IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 26, 2020, accepted January 4, 2021, date of publication January 12, 2021, date of current version January 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051174

An Ensemble Model for Fake Online Review Detection Based on Data Resampling, Feature Pruning, and Parameter Optimization

JIANRONG YAO, YUAN ZHENG, AND HUI JIANG
School of Information Engineering and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou 310018, China
Corresponding author: Hui Jiang (fidojianghui@hotmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 71704153, Grant 71701180, and Grant 71704020; and in part by the China Postdoctoral Science Foundation under Grant 2018M642472.

ABSTRACT With the widespread of fake online reviews, the detection of fake reviews has become a hot research issue. Despite the efforts of existing studies on fake review detection, the issues of imbalanced data and feature pruning still lack sufficient attention. To address these gaps, the present study proposes an ensemble model for the detection of fake online reviews. The model consists of four steps, and the first three steps are proposed to optimize the base classifiers: (i) Data resampling: We propose a novel way to address the data imbalance problem by combining the resampling and the grid search technique. (ii) Feature pruning: We propose an ablation study to drop unimportant features. (iii) Parameters optimization: We apply the grid search algorithm to determine suitable values of the relevant parameters for each base classifier. (iv) Classifier ensembling: We apply majority voting and stacking strategies to integrate the optimized base classifiers. The proposed data resampling method is also applied for the meta-classifier in the stacking ensemble model. This study produces advances in terms of combining different methods or algorithms into a model and the results show that the proposed ensemble model outperforms some existing techniques, thereby providing a new way to solve the data imbalance and feature pruning issues in the field of fake review detection.

INDEX TERMS Data resampling, ensemble model, fake review, feature pruning, parameter optimization.

I. INTRODUCTION

Currently, with the development of e-commerce, online shopping is becoming increasingly prevalent. Researchers have demonstrated that online reviews have a significant impact on consumers' purchase decisions, thus influencing the sale of products [1], [2]. Unfortunately, some merchants or consumers manipulate product ratings by writing fake reviews, which aim to mislead consumers in making their purchase decision making [3], [4]. Studies have found that fake reviews widely exist on online websites [5], [6]. For instance, one study estimated that 16 percent of restaurant reviews on Yelp (one of the most famous review websites in America) are spam [5].

The prevalence of fake reviews is becoming a severe problem, as it misleads consumers when making their purchase decisions [7] and results in great damage to the sustainable development of online review systems. Some websites allow consumers to report reviews that they suspect to be fake. However, it is difficult for consumers to identify fake reviews [8] because some of them are written carefully and resemble authentic reviews. Because of the difficulty of identifying fake reviews manually, searching for an automatic detection method is the main direction of related research. Among various types of fake review detection methods, machine learning methods have been widely used [9]–[14]. However, some problems remain understudied.

First, fake reviews represent only a small proportion of the total reviews, and this causes a data imbalance problem. To reduce the negative impact of imbalanced data, some researchers select only a subset of truthful reviews [3], [15], [16], but some useful information might be removed in this way. In addition, the performances of

The associate editor coordinating the review of this manuscript and approving it for publication was Haiyong Zheng.

machine learning algorithms greatly rely on the selection of features; however, the usefulness of the selected features is uncertain. Hence, a feature pruning process needs to be considered. Although there are various methods for calculating the importance of features, the actual performance loss incurred when a feature is dropped remains unknown. Therefore, it is not easy to conclude which features are not important. Moreover, ensemble strategies still have not been sufficiently considered.

To address these gaps, the current study proposes an ensemble model for fake review detection. Five supervised machine learning classifiers are selected as the base classifiers. Initially, three steps are conducted to optimize the selected base classifiers. First, we propose a novel approach to alleviate the impact of data imbalances by combining data resampling with the grid search method. We select two representative data resampling methods. For each resampling method and classifier, we apply the grid search method to search for the best sampling ratio for each classifier because for different classifiers, the most suitable sampling ratios might be different [17]. However, some existing studies [16], [18], [19] neglect the impact of the sampling ratio and completely rebalance the dataset, and this decreases the robustness of the model against imbalanced datasets. Since there are several different classifiers, it is time-consuming and inaccurate to manually find the best sampling ratio for each classifier. The proposed approach can find the best sampling ratio for each classifier effectively and accurately. Next, we conduct an ablation study for feature pruning. Specifically, we calculate the influence of each predictive feature by dropping one of them at a time. The basic idea is that since the performance of a classifier can be evaluated by its F1-score, the usefulness of a feature can be calculated by recording how the F1-score changes when this feature is dropped. Next, the grid search method is applied once again to optimize the parameters for each classifier. To avoid overfitting, 10% of the data are selected as the validation data for the optimization task. After we optimize the base classifiers, we apply the voting and stacking ensemble strategy to integrate them. The ensemble model can reduce the relative weaknesses of single classifiers as they are compensated by the advantages of other classifiers [20]. And this approach has been proven to be effective in improving the performance of classification [21], [22]. And two Yelp datasets [16] are used to evaluate the model. To the best of our knowledge, few studies on fake review detection have combined these processes in a model.

The present study contributes to the literature by providing a different way to effectively detect fake reviews. (i) Specifically, we initially look at a very novel approach by combining data resampling with the grid search method to address the data imbalance problem, as this can effectively improve the performance of the model to a large extent on an imbalanced dataset. The results show that the proposed approach can effectively improve the performance for each classifier, especially for the random forest (RF), whose F1-scores are improved by 4.65% and 2.98% on the two Yelp datasets, respectively. (ii) Prior studies proposed many features for fake review detection [3], [16], [18]. However, few studies have focused on the usefulness of the selected features, and there is no definition of which features are useless. We propose an ablation study, and the results show that it can effectively identify useless features for each classifier. Therefore, the performances of the classifiers are improved. (iii) We propose two ensemble strategies for integrating the base classifiers, and these help to compensate for the weaknesses and instabilities of the base classifiers. Additionally, unlike those in the basic ensemble method, the base classifiers in our ensemble model are optimized. According to the results, the F1-scores increase by approximately 3% when the base classifiers are optimized.

The remainder of the current study is structured as follows. Section 2 presents a literature review with regard to fake review detection. Section 3 shows the modeling process. Section 4 demonstrates the results of the experiments and our analysis. Finally, we present our conclusions, limitation, and future work in section 5.

II. LITERATURE REVIEW

In this section, we briefly review several related works on fake review detection, including classification methods, approaches for addressing data imbalances, and feature selection methods. We also present the problems with existing studies.

A. CLASSIFICATION METHODS

Regarding classification methods, machine learning methods are the most frequently used subtype for fake review detection. Machine learning can be classified into two categories: supervised learning and unsupervised learning. Supervised learning is the dominant approach in the field of fake review detection [8], [10], [12], [15], [23]. There are many supervised learning algorithms, and it is not easy to decide which one is the best [10]. Apart from supervised learning, some researchers use unsupervised learning methods [24], [25] or deep learning methods [26] to identify fake reviews due to the difficulty of labeling data. However, considering the lack of large scale datasets, deep learning methods might not be effective. In addition, ensemble learning methods have also been proposed. Ruan *et al.* [27] proposed an ensemble model using the geolocation information of users. The results showed that the ensemble model could enhance the stability over those of the base models. However, ensemble strategies are still understudied.

B. APPROACHES FOR ADDRESSING IMBALANCED DATA

In the field of fake review detection, the issue of imbalanced data should be considered. Imbalanced data refers to a dataset within which one or some of the classes have a larger number of examples than others. The most prevalent class is called the majority class, while the less prevalent classes are the minority classes [28]. In most cases, the number of fake reviews is much smaller than that of truthful reviews.

Trained with such an imbalanced dataset, a model often provides suboptimal classification results in which the majority examples are converged while the minority examples are discarded [17]. To solve this problem, some researchers have randomly selected the same quantity of non-fake reviews as that of fake reviews [3], [15], [16], but in this process, some important information may be removed. Some other researchers have used Amazon Mechanical Turk (AMT) to generate equivalent numbers of fake reviews and non-fake reviews [8], [23], [29]. However, pseudo fake reviews are quite different from fake reviews written by real spammers; fake reviews written by real spammers are carefully crafted to resemble truthful reviews, while pseudo fake reviews are quite different from truthful reviews. Thus, the pseudo fake reviews are much easier to detect. Hence, it might not be effective to use a model trained with pseudo fake reviews to identify fake reviews written by real spammers [12].

Among several methods, resampling techniques are frequently used to rebalance imbalanced datasets because they are independent of the selected classifier [30]. There are three types of resampling methods that have been used by researchers: over-sampling methods [31]–[33], under-sampling methods [34], [35], and hybrid methods [36]. Furthermore, according to a recent study [17], fraud detection is one of the most researched imbalanced learning topics. However, only a small number of papers [12] have proposes resampling methods for fake review detection.

C. FEATURE SELECTION METHODS

To improve the performances of fake review detection models, some researchers have studied the selection of predictive features. Features can be divided into two types: review-centric features and user-centric features [18].

Review-centric features refer to features based on the textual content of reviews. Research has shown that the textual content of a review is essential for consumers to judge whether the review is fake [37]. A typical model was proposed by Ott *et al.* [8], who applied an n-gram term frequency method to detect fake reviews. Using SVM, the results achieved an accuracy of 89.6%. However, Mukherjee *et al.* [16] applied the same methods [8], [23] to the Yelp dataset, where the fake reviews are written by real spammers. The accuracy decreased to 67.8%, which is significantly lower than 89.6%. Thus, the researchers [16] concluded that it might be difficult to detect fake reviews written by real spammers merely by using review-centric features. Furthermore, the study [16] proves that adding reviewer-centric features significantly improves the performance of the classification task.

User-centric features (or reviewer-centric features) refer to the features collected from the reviewer's profile characteristics and behavioral patterns [10], such as the reviewer's number of friends and the number of reviews a reviewer has posted. Although most researchers focus on review-centric features, some researchers [16] have found that fake reviews use languages similar to those used in authentic reviews, thereby increasing the difficulty of identifying fake reviews merely by using review-centric features [3]. Thus, some studies have combined both review-centric and reviewer-centric features to identify fake reviews. Zhang *et al.* [3] divided features into two categories: verbal and nonverbal features. The results showed that after adding nonverbal features, the accuracy of the model increased remarkably (5%-10%). Barbado *et al.* [18] proposed a Fake Feature Framework (F3) model based on machine learning classifiers; this model further subdivides user-centric features into four types: personal, social, review activity, and trust features.

In the studies mentioned above, the selection of features was mainly based on the prediction of their usefulness. To further explore the usefulness of each feature, Zhang *et al.* [3] conducted a sensitivity analysis to calculate the importance of each feature. The results showed that the model trained with the top twelve important features outperforms the model trained with all features. Additionally, Mukherjee *et al.* [12] dropped one feature at a time to calculate the actual performance change when a feature was removed.

Based on the previous studies mentioned above, we find that even though various fake review detection methods have been proposed, and some of them have proven to be effective, there still exist some problems. First, to solve the problem of imbalanced data, some studies [3], [15], [16] merely select a subset of non-fake reviews randomly to balance the distributions of the two classes, but this may result in the loss of some vital information. In addition, the performances of these methods [3], [15], [16] are estimated using a balanced testing dataset. Hence, these methods may not be effective in detecting fake reviews on online websites because the distributions of their reviews are imbalanced [12]. Furthermore, since the selection of features is based on the subjective prediction of their usefulness, the actual usefulness of each feature remains unknown. Thus, a feature pruning process should be considered. Moreover, ensemble strategies still lack enough attention.

III. RESEARCH FRAMEWORK

The research framework followed in the current study is shown in Fig. 1. We first select the predictive features that we need. We divide all the features into two categories: review-centric and reviewer-centric features. Most of the reviewer-centric features are already contained in the dataset. For textual features, we need several stages (text preprocessing) and tools (NLTK, etc.) before extracting features from the content of reviews to attain improved input data. Next, we select five supervised machine learning classifiers as the base classifiers, and three steps: data resampling, feature pruning, and parameter optimization are applied to optimize the base classifiers. Finally, we apply two ensemble strategies to integrate the base classifiers.

A. NOMENCLATURE

The notions that appear in this study are listed in Table 1.

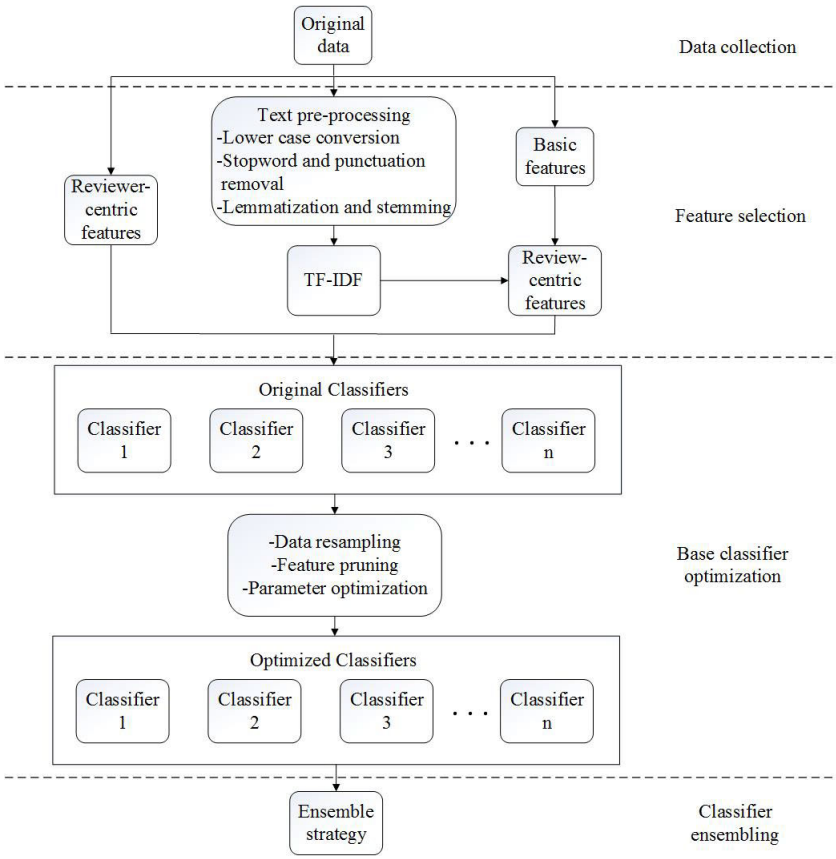


FIGURE 1. Research framework.

TABLE 1. Nomenclature.

Notion	Description
TF (w, D)	The term frequency (TF) value of a word/term
f_w	The ratio at which a word/term appears in a document
IDF (t)	The inverse document frequency (IDF) value of a word//term
C	The number of documents
df(t)	The number of documents containing word t
C	The set of selected classifiers
C_i	Classifier *i*
F	The set of selected features
F_i	Feature *i*
U	The set of the usefulness values of features
U_{ij}	The usefulness of F_j for C_i

B. FEATURE SELECTION

In this study, we divide the features into two types: review-centric features and reviewer-centric features. The reviewer-centric features reflect the social interaction of reviewers, such as the total number of friends a reviewer has and so on. Social interaction reflects a reviewer's preference for sharing his/her personal experiences or feelings. We believe that reviewers who are willing to share their feelings and interact with others are less likely to write fake reviews. The selection of reviewer-centric features is based on their availability and our prediction of their influence on our model.

For review-centric features, some studies indicate that it is not effective to classify fake reviews with only these review-centric features [3], [16], [18] because spammers are becoming increasingly sophisticated. However, we think there still exist some differences between the textual content of fake and non-fake reviews. Some basic review-centric features, such as the number of words and the rating of a review, are easy to obtain. However, it is difficult to extract features from textual content. To extract features from textual content, a text pre-processing step and a textual feature extraction method are needed.

1) TEXT PRE-PROCESSING

To extract features from the textual content of reviews accurately and effectively, we first conduct a text pre-processing stage, including lowercase conversion, stopword and punctuation removal, word lemmatization, word stemming, and spelling mistake correction. By performing lowercase conversion and punctuation removal, the text is homogenized [10], which facilitates further processing. Stopwords are words that do not carry useful information, such as pronouns and prepositions, and removing stopwords contributes to improving the text processing performance [38]. Word lemmatization and stemming are quite similar: they can transform words into their roots. By word lemmatization and stemming, most of the derivational affixes are transformed to their original

TABLE 2. An example of a review after pre-processing.

Original text	Pre-processing
Being a bit of a Japonais snob, I was somewhat reluctant to try this place, but after checking out their bar on Saturday and then dinner on Sunday, I may have to change my ways. The lychee martinis (don't get the divine, just the regular) are simply delicious and not too sweet. ?The oxtail potstickers were super tasty, and the entrees that I ordered (yes, I ordered two entrees) -- the cod and the pork bell -- were simply heaven. ?Both dishes were so tender and mouth-wateringly good. ?To top it off, I had a dessert known simply as "Ridiculous" and, indeed, it lived up to the name. ?Unbelievable drinks, sublime food and dessert, and mesmerizing decor... this place deserves 5 stars. I've been told that I need to come back for brunch to try their famous Indonesian Nasi Goreng. ?I can't wait.	bit japonais snob somewhat reluctant try place check bar saturday dinner sunday may change way lychee martini get divine regular simply delicious sweet oxtail potstickers super tasty entree order yes order two entree cod pork bell simply heaven dish tender mouth wateringly good top dessert know simply ridiculous indeed live name unbelievable drink sublime food dessert mesmerize decor place deserve 5 star tell need come back brunch try famous indonesian nasi goreng wait

forms. For example, ''stays'', ''stayed'', and ''staying'' are transformed to ''stay'' after lemmatization.

The pre-processing step can be achieved using the Natural Language Toolkit (NLTK), which provides interfaces for various corpora and other resources such as lexical resources. Table 2 shows an example of a review after text pre-processing.

2) TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

TF-IDF is a statistical metric used to measure the importance of a term to a text in a dataset or corpus. The importance of a term is positively correlated with the frequency the term appears in a text, and it is negatively correlated with the frequency with which the term appears in the whole dataset. Thus, TF-IDF reduces the influence of some useless but frequently appearing words (which is one of the disadvantages of the n-gram model).

The value of TF-IDF equals the product of two terms: term frequency (TF) and inverse document frequency (IDF). The term frequency equals the ratio by which a term appears in a certain document, and it can be computed as follows:

$$TF\ (w, D) = f_w \tag{1}$$

The IDF equals the logarithm of the total number of times that a term appears in the whole corpus divided by the number of documents where the term appears. The IDF value of word t can be computed as follows:

$$IDF\ (t) = 1 + log(C/1 + df(t)) \tag{2}$$

where C is the number of documents and $df(t)$ is the number of documents that contain word t.

Researchers [16] have proven that for the n-gram model, bigrams slightly outperform unigrams in both restaurant and hotel domains, so in this paper, we use TF-IDF with bigrams. All the selected features are presented in Table 3.

C. CLASSIFICATION ALGORITHMS

Among various classification algorithms, supervised algorithms are still the most frequently used methods for fake review detection, such as supervised machine learning algorithms and deep learning methods. Although deep learning methods have been proven to be effective, they require sufficient training data, so they might not be suitable for fake review detection due to the difficulty of labeling data manually. Given such conditions, we select five machine learning algorithms as the base classifiers: random forest (RF), Xgboost, Lightgbm, Catboost and gradient boosting decision tree (GBDT). The reason for choosing several different classifiers is because of the famous no free lunch theorem [39], which states that there is no guarantee that a single classifier can perform best for all datasets.

D. CLASSIFIER OPTIMIZATION

In this section, we introduce the process for optimizing the selected base classifiers, and this consists of three steps: data resampling, feature pruning, and parameter optimization. For the first two steps, the parameters of the classifiers are set to default values since their goals are addressing the data imbalance and feature pruning problems. The main parameters of the classifiers are optimized in the parameter optimization step using the grid search method.

1) DATA RESAMPLING

In most cases, fake reviews only represent a small proportion of the total reviews, resulting in a data imbalance problem. It is recognized that the performance of a model trained with highly imbalanced data is usually poor. To solve this problem, creating a balanced dataset by crowdsourcing [8], [23], [29] or simply deleting a portion of the truthful reviews [3], [15], [16] is frequently applied to balance the distributions of two classes. However, these models may not be practical for detecting fake reviews on review websites, as their reviews are highly imbalanced.

One study by Guo *et al.* [17] proved that resampling techniques are feasible for alleviating the influence of imbalanced data. There exist two widely used resampling methods: SMOTE and Random Under Sampler (RUS).

SMOTE is an over-sampling method that eliminates the influence of a imbalanced dataset by generating samples belonging to minority classes. Rather than merely duplicating the minority samples, the SMOTE method uses the K-nearest neighbors (K-NN) algorithm to generate synthetic minority samples, thus enhancing the stability of resampling. RUS

TABLE 3. Selected features of our model.

Categories	Features	Description
Review-centric features	Rating	The rating related to a review (5 means most satisfying, and 1 means least satisfying)
	Word count	The number of words a review has
	TF-IDF	A method that extracts features from texts
Reviewer-centric features	Friend count	The number of friends a reviewer has
	Review count	The number of reviews a reviewer has ever posted
	First review count	The number of first reviews a reviewer has posted about any restaurant
	Useful votes received	The number of "useful"" votes a reviewer has received
	Cool votes received	The number of "cool" votes a reviewer has received
	Funny votes received	The number of "funny" votes a reviewer has received
	Compliments received	The number of "compliments" a reviewer has received
	Tip count	The number of tips
	Follower count	The number of followers a reviewer has
	Active window	The time interval between the first review and the last review
	Average rating	The average rating of reviews from a reviewer

is one of the simplest yet most effective under-sampling methods; it balances a dataset by randomly eliminating the majority class samples [40].

However, it should be noted that it is not always necessary to balance the number of majority and minority samples: for different datasets or classifiers, the best sampling ratios might be different [17]. However, this has been neglected by some previous studies on fake review detection [3], [15], [16], which may decrease the robustness of the model when the dataset is imbalanced. To address this gap, we propose a novel approach by combining the resampling method and the grid search method [41]. Grid search is a representative method for parameter optimization that makes a complete search over the given subset of the parameter space of an algorithm [41]. When there are few parameters that need to be optimized, the grid search method can optimize the parameters accurately and efficiently. Since the best sampling ratios for data resampling with different classifiers might be different, it is time consuming to manually search for a suitable sampling ratio for each classifier. Therefore, applying grid search can efficiently and accurately to automatically find the best sampling ratio for each classifier.

For each resampling method, we gradually increase the sampling ratio and record the performance (F1-score) of the model, and the sampling ratio that produces the highest F1-score is regarded as the best sampling ratio for a given classifier. To avoid overfitting, this process is conducted only on training data. We do not apply resampling methods to the testing data since some bias may result when some of the testing data are eliminated.

2) FEATURE PRUNING

Although the results of the tested classifiers trained with all features may be satisfactory, we believe the performance can be further improved by pruning unimportant features. On the one hand, the selection of features is based on previous studies and our subjective predictions of their usefulness. Hence, some features might be regarded as noise and are not useful to a classifier. On the other hand, for different classifiers, the influences of a feature could be different. However, it is difficult to determine which features are unimportant.

Although there are various feature selection methods, they only calculate the importance of a feature, while the actual performance loss incurred when the feature is dropped remains unknown [12]. Under such circumstances, we apply an ablation study to determine and prune the unimportant features. The ablation study typically refers to removing some part of the model or algorithm, and determine how the performance of the model changes. In most cases, the ablation study is applied to evaluate the performance of a submodel, and it has rarely been used for calculating the usefulness of features.

In this model, we apply an ablation study to calculate the usefulness of features. For each classifier, we drop one feature at a time from the full feature set and record how the F1-score changes. A feature is regarded to be unimportant and is then pruned if the F1-score increases or remains unchanged when this feature is dropped; otherwise the feature is retained. Algorithm 1 shows the detailed process of the ablation study.

Algorithm 1 Ablation Study for Feature Pruning

Input: Set of selected classifiers $\mathbf{C} = \{C_1, C_2, \ldots, C_m\}$;
Set of selected features $\mathbf{F} = \{F_1, F_2, \ldots, F_n\}$;
Output: The usefulness of F_j for C_i $\mathbf{U} = \{U_{ij}\}$ (i $\in$ [1,m],j $\in$ [1,n]) // $U_{ij} = 1$ represents that feature j is useful for classifier i, and $U_{ij} = 0$ represents that feature j is unimportant or not useful for classifier i and should be pruned in classifier i

```
for i = 1, 2, ..., m do
    for j = 1, 2, ..., n do
      drop F_j in C_i;
      if F1-score (without F_j) < F1-score (with F_j) then
        U_ij = 1
      else
        U_ij = 0
      end if
    end for
end for
```

3) PARAMETER OPTIMIZATION

After pruning the unimportant features, we optimize the parameters for each classifier since the performances of machine learning methods can be affected by the settings of some parameters [21], [22]. Additionally, since there are five classifiers in our ensemble model, it is time-consuming to manually optimize the parameters. To ensure the efficiency of the model, we apply the grid search method to automatically find suitable values for the main parameters of each classifier, and the other parameters are set to their default values since their influence is limited. To decrease the time-consumption of the proposed approach, for each optimized parameter, the alternative values are first set from a large range of values. Then, the range is reduced according to the results. This process is also proposed for the training data to avoid overfitting. By using the grid search method, the proposed model can be computationally efficient to be deployed in an online review system.

The reason why the optimization of the parameters is proposed as the last step is that for different features, the suitable values of a given parameter may be different. Therefore, if we conduct this process before feature pruning, the results may not be effective for classifiers trained with retained features.

E. CLASSIFIER ENSEMBLING

The use of an ensemble strategy that combines several single classifiers is effective to guarantee the robustness and stability of the model [42] because the ensemble model can compensate for the weakness of an individual classifier. In the current study, we propose two types of ensemble strategies: majority voting and stacking.

The majority voting strategy is widely used due to its simplicity. In the majority voting strategy, each base classifier predicts the labels of samples, and the labels with the highest numbers of votes are regarded as the final outputs. Another strategy, stacking, trains a meta-classifier that receives the results of several low-level classifiers [43]. The performance of the meta-classifier is usually better than that of any low-level classifier. In the present study, we select the Gaussian kernel (because it performs best on the proposed datasets) support vector machine (SVM) classifier as the meta-classifier since it is found to outperform other meta-classifiers in the proposed model, and the parameters are also optimized using the grid search method. We also apply the proposed data resampling method on the meta-classifier to optimize the model to the largest extent.

IV. RESULTS AND ANALYSIS

In this section, we present the performance of the proposed model. For all the experiments, we use the F1-score as the metric to evaluate the performances of the classifiers, and the fake reviews are set as positive samples to ensure that the F1-score reflects the performance in terms of detecting fake reviews. All experiments are conducted using the Python 3.7 environment on a PC with 4 Intel (R) Core (i5) quad core CPUs (3.10 GHz) and 8 GB of RAM.

A. DATASET

Due to the inaccuracy of humans with regard to identifying fake reviews, there are only a few datasets that contain labels [8], [16], [29]. Initially, researchers [8], [23] created the first public dataset that contains fake reviews using crowdsourcing. The fake reviews were written by a group of workers recruited from Amazon Mechanical Turk, and the legitimate reviews were collected from the 20 most popular hotels on TripAdvisor. Later, researchers [16] conducted an experiment on the Yelp dataset, where all the fake reviews were written by real spammers. Yelp uses its filtering algorithm to filter fake reviews automatically. However, Yelp does not reveal the process of the algorithm. The study [16] concluded that pseudo fake reviews are quite dissimilar to real-life fake reviews. Thus, classifiers trained with pseudo fake reviews may not be effective in detecting fake reviews written by real spammers.

Given the above concerns, we apply two representative Yelp datasets [12], [16] from the hotel and restaurant domains, respectively that have been used by many studies, such as [3], [12], [15], [16], [27]. The hotel dataset contains reviews across 85 hotels in Chicago. The restaurant dataset contains reviews across 130 restaurants in Chicago. In addition to the contents of reviews, the dataset contains additional features about the reviewers, such as the number of friends of each reviewer. The reason why we choose these datasets is that they contain not only the contents of the reviews but also abundant features about the reviewers, which might be useful for fake review detection. Although the datasets have been widely used by many studies, most of them simply chose some of the truthful data or rebalanced the distributions of two samples using the RUS method. A model trained with balanced training data might not be effective when the testing data are imbalanced [12]. To enhance the stability of the model, we retain the data imbalance. Only the data that contain null values or abnormal values are eliminated because these data could be seen as noise and are thus useless to our model. The details of the Yelp datasets are shown in Table 4. From Table 4, we can see that the distribution of data is imbalanced.

TABLE 4. The components of our dataset.

Domain	Total	Truthful	Fake	Fake
Hotel	5778	5009	769	13.3%
Restaurant	26300	20231	6069	23.1%

B. DATA RESAMPLING

In this experiment, we train and test five classifiers with all the selected features (shown in Table 3). Most of the parameters of the tested algorithms are set to their default values. The range of the sampling ratio is 0.2 to 1 for the hotel dataset and 0.35 to 1 for the restaurant dataset. The sampling

TABLE 5. F1-scores of 5 classifiers trained with all features.

Resampling method	Classifier	F1-score (%)	
		Hotel	Restaurant
No	RF	66.37	72.18
	Xgboost	69.24	74.11
	Lightgbm	69.60	75.66
	Catboost	68.70	75.30
	GBDT	69.31	74.46
SMOTE	RF	70.70	73.57
	Xgboost	69.88	74.93
	Lightgbm	70.02	75.45
	Catboost	69.49	75.57
	GBDT	69.38	74.88
RUS	RF	71.02	75.16
	Xgboost	70.98	75.47
	Lightgbm	70.36	76.36
	Catboost	70.92	76.65
	GBDT	70.33	75.64

TABLE 6. The performance of different sampling ratios on the restaurant dataset.

Sampling ratio	Precision (%)	Recall (%)	F1-score (%)
Original	76.39	74.24	75.30
0.35	75.08	76.86	75.95
0.4	73.66	79.04	76.24
0.45	72.51	80.99	76.50
0.5	71.58	82.52	**76.65**
0.55	70.45	83.83	76.56
0.6	69.52	85.12	76.52
0.65	68.93	85.98	76.52
0.7	68.05	87.11	76.40
0.75	67.29	87.84	76.19
0.8	66.34	88.46	75.81
0.85	65.87	88.89	75.65
0.9	65.32	89.31	75.45
0.95	64.85	90.15	75.42
1	64.22	90.28	75.04

ratio is regarded as the best sampling ratio if it yields the highest F1-score. The results of the experiment are shown in Table 5.

From Table 5, we can see that although the data are imbalanced, the five classifiers perform well on the datasets. The Lightgbm classifier performs best on both datasets when no resampling method is applied.

We also find that compared to the results obtained without using any resampling method, both the SMOTE and RUS algorithms are effective for solving the data imbalance problem, especially with regard to the RF classifier, whose F1-scores are significantly improved after applying the RUS method.

Moreover, we can observe that the RUS method outperforms the SMOTE method. We think the potential reason for this phenomenon is that when adequate training data are available, synthetic samples generated by over-sampling methods are quite different from real samples, which could produce noise for the classifiers. Considering the above analysis and previous research [44], [45], we conclude that when the dataset is large, under-sampling methods (such as RUS) outperform over-sampling methods (such as SMOTE) in most cases. Given that the RUS method outperforms SMOTE on our datasets, we apply the RUS method in the subsequent experiments.

To determine how the resampling method reduces the negative impact of imbalanced data, we take the Catboost classifier with RUS as an example and present the precision, recall, and F1-score values for different sampling ratios on the restaurant dataset. The results are shown in Table 6. It can be observed that the recall initially increases rapidly when we increase the sampling ratio of the RUS method, and this indicates that many more fake reviews are correctly identified than before. At the same time, the precision decreases, but the speed of this decrease is slower, which explains why the F1-score increases at the beginning. The classifier produces the highest F1-score when the sampling ratio is 0.5. When the sampling ratio is larger than 0.5, although the recall still increases when we increase the sampling ratio, it increases much more slowly than at the start. Therefore, the F1-score decreases simultaneously. The performance of the classifier worsens when the sampling ratio is set to 1, and this reinforces the conclusion that it is not always necessary to completely rebalance the two samples [17].

C. FEATURE PRUNING

After the data resampling step is completed, we conduct an ablation study for feature pruning. We drop each feature from the whole feature set (one at a time) and calculate the change in the F1-score. If the F1-score increases when a feature is dropped, then this feature is regarded as not useful and is removed. Here, we note that although the output of the TF-IDF is a vector, rather than an individual feature, we also regard it as a feature because it contains textual information. The results of the ablation study are shown in Table 7. From Table 7, we find that for different classifiers, the unimportant features are different, and this proves that some feature pruning methods might not be effective for some classifiers because they merely show the significance of each feature universally, rather than aiming at a specific classifier. Moreover, we notice that although TF-IDF is a well-known method for extracting features from textual content, it is not important for half of the selected classifiers. We believe the reason for this is that fake reviews are cautiously crafted, thus, both the writing style and the usage of words in fake reviews are similar to the approaches used in truthful reviews. In addition, we also find that the word count of a review is not important for most of the classifiers on the restaurant dataset, and this indicates that the length of a review is not a good indicator for fake review detection. These findings consolidate the idea that it is hard to detect fake reviews merely by textual content [12], [16], [18].

After pruning the unimportant features for each classifier, we train the classifiers with the retained features, and the results are shown in Table 8. From Table 8, we find that the F1-scores increase for all the selected classifiers, especially for the RF classifier. The results prove that it is feasible to improve the performance of a classifier by using an ablation study. Therefore, the removal of unimportant features should

TABLE 7. Unimportant features for each classifier.

Classifier	Unimportant features	
	Hotel	Restaurant
RF	Follower count	Rating
		Word count
		TF-IDF
Xgboost	First review count	"Cool" votes received
		Follower count
		Average rating
Lightgbm	TF-IDF	Word count
Catboost	First review count	Word count
	"Cool" votes received	TF-IDF
GBDT	Rating	Average rating
	"Cool" votes received	
	TF-IDF	

TABLE 8. F1-scores of 5 classifiers trained with pruned features.

Resampling method	Classifier	F1-score (%)	
		Hotel	Restaurant
RUS	RF	71.08	78.25
	Xgboost	71.10	75.79
	Lightgbm	70.73	77.02
	Catboost	71.16	77.12
	GBDT	71.34	75.82

TABLE 9. Main parameters and their values calculated by grid search.

Classifier	Parameters and their values	
	Hotel	Restaurant
RF	n_estimators = 85	n_estimators = 78
	max_features = 15	max_features = 3
	max_depth = 24	max_depth = 15
Xgboost	learning_rate = 0.1	learning_rate = 0.08
	n_estimators = 70	n_estimators = 50
	max_depth = 3	max_depth = 10
	min_child_weight = 1	min_child_weight = 1
Lightgbm	learning_rate = 0.04	learning_rate = 0.1
	n_estimators = 140	n_estimators = 85
	max_depth = 20	max_depth = 23
	num_leaves = 120	num_leaves = 120
Catboost	learning_rate = 0.1	learning_rate = 0.1
	depth = 6	depth = 7
	leaf_reg = 10	leaf_reg = 5
	iterations = 1000	iterations = 750
	one_hot_max_size = 3	one_hot_max_size = 10
GBDT	learning_rate = 0.08	learning_rate = 0.1
	max_depth = 3	max_depth = 9
	max_features = 11	max_features = 10
	n_estimators = 100	n_estimators = 35

be given attention, because these features may produce noise and decrease the performances of some classifiers.

D. PARAMETER OPTIMIZATION

After pruning the unimportant features, we apply the grid search method to find suitable values for the main parameters of each classifier. The other parameters are set to their default values. Table 9 presents the parameters and their appropriate values calculated by the grid search method.

After optimizing the main parameters, we train the classifiers with the retained features, and the results are demonstrated in Table 10. Comparing the performances before and after the parameter optimization step, the F1-scores improve for all the classifiers, and this reinforces the conclusion that parameter optimization can effectively improve the classification performance of a given classifier [21], [22]. Fig. 2 presents the comparison of the results obtained before and after optimizing the classifiers. From Fig. 2, we find that all the base classifiers are optimized effectively. Among all the classifiers, the F1-scores of the RF classifier are the highest after the optimization process on both datasets, and this proves that the proposed optimization process is effective for datasets with different scales. Therefore, the scalability of the optimization process is proven.

TABLE 10. F1-scores of 5 classifiers after parameter optimization.

Resampling method	Classifier	F1-score (%)	
		Hotel	Restaurant
RUS	RF	71.94	78.72
	Xgboost	71.21	76.76
	Lightgbm	71.68	78.09
	Catboost	71.26	78.10
	GBDT	71.78	76.59

We also find that compared with the results on the hotel dataset, the classifiers perform much better on the restaurant dataset. We believe the reason for this is that the restaurant dataset contains more data for training the model. Additionally, the ratio of fake reviews is higher.

E. CLASSIFIER ENSEMBLING

After data resampling, feature pruning, and parameter optimization, all the base classifiers are optimized. Thus, we can integrate the base classifiers using the majority voting and stacking strategies. According to Table 10, we can observe that the performances of Xgboost and GBDT are poor compared with those of other classifiers on the restaurant dataset. The poor performance of the base classifiers might influence the performance of the ensemble models. Thus, we also present the performance for each ensemble strategy, where only RF, Lightgbm, and Catboost are selected as the base classifiers. For the hotel dataset, the performances of ensemble models that only integrate RF, GBDT, and Lightgbm are also presented since they outperform the other two classifiers. Additionally, to prove that optimizing the base classifiers can effectively improve the performances of the ensemble models, the results of the ensemble models with nonoptimized base classifiers are also presented. The results are shown in Table 11. From Table 11 we can observe that all the ensemble models outperform the RF classifier, which performs best among the five base classifiers. We can also find that the stacking strategy is a better choice for integration than the majority voting strategy. The stacking model with RF, Lightgbm, and Catboost classifiers is the best choice for the optimal model. We also find that optimizing the base classifiers can significantly improve the performance of the ensemble models.

Furthermore, we find that the computational costs of the proposed ensemble models increase when all the base

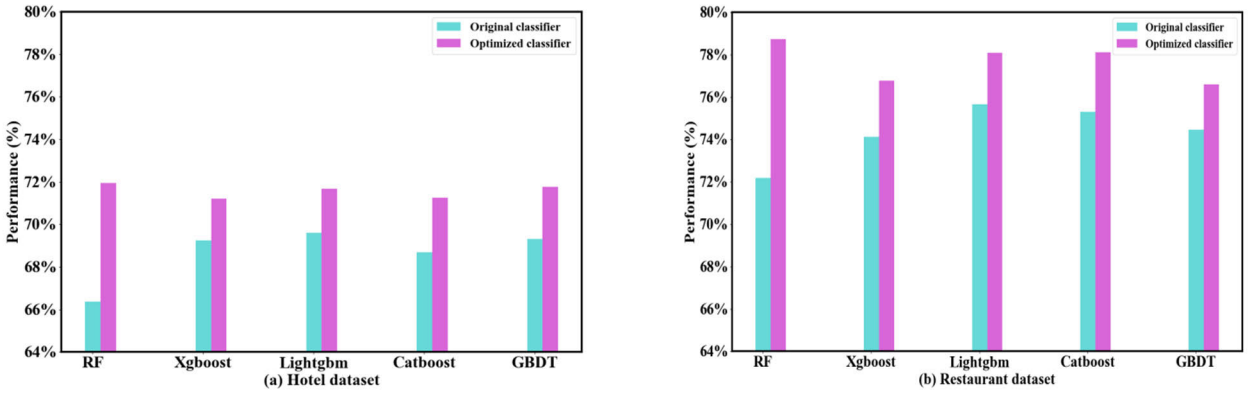


FIGURE 2. Performance comparison before and after optimizing the classifiers.

classifiers are integrated. However, the proposed model is less time consuming when only the best three classifiers are integrated. We also find that the majority voting strategy is much more efficient than the stacking strategy, while the performance of majority voting is very similar to that of stacking, thereby guaranteeing the feasibility of the model to be deployed in a real system. Overall, we can conclude that the ensemble strategy with optimized base classifiers is effective for building a highly robust and stable model.

F. MODEL COMPARISONS

To show the advantages of the proposed model, in this section, we compare the model with other approaches. Several popular supervised classifiers are selected and presented below since our model is also a supervised model. The experiment uses 10-fold cross validation to ensure the evaluation results.

(1) **RF + feature ranking** [3]: A random forest classifier using the varImp function to rank the importance of features. The top-twelve features are selected for training the classifier.
(2) **Fake Feature Framework (F3)** [18]: A framework using the Adaboost classifier that receives four types of user-centric features (personal, social, review activity, and trust features).
(3) **SVM (bigram + BF) [16]**: A support vector machine classifier with a linear kernel that receives both bigram and behavioral features.
(4) **CNN (convolutional neural network)**: a popular deep learning method that receives both review-centric and reviewer-centric features.
(5) **LSTM (long short term memory)**: a kind of recurrent neural network that receives both review-centric and reviewer-centric features.

For the first three models, we train them with a balanced dataset, making them the same as those proposed in their respective previous studies. The testing data remain imbalanced because we want to prove that the models trained with balanced training data are not effective for predicting imbalanced testing data. The evaluation is based on four performance measures: accuracy, precision, recall, and F1-score.

1) RESTAURANT DATASET

Fig 3 demonstrates the results of the models on the restaurant dataset and shows that the proposed model performs best in terms of the accuracy, recall, and F1-score measures. The accuracies of the six classifiers or models are quite similar. The LSTM performs best in precision but yields the lowest recall result. The CNN method performs poorly in both metrics, indicating that these two deep learning methods fluctuate on these measures. We think the reason for this is that deep learning methods need sufficient labeled data for training, which is hard to obtain due to the difficulty of manually identifying fake reviews [8]. For the F3 [18], RF + Feature ranking [3], and SVM (bigram + BF) [16] models, although researchers proposed them for balanced datasets and they performed well (F3 reached 81%, RF + Feature ranking reached 90%, and SVM (bigram + BF) reached 85.7% in terms of F1-scores on the balanced dataset), their performances decrease significantly on our imbalanced dataset, and they fluctuate greatly with respect to the precision and recall measures; this proves that a model trained with balanced data may not be effective in dealing with imbalanced samples. However, the proposed model can find a balance between precision and recall, therefore, the F1-score of the proposed model is the highest. This finding is reasonable because the proposed model applies a novel approach to address the data imbalance problem, and this guarantees the robustness of the model against imbalanced datasets.

2) HOTEL DATASET

We also compare the models on the hotel dataset, and the results are presented in Fig. 4. It can be observed that the proposed model can still find a compromise between precision and recall on a dataset with few training data. Therefore, the proposed model still obtains the highest F1-score. However, the traditional methods [3], [16], [18] fluctuate greatly on these metrics. We believe there are two major reasons for

TABLE 11. F1-scores of ensemble models.

Dataset	Ensemble Strategy	Base classifier	F1-score (%)		Training time (s)	
			Not optimized	Optimized	Not optimized	Optimized
Hotel	Majority voting	All	67.71	71.51	21.45	11.84
		RF + GBDT + Lightgbm	68.10	72.14	3.27	1.38
	Stacking	All	69.69	72.06	205.15	118.17
		RF + GBDT + Lightgbm	70.05	72.25	33.37	13.42
Restaurant	Majority voting	All	75.24	78.97	29.96	38.73
		RF + Lightgbm + Catboost	75.82	79.07	21.45	9.32
	Stacking	All	77.45	79.46	302.14	444.56
		RF + Lightgbm + Catboost	77.33	79.65	217.40	139.55

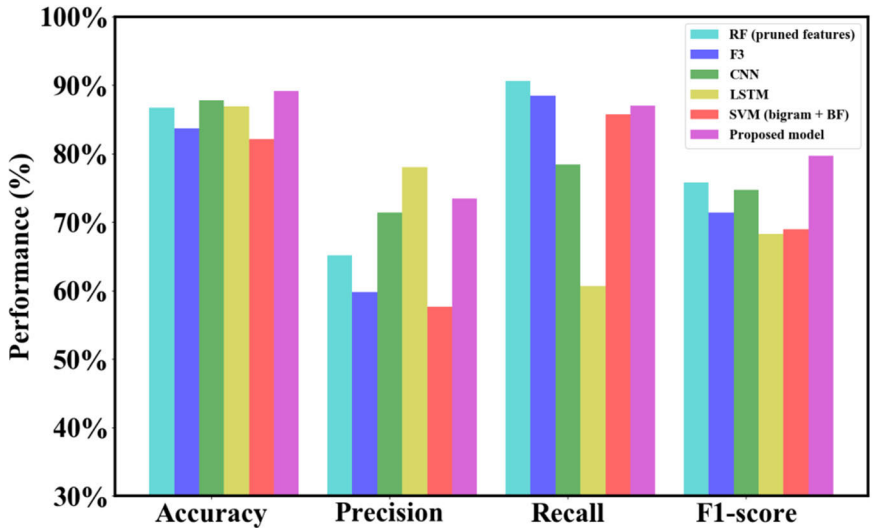


FIGURE 3. Model evaluations on the restaurant dataset.

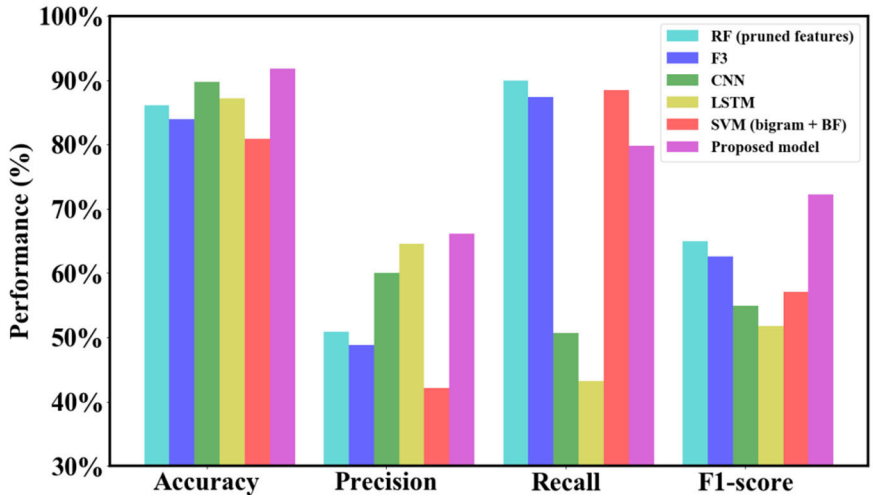


FIGURE 4. Model evaluations on the hotel dataset.

this phenomenon: first, the hotel dataset contains less data for training the model than the restaurant dataset; thus, the performances of all the models decrease significantly, especially those of traditional models that lack stability. Under such circumstances, the proposed ablation study can significantly eliminate the noise by pruning useless features and improve the performance, and this explains why the F1-scores of the models differ more greatly than those on restaurant dataset. Second, the hotel dataset is more imbalanced, which poses great challenges for a model. Traditional models are most likely to be influenced by imbalanced datasets. However, our model can still maintain a balance between these metrics.

Overall, the proposed model performs well in terms of four measures on the imbalanced datasets, thereby proving the feasibility of the proposed approach for addressing the data imbalance problem. Additionally, the performance of our model remains comparatively stable on datasets with different data distributions, while this is almost impossible for some traditional models or techniques. Furthermore, the model is proven to be effective on datasets with different scales; therefore, the scalability of the proposed model is also guaranteed. According to the above findings, the proposed model outperforms several other techniques or methods, and it thus provides a new option for fake review detection. The effectiveness of the proposed resampling approach and ablation study for feature pruning is also proven in the experiments. Moreover, two representative ensemble strategies are applied to enhance the robustness of the model, while this technique has been frequently ignored by previous studies. Overall, the study proves that the proposed model performs more stably and effectively than several traditional methods.

V. CONCLUSION

Identifying fake reviews is challenging for researchers, and there are several primary challenges for fake review detection. One problem is that fake reviews represent only a small proportion of the total reviews. Thus, the dataset is imbalanced, which may influence the performance of a model. In addition, feature selection for training machine learning classifiers is based on subjective prediction. Therefore, some of the features may not be useful for some classification methods.

The current study proposes an ensemble fake review detection model using for the detection of fake reviews. The model consists of four steps: data resampling, feature pruning, parameter optimization, and classifier ensembling. The first three steps are proposed to optimize the base classifiers. By optimizing the base classifiers, the performance of the ensemble model improves significantly on two representative Yelp datasets whose distributions are imbalanced; this performance is noticeably better than those of some traditional models and techniques. The study offers several implications from both theoretical and practical perspectives.

A. THEORETICAL IMPLICATIONS

The current study offers theoretical implications in several aspects. First, we explore a novel approach to address the data imbalance problem by combining data resampling and the grid search technique, which together can accurately find the best sampling ratio for data resampling and effectively improve the performance of each classifier. However, some studies [3], [16], [18] failed to consider the setting of the sampling ratio, and they simply rebalanced the distributions of samples. Therefore, the current findings provide a new way of dealing with imbalanced datasets.

Second, feature pruning has received little attention in the field of fake review detection. The current study offers insights into the pruning of individual features. Although there are various methods that can calculate the importance of features, such as the random forest and the VarImp function, they cannot calculate the actual performance change induced when a feature is dropped, and there is no definition that regulates which features can be regarded as useless. However, the proposed ablation study can determine the actual performance change, thereby significantly enhancing the robustness of the model. Ablation studies are usually proposed to evaluate a subset of a model or a method, and they have rarely been used for feature pruning. Therefore, our findings provide a novel method for feature pruning. Additionally, by pruning the unimportant features, the computational cost of the model is significantly decreased.

Third, the study proposes two representative ensemble strategies that can compensate for the weaknesses of the base classifiers. Different from those in basic ensemble models, the base classifiers in the proposed model are optimized, and this can significantly improve the performances of the ensemble models, as shown in Table 11. The proposed method for improving the performance of the ensemble strategy is applicable for almost all ensemble methods in any field. Overall, the proposed ensemble is proven to outperform several representative techniques.

B. PRACTICAL IMPLICATIONS

The findings also provide multiple practical implications for online websites. Considering the ubiquity of imbalanced data distributions, designers of fake review detection models should pay attention to data resampling techniques. It should be noted that it is not always useful to completely rebalance the distributions of truthful and fake reviews. In addition, for operators who are occupied with developing models for fake online review detection, the findings of our research indicate that incorporating features does not necessarily improve the classification performance of a given model. What truly matters is the relevance and usefulness of the chosen features. Hence, the pruning of features should not be despised as it helps to eliminate unimportant features, and the proposed ablation study is a simple yet effective method for feature pruning.

From the perspective of consumers, with the development of technology, online websites are actively working on exploiting their social functions. Consumers on websites are encouraged to interact with others, as this facilitates the detection of fake reviews. On the one hand, consumers are encouraged to write reviews on the products or services they have experienced. Although the findings suggest that features related to the textual contents of reviews are of little use, some other review-centric features, such as the number of ''useful'' votes a review received, are still effective. On the other hand, the exploitation of social functions enables consumers to respond to other users, such as being a follower of a user or giving a ''useful'' vote to a review, which produces abundant features for fake review detection. To make these features useful and relevant for fake review detection, consumers should be engaged in interacting with other reviewers on

websites, as spammers are unable to hide their footprints on such websites

C. LIMITATIONS AND FUTURE RESEARCH

The current study has some limitations. The dataset used in our study only contains data from two Yelp datasets. Another limitation is that the proposed ablation study is time-consuming when there are too many features.

For future research, we intend to evaluate the models with various datasets to analyze the robustness of the proposed model. Additionally, considering the difficulty of obtaining labeled fake review datasets, semi-supervised, and unsupervised learning methods will be investigated since they are understudied in the domain of fake review detection.

ACKNOWLEDGMENT

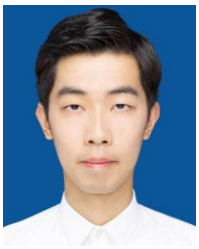
The authors express great thanks to the anonymous editors and reviewers of this paper for providing constructive comments. They also thank Lu Wang for revising the final version of the manuscript.

REFERENCES

[1] M. L. Jensen, J. M. Averbeck, Z. Zhang, and K. B. Wright, "Credibility of anonymous online product reviews: A language expectancy perspective," *J. Manage. Inf. Syst.*, vol. 30, no. 1, pp. 293–324, Jul. 2013.
[2] G. Cui, H.-K. Lui, and X. Guo, "The effect of online consumer reviews on new product sales," *Int. J. Electron. Commerce*, vol. 17, no. 1, pp. 39–58, Oct. 2012.
[3] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic, "What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews," *J. Manage. Inf. Syst.*, vol. 33, no. 2, pp. 456–481, Apr. 2016.
[4] A. U. Akram, H. U. Khan, S. Iqbal, T. Iqbal, E. U. Munir, and M. Shafi, "Finding rotten eggs: A review spam detection model using diverse feature sets," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 10, pp. 5120–5142, Oct. 2018.
[5] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," *Manage. Sci.*, vol. 62, no. 12, pp. 3412–3427, Dec. 2016.
[6] M. Ott, C. Cardie, and J. Hancock, "Estimating the prevalence of deception in online review communities," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, Lyon, France, Apr. 2012, pp. 201–210.
[7] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi, and X. Xiao, "Fake review detection based on multiple feature fusion and rolling collaborative training," *IEEE Access*, vol. 8, pp. 182625–182639, 2020.
[8] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol. (HLT)*. Portland, OR, USA: Association for Computational Linguistics, Jun. 2011, pp. 309–319.
[9] T. C. Alberto, J. V. Lochter, and T. A. Almeida, "Post or block? Advances in automatically filtering undesired comments," *J. Intell. Robot. Syst.*, vol. 80, no. S1, pp. 245–259, Dec. 2015.
[10] M. R. Martinez-Torres and S. L. Toral, "A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation," *Tourism Manage.*, vol. 75, pp. 393–403, Dec. 2019.
[11] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, pp. 27–34, Jul. 2015.
[12] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "Fake review detection: Classification and analysis of real and pseudo reviews," Univ. Illinois Chicago, Chicago, IL, USA, Tech. Rep. UIC-CS-2013-03, 2013.
[13] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli, "Support vector machines under adversarial label contamination," *Neurocomputing*, vol. 160, pp. 53–62, Jul. 2015.
[14] H. A. Najada and X. Zhu, "ISRD: Spam review detection with imbalanced data distributions," in *Proc. IEEE 15th Int. Conf. Inf. Reuse Integr. (IEEE IRI)*, Redwood City, CA, USA, Aug. 2014, pp. 553–560.
[15] E. F. Cardoso, R. M. Silva, and T. A. Almeida, "Towards automatic filtering of fake reviews," *Neurocomputing*, vol. 309, pp. 106–116, Oct. 2018.
[16] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing," in *Proc. 7th Int. Conf. Weblogs Social Media (ICWSM)*, Cambridge, MA, USA, Jul. 2013, pp. 409–418.
[17] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.
[18] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1234–1244, Jul. 2019.
[19] L. Chen, W. Li, H. Chen, and S. Geng, "Detection of fake reviews: Analysis of Sellers' manipulation behavior," *Sustainability*, vol. 11, no. 17, p. 4802, Sep. 2019.
[20] D. Liang, C.-F. Tsai, and H.-T. Wu, "The effect of feature selection on financial distress prediction," *Knowl.-Based Syst.*, vol. 73, pp. 289–297, Jan. 2015.
[21] Y. Li and Z. Yang, "Application of EOS-ELM with binary jaya-based feature selection to real-time transient stability assessment using PMU data," *IEEE Access*, vol. 5, pp. 23092–23101, 2017.
[22] H. Ullah, M. Uzair, A. Mahmood, M. Ullah, S. D. Khan, and F. A. Cheikh, "Internal emotion classification using EEG signal with sparse discriminative ensemble," *IEEE Access*, vol. 7, pp. 40144–40153, 2019.
[23] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technologie (HLT-NAACL)*, Atlanta, GA, USA, Jun. 2013, pp. 497–501.
[24] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Identify online store review spammers via social review graph," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1–21, Sep. 2012.
[25] D. H. Fusilier, M. Montes-y-Gómez, P. Rosso, and R. G. Cabrera, "Detecting positive and negative deceptive opinions using PU-learning," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 433–443, Jul. 2015.
[26] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network," *Inf. Process. Manage.*, vol. 54, no. 4, pp. 576–592, Jul. 2018.
[27] N. Ruan, R. Deng, and C. Su, "GADM: Manual fake review detection for O2O commercial platforms," *Comput. Secur.*, vol. 88, Jan. 2020, Art. no. 101657.
[28] L. Yijing, G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, "Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data," *Knowl.-Based Syst.*, vol. 94, pp. 88–104, Feb. 2016.
[29] C. G. Harris, "Detecting deceptive opinion spam using human computation," in *Proc. Workshops 26th AAAI Conf. Artif. Intell.*, Toronto, ON, Canada, Jul. 2012, pp. 1–7.
[30] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.
[31] H. Cao, V. Y. F. Tan, and J. Z. F. Pang, "A parsimonious mixture of Gaussian trees model for oversampling in imbalanced and multimodal time-series classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2226–2239, Dec. 2014.
[32] J. Zhai, S. Zhang, and C. Wang, "The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers," *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 3, pp. 1009–1017, Jun. 2017.
[33] I. Nekooeimehr and S. K. Lai-Yuen, "Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets," *Expert Syst. Appl.*, vol. 46, pp. 405–416, Mar. 2016.
[34] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, May 2015.
[35] A. D'Addabbo and R. Maglietta, "Parallel selective sampling method for imbalanced and large data classification," *Pattern Recognit. Lett.*, vol. 62, pp. 61–67, Sep. 2015.
[36] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*, vol. 135, pp. 32–41, Jul. 2014.
[37] L. Huang, C.-H. Tan, W. Ke, and K.-K. Wei, "Comprehension and assessment of product reviews: A review-product congruity proposition," *J. Manage. Inf. Syst.*, vol. 30, no. 3, pp. 311–343, Dec. 2013.

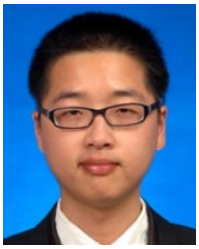
[38] B. Yee Liau and P. Pei Tan, "Gaining customer knowledge in low cost airlines through text mining," *Ind. Manage. Data Syst.*, vol. 114, no. 9, pp. 1344–1359, Oct. 2014.
[39] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.
[40] M. A. Tahir, J. Kittler, K. Mikolajczyk, and F. Yan, "A multiple expert approach to the class imbalance problem using inverse random under sampling," in *Proc. Int. Workshop Multiple Classifier Syst.* Berlin, Germany: Springer, 2009, pp. 82–91.
[41] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: A big comparison for NAS," 2019, *arXiv:1912.06059*. [Online]. Available: http://arxiv.org/abs/1912.06059
[42] S. Wei, D. Yang, W. Zhang, and S. Zhang, "A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning," *IEEE Access*, vol. 7, pp. 99217–99230, 2019.
[43] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992.
[44] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases," *Neurocomputing*, vol. 175, pp. 935–947, Jan. 2016.
[45] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *J. Intell. Inf. Syst.*, vol. 46, no. 3, pp. 563–597, Jun. 2016.

JIANRONG YAO is currently a Professor with the School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, China. He is also a Council Member of the China Management Science and Engineering Society. As an academic, he has been involved in research and teaching in various disciplines of management science and engineering, including artificial intelligence, data mining, and e-commerce. He has published articles in journals related to information security recently.



YUAN ZHENG received the B.S. degree in management from Zhejiang Gongshang University, Hangzhou, China. He is currently pursuing the master's degree with the School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou.

His major is management science and engineering. He is also working on online fake review detection using machine learning, deep learning and other computational intelligence methods. His research interests include fraud detection, especially fraud detection of online reviews using machine learning, deep learning, and other computational intelligence methods.



HUI JIANG received the B.S. and M.S. degrees from Harbin Engineering University, Harbin, China, in 2010, and the Ph.D. degree in management science and engineering from the Harbin Institute of Technology, Harbin.

He is currently an Assistant Professor with the Department of E-commerce, School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou, China. His research has been published in the *Journal of Information Science*, *Computers in Human Behavior*, and others. His research interests include fake review detecting, consumer creativity, and new product adoption. He serves as an Anonymous Reviewer for academic journals, such as the *Journal of Creative Behavior* (JCB) and *Creativity Research Journal* (CRJ).