Multimedia Tools and Applications (2024) 83:15591–15615
https://doi.org/10.1007/s11042-023-15173-8

Identifying fake job posting using selective features and resampling techniques

Hina Afzal[1] · Furqan Rustam[2] · Wajdi Aljedaani[3] · Muhammad Abubakar Siddique[4] · Saleem Ullah[1] · Imran Ashraf[5]

Received: 24 September 2021 / Revised: 22 March 2023 / Accepted: 27 March 2023 /
Published online: 15 July 2023

Abstract
The fake job posting has emerged as an alarming cyber-crime during the last few years which affects both job seekers and companies alike. Fraudulent companies and individuals lure job-seekers using multifarious methods on digital media platforms. Although several machine learning-based approaches exist for the automatic detection of fake job posts, they lack high accuracy and show skewed performance on imbalanced data. In addition, the influence of feature selection is not very well studied. This study overcomes these limitations using selective features through Chi-square and principal component analysis (PCA). The influence of dataset imbalance is also investigated through the synthetic minority oversampling technique (SMOTE). The performance of the proposed model is compared with individual machine learning models, as well as, existing state-of-the-art models. Results indicate that using SMOTE with Chi-square-based selective features yields the best results with a 0.99 accuracy using the proposed model. K-fold cross-validation further corroborates these results.

Keywords Fake job posts detection · Principal component analysis · Feature selection · Text classification

1 Introduction

The fake job posting has emerged as one of the major cyber-crimes which is drastically increased in recent years. Fake job posting affects job-seekers and companies and causes tension, depression, and frustration among individuals and a bad reputation for companies. Fraudulent individuals and companies attract job-seekers through different lucrative methods on digital media platforms. With the wide availability of online content, online job

Hina Afzal and Furqan Rustam have equal contributions.

✉ Furqan Rustam
furqan.rustam1@gmail.com
✉ Imran Ashraf
imranashraf@ynu.ac.kr

Extended author information available on the last page of the article.

posting and job application have become extremely easy for companies and candidates, respectively. However, With this evolution of the internet and the advancement of web technologies, anyone can post online without any restriction. Unlike traditional media with more diverse catalytic effects, social media has been used for posting both real and fake jobs [25].

During the past few years, numerous organizations opted for online platforms for a job posting and candidate selection due to the ease of the recruitment process, wide coverage of online media, and simple process of job application [10]. However, the online job posting is not without demerits and can be leveraged by scammers and fraudulent companies who take benefit from the cost of both individuals and companies. Especially, during the Covid-19 era when a large population experienced job loss, luring people for jobs has become an easy task [9]. This provides a great opportunity for scammers to extort money, as well as, the personal information of individuals that apply online. The content used to post the fake job is crafted well to make the information believable which either convinces the job seeker or makes them confused. Consequently, over the past few years, the number of victims of fake job postings has been increasing drastically, and fake job posting has become an issue of great concern.

Cybercrimes pose a real threat to the security of both individuals and organizations [37]. In 2012, about 600 resumes were received by a job seeker in a single day to recognize his competitor on Craigslist after posting a fraudulent job post. According to [44] reports that approximately 60% of online posted jobs are fake. However, only 48% of job seekers are mindful of these kinds of scams while looking for new opportunities. Furthermore, around 7% of online job application candidates have been victims of these scams even after knowledge of these scams. Similarly, according to CNBC, employment fraud in 2018 doubled as compared to 2017. So fake job posting has become a matter of significant importance and devising automatic models for identifying fake job posts is not a trivial job for several reasons. Text classification is a challenging task, from both theoretical and empirical points of view [2]. Often, classification has to be performed on a highly imbalanced dataset which makes this process difficult and accuracy degraded. Selecting the proper data structure for the representation of a document is yet another problem. Similarly, the selection of a suitable objective function to overcome the problem of overfitting due to imbalanced data poses an extra challenge. The objective function is also important regarding generalization. It is also important to deal with the high dimensionality of feature space to optimize the performance of a selected algorithm.

Machine learning models have proven their importance in several fields including image processing, prediction, text analysis, text classification, sentiment analysis, etc. [7, 11, 12, 22]. This study worked on the fake job posting detection from online platforms which is significant because it can prevent lots of scams and frauds from the internet world. Lots of scammers are looting money from people by posting fake job ads. We proposed an approach to detect fake job postings to avoid people from such scams. Our approach automatically predicts whether a post is real or fraudulent using its description and other features. The proposed approach works in three steps. First, feature engineering is performed to select the best features from input features. These important features help to train models effectively which increases the accuracy of models. Second, the problem of the imbalanced dataset is handled to reduce the model's overfitting problem. The target ratio is unequal in the dataset and we resolve this problem using SMOTE technique which generates data artificially to equal the number of samples of different classes. Third, an ensemble model is proposed to obtain higher performance for fake job postings. The ensemble model uses majority voting criteria. In a nutshell, this study makes the following main contributions.

- An ensemble model is proposed to obtain higher prediction accuracy for fake job postings from online platforms. For this purpose, three models logistic regression (LR), random forest (RF), and extra tree classifier (ETC) are combined under the hard voting criterion.
- The impact of dataset imbalance is investigated regarding the accuracy, precision, recall, and F1 score. In addition, the efficacy of using the synthetic minority oversampling technique (SMOTE) for data balance and its influence on classification accuracy is studied.
- The role of selective features is analyzed using extensive experiments without and with selective features with the Chi-square and principal component analysis (PCA) methods. Term frequency-inverse document frequency (TF-IDF) is utilized as the feature extraction approach for experiments.
- Performance analysis of the proposed model is carried out in comparison to several individual machine learning models including RF, ETC, LR, k-nearest neighbor (k-NN), and Naive Bayes (NB). In addition, performance is compared to existing state-of-the-art studies.

The rest of the paper is organized as follows. Section 2 describes research papers related to the current study. The proposed methodology, the dataset used for experiments, and machine learning classifiers are presented in Section 3. Results and discussions are provided in Section 4 while Section 5 concludes this study with limitations and probable future work.

2 Related work

Due to the importance of fake jobs post-identification, several studies have presented models based on machine learning. For example, study [45] analyzed an employment scam on the EMSCAD dataset by applying text mining and different machine learning algorithms such as ZeroR, OneR, NB, J48 DTs, RF, and LR and attained an accuracy of 90.6%. Similarly, the authors used J48 DT, JRip, and NB for the same task in [28]. A feature space is designed to utilize these machine learning algorithms to classify fake and real jobs. A well-structured feature space has different feature classes that make it easy and manageable and improves the accuracy, precision, and recall of the classification system. Study [13] introduced a semantic method using similarity measures and wordnet ontology to decrease the number of extracted features. For calculating the similarity measures, a path length measure has been chosen which reduces the time and space complexity. The proposed approach achieves an accuracy of 90% with the reduction of feature space.

An ensemble model is proposed in [41] that uses LR, SVM, and NB to make the final classification. These classifiers are combined into a complex model by considering the probability of each class from these classifiers. The ensemble model achieves an accuracy of 0.79 on the misogynistic tweets dataset. Along the same lines, [3] introduced an ensemble approach to detect online recruitment frauds. In the proposed model, SVM is used as a feature selection method while RF is trained on the extracted features. The ensemble classifiers show better performance than individual classifiers with an accuracy score of 0.9741. Similarly, [24] developed an ensemble of LR and RF for fake job post prediction. Experimental results on the EMSCAD dataset indicate 95.4% and 94.4% for accuracy and F1 score, respectively.

Machine learning models have been utilized for different applications and proved to show promising results. For example, [6] uses CNN and LSTM for citywide traffic prediction.

Ali et al. [4] adds spatiotemporal patterns to enhance traffic flow prediction using neural networks. Similarly, [5, 8] also employs hybrid machine learning approaches and obtains better results. The performance of different machine learning classifiers is investigated in [16] to detect fake job posts. The k-NN, NB, multilayer perceptron, DT, RF, AdaBoost, and gradient boosting are tested with the highest accuracy of 98.27%. For the fake job text classification problem, [29] used different classifiers such as k-NN, DT, SVM, RF, and MNB. The reported accuracy is 98.2% which is achieved when TF-IDF is used for feature extraction. A recent study [17] used an ensembled approach for fake news classification. By applying DT, ETC, and RF classifiers on the liar and ISOT datasets, testing accuracy of 100% and 44.15% are achieved, respectively. The study [40] used Microsoft azure machine learning studio and designed a model in which two-class boosted DT and two-class decision forest algorithms were used. Results show that the two-class boosted DT performed better as compared to the other algorithms concerning the accuracy, precision, recall, and F1 score.

Besides machine learning classifiers, deep learning models have also been utilized such as [36]. The study used deep learning, as well as, machine learning approaches for detecting fake news on Facebook. A bidirectional GRU and bidirectional LSTM are used which yields an accuracy of 99.4%. Deep learning models show better results than machine learning models. In the same way, the authors propose a fake news detection model in [27]. The authors show that the sentiments of tweets are important for the identification of legitimate and fake news. The model comprises natural language processing, machine learning, and deep learning approaches along with Apache spark. Using RF for the said task, a 79% accuracy can be achieved.

The study [10] proposed an approach for fake job posting prediction. They used machine learning techniques and methods. They used TF-IDF features with the ADASYN re-sampling technique to achieve significant 99.9% accuracy using the ETC model. The study [15] proposed an approach for fake job posting using SVM models and TF-IDF features. They achieved a significant 97.6 accuracy score. Another study [42], also proposed an approach for the classification of fake job postings and they deployed state of the arts artificial neural networks (ANN) and random forest (RF). RF achieved a significant 95.2% accuracy score in study [42].

Despite the results reported in the above-discussed research works, these approaches lack in several aspects. First of all, several of the approaches where high accuracy is reported to use highly imbalanced datasets such as [3, 24]. As a result, there is a high difference in the reported accuracy and F1 scores which shows the model overfitting. The results of these approaches can not be generalized. Secondly, many approaches focus on fake news detection which may and may not contain fake job news. Thirdly, the impact of using selective features for fake jobs posts identification is not been extensively studied. Similarly, the performance of many well-known machine learning classifiers is not studied very well and the performance of ensemble and individual classifiers is not compared. The feature engineering part is also not investigated well. Empirical results from the existing literature indicate the superior performance of ensemble models for text classification. Keeping in view these results, this study considers the ensemble model for fake job post-detection. However, in addition, we adopt the use of SMOTE which balances the dataset and reduces the probability of model overfitting. In this regard, this study utilizes both individual and ensemble classifiers with and without selective features to analyze their performance for the task at hand. A summary of the discussed literature on fake job postings is presented in Table 1.

Table 1 Summary of the systematic analysis studies in related work

Reference	Models	Dataset	Results	Pros	Cons
Vidros et al. [45]	ZeroR, OneR, NB, J48 DTs, RF, LR	EMSCAD dataset, EMSCAD dataset	F1-score: ZeroR:0.33, OneR:0.84, NB:0.88, J48:0.90, RF:0.906, LR: 0.906	Low computational cost because of less complex models	Low Accuracy
Mahbub and Pardede [28]	J48 DT, JRip, NB	EMSCAD dataset	Accuracy:J48: 94.29%, JRip:96.19%, NB:83.42%	Low Computational Cost	Low Accuracy
Bahgat iet al. [13]	NB, SVM, LR, J48, RF, RBF network	Enron-Spam dataset	Accuracy:NB:94%, SVM:94%, LR:0.95%, RF:92%, RBF:93%	Good use of state of the arts techniques	Imbalanced dataset
Shushkevich and Cardiff [41]	LR, NB, SVM, blended model	3251 tweets in English	blended model: 0.79%	Ensemble learning to improve performance	High computational Cost and Small dataset.
Alghamdi et al. [3]	RF	EMSCAD dataset	Accuracy:97.41%	Low computational cost	Highly imbalanced dataset
Lal et al. [24]	J48, LR, RF, ensemble	Fake job	Accuracy: 95.4%, F1-score:94.4%	Ensemble learning to reduce over-fitting chances	Imbalanced dataset
Nasser and Alzaanin [29]	MNB, SVM, DT, KNN, RF	Fake job	Accuracy:MNB: 95.6%, SVM: 97.7%, DT:97.4%, KNN:97.8%, RF:98.2%	Low computational cost because of simple techniques	Imbalanced dataset
Hakak et al. [17]	DT, RF, ETC	ISOT and Liar datasets	Accuracy on ISOT:44.15%, Liar:100%	Low computational cost because of simple techniques	Imbalanced dataset & Low performance score

Table 1 (continued)

Reference	Models	Dataset	Results	Pros	Cons
Sahoo and Gupta [36]	KNN, SVM, LR, DT, NB, LSTM	Fake news on Facebook	Accuracy on LSTM:99.4%, KNN: 99.3%, SVM:99.3%, LR:99.0%, DT:99.1%, NB:98.6%	High performing because of significant approach	High computational cost
Ablel-Rheem et al. [1]	NB, DT, ensemble, hybrid ensemble	UCI based spam dataset	Accuracy on: NB:79.28%, DT:92.79%, ensemble90.06%, hybrid ensemble:94.41%	Ensemble learning to improve the performance	High computational cost as compared to the state of the art models
Shibly et al. [40]	Two class decision boosted tree, two-class decision forest algorithms	Fake job	accuracy 93.8%, F1-score: 0.73%	Low computational cost	Model over-fitting
Madani et al. [27]	LR, DT, RF, NB, GBT, SVM, MLP	Covid-19 dataset	Accuracy on RF: 79%	Simple techniques	low performance by used models
Amaar et al. [10]	ETC, LR, SVM, CNN, LSTM, GRU	Fake Job	ETC: 99.9%	Low computational cost and no over-fitting	No validation, only training set resampling
Chiraratanasopha and Chay-intr [15]	SVM, TF-IDF	Fake Job	SVM : 97.6%	Simple approach	Low accuracy

3 Material and methods

3.1 Data collection

Experiments are performed on the 'Fake JobPosting' dataset that contains 17,880 job posts from which 17,014 posts are real while the rest 866 are fake (dataset is available at [14]. Every entry is represented as a group of structured and unstructured data. Figure 1 shows the distribution of records for real and fake jobs which indicates that the dataset is highly imbalanced. There are four types of fields in our dataset which are listed below in Table 2.

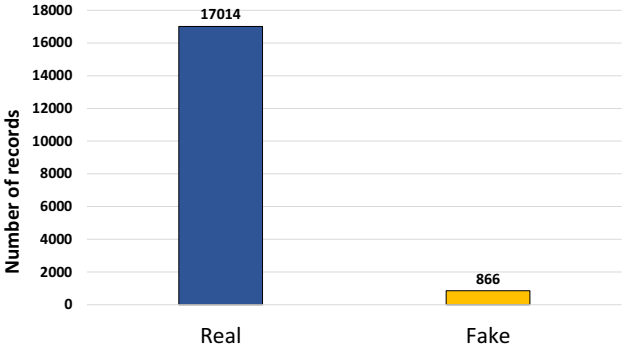

Fig. 1 Distribution of records for real and fake job posts

Table 2 List of field types in a fake job posting dataset

Type	Name	Description
String	Title	Job ad entry-title.
	Location	Geographical location.
	Salary range	Suggested salary like $40,000-$50,000
	Department	Job related departments like telecommunication.
HTML fragments	Company profile	Company's brief description.
	Description	Details of the job ad.
	Requirements	Requirement list for job.
	Benefits	Enlisted Benefits offered by the company.
Binary	Company logo	True if the company logo exists.
	Questions	True for screening questions.
	Telecommunication	True for telecom. Positions.
	Fraudulent	Classification attributes.
Nominal	Required education	Doctorate, Bachelor, Master's Degree, etc.
	Required experience	Executive, Entry level, Intern, etc.
	Employment type	Full-type, Part-time, Contract, etc.
	Function	Function Consulting, Engineering, Research, Sales, etc.
	Industry	Industry Automotive, IT, Health care, Real estate, etc.

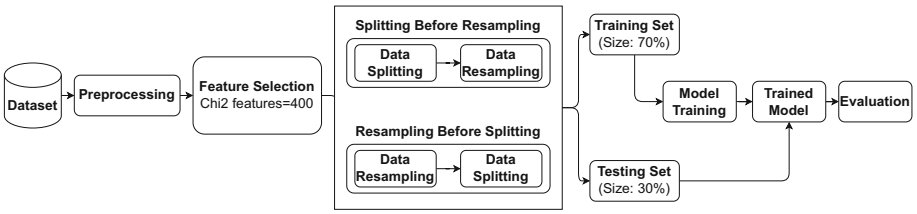

Fig. 2 The architecture of the proposed framework

3.2 Proposed framework

The main purpose of the proposed methodology is to predict fake job posts using machine learning algorithms with dataset balancing and selective features. The architecture of the proposed model is shown in Fig. 2. A preprocessing phase is carried out before feature selection with Chi-2. For training and testing, two courses of action are followed where 'splitting before sampling' and 'sampling before splitting' are performed. The objective of this strategy is to analyze the influence of using SMOTE before and after the train-test split. The dataset contains data related to two classes; 'real' job posts and 'fake' job posts where the 'real' class has a higher number of records which makes the dataset highly imbalanced. Machine learning algorithms tend to fit toward the majority class which leads to overfitting. Consequently, for the minor class, the number of correct predictions is low than in the major class. To resolve this issue, data balance is one of the potential approaches used today. Several approaches are available for data resampling; this study utilizes SMOTE. The steps of the proposed framework are discussed in the following sections.

3.3 Preprocessing

Data preprocessing is the first step in fake jobs posts prediction. In this step, the text is cleaned with meaningless information that is not suitable for training the classification algorithms [35]. It involves removing special characters, numbers, multiple spaces, punctuation marks, and non-English words. Similarly, stop word removal is also part of preprocessing. Stemming and converting to lowercase techniques are applied to remove uncertainty in the feature set

- Removing special characters, numbers, and punctuation: In this step remove all kinds of characters numbers, and punctuation marks using the regular expression in Python language. These features are not important in text data to train machine learning models [10].
- Removing Stopwords: Stopwords are part of the text to make the text significant but didn't contain too much information so are not good to use for machine learning model training. We remove stopwords using the NLTK library.
- Concert to lowercase: This step will convert each character into lowercase which will help to reduce the feature set size and improve the feature weightage. For explain, Go and go are the same words but because of differences in case but are separate features with low weightage. Conversion to the lowercase will convert both of them to go so both can be taken as a single feature with more weightage [33].

– Stemming: Stemming convert each word in the text to its root form so uncertainty in the feature set can be controlled. For example, the words, like go and gone will be converted into its root for go. We have done stemming using the PorterStemmer library [20].

3.4 Feature extraction

We used the TF-IDF feature extraction technique and feature selection techniques including Chi2 and PCA. TF-IDF is the predominantly used feature extraction technique for text analysis. TF-IDF maps text data into a numerical form which can be directly fed into machine learning models for training. All features are not equally important and the use of selective features often proves more fruitful to obtain high performance. For feature selection, this study adopts Chi2 and PCA which are the most widely used approaches for the text analysis domain. Chi2 and PCA are used to select the most important features from the data. First, the TF-IDF features are extracted, and later Chi2 and PCA are applied to select the most important features for training machine learning models, as shown in Fig. 3.

3.4.1 TF-IDF

TF-IDF is one of the most commonly used approaches for text classification. It converts the entire document into an appropriate representation of weights [46]. For calculated TF-IDF for a word in a document, two different metrics are multiplied including TF and IDF.

TF is an estimation of how often a term occurs in a given document. The main purpose for using it is that the words that occur repeatedly are more significant than the words that rarely occur. TF is calculated as

$$tf(t, d) = \frac{\text{count of } t \text{ in } d}{\text{no. of words in } d} \tag{1}$$

where t, and d shows the term(word) and document for which tf is calculated.

IDF is the calculation of how significant a term is by taking the total number of documents and dividing it by the number of documents in which the term occurs. Contrary to

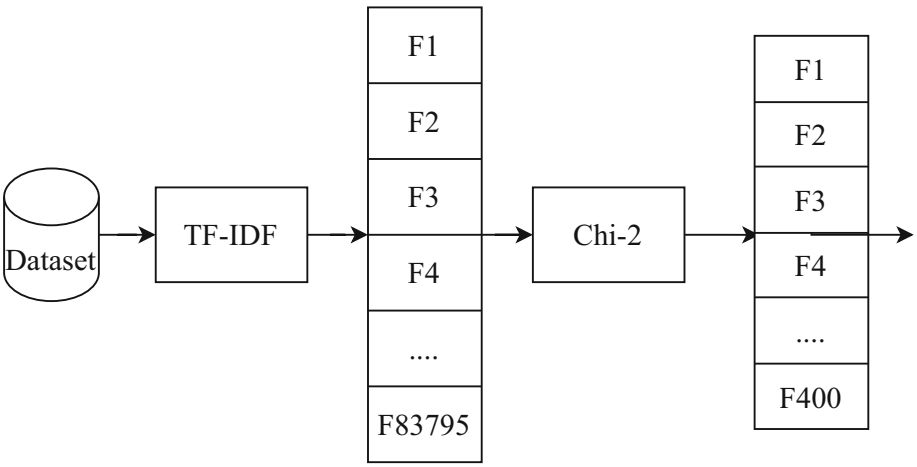

Fig. 3 Feature extraction and feature selection

TF which gives higher weight to the most frequent words, IDF considers rare words more important. It can be calculated using

$$idf(t) = \frac{N}{df} \tag{2}$$

where N shows the count of the corpus, while df is the document frequency of term t and it shows how many times t has appeared in the set of documents.

Using TF and IDF, TF-IDF can be computed

$$TF - IDF(t, d) = tf(t, d) \times log(\frac{N}{df + 1}) \tag{3}$$

Contrary to TF-IDF, BoW is a simple yet effective feature technique. Also known as count-vector, it transforms arbitrary text into fixed-length vectors that contain words and their occurrence. For making vectors, it determines a vocabulary where all words that appeared in the text are gathered. In the vector, each word and its frequency are counted. BoW maintains the order for words' appearance in the text [49].

3.4.2 Chi-2

The objective of feature selection is to choose an optimal set of features from the original dataset to achieve optimal results. It is used to remove features that do not contribute or have smaller contributions in the classification process. By eliminating the redundant and irrelevant/less contributing features, it reduces the dimensionality of the feature vector. It also avoids overfitting and helps in reducing storage requirements. Generally, the feature selection method consists of four steps; feature subset generation, subset evaluation, stopping criterion, and result from validation. This study uses Ch2 for feature selection.

Chi-2 is one of the most popular and frequently used filter-based feature selection methods. It is a supervised method used for selecting features from a categorical dataset. It ranks the features concerning their importance which is further used to reduce the number of features by removing the features with low ranks.

In statistics, the Chi-square test is used to examine the independence of two events. The events are supposed to be independent if:

$$p(X, Y) = p(X)p(Y) \tag{4}$$

In-text feature selection, these two events are correlated with the occurrence of a specific term and a class, respectively. Chi-2 can be calculated using

$$Chi - 2(t, C) = \sum_{t \in 0,1} \sum_{C \in 0,1} \frac{(N_t, C - E_t, C)^2}{(E_t, C)} \tag{5}$$

Here, E represents the expected frequency and N is the observed frequency for both C and t. Chi-2 is an evaluation of how much-observed count N and expected count deviate from each other. The high value of Chi-2 shows that the hypothesis of independence is not correct. For individual classes, the Chi-2 value of a term is calculated which is then globalized all over the classes in two different ways. The first way is to calculate the weighted average score and the second way is to select the highest score between all classes. In this work, a prior approach is selected to globalize the Chi-2 value for all classes.

$$\sum P(C_i).Chi - 2(t, C_i) \tag{6}$$

3.5 Synthetic minority oversampling technique

SMOTE is an oversampling technique, used to increase the number of minority class samples by randomly copying the samples of the minority to balance the minority and majority classes. SMOTE is considered one of the efficient approaches that are widely applied. It produces a synthetic training model for minority class using linear interpolation [47]. For each minority class sample, k number of nearest neighbors is selected randomly and synthetic training models are generated. The data is reconstructed after the process of oversampling which can be later used with different classification techniques. SMOTE follows these steps to oversample data.

Step 1: From the dataset, the total number of majority (i.e., 'Real' class of the dataset) and minority (i.e., 'Fake' class of the dataset) are captured, respectively. Then the threshold value d^{th} is presented for the maximum degree of class imbalance. The total number of synthetic samples to be generated is, $G = (Real - Fake) \times \beta$, where $\beta = (Fake/Real)$.

Step 2: For every minority sample x_i, k-NNs are obtained using Euclidean distance. Next ratio r_i is calculated using $\Delta i/k$ which is normalized using $r_x \leq r_i / \sum r_i$.

Step 3: Later, the total synthetic samples for each x_i are generated using $g_i = r_x \times G$. The process is iterated from 1 to g_i to generate samples.

We used the re-sampling technique to make the dataset balanced because imbalanced data models can be overfitted on majority class data which can lead to wrong predictions. So we used SMOTE to reduce this over-fitting problem.

3.6 Machine learning models

This section describes the machine learning classification algorithms which are used for fake jobs post-identification. For this purpose, five classifiers RF, ETC, LR, k-NN, and NB, and one hybrid classifier is used in this study (Table 3). A brief description of machine learning classifiers is given in Table 4. We used machine learning models with several hyper-parameters settings. These hyper-parameters settings we find using the grid search method as we tune models between a specific range. All hyper-parameters and their tuning range are shown in Table 3.

3.7 Ensemble classifier

Finally, the above-mentioned classifiers are combined in a hybrid classifier i.e., classification is not based on a single classifier but based on the decision of multiple classifiers (Fig. 4). Hybrid models have a significant contribution to automated predictions [4, 32]. By using

Table 3 Machine learning models hyperparameter settings

Model	Hyperparameters	Tuning range
ETC	n_estimators=300, max_depth=300	n_estimators= {10 to 500}, max_depth={50 to 500}
RF	n_estimators=300, max_depth=300	n_estimators={10 to 500}, max_depth={50 to 500}
LR	Solver=liblinear, C=3.0, multi_class=ovr	Solver={liblinear,saga}, C={1.0 to 5.0}, multi_class=ovr
NB	Default Setting	Default Setting
k-NN	n_neighbor=5	n_neighbor={1 to 10}
Ensemble	Voting=Hard, Models=ETC, RF, LR	Voting=Hard, Soft Models=ETC, RF, LR

Table 4 Brief description of machine learning models

Model	Description
Random Forest	RF is a supervised learning classifier. It is used for classification and regression problems [35]. RF performs well for lots of text classification tasks according to literature [33, 39]. For model calibration, the bagging technique is used in which training data is frequently used as a bootstrap sample. RF creates multiple decision trees and finally joins them to achieve higher accuracy. Because of its flexibility and diversity, it is one of the most commonly used algorithms today [31].
Extra Tree Classifier	ETC combines the results of many de-correlated decision trees. Generally, it uses the initial training set to construct the decision tree. At each node, k-features are randomly picked to make decision trees. The Gini index is used to select the important features which help to determine tree split criteria. Various de-correlated decision trees are constructed based on these selected features. ETC controls the overfitting and improves the accuracy [21]. ETC is used by previous studies to achieve significant results for predictions task as study [35] recently used for sentiment prediction and achieved the highest accuracy.
Logistic Regression	Logistic means logit function which is a probabilistic function having 0 or 1 value. LR is a statistical technique used for predicting categorical variables. It is used to anticipate the classes which have dichotomous nature. Generally, multiple feature values are used to calculate the probability of a binary response. It is generally used in regression and classification problems [43]. LR is a linear model and considers a favorite for binary classification problems as in our study.
k-Nearest Neighbor	k-NN is an important and popular machine learning approach that is used for classification and regression problems. The k-NN focuses on keeping similar data nearest to each other. Here the number of neighbors is represented by k. The k-NN approach depends on the similarity learning that is involved in many text classifications and data analytic approaches [26]. KNN is less complex and performs very well for prediction and detection tasks when the feature set will small.
Naive Bayes	NB classifier is a popular statistical technique that belongs to the group of probabilistic classifiers which follow the Bayes theorem. It has the basic assumption that all features are independent of each other (i.e., no words are associated with each other). NB classifiers vary mainly by the assumption they make about the distribution of features. Training data relies on a limited number of data points. Datasets of high dimensionality tend to show better performance with NB classifiers [48].

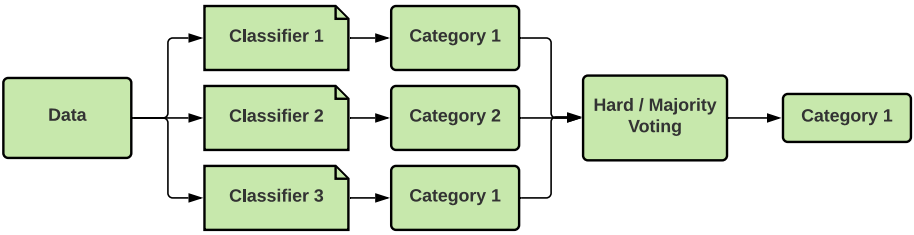

Fig. 4 Hard voting method used in the hybrid model

this approach, we can eliminate the weakness of different classifiers while utilizing their strengths in classification. Typically, this provides better accuracy and robust classification performance. In the current work, the classification is done using hard voting (majority vote), in which each classifier had one vote. The implementation of all classifiers is done using the Sci-kit library [19].

In this study, we combined the three best performers of this study which are LR, RF, and ETC under majority voting criteria. The combination of two tree-based models and a linear model LR helps to create a strong ensemble model. We used hard voting criteria (majority voting) which makes a final prediction by taking votes from each model.

3.8 Performance measures

To evaluate the performance of machine learning models, accuracy, recall, F1-score, and precision are used and their brief description is provided in Table 5. To calculate these measures confusion matrix is used which consists of four terms that are explained in Table 6.

True positive (TP) are observations that are classified correctly in the positive class, and true negative (TN) are observations that are classified correctly in the negative class. False-positive (FP) refers to samples of negative class incorrectly classified as positive while a false negative (FN) indicates samples of positive class incorrectly classified as negative.

4 Results and discussion

This section contains the results and discussion of machine learning classifiers. Experiments are performed on Jupyter notebook with Python using pandas, NumPy, and Scikit-learn libraries.

4.1 Results on original dataset

Initially, machine learning classifiers are applied to the original dataset without oversampling. ETC, RF, and hybrid models performed way better than other models. With ETC, the highest accuracy of 97.33% is achieved while RF and Hybrid models achieved 97.22% and 97.14% accuracy, respectively. In terms of F1-score, ETC achieved 0.81 while RF and hybrid model achieved an F1-score of 0.80 each as shown in Table 7. Despite the high classification accuracy for ETC, RF, and hybrid models, accuracy is no longer a proper evaluation metric. Due to the highly imbalanced dataset, machine learning models show a biased attitude towards the samples of the majority class and experience an overfit. Consequently, a large gap between theaccuracy and F1 score is observed.

Table 5 Summary of performance measures, definitions, and formulas

Measures	Definition	Formula
Accuracy	For text classification, the best and most widely used classification measure is accuracy. It is considered a good criterion for classification, it is always desired to be higher [38].	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	Recall generally refers to the number of correct predictions divided by the number of discarded predictions. It is also known as sensitivity [23].	$\frac{TP}{TP+FN}$
Precision	Precision estimates the ratio of correct positive values over the total number of positive values [18].	$\frac{TP}{TP+FP}$
F1-score	F1-score is considered to be another important metric for both precision and recall and it evaluates the harmonic mean of recall and precision [30].	$2 \times \frac{Precision \times Recall}{Precision+Recall}$
AUC-ROC	AUC refers to the area under the ROC curve. It is the measure of the absolute two-dimensional area under the entire ROC curve from (0,0) to (1,1). It is considered a benefit criterion. AUC is a diverse measure of classification performance. The higher the accuracy of AUC, the better the performance of model [47].	

Table 6 Definition of TP, FP, TN, and FN in a confusion matrix

	Predicted As Positive	Predicted As Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Table 7 Machine learning performance on the original dataset

Models	Accuracy	Precision	Recall	F1	AUC
ETC	97.33	0.99	0.74	0.81	0.73
RF	97.22	0.99	0.73	0.80	0.72
LR	96.14	0.91	0.64	0.70	0.64
NB	86.18	0.56	0.65	0.58	0.64
k-NN	97.09	0.88	0.79	0.83	0.78
Ensemble	97.14	0.99	0.72	0.80	0.71

The confusion matrix for classifiers is given in Table 8 which shows the number of correct predictions (CP) and wrong predictions (WP), in addition to, TP, TN, FP, and FN. ETC, RF, and hybrid models each have the highest number of TP, i.e., 5,093 while the highest TN is from ETC. As a result, ETC has the highest CP which is 5,221, and the lowest number of WP which is 143. NB shows the worst performance with only 4,623 CP and 741 WP. The performance of the hybrid model is 2nd to ETC with 5,211 CP and 153 WP.

4.2 Performance of classifiers on original dataset with Chi-2

In this approach, machine learning classifiers are trained on selective features using Chi-2. For this purpose, the 400 best features from the data are selected to train the classifiers. Results show that the performance of machine learning models has slightly improved concerning accuracy and F1-score. ETC achieved the highest accuracy of 97.78% which is slightly better than its 97.33% without Chi-2. On the other hand, its F1 score has substantially improved from 0.81 to 0.85 when used with Chi-2 features as shown in Table 9.

The confusion matrix given in Table 10 indicates that the number of CP has been improved as well. ETC has 5245 correct predictions than 5221 without Chi-2 and the number of wrong predictions has been reduced to 119 from 143. Similarly, the performance of other classifiers has been enhanced. The worst performance is still by NB with 724 wrong predictions.

Table 8 Confusion matrix for original dataset

Models	TP	TN	FP	FN	CP	WP
ETC	5,093	128	143	0	5,221	143
RF	5,093	122	149	0	5,125	149
LR	5,080	77	194	13	5,157	207
NB	4,512	111	160	581	4,623	741
k-NN	5,049	159	112	44	5,208	156
Ensemble	5,093	118	153	0	5,211	153

Table 9 Performance of machine learning classifiers using Chi-2 on the original dataset

Models	Accuracy	Precision	Recall	F1	AUC
ETC	97.78	0.99	0.78	0.85	0.78
RF	97.61	0.99	0.76	0.84	0.76
LR	96.43	0.93	0.67	0.74	0.66
NB	86.50	0.57	0.65	0.58	0.65
k-NN	97.11	0.88	0.79	0.83	0.78
Ensemble	97.44	0.99	0.75	0.82	0.74

Table 10 Confusion matrix of machine learning classifiers using Chi-2 on the original dataset

Models	TP	TN	FP	FN	CP	WP
ETC	5,093	128	143	0	5,221	143
RF	5,093	122	149	0	5,125	149
LR	5,080	77	194	13	5,157	207
NB	4,512	111	160	581	4,623	741
k-NN	5,049	159	112	44	5,208	156
Ensemble	5,093	118	153	0	5,211	153

4.3 Results using resampled dataset before splitting

In this approach, an oversampling technique SMOTE is used before splitting the data into training and testing. It increases the sample size which helps to improve the learning performance of the models. The results shown in Table 11 suggest that the ETC, RF, and hybrid model has significant performance improvement with the highest accuracy of 99.53%, 99.44%, and 98.88%, respectively. F1-score reaches 1 with ETC and 0.99 with RF and hybrid model.

From the above-mentioned results and the confusion matrix, as shown in Table 12, it can be seen that using the proposed framework, ETC performs best when trained on re-sampled data. Out of 10209 total predictions, ETC makes the highest number of correct predictions, i.e., 10162, and the lowest number of wrong predictions at only 47. RF also performs well with 10152 correct and 57 wrong predictions.

Table 11 Results of machine learning classifiers on the resampled dataset before splitting

Models	Accuracy	Precision	Recall	F1	AUC
ETC	0.9953	1.00	1.00	1.00	0.99
RF	0.9944	0.99	0.99	0.99	0.99
LR	0.8380	0.84	0.84	0.84	0.83
NB	0.7051	0.71	0.71	0.71	0.70
k-NN	0.9591	0.96	0.96	0.96	0.95
Ensemble	0.9888	0.99	0.99	0.99	0.98

Table 12 Confusion matrix for resampled dataset before splitting

Models	TP	TN	FP	FN	CP	WP
ETC	5,072	5090	14	33	10,162	47
RF	5,074	5078	26	31	10,152	57
LR	4,259	4297	807	846	8,556	1653
NB	3,660	3539	1565	1445	7,199	3010
k-NN	4,713	5079	25	392	9,792	417
Ensemble	5,006	5089	15	99	10,095	114

4.4 Performance using Chi-2 on re-sampled dataset before splitting

Additional experiments are performed using Chi-2 on the re-sampled dataset to improve the performance of the models. For this purpose, the best 400 features are selected, then data is split into training and testing. Experimental results are given in Table 13. Results indicate that the performance of the machine learning classifiers has been slightly improved. The accuracy of ETC increases from 99.53% to 99.76%. Similarly, the performance of other classifiers has been improved as well. The ensemble model achieved significant accuracy with the approach which is 0.999. This significant result is just of ensemble architecture in which tree-based and linear models perform significantly in combination with each other.

The confusion matrix for these experiments is given in Table 14 which shows that the number of wrong predictions has decreased when the classifiers are trained on Chi-2 selected features. For example, the number of wrong predictions for ETC, RF, NB, LR, k-NN, and the hybrid model has been reduced from 47, 57, 1653, 3010, 417, and 114 to 24, 34, 1144, 2276, 387, and 18, respectively which shows the significance of Chi-2 feature selection method. However, these experiments are performed with data resampling before the train-test data split. As a result, it involves the risk of having similar data samples in both training and testing datasets which overstate the accuracy of classifiers. Because of this opinion, additional experiments are performed where the resampling is performed after the dataset is split into training and test sets.

4.5 Results for re-sampled dataset after train-test split

Following this approach, first, the dataset is divided into training and test sets. Later, SMOTE is applied only to the training data samples while the testing dataset comprises the original data samples. Table 15 summarizes the results using the above-stated approach. Results indicate that the classification accuracy of the machine learning model is degraded with this approach. ETC achieves the highest accuracy of 97.87%, followed by RF and hybrid models with accuracy scores of 97.74% and 96.70%, respectively.

Table 13 Results of machine learning classifiers using Chi-2 on the resampled dataset before splitting

Models	Accuracy	Precision	Recall	F1	AUC
ETC	0.9976	1.00	1.00	1.00	0.99
RF	0.9966	1.00	1.00	1.00	0.99
LR	0.8879	0.89	0.89	0.89	0.88
NB	0.7770	0.78	0.78	0.78	0.77
k-NN	0.9621	0.96	0.96	0.96	0.96
Ensemble	0.9990	0.99	0.99	0.99	0.99

Table 14 Confusion matrix for Chi-2 on resampled dataset before splitting

Models	TP	TN	FP	FN	CP	WP
ETC	5,087	5,098	06	18	10,185	24
RF	5,090	5,085	19	15	10,175	34
LR	4,505	4,560	544	600	9,065	1144
NB	4,059	3,874	1230	1046	7,933	2276
k-NN	4,730	5,092	12	375	9,822	387
Ensemble	5,096	5,095	09	09	10,191	18

Table 15 Results of machine learning classifiers on resampled dataset after splitting

Models	Accuracy	Precision	Recall	F1	AUC
ETC	0.9787	0.94	0.83	0.87	0.82
RF	0.9774	0.95	0.80	0.86	0.80
LR	0.8199	0.59	0.82	0.61	0.81
NB	0.6866	0.54	0.68	0.49	0.68
k-NN	0.9209	0.68	0.87	0.73	0.87
Ensemble	0.9670	0.82	0.85	0.83	0.84

The confusion matrix shown in Table 16 suggests that the number of wrong predictions has been increased for all the classifiers using this approach. Despite that ETC, RF, and hybrid model has the lowest number of wrong predictions with 114, 121, and 177 WP, respectively.

4.6 Performance of classifiers using Chi-2 on re-sampled dataset after splitting

Table 17 shows the results obtained by using Chi-2 with resampling after the train-test split. This technique allows for avoiding possible over-fitting while selecting the important features and also helps to achieve the best accuracy. Consequently, a smaller gap between accuracy and F1-score is seen in Table 17. ETC achieves the highest accuracy of 98.41%, as well as, the highest F1 score of 0.91. Similarly, RF and hybrid model achieves 2nd and 3rd best accuracy with scores of 98.32% and 97.46%, respectively.

The confusion matrix shown in Table 18 shows that ETC yields the highest number of correct predictions which are 5279 out of 5364 with only 85 wrong predictions. The worst performance is by NB with 4143 correct and 1221 wrong predictions.

We used the PCA feature selection technique in comparison with Chi-2. The performance of models is also good with PCA as well as Chi-2. There is little difference in

Table 16 Confusion matrix for resampled dataset after splitting

Models	TP	TN	FP	FN	CP	WP
ETC	5,072	178	93	21	5,250	114
RF	5,077	166	105	16	5,243	121
LR	4,177	221	50	916	4,398	966
NB	3,498	185	86	1595	3,683	1681
k-NN	4,717	223	48	376	4,940	424
Ensemble	4,994	193	78	99	5,187	177

Table 17 Results of machine learning classifiers using Chi-2 on resampled dataset after splitting

Models	Accuracy	Precision	Recall	F1	AUC
ETC	0.9841	0.97	0.86	0.91	0.85
RF	0.9832	0.97	0.85	0.90	0.85
LR	0.8711	0.63	0.88	0.67	0.87
NB	0.7723	0.57	0.77	0.56	0.77
k-NN	0.9652	0.69	0.88	0.75	0.88
Ensemble	0.9746	0.86	0.89	0.87	0.89

Table 18 Confusion matrix for Chi-2 on resampled dataset after splitting

Models	TP	TN	FP	FN	CP	WP
ETC	5,086	193	78	07	5,279	85
RF	5,082	192	79	11	5,274	90
LR	4,433	240	31	660	4,673	691
NB	3,935	208	63	1158	4,143	1221
k-NN	4,736	227	44	375	4,963	419
Ensemble	5,010	218	53	83	5,228	136

Table 19 Results using PCA-based feature selection

Approach	Model	Accuracy	Precision	Recall	F1 Score
PCA + Resampling	ETC	0.990	0.99	0.99	0.99
	RF	0.984	0.98	0.98	0.98
	LR	0.980	0.98	0.98	0.98
	NB	0.930	0.93	0.93	0.93
	k-NN	0.960	0.96	0.96	0.96
	HYBRID	0.992	0.99	0.99	0.99
PCA	ETC	0.970	0.99	0.73	0.81
	RF	0.970	0.99	0.70	0.77
	LR	0.970	0.84	0.80	0.82
	NB	0.970	0.95	0.68	0.74
	k-NN	0.970	0.82	0.94	0.87
	HYBRID	0.970	0.84	0.83	0.82

both approaches as in some cases Chi-2 outperforms and with some models, PCA helps to achieve significant accuracy but has the highest accuracy with the Chi-2 approach which is 0.999. The results of machine learning models using PCA as feature selection are shown in Table 19.

4.7 Results using different data splitting ratios

We have done experiments with different splitting ratios to show the significance of the proposed approach. We perform experiments with 80:20 and 90:10 ratios. These experiments

Table 20 Results using 80:20 ratio

Model	Accuracy	Precision	Recall	F1 Score
ETC	0.992	0.99	0.99	0.99
RF	0.990	0.99	0.99	0.99
LR	0.980	0.98	0.98	0.98
NB	0.942	0.94	0.94	0.94
k-NN	0.941	0.94	0.94	0.94
HYBRID	0.994	0.99	0.99	0.99

Table 21 Results using 90:10 ratio

Model	Accuracy	Precision	Recall	F1 Score
ETC	0.996	0.99	0.99	0.99
RF	0.991	0.99	0.99	0.99
LR	0.978	0.98	0.98	0.98
NB	0.942	0.94	0.94	0.94
k-NN	0.941	0.94	0.94	0.94
HYBRID	0.996	0.99	0.99	0.99

we only analyzed these with our proposed approach (resampling and Chi-2). Tables 20 and 21 show the results of machine learning models with 80:20 and 90:10 splitting ratios respectively.

We observe that as we increase the training set size there is little increase in accuracy scores. The HYBRID model achieved a significant accuracy score of 0.994 with 80:20 splitting ratios. While with a 90:10 splitting ratio, we have 0.996 accuracy which is more compared to others, this significant improvement in accuracy is because of an increase in the training set.

4.8 Results of 10-fold cross-validation

We also performed K-Fold cross-validation. The results of K fold cross validation are also significant as well as the models with train testing methods. HYBRID models achieved a significant mean accuracy score of 0.99 with a +/-0.01 standard deviation with the proposed approach. This high accuracy with 10-fold cross-validation shows the significance of our proposed approach. Similarly ETC is also significant with a 0.99 means accuracy score with the proposed approach. Table 22 shows the results of 10-fold cross-validation and according to the results, we can see that only oversampling is not enough to improve the performance of learning models as feature selection can also help to improve the accuracy. That's we combine both Chi-2 and resampling techniques to achieve significant results.

4.9 Comparison with existing studies

Table 23 shows the comparison between the current study and other state-of-the-art approaches. We have done a comparison with those studies that utilized the same dataset to evaluate the proposed approach. We find that the majority of the studies did not work on data balancing which leads models to over-fitting towards the majority class. Our previous study [10] used the data re-sampling technique, however, two important aspects were

Table 22 Models results using 10-fold cross-validation

Model	Original	Chi-2	Oversampled	Oversampled + Chi-2
ETC	0.96(+/-0.01)	0.98(+/-0.02)	0.98(+/-0.01)	0.99(+/-0.01)
RF	0.96(+/-0.01)	0.98(+/-0.01)	0.97(+/-0.03)	0.99(+/-0.02)
LR	0.96(+/-0.02)	0.94(+/-0.04)	0.98(+/-0.02)	0.99(+/-0.02)
NB	0.91(+/-0.04)	0.89(+/-0.01)	0.81(+/-0.03)	0.89(+/-0.01)
k-NN	0.95(+/-0.03)	0.94(+/-0.01)	0.85(+/-0.02)	0.95(+/-0.01)
HYBRID	0.96(+/-0.00)	0.99(+/-0.02)	0.98(+/-0.01)	0.99(+/-0.01)

Table 23 Comparison with other state-of-the-art studies

Ref.	Year	Model	Resampling	Results
Lal et al. [24]	2019	Ensemble Model	-	Accuracy: 95.4%
Nasser and Alzaanin et al. [29]	2020	RF	-	Accuracy: 98.2%
Shibly et al. [40]	2021	Two-class decision forest	-	Accuracy 93.8%, F1-score: 73%
Amaar et al. [10]	2022	ETC	ADASYN (Before Data Splitting)	Accuracy: 99.9%, F1 Score: 99%
Chiraratanasopha and Chay-intre [15]	2022	SVM	-	Accuracy : 97.6%
Our	2022	HYBRID	SMOTE (Before Data Splitting)	Accuracy: 99.9%, F1 Score: 99%
	2022	ETC	SMOTE (After Data Splitting)	Accuracy: 98.4%, F1 Score: 85%

ignored. First, the authors did not investigate the role of feature selection. Second, the performance is only validated when data is split after data re-sampling which can cause data leakage. In the current approach, we have done both experiments. Table 23 shows the comparison results of the current study with existing studies which shows that the proposed ensemble model shows better results.

4.10 Statistical significance test

The proposed approach's significance is evaluated using the statistical T-test. T-test compared the proposed approach results with other state-of-the-art techniques results as shown in Table 24. T-test rejects or accepts the null hypothesis (N_h) in output if the compared results are not statistically significant it accepts the N_h and if they are statically significant then rejects the N_h and accepts the alternative hypothesis (A_h) [34]. In most of the cases, the P-value score is less that the alpha value (0.05) which shows that the compared results are statistically significant. When we compared our proposed approach results with the results on the original dataset T-test reject N_h in all cases which means the proposed approach accepted A_h and show statistical significance.

Table 24 Results for statistical T-test

Scenarios		Statistic	P-Value	N_h
Results using Chi-2 on resampled dataset before splitting	Ensemble Vs KNN	17.02	1.43 e^{-07}	Reject
	Ensemble Vs NB	70.99	1.72 $e^{-}12$	Reject
	Ensemble Vs LR	39.45	1.87 e^{-10}	Reject
	Ensemble Vs RF	2.02	0.07	Accept
	Ensemble Vs ETC	2.11	0.06	Accept
Results using Chi-2 on resampled dataset before splitting Vs Results using Resampled Dataset Before Splitting	Ensemble Vs KNN	12.77	1.32 e^{-06}	Reject
	Ensemble Vs NB	105.84	7.09 e^{-14}	Reject
	Ensemble Vs LR	58.28	8.34 e^{-12}	Reject
	Ensemble Vs LR	0.50	0.62	Accept
	Ensemble Vs LR	1.96	0.08	Accept
Results using Chi-2 on resampled dataset before splitting Vs Results on Original Dataset	Ensemble Vs KNN	4.07	0.003	Reject
	Ensemble Vs NB	6.25	0.002	Reject
	Ensemble Vs LR	3.22	0.012	Reject
	Ensemble Vs LR	2.57	0.031	Reject
	Ensemble Vs LR	2.57	0.032	Reject

5 Conclusion and future work

This study proposes a framework to evaluate the performance of machine learning classifiers, both individual, as well as, ensemble models to identify fake job posts. Two problems are investigated in essence including data imbalance and the use of selective features on the performance of classifiers. Several experiments are performed with the original dataset with all features and Chi-2 selected features, SMOTE resampled dataset with resampling before and after the train-test split, and resampled dataset with all and Chi-2 selected features. ETC, RF, and the ensemble model achieve the best performance regarding the accuracy, precision, recall, and F1 score. Results indicate that with the imbalanced dataset, a higher gap between accuracy and F1 score exists which shows models' overfitting. Data balancing with SMOTE tend to solve this problem and elevates the performance of the classifiers. Similarly, the use of selective features with Chi-2 slightly improves the performance, and the number of wrong predictions is reduced. The sequence of data resampling also influences the performance where the resampling before the train-test split shows the highest accuracy of 99.53% without Chi-2 features and 99.76% when ETC is used with Chi-2 features. On the other hand, resampling after the train-test split and only on the training test shows the highest accuracy of 97.87% without Chi-2 features and 98.41% when ETC is trained on Chi-2 selected features. Experiments are performed using TF-IDF, Chi-2, and SMOTE only, for feature extraction, feature selection, and resampling, we intend to utilize more approaches in the future for performance analysis. The use of deep learning approaches is also under consideration.

Funding This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2021-2016-0-00313) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

Data Availability The dataset used in this study is available at https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Ablel-Rheem DM, Ibrahim AO, Kasim S, Almazroi AA, Ismail MA (2020) Hybrid feature selection and ensemble learning method for spam email classification. Int J 9(1):4
2. Agarwal B, Mittal N (2014) Text classification using machine learning methods-a survey. In: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28–30, 2012. Springer, pp 701–709
3. Alghamdi B, Alharby F et al (2019) An intelligent model for online recruitment fraud detection. J Inf Secur 10(03):155
4. Ali A, Zhu Y, Chen Q, Yu J, Cai H (2019) Leveraging spatio-temporal patterns for predicting citywide traffic crowd flows using deep hybrid neural networks. In: 2019 IEEE 25th international conference on parallel and distributed systems (ICPADS). IEEE, pp 125–132
5. Ali A, Zhu Y, Zakarya M (2021) Exploiting dynamic spatio-temporal correlations for citywide traffic flow prediction using attention based neural networks. Inf Sci 577:852–870
6. Ali A, Zhu Y, Zakarya M (2021) A data aggregation based approach to exploit dynamic spatio-temporal correlations for citywide crowd flows prediction in fog computing. Multimed Tools Appl, pp 1–33
7. Ali A, Zhu Y, Zakarya M (2021) Exploiting dynamic spatio-temporal correlations for citywide traffic flow prediction using attention based neural networks. Inform Sci 577:852–870. Available Online: https://www.sciencedirect.com/science/article/pii/S0020025521008483
8. Ali A, Zhu Y, Zakarya M (2022) Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. Neural Netw 145:233–247
9. Allas T, Canal M, Hunt V (2020) Covid-19 in the united kingdom: Assessing jobs at risk and the impact on people and places. McKinsey and Company Article, vol. 11
10. Amaar A, Aljedaani W, Rustam F, Ullah S, Rupapara V, Ludi S (2022) Detection of fake job postings by utilizing machine learning and natural language processing approaches. Neural Process Lett 54(3):2219–2247
11. Ashraf I, Hur S, Shafiq M, Kumari S, Park Y (2019) Guide: Smartphone sensors-based pedestrian indoor localization with heterogeneous devices. Int J Commun Syst 32(15):e4062
12. Ashraf I, Hur S, Shafiq M, Park Y (2019) Floor identification using magnetic field data with smartphone sensors. Sensors 19(11):2538
13. Bahgat EM, Rady S, Gad W, Moawad IF (2018) Efficient email classification approach based on semantic methods. Ain Shams Eng J 9(4):3259–3269
14. Bansal S (2020) Real or fake fake jobposting prediction, https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction
15. Chiraratanasopha B, Chay-intr T (2022) Detecting fraud job recruitment using features reflecting from real-world knowledge of fraud, Current Applied Science And Technology, pp 12
16. Dutta S, Bandyopadhyay SK (2020) Fake job recruitment detection using machine learning approach. Inter J Eng Trends Technol 68(4):48–53
17. Hakak S, Alazab M, Khan S, Gadekallu TR, Maddikunta PKR, Khan WZ (2021) An ensemble machine learning approach through effective feature extraction to classify fake news. Futur Gener Comput Syst 117:47–58
18. Hossin M, Sulaiman M (2015) A review on evaluation metrics for data classification evaluations. Int J Data Mining Knowl Manag Process 5(2):1
19. Hug N (2020) Surprise: a python library for recommender systems. J Open Source Softw 5(52):2174
20. Jamil R, Ashraf I, Rustam F, Saad E, Mehmood A, Choi GS (2021) Detecting sarcasm in multi-domain datasets using convolutional neural networks and long short term memory network model. PeerJ Comput Sci 7:e645

21. Kaur K, Mittal S (2020) Classification of mammography image with cnn-rnn based semantic features and extra tree classifier approach using lstm, Materials Today: Proceedings
22. Khalid M, Ashraf I, Mehmood A, Ullah S, Ahmad M, Choi GS (2020) Gbsvm: Sentiment classification from unstructured reviews using ensemble classifier. Appl Sci 10(8):2788
23. Kynkäänniemi T, Karras T, Laine S, Lehtinen J, Aila T (2019) Improved precision and recall metric for assessing generative models, arXiv preprint arXiv: 1904.06991
24. Lal S, Jiaswal R, Sardana N, Verma A, Kaur A, Mourya R (2019) Orfdetector: ensemble learning based online recruitment fraud detection. In: 2019 Twelfth International Conference on Contemporary Computing (IC3). IEEE, pp 1–5
25. Liu B, Fraustino JD, Jin Y (2013) Social media use during disasters: A nationally representative field experiment, College Park, MD. Tech. Rep
26. Luo X (2021) Efficient english text classification using selected machine learning techniques. Alexandria Eng J 60(3):3401–3409
27. Madani Y, Erritali M, Bouikhalene B (2021) Using artificial intelligence techniques for detecting covid-19 epidemic fake news in moroccan tweets. Results Phys 25:104266
28. Mahbub S, Pardede E (2018) Using contextual features for online recruitment fraud detection
29. Nasser I, Alzaanin AH (2020) Machine learning and job posting classification: a comparative study. In: International Journal of Engineering and Information Systems (IJEAIS) ISSN, pp 6–14
30. Novaković JD, Veljović A, Ilić SS, Papić Ž, Milica T (2017) Evaluation of classification models in machine learning. Theory Appl Math Comput Sci 7(1):39–46
31. Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens 67:93–104
32. Rupapara V, Rustam F, Aljedaani W, Shahzad HF, Lee E, Ashraf I (2022) Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model. Sci Rep 12(1):1000
33. Rupapara V, Rustam F, Shahzad HF, Mehmood A, Ashraf I, Choi GS (2021) Impact of smote on imbalanced text features for toxic comments classification using rvvc model. IEEE Access 9:78 621–78 634
34. Rustam F, Ashraf I, Mehmood A, Ullah S, Choi GS (2019) Tweets classification on the base of sentiments for us airline companies. Entropy 21(11):1078. Available Online: https://www.mdpi.com/1099-4300/21/11/1078
35. Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS (2021) A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis. Plos one 16(2):e0245909
36. Sahoo SR, Gupta BB (2021) Multiple features based approach for automatic fake news detection on social networks using deep learning. Appl Soft Comput 100:106983
37. Scanlon JR, Gerber MS (2014) Automatic detection of cyber-recruitment by violent extremists. Secur Inform 3(1):1–10
38. Sebastiani F (2002) Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1):1–47
39. Shah K, Patel H, Sanghvi D, Shah M (2020) A comparative analysis of logistic regression, random forest and knn models for the text classification. Augmented Human Res 5(1):1–16
40. Shibly F, Uzzal S, Naleer H (2021) Performance comparison of two class boosted decision tree snd two class decision forest algorithms in predicting fake job postings
41. Shushkevich E, Cardiff J (2018) Classifying misogynistic tweets using a blended model: The ami shared task in ibereval 2018. In: IberEval@ SEPLN, pp 255–259
42. Srivastava R (2022) Identification of online recruitment fraud (orf) through predictive models. Emirati Journal of Business Economics and Social Studies
43. Sur P, Candès EJ (2019) A modern maximum-likelihood theory for high-dimensional logistic regression. Proc Nat Aca Sci 116(29):14516–14525
44. Vidros S, Kolias C, Kambourakis G (2016) Online recruitment services: Another playground for fraudsters. Comput Fraud Secur 2016(3):8–13
45. Vidros S, Kolias C, Kambourakis G, Akoglu L (2017) Automatic detection of online recruitment frauds: characteristics, methods, and a public dataset. Future Intern 9(1):6
46. Wu HC, Luk RWP, Wong KF, Kwok KL (2008) Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS) 26(3):1–37
47. Xie W, Liang G, Dong Z, Tan B, Zhang B (2019) An improved oversampling algorithm based on the samples^{TM} selection strategy for classifying imbalanced data. Math Problems Eng 2019:1–13
48. Xu S (2018) Bayesian naïve bayes classifiers to text classification. J Inf Sci 44(1):48–59
49. Zhang Y, Jin R, Zhou Z-H (2010) Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybern 1(1-4):43–52

Affiliations

Hina Afzal[1] · Furqan Rustam[2] · Wajdi Aljedaani[3] · Muhammad Abubakar Siddique[4] · Saleem Ullah[1] · Imran Ashraf[5]

Hina Afzal
hinaafzal770@gmail.com

Wajdi Aljedaani
wajdialjedaani@my.unt.edu

Muhammad Abubakar Siddique
abubakar.ahmadani@gmail.com

Saleem Ullah
saleem.ullah@kfueit.edu.pk

1 Khwaja Fareed University of Engineering and Information Technology, Abu Dhabi Rd, Rahim Yar Khan, Punjab, Pakistan

2 School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland

3 University of North Texas, 155 Union Cir, Denton, TX 76203 USA

4 Department of Computer Science & IT, Ghazi University, Dera Ghazi Khan, Punjab, Pakistan

5 Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea