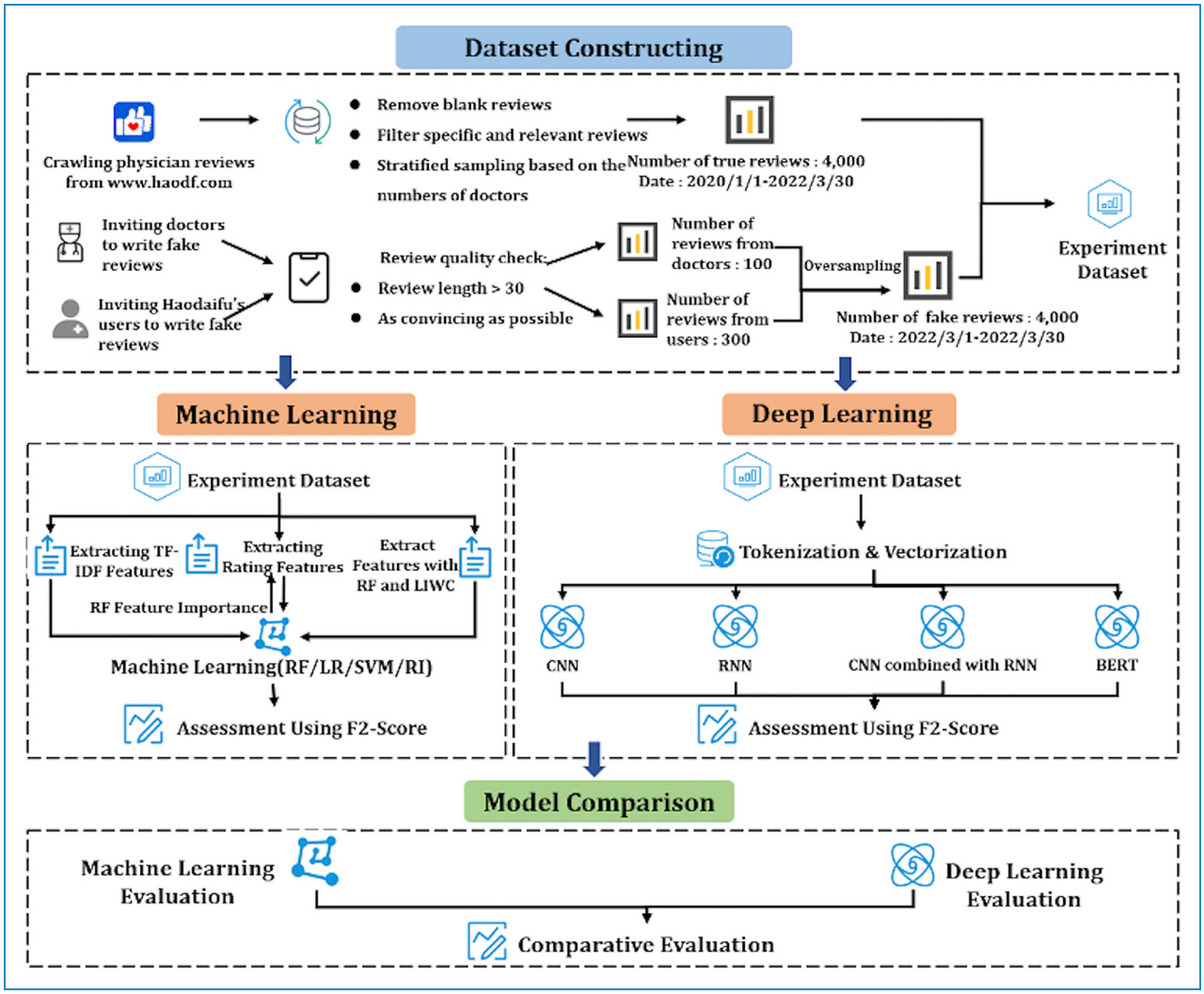

Figure 1. Technical route of this study.

the true review dataset, we used a web crawler to collect reviews of physicians in the four departments of internal medicine, surgery, dentistry, and oncology from the homepage of an online medical platform. We randomly selected 4000 records from the crawled physicians' reviews to construct the true review dataset, using a stratified sampling method to ensure balance between positive and negative cases. The number of reviews for each department in the sample set (465 reviews of internal medicine, 1092 reviews of surgery, 1114 reviews of dentistry, and 1299 reviews of oncology) was assigned based on the ratio of the number of physicians in each department provided by the platform (internal medicine: surgery: dentistry: oncology = 9700: 22764: 23840: 27079).

To construct the dataset of fake online physician reviews, we drew inspiration from Li et al.'s[7] study. We established several rules to recruit qualified fake review writers. Only experienced users of online medical platforms were invited to write fake reviews. We asked them to create several convincing reviews for online physicians, similar to the approach taken by writers hired to produce fake reviews. This enabled them to simulate fake medical reviews on the platforms as closely as possible.

We solicited both ordinary and expert users to write fake reviews for physicians in four common departments: internal medicine, surgery, dentistry, and oncology. Expert users, who were practicing physicians, produced 100 of the 400 fake reviews, with the remaining 300 written by general users. We screened the fake reviews for review length, detail, and relevance. The threshold for the review length was set at 30 words, based on the average review length of 29.325 words calculated from the reviews of physicians in the four departments on the platform. We screened out fake reviews that were shorter than the threshold. Table 2 presents examples of genuine reviews collected from the platform alongside fake reviews created by writers.

We screened the fake reviews based on their sentiment tendency (positive or negative) as well. The ratio of positive reviews of each physician counted by the platform was used to determine the number of positive and negative reviews in the fake review dataset. The weighted average of positive reviews was calculated to be 99.28%.

After constructing the fake and true review datasets separately, we assigned labels to create a dataset for subsequent classification learning. The composition of experiment datasets is presented in Table 3.

Table 2. Examples of true and fake reviews.

Review type	Examples
True review	"我之前做过矫正，因偏侧咀嚼，想要再次调整，程医生告诉我好好佩戴保持器，维持现状就可以了。感谢医生的讲解！" *"I have had orthodontic treatment before and wanted another adjustment due to partial chewing, Dr Ching told me to just wear my retainer properly and maintain the status quo. Thank you doctor for your explanation!"*
	"头疼16年，头疼前有视觉先兆或伴随肢体麻木，经多家医院治疗无好转，后经熟人推荐袁大夫，经入院治疗，目前，症状有好转。" *"Headache for 16 years, headache before a visual aura or accompanied by numbness of the limbs, by a number of hospitals to treat no improvement, and then recommended by an acquaintance of Dr Yuan, after admission to the hospital for treatment, at present, the symptoms have improved."*
Fake review	"医生医术高明，态度和蔼可亲，对病人很有耐心，亲切询问病情。我对医生诊治及态度非常满意。" *"The doctor is highly skilled, kind and patient with the patient, asking questions about his condition. I am very satisfied with the doctor's treatment and attitude."*
	"医术精湛，讲解清晰，认真，态度和蔼，很能耐心的帮助患者看好病。" *"Excellent medical skills, clear explanations, serious, kind attitude, very patient to help patients to recover."*

Table 3. Experiment datasets composition.

Sources	True review dataset	Fake review dataset
Expert users	/	100
Platform users	4000	300
Total	4000	400

Feature extraction

The feature extraction method employed in this study is Term Frequency-Inverse Document Frequency (TF-IDF), which is a commonly used weighting technique in the fields of information retrieval and text mining.[51] Term Frequency-Inverse Document Frequency assigns importance to a term based on its frequency within a specific document and inversely proportional to its frequency across the entire corpus. This technique is derived from TF and IDF values, where TF represents the frequency of

a term's occurrence within a given document, as shown in equation (1):

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

In this equation, n_{ij} denotes the frequency of term i in document D_j, while the denominator represents the sum of the frequencies of all terms in document D_j. The IDF value of a term is obtained by dividing the total number of documents in the corpus by the number of documents containing that specific term and taking the logarithm of the resulting quotient. The calculation is presented in equation (2):

$$idf_j = \log \frac{|D|}{|\{j{:}t_i \in d_j\}|} \tag{2}$$

Here, $|D|$ represents the total number of documents in the corpus, and $\frac{|D|}{|\{j:t_i \in d_j\}|}$ signifies the number of documents containing the term t_i. If the term t_i is absent from the corpus, the denominator becomes zero. To prevent this, a modified formula is often used, where the denominator is represented as $1 + |\{j{:}t_i \in d_j\}|$, thus ensuring a nonzero value. The modified formula is shown in equation (3):

$$\mathrm{IDF} = \log \frac{|D|}{|\{j{:}t_i \in d_j\}| + 1} \tag{3}$$

Term Frequency-Inverse Document Frequency is the product of TF and IDF, expressed as TF-IDF = TF * IDF. In this study, after importing the dataset into Python, the text was first segmented into individual words. Subsequently, the TfidfVectorizer[52] was employed to construct the word vector matrix for the documents. A training session utilizing RF was then conducted. As the dimensionality of the input TF-IDF feature set was too high, resulting in a low accuracy of the RF model, we drew on previous study[53] to filter the top 10 features in terms of importance. For our classification task, the feature importance assessment method provided by RF is Mean Decrease Accuracy (MDA), which evaluates the importance of features by calculating their impact on model accuracy in each decision tree.[54] Specifically, the MDA method randomly exchanges the values of a feature in the Out-of-Bag (OOB) dataset, and then reruns the prediction, calculating the significance of the feature by measuring the degree of degradation in the accuracy of the classification, as shown in equation (4):

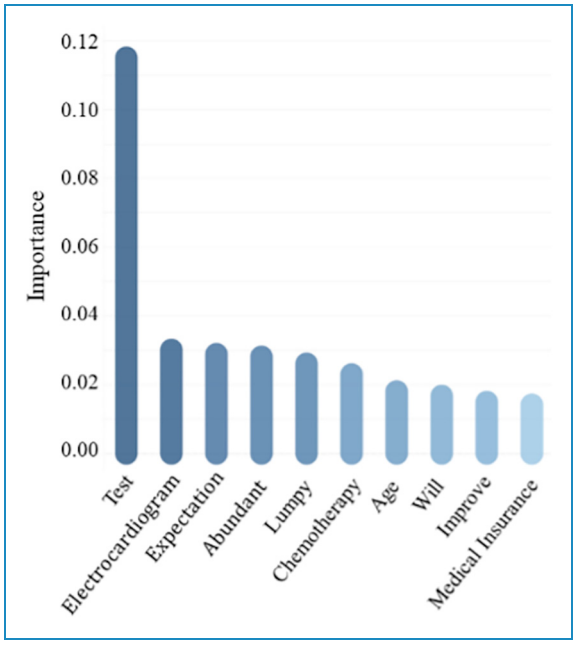

Figure 2. Filtered TF-IDF features.

$$\mathrm{MDA}_c(j) = \frac{1}{T}\sum_{1}^{T}\left[\frac{1}{|D_t|}\left(\sum_{X_i \in D_t} I(P(X_i) = = y_i) - \sum_{X_i^j \in D_t^j} I(P(X_i^j) = = y_i)\right)\right] \tag{4}$$

In the formula, T is the number of random trees in RF; (X_i, y_i) is the sample, and for classification, y_i is the category; X_i^j is the sample after random swapping of the jth dimension (feature) of X_i; D_t is the set of OOB samples of the random tree t, and D_t^j is the set of samples formed after swapping of the j dimension; $P(X_i)$ is the prediction result (category) of the sample X_i; and I is the indicator function, which returns 1 if the prediction result is the same as the true category, otherwise it returns 0; and $I(P(X_i) = = y_i)$ is the indicator function. True category is the same, then return 1, otherwise return 0.

Based on the feature importance provided by RF, 10 textual features were selected, specifically the TF-IDF features. This process preserves TF-IDF features with high value for model recognition while reducing dimensionality. These important terms include test (e.g., ordinary blood test), electrocardiogram, expectation, abundant, lumpy,

chemotherapy, age, will, improve, and medical insurance. Figure 2 shows the importance of each TF-IDF feature provided by RF. As illustrated in Figure 2, terms such as "test," "electrocardiogram," "expectation," "abundant," "lumpy," "chemotherapy," "age," "improve," and "medical insurance" emerge as highly important. These terms typically represent more detailed evaluations of physicians and thus appear more frequently in comments written by actual patients. Additionally, the length of each review was calculated, normalized, and combined with the TF-IDF features to create the final set of text features.

The patient ratings obtained from the platform include evaluations of treatment effectiveness and the physician's attitude during the consultation. To facilitate model input requirements, we assigned noncontinuous values ranging from 0 to 1 to represent the rating levels. Specifically, "very satisfied" corresponds to 0.9, "satisfied" corresponds to 0.7, "average" corresponds to 0.5, "unsatisfactory" corresponds to 0.3, and "very unsatisfactory" corresponds to 0.1.

In this study, we utilized two sets of features: the filtered TF-IDF features (referred to as filtered TF-IDF features) obtained through RF, and the comprehensive feature set (referred to as comprehensive features) that incorporates the filtered TF-IDF features, review length, and patient rating feature and features extracted by LIWC package. Table 4 provides a detailed description of the features used in this study, and the set comprising all the features listed in Table 4 is defined as comprehensive features.

Deep learning methods

The application of CNN in text classification research by Kim[55] was a major breakthrough in the application of deep learning techniques in text classification tasks. His proposed CNN model consists of four parts: input layer, convolutional layer, pooling layer, and fully connected layer. Convolutional Neural Network has been widely used in text classification tasks, and its efficacy and maturity make it an ideal choice as a detection model. Recurrent Neural Networks have a more versatile application than ordinary neural networks, as its basic structure includes an input layer, hidden layer, and output layer. The value of its hidden layer is determined by the input layer of this moment and the hidden layer of the last moment, which takes into account the role of context. Consequently, RNN is chosen as one of the comparative models in this paper.

Bidirectional Encoder Representations from Transformer is a pretrained linguistic representation model that achieved state-of-the-art performance in 11 distinct natural language

Table 4. Comprehensive feature set.

Feature name	Composition	Description
Filtered TF-IDF features	/	The set of important TF-IDF features
Review length	/	Normalized text length
Patient rating feature	Efficacy satisfaction	Very satisfied/satisfied/average/ unsatisfactory/very unsatisfactory
	Attitude satisfaction	Very satisfied/satisfied/average/ unsatisfactory/very unsatisfactory
LIWC features	Function	Total function words (e.g., the, to, and, I)
	I-pronoun	Impersonal pronouns (e.g., that, it, this, what)
	You-pronoun	2nd person (e.g., you, your, yourself)
	He-pronoun	3rd person (e.g., he, she, they)
	Health	Health related (e.g., medic, patients, physician)
	Informal	Informal terms
	Compare	Extent to which the text contains comparative language
	Positive emotion	posemo from LIWC (e.g., good, love)
	Negative emotion	negemo from LIWC (e.g., bad, hate)

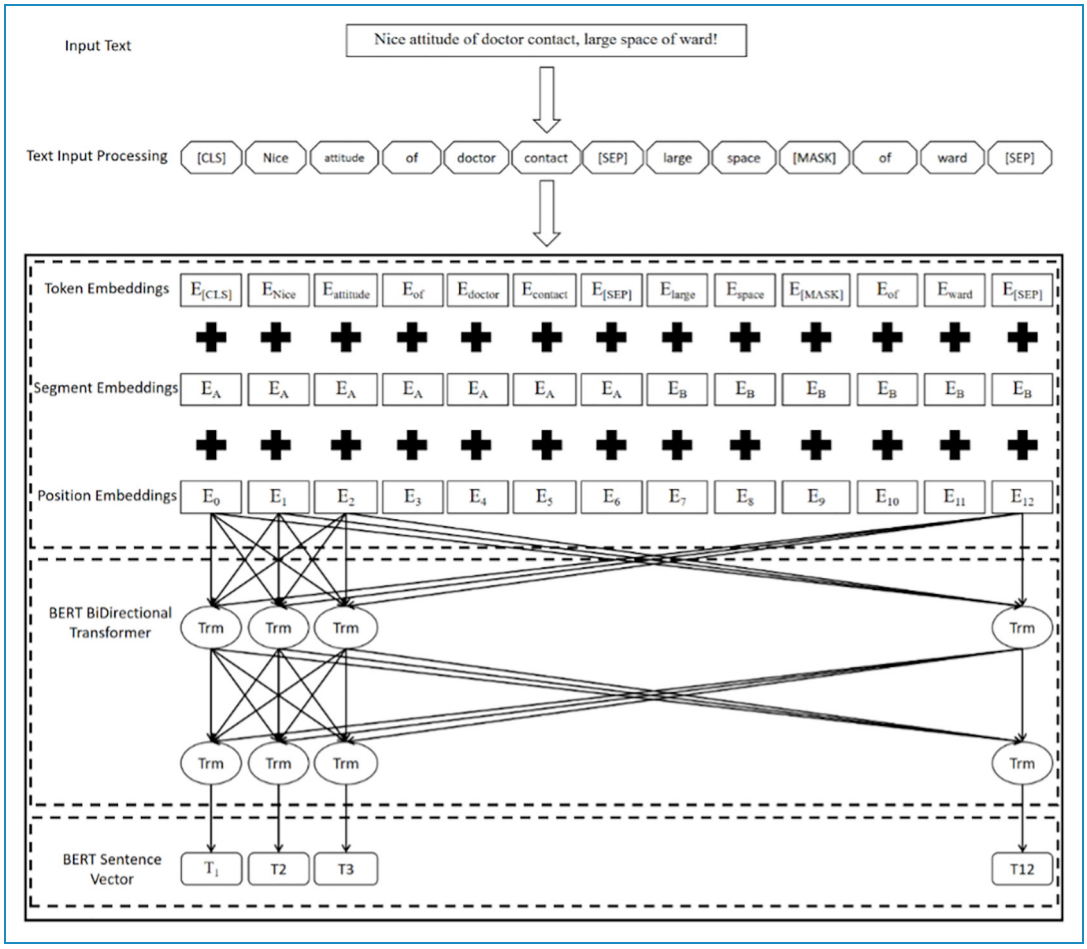

Figure 3. Structure of BERT-based text classification model.

processing tasks.[56] The BERT model's input layer consists of three types of embeddings: Token Embeddings, Segment Embeddings, and Position Embeddings. These embeddings serve the purpose of converting words into fixed-dimensional vectors, distinguishing sentence pairs, and encoding the sequential nature of input sequences, respectively (Figure 3).

Experimental design and analysis

Experiment preparation. The experiment primarily relied on Python 3.8, with the integrated development environment Jupyter Notebook employed for development purposes. The model construction was accomplished using the scikit-learn and Pytorch frameworks. The hardware environment featured an Intel Core i5-7200U 2.50 GHz CPU and 8GB of memory.

In evaluating binary classification models, Accuracy, Precision, Recall, and F-Score are commonly used. These metrics can be calculated using a Confusion Matrix.[57] Given the objective of identifying fake medical reviews in this study, the fake reviews were assigned as positive cases, while the true reviews were designated as negative cases for constructing the confusion matrix. In this study, the detection of fake reviews focused on identifying as many fake reviews as possible to minimize the impact of fake reviews on patients' choice of physicians. As a result, the F2-Score evaluation index was constructed with a higher weight value of recall R to better suit the actual task. The weighted F2-Score is calculated accordingly. The original F-Score is shown in equation (5):

$$F - \text{Score} = (1 + F^2) * \frac{\text{Precision} * \text{Recall}}{F^2 * \text{Precision} + \text{Recall}} \quad (5)$$

The F value was set to 2, and the formula used in this work is shown in equation (6):

$$F2 - \text{Score} = 5 * \frac{\text{Precision} * \text{Recall}}{4 * \text{Precision} + \text{Recall}} \quad (6)$$

Both machine learning and deep learning methods were utilized to construct models, incorporating both filtered TF-IDF features and comprehensive features. The dataset used in this study consisted of a true review dataset comprising 4000 reviews collected from an online medical platform, and a fake review dataset containing 400 reviews

constructed using the crowdsourcing method. To address the imbalance between the true and fake review datasets, we employed oversampling to increase the number of fake reviews. This approach was chosen to mitigate the class imbalance issue, which can adversely affect the performance of machine learning models. Oversampling helps to ensure that the models are not biased toward the majority class, thus improving their ability to accurately detect fake reviews.

To evaluate the models, a 10-fold cross-validation approach was employed. In each validation iteration, 90% of the data served as the training set, while the remaining 10% constituted the test set. This method ensures that each review is used for both training and testing, providing a robust evaluation of the model's performance.

Experimental construction. Both machine learning and deep learning models were used in this study to compare and evaluate their performance. Four machine learning methods, namely, LR, SVM, RF, and Ri, were applied to the fake physician review detection task. Each machine learning model was tested with multiple sets of feature set inputs.

Convolutional Neural Network models were constructed with an input layer, embedding layer, 1D convolutional layer, 1D pooling layer, and fully connected layer using Python. The training epochs were set to 6, and 30% (1920/6400) of the data from the training set were discarded to prevent overfitting. The CNN model was fine-tuned in the pooling layer using two methods, MaxPool1D and AveragePool1D, for pooling. The pooling window size was adjusted to the number of rows of the word vector matrix minus 3, 4, and 5, respectively. The experimental settings are as follows: workers (the number of CPU compute cores used for parallelized training) = 4, vector_size(word vector dimension) = 100, min_count(minimum frequency of the considered words) = 3, and window(word context window size) = 4.

Additionally, we constructed standard bidirectional RNN and CNN models. The CNN model consisted of an input layer, an embedding layer, a 1D convolutional layer, a 1D pooling layer, and a fully connected layer. The RNN model comprised an input layer, an embedding layer, a bidirectional GRU layer, and a fully connected layer. These models were combined using the concatenate method.

For the BERT model, only certain parameters were fine-tuned, and the adjustable parameters based on the Chinese pretrained BERT model are listed as follows. Considering the mean value and distribution of review text length in the dataset, the max_len parameter was set to 256, and the fill_paddings method was used to complete the sentence length. The batch_size parameter was set to 16, the learning_rate parameter was set to 2e-5, and the epochs parameter was set to 3.

Results

Fake review detection models

The detection results of the machine learning models using filtered TF-IDF features (FF) and comprehensive features (CF), as well as the performance of the deep learning models, are summarized in Table 5. Models with better performance are highlighted in italics.

Table 5 demonstrates that the BERT model achieved impressive results, with significantly higher precision and accuracy compared to other methods. This highlights the feasibility and effectiveness of the BERT model in detecting fake physician reviews. Among the machine learning models, the RF model exhibited the best performance (90.27%), while the remaining models achieved moderate performance (>70%) as evaluated by the F2-Score. However, most models demonstrated poor performance (<65%) in terms of precision and accuracy when using only the filtered TF-IDF features. The incorporation of constructed features significantly enhanced the performance of the machine learning model across all metrics, albeit resulting in a slight decrease in precision for the RF model.

Compared to the machine learning models, the deep learning models exhibited a more balanced performance across all metrics, without any excessively low scores. Among the deep learning models, BERT and CNN achieved the highest performance (F2-Score > 90%), while the remaining deep learning models showed more moderate performance. Although machine learning models required less training time and yielded satisfactory results, there is still considerable room for improvement in their detection performance (F2-Score ranging from 79.82% to 90.27%). In contrast, BERT showcased excellent performance across all metrics and maintained a clear lead of approximately 7.7–18.15% over other machine learning algorithms in the overall evaluation metric, F2-Score. When compared to other deep learning models, BERT consistently outperformed them across all metrics.

Feature importance

Fake review detection studies commonly employ two main types of features: linguistic features of reviews and behavioral features of reviewers. However, since the platform hides users' personal information, we were unable to obtain the behavioral features of reviewers. Consequently, in this study, we constructed two feature sets primarily based on the linguistic features of reviewers as inputs for the machine learning models. These feature sets include the TF-IDF feature set filtered by RF (filtered TF-IDF features) and the comprehensive feature set, which incorporates patient rating features, review length feature and linguistic features extracted by LIWC into the filtered TF-IDF features.

Table 5. Model performance.

	Model	Precision	Accuracy	Recall	F2-score
Machine Learning Methods	SVM(FF)	54.78%	57.36%	76.82%	72.47%
	SVM(CF)	81.79%	81.93%	80.57%	80. 80%
	LR(FF)	61.94%	62.45%	76.75%	72.43%
	LR (CF)	82.15%	81.05%	79.27%	79.82%
	RF(FF)	90.20%	82.19%	71.45%	72.72%
	RF(CF)	*89.51%*	*90.19%*	*90.64%*	*90.27%*
	RI(FF)	61.85%	62.40%	76.82%	72.47%
	RI(CF)	81.84%	81.93%	81.98%	81.94%
Deep Learning Methods	CNN(maxpool,maxlen-4)	*88.59%*	*88. 66%*	*90.58%*	*90.17%*
	CNN(maxpool,maxlen-5)	87.25%	86.43%	86.16%	86.38%
	CNN(maxpool,maxlen-6)	*92.41%*	*91.25%*	*90.09%*	*90.54%*
	CNN(averagepool,maxlen-4)	82.56%	81.25%	79. 94%	80.45%
	CNN(averagepool,maxlen-5)	80.87%	83.48%	88.77%	87.07%
	CNN(averagepool,maxlen-6)	82.04%	83. 75%	87.86%	86.63%
	RNN	86.58%	81. 88%	76.02%	77.92%
	Bi-GRU	76.48%	79.20%	83.23%	81.79%
	RNN + CNN(parallel)	83.78%	83.57%	82.65%	82.87%
	RNN + CNN(series)	82.71%	84.55%	88.28%	87.11%
	BERT	*98.36%*	*97. 88%*	*97.87%*	*97.97%*

For each machine learning model, we inputted these two feature sets and adjusted the model parameters to obtain the best-performing versions. Our experiments revealed that when using the comprehensive feature set as input, each machine learning model exhibited better performance. To further examine the impact of each feature on the detection performance of machine learning models, we employed the SHAP plot[58] for the RF model on the comprehensive features. We employed the Python package *SHAP* to generate value representation figures (Figure 4) for each feature in the RF model. Figure 4 presents the variables listed in descending order of importance. Each review is represented as a data point in the plot, with the horizontal position of the point indicating the correlation between the magnitude of the feature value and the superiority of the model's performance. Notably, the RF model identified Review Length as the most crucial variable, followed by Attitude Satisfaction and the term "test."

Distribution disparities between true and fake reviews

To provide further insights into the differences between true and fake reviews, we visually depicted the distribution of all constructed features, as shown in Figure 5. Notably, the distribution of text lengths exhibited a distinct pattern, with the peak of fake reviews skewed toward longer text lengths compared to true reviews (Figure 5(d) and (e)). This finding corroborates previous research conducted by Jindal and Liu,[6] who discovered that spam tends to have significantly longer lengths than normal emails. Similarly, Rout et al.[59] found

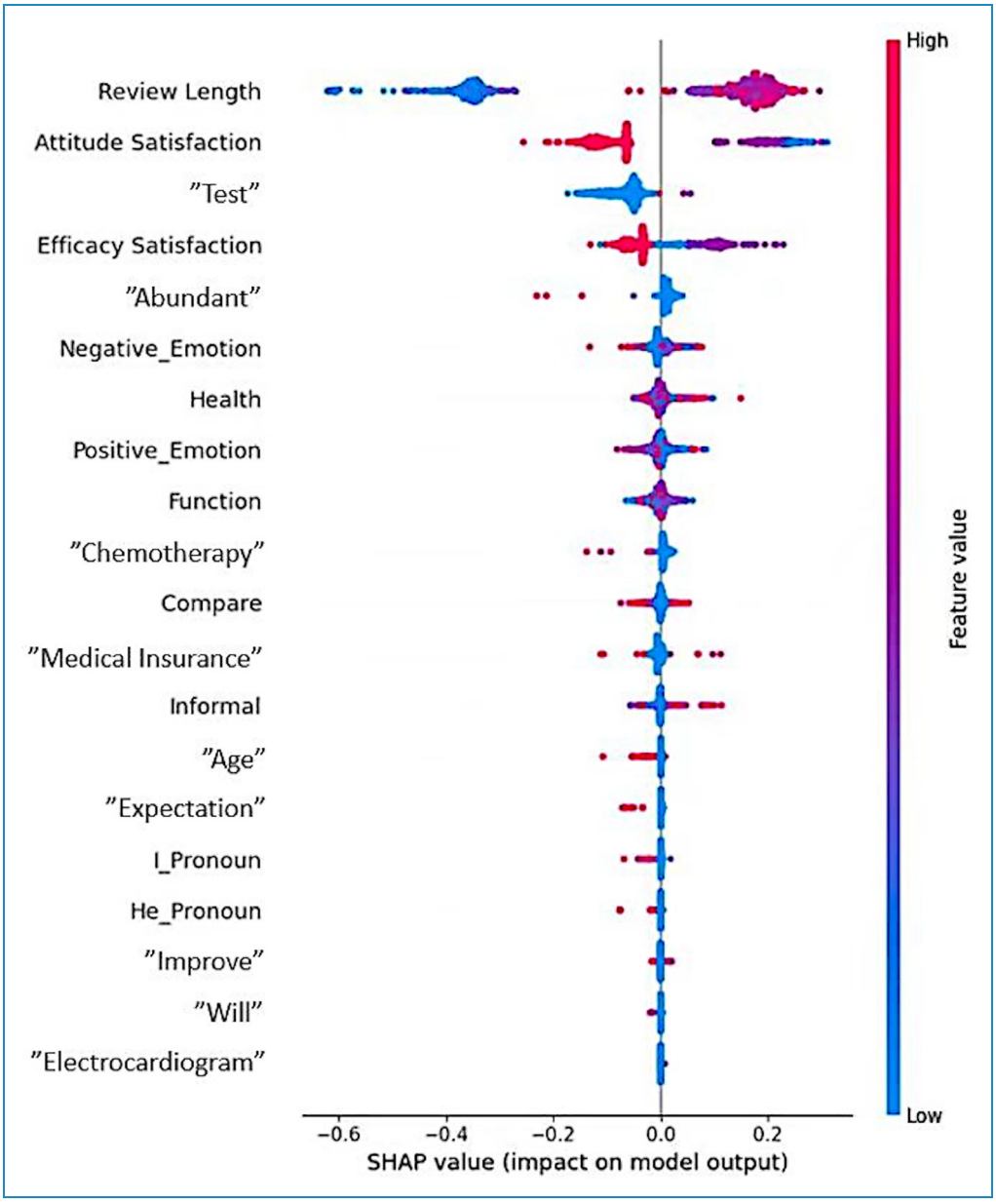

Figure 4. Feature importance by SHAP.

that deceptive reviews tend to be lengthier than truthful reviews. In our study, fake reviews exhibited text lengths ranging between 30 and 300, while true reviews had a broader range spanning from 1 to 4500 characters. Text length is a widely employed feature in the detection of fake reviews[42] and has demonstrated its significance in the task of distinguishing between genuine and deceptive reviews.[60] Moreover, text length assists consumers in making more informed judgments regarding the authenticity of reviews.

Furthermore, we examined the distribution disparities in other characteristics. Specifically, in terms of Efficacy Satisfaction and Attitude Satisfaction indicators, true reviews displayed a higher concentration of the highest ratings, while fake reviews exhibited a more even distribution (Figure 5(a) and (b)). This observation can be attributed to the tendency of patients to hold doctors in high regard and to post their reviews on the platform only when they are highly satisfied with their experience.[61] Additionally, the distribution of other features revealed notable variations. For instance, we observed a wider distribution interval for third-person pronouns in true reviews, indicating a higher frequency of their usage. This usage pattern may be positively correlated with the authenticity of the comments.[62] It is reasonable to infer that an increased utilization of third-person pronouns signifies

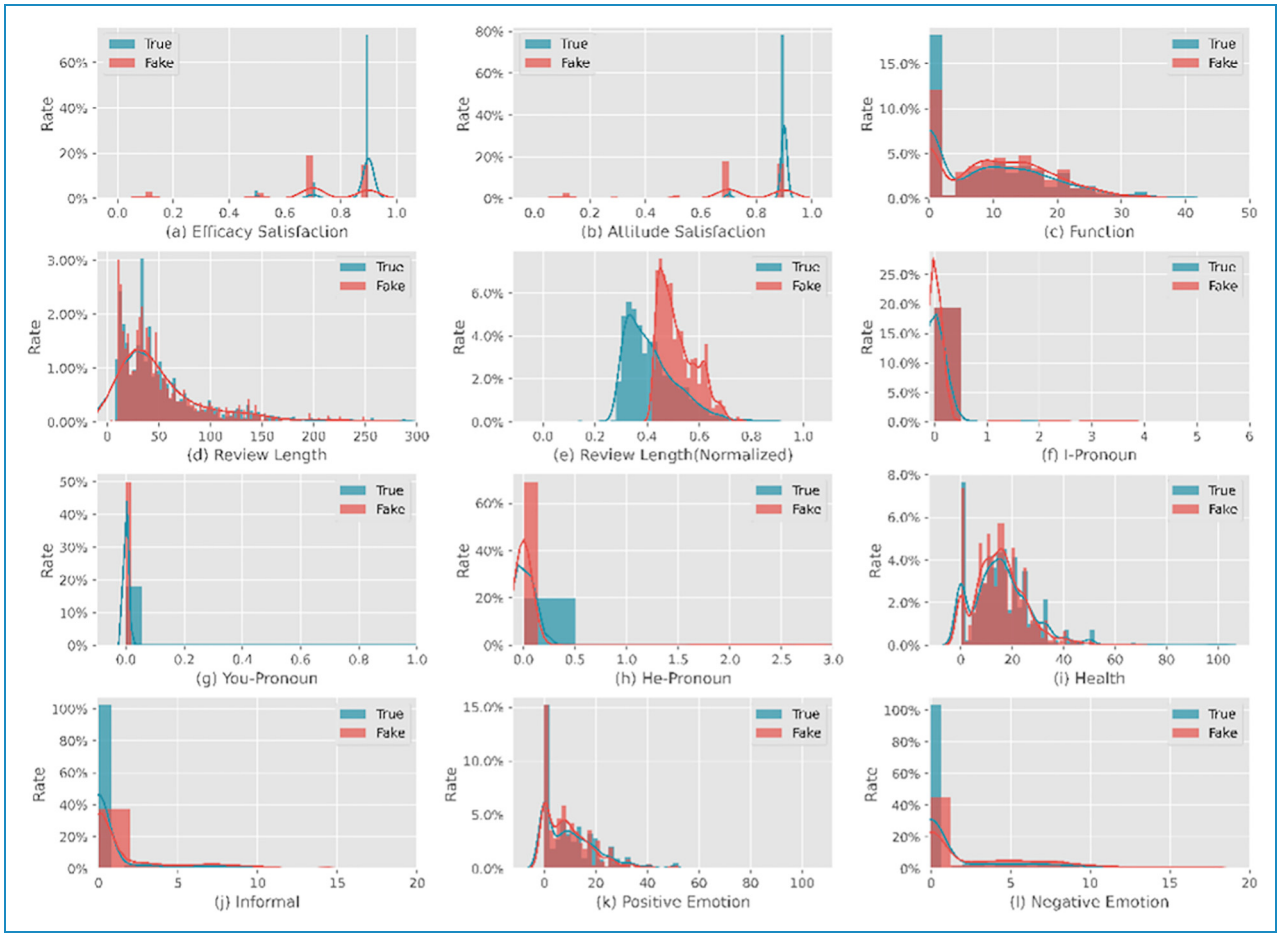

Figure 5. Distribution of each feature.

a closer proximity to an objective statement and a higher degree of truthfulness.

Discussion

Dataset construction

Table 1 demonstrates that previous studies on fake review detection have primarily been centered around fake product reviews or fake hotel reviews, with limited research conducted on fake physician reviews due to the scarcity of established datasets in this domain. This issue was addressed in 2014 when Li et al.[7] constructed a dataset specific to fake physician reviews, a resource that has since been utilized in numerous studies. However, a study by Hao et al.[24] highlighted the differences in review characteristics between Chinese and American PRWs, implying that the findings from existing studies may not be directly applicable to the Chinese context. Consequently, the fake online physician review dataset we have constructed offers a valuable complement to research in this area.

With respect to dataset construction methods, many studies have employed manual annotation to create fake review datasets. However, identifying fake physician reviews through manual annotation could be challenging. To address this concern, we opted for the use of the crowdsourcing method. Furthermore, in terms of dataset composition, we invited both platform users and physicians to participate during dataset construction, a practice that sets our approach apart from some related studies that did not specifically require participant expertise.[36] This strategy serves to enrich the data sources and enhances the dataset's resemblance to the actual scenario of reviews on PRWs.

Classification methods for online fake physician review detection

Previous studies have extensively utilized both machine learning methods and deep learning methods for fake review detection tasks.[63,64] Notably, RF, SVM, and CNN have demonstrated commendable performance in fake review detection tasks.[65,66] RF is advantageous due to its robustness and ability to handle large datasets with higher dimensionality.[67] It is less likely to overfit compared to

individual decision trees, and it provides insights into feature importance. However, RF can be computationally intensive and may not perform well with very high-dimensional sparse data, typical in text analysis. SVM is known for its effectiveness in high-dimensional spaces and its ability to find the optimal separating hyperplane between classes. SVM is particularly effective when the number of dimensions exceeds the number of samples.[68] Nonetheless, it can be less effective with large datasets and may require significant tuning of hyperparameters.

In this study, we selected four machine learning methods (RF, LR, RI, and SVM) and three deep learning methods (BERT, CNN, and RNN). The experimental results revealed that BERT exhibited the best classification performance in this task, surpassing the other methods employed in this experiment. It aligns with previous studies on fake review and fake news detection which have also demonstrated the excellent performance of BERT,[69,70] underscoring its ability to effectively extract relevant features from text.

Conclusion

The rapid advancement of online healthcare platforms has revolutionized the process of seeking and receiving medical advice for patients. However, with the increasing number of online reviews for physicians, the prevalence of fake reviews has also risen, posing a threat to the perception of physicians' quality and undermining trust in online healthcare platforms. Therefore, the detection of fake reviews is of utmost importance in maintaining the credibility of these platforms.

To address the need for effective fake review detection, we employed a crowdsourcing approach to construct a dataset specifically designed for online medical platforms. The dataset created through crowdsourcing exhibited higher accuracy compared to datasets manually labeled or identified through pattern detection methods. Subsequently, we developed a detection model utilizing both classical machine learning and deep learning models. For evaluating the performance of the models in the context of online medical platforms, we utilized the F2-Score evaluation index, which is particularly suitable for fake review detection. Our experimental results revealed that the deep learning model outperformed traditional machine learning models in identifying fake physician reviews on online medical platforms. The deep learning model achieved a remarkable precision of 98.36% and an impressive F2-Score of 97.97%. This represents an 8.16% improvement in precision and a 7.7% increase in F2-Score compared to the traditional machine learning model. Therefore, our study makes a valuable contribution by providing an effective fake physician review detection model tailored for online medical platforms.

While this study advances the development of an effective fake physician review detection model, there still remains some limitations that need to be acknowledged. Firstly, one primary limitation is the relatively small size of the fake review dataset. Although oversampling techniques were employed to mitigate this issue, the synthetic data may not fully capture the complexity of real-world fake reviews. Expanding the scale of the dataset would enhance the generalizability of the model and its ability to detect fake reviews across a wider range of scenarios. Additionally, the study's dataset is sourced from a single online medical platform, which may limit the model's applicability to other platforms with different user behaviors and review characteristics. Future research should consider collecting and integrating datasets from multiple platforms to ensure the model's robustness and generalizability across diverse contexts. Investigating the differences in user behavior and review characteristics across platforms can also provide valuable insights into how these factors influence the detection of fake reviews. These future endeavors hold the potential to further refine and enhance the capabilities of the fake physician review detection model on online medical platforms.

Acknowledgements: The authors thank the participants in this study.

Contributorship: Yuehua Zhao: Conceptualization, Methodology, Writing, Funding acquisition. Tianyi Li: Formal analysis, Writing. Qinjian Yuan: Conceptualization, Writing. Sanhong Deng: Conceptualization, Writing.

Consent to participate: All participants' information was de-identified, and participant consent was not required.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The study protocol was approved by the Ethnic Committee of School of Information Management, Nanjing University (approval no. IM-23-0115).

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China, (grant number 72004091, 72474098).

Guarantor: YZ

ORCID iD: Yuehua Zhao https://orcid.org/0000-0002-8412-2878

References

1. Fox S and Duggan M. Health Online 2013, https://www.pewresearch.org/internet/2013/01/15/health-online-2013/ (2013, accessed 23 October 2023).

2. Silver L and Huang C. In Emerging Economies, Smartphone and Social Media Users Have Broader Social Networks, https://www.pewresearch.org/internet/2019/08/22/in-emerging-economies-smartphone-and-social-media-users-have-broader-social-networks/ (2019, accessed 23 October 2023).
3. Yang Y, Zhang X and Lee PKC. Improving the effectiveness of online healthcare platforms: an empirical study with multi-period patient-doctor consultation data. *Int J Prod Econ* 2019; 207: 70–80.
4. Wu G, Greene D, Smyth B, et al. Distortion as a validation criterion in the identification of suspicious reviews. In: Proceedings of the First Workshop on Social Media Analytics, Washington D.C., District of Columbia, USA, 25–28 July 2010, pp.10–13. New York: ACM.
5. Mukherjee A, Venkataraman V, Liu B, et al. What yelp fake review filter might be doing? In: Proceedings of the 7th International Conference on Weblogs and Social Media, Massachusetts, USA, 8–11 July 2013, pp.409–418. Menlo Park: AAAI.
6. Jindal N and Liu B. Review spam detection. In: Proceedings of the 16th international conference on World Wide Web, Alberta, Canada, 8–12 May 2007, pp.1189–1190. New York: ACM.
7. Li J, Ott M, Cardie C, et al. Towards a general rule for identifying deceptive opinion spam. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, Baltimore, Maryland, 2014, pp.1566–1576. Pennsylvania: ACL.
8. Liu Y, Wang L, Shi T, et al. Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. *Inf Syst* 2022; 103: 101865.
9. Ziebland S, Chapple A, Dumelow C, et al. How the internet affects patients' experience of cancer: a qualitative study. *Br Med J* 2004; 328: 564. PMID: 15001506.
10. Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res* 2008; 10: e1030. PMID: 18725354.
11. Hughes B, Joshi I and Wareham J. Health 2.0 and medicine 2.0: tensions and controversies in the field. *J Med Internet Res* 2008; 10: e1056. PMID: 18682374.
12. Liu X, Guo X, Wu H, et al. The impact of individual and organizational reputation on physicians' appointments online. *Int J Electron Commer* 2016; 20: 551–577.
13. Nambisan P. Information seeking and social support in online health communities: impact on patients' perceived empathy. *J Am Med Inform Assoc* 2011; 18: 298–304. PMID: 21486888.
14. Chen Q, Xu D, Fu H, et al. Distance effects and home bias in patient choice on the internet: evidence from an online healthcare platform in China. *China Econ Rev* 2022; 72: 101757.
15. Guo S, Guo X, Fang Y, et al. How doctors gain social and economic returns in online health-care communities: a professional capital perspective. *J Manag Inf Syst* 2017; 34: 487–519.
16. Angst CM, Devaraj S and D'Arcy J. Dual role of IT-assisted communication in patient care: a validated structure-process-outcome framework. *J Manag Inf Syst* 2014; 29: 257–292.
17. Guo S, Guo X, Zhang X, et al. Doctor–patient relationship strength's impact in an online healthcare community. *Inf Technol Dev* 2017; 24: 279–300.
18. Lagu T, Metayer K, Moran M, et al. Website characteristics and physician reviews on commercial physician-rating websites. *JAMA* 2017; 317: 766–768. PMID: 28241346.
19. Nabi MNU, Zohora FT and Akther F. Influence of word of mouth (WOM) in physician selection by the patients in Bangladesh. *Int J Pharm Healthc Mark* 2022; 16: 542–560.
20. Hall MA. The legal and historical foundations of patients as medical consumers. *Geo LJ* 2007; 96: 583.
21. SG. B. To quell criticism, some doctors require patients to sign 'gag orders', https://www.washingtonpost.com/wp-dyn/content/article/2009/07/20/AR2009072002335.html (2009, accessed 23 October 2023).
22. Reimann S and Strech D. The representation of patient experience and satisfaction in physician rating sites. A criteria-based analysis of English- and German-language sites. *BMC Health Serv Res* 2010; 10: 1–14. PMID: 21138579.
23. Emmert M, Sander U and Pisch F. Eight questions about physician-rating websites: a systematic review. *J Med Internet Res* 2013; 15: e2360. PMID: 23372115.
24. Hao H, Zhang K, Wang W, et al. A tale of two countries: international comparison of online doctor reviews between China and the United States. *Int J Med Inform* 2017; 99: 37–44. PMID: 28118920.
25. Rothenfluh F, Germeni E and Schulz PJ. Consumer decision-making based on review websites: are there differences between choosing a hotel and choosing a physician? *J Med Internet Res* 2016; 18: e129. PMID: 27311623.
26. Comscore. Online Consumer Generated Reviews Have Significant Impact on Offline Purchase Behavior, https://www.comscore.com/Insights/Press-Releases/2007/11/Online-Consumer-Reviews-Impact-Offline-Purchasing-Behavior (2007, accessed 23 October 2023)
27. Emmert M and Meier F. An analysis of online evaluations on a physician rating website: evidence from a German public reporting instrument. *J Med Internet Res* 2013; 15: e157. PMID: 23919987.
28. Hanauer DA, Zheng K, Singer DC, et al. Parental awareness and use of online physician rating sites. *Pediatrics* 2014; 134: e966–e975. PMID: 25246629.
29. McBride DL. Parental use of online physician rating sites. *J Pediatr Nurs* 2015; 30: 268–269. PMID: 25450443.
30. Zusman EE. Managing an online reputation. *Neurosurgery* 2013; 72: N11–N14. PMID: 23511827.
31. Land M. How your content strategy is critical for reputation management, https://martech.org/how-your-content-strategy-is-critical-for-reputation-management/ (2012, accessed 23 October 2023).
32. Escoffery RM and Bauer JG. Manage your online reputation–or someone else will. *Aesthet Surg J* 2012; 32: 649–652. PMID: 22628897.
33. McGrath RJ, Priestley JL, Zhou Y, et al. The validity of online patient ratings of physicians: analysis of physician peer reviews and patient ratings. *Interact J Med Res* 2018; 7: e9350. PMID: 29631992.
34. Ramachandran S, Ring D, Langerhuizen D, et al. A large number of reviews on physician rating websites may reflect reputation management. *Iowa Orthop J* 2022; 42: 283–286.
35. Ott M, Choi Y, Cardie C, et al. Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Oregon, USA, 19–24 June 2011, pp.309–319. Pennsylvania: ACL.

36. Rayana S and Akoglu L. Collective opinion spam detection: bridging review networks and metadata. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, Sydney, Australia, 10–13 August 2015, pp.985–994. New York: ACM.
37. Li H, Fei G, Wang S, et al. Bimodal distribution and co-bursting in review spam detection. In: Proceedings of the 26th international conference on world wide web, Perth, Australia, 3–7 April 2017, pp.1063–1072.
38. Ott M, Cardie C and Hancock JT. Negative deceptive opinion spam. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies, Atlanta, Georgia, June 2013, pp.497–501. Pennsylvania: ACL.
39. Hammad AA and El-Halees AM. An approach for detecting spam in Arabic opinion reviews. *Int Arab J Inf Technol* 2015; 12: 9–16.
40. Barbado R, Araque O and Iglesias CA. A framework for fake review detection in online consumer electronics retailers. *Inf Process Manag* 2019; 56: 1234–1244.
41. Li F, Huang M, Yang Y, et al. Learning to identify review spam. In: Proceedings of the twenty-second international joint conference on artificial intelligence, Catalonia, Spain, 16–22 July 2011, pp.2488–2493. California: AAAI.
42. Mukherjee A, Liu B and Glance N. Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st international conference on world wide web, Lyon, France, 16–20 April 2012, pp.191–200. New York: ACM.
43. Fornaciari T and Poesio M. Identifying fake Amazon reviews as learning from crowds. In: Proceedings of the 14th conference of the European chapter of the association for computational linguistics, Gothenburg, Sweden, April 2014, pp.279–287. Pennsylvania: ACL.
44. Shojaee S, Murad MAA, Azman A, et al. Detecting deceptive reviews using lexical and syntactic features. In: 2013 13th international conference on intelligent systems design and applications, Salangor, Malaysia, 8–10 December 2013, pp.53–58. Piscataway: IEEE.
45. Lau RYK, Liao SY, Kwok RC-W, et al. Text mining and probabilistic language modeling for online review spam detection. *ACM Trans Manage Inf Syst* 2012; 2: 1–30.
46. Liu W, Jing W and Li Y. Incorporating feature representation into BiLSTM for deceptive review detection. *Computing* 2019; 102: 701–715.
47. Rastogi A, Mehrotra M and Ali SS. Effect of Various factors in context of feature selection on opinion spam detection. In: 2021 11th international conference on cloud computing, data science & engineering (Confluence), Noida, India, 28–29 January 2021, pp.778–783. Piscataway: IEEE.
48. Shang Y, Liu ML, Zhao TJ, et al. T-Bert: a spam review detection model combining group intelligence and personalized sentiment information. In: 30th international conference on artificial neural networks, Bratislava, Slovakia. Cham: Springer International Publishing, 14–17 September 2021, pp.409–421.
49. Zhang W, Du Y, Yoshida T, et al. DRI-RCNN: an approach to deceptive review identification using recurrent convolutional neural network. *Inf Process Manage* 2018; 54: 576–592.
50. Jain N, Kumar A, Singh S, et al. Deceptive reviews detection using deep learning techniques. In: 24th international conference on applications of natural language to information systems, Salford, UK. Cham: Springer International Publishing, 26–28 June 2019, pp.79–91.
51. Al-Obaydy WNI, Hashim H, Najm Y, et al. Document classification using term frequency-inverse document frequency and K-means clustering. *Indones J Electr Eng Comput Sci* 2022; 27: 1517–1524.
52. Qaiser RA. Text mining: use of TF-IDF to examine the relevance of words to documents. *Int J Comput Appl* 2018; 181: 25–29.
53. Ikemura K, Bellin E, Yagi Y, et al. Using automated machine learning to predict the mortality of patients with COVID-19: prediction model development study. *J Med Internet Res* 2021; 23: e23458. PMID: 33539308.
54. Biau G and Scornet E. A random forest guided tour. *TEST* 2016; 25: 197–227.
55. Kim Y. Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, October 2014. Pennsylvania: ACL.
56. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
57. Derczynski L. Complementarity, F-score, and NLP evaluation. In: Proceedings of the 10th international conference on language resources and evaluation (LREC'16), Portorož, Slovenia, May 2016. pp.261–266. Paris: ELRA.
58. Molnar C. Interpretable machine learning: A guide for making black box models explainable, https://christophm.github.io/interpretable-ml-book/ (2021, accessed 23 October 2023).
59. Rout JK, Singh S, Jena SK, et al. Deceptive review detection using labeled and unlabeled data. *Multimed Tools Appl* 2016; 76: 3187–3211.
60. Zhao Y, Da J and Yan J. Detecting health misinformation in online health communities: incorporating behavioral features into machine learning based approaches. *Inf Process Manag* 2021; 58: 102390.
61. Hao H and Zhang K. The voice of Chinese health consumers: a text mining approach to web-based physician reviews. *J Med Internet Res* 2016; 18: e108. PMID: 27165558.
62. Alsubari SN, Shelke MB and Deshmukh SN. Fake reviews identification based on deep computational linguistic, https://www.researchgate.net/publication/342154025_Fake_Reviews_Identification_Based_on_Deep_Computational_Linguistic_Features (2020, accessed 23 October 2023).
63. Banerjee S, Chua AYK and Kim J-J. Using supervised learning to classify authentic and fake online reviews. In: Proceedings of the 9th international conference on ubiquitous information management and communication, Bali, Indonesia, 8–10 January 2015, pp.1–7. New York: ACM.
64. Li Y, Feng X and Zhang S. Detecting fake reviews utilizing semantic and emotion model. In: 2016 3rd international conference on information science and control engineering (ICISCE), Beijing, China, 8–10 July 2016, pp.317–320. Piscataway: IEEE.
65. Wang JD, Kan HT, Meng FQ, et al. Fake review detection based on multiple feature fusion and rolling collaborative training. *IEEE Access* 2020; 8: 182625–182639.
66. Mohawesh R, Xu SX, Tran SN, et al. Fake reviews detection: a survey. *IEEE Access* 2021; 9: 65771–65802.
67. Capitaine L, Genuer R and Thiébaut R. Random forests for high-dimensional longitudinal data. *Stat Methods Med Res* 2021; 30: 166–184.

68. Iliev AI, Nimmala A, Rahiman RA, et al. Fake review recognition using an SVM model. In: Intelligent computing, SAI 2023. Cham: Springer International Publishing, 13–14 July 2023, pp.885–893.
69. Liu C, Wu X, Yu M, et al. A two-stage model based on BERT for short fake news detection. In: Knowledge science, engineering and management: 12th International Conference, KSEM 2019, Athens, Greece. Cham: Springer International Publishing, 28–30 August 2019, pp.172–183.
70. Refaeli D and Hajek P. Detecting fake online reviews using fine-tuned BERT. In: Proceedings of the 2021 5th international conference on e-business and internet, Singapore, Singapore, 15–17 October 2021, pp.76–80. New York: ACM.