IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

A DEEP LEARNING-BASED APPROACH TO ONLINE RECRUITMENT FRAUD DETECTION

[1] *D SHARATH KUMAR, MCA Student, Department of MCA*

[2] *Dr. C. GULZAR ,Ph.D, Associate Professor,Department of MCA*

[12]*Dr KV Subba Reddy Institute of Technology,Dupadu, Kurnool*

ABSTRACT

These days, the majority of businesses use digital platforms to seek new hires in order to streamline the recruiting process. Fraudulent advertising is a consequence of the sharp rise in the usage of online job posting platforms. The fraudsters use phoney job advertisements to get revenue. Fraud in online hiring has become a significant problem in cybercrime. Therefore, to eliminate online job frauds, it is essential to identify phoney job ads. The goal of this research is to employ two transformer-based deep learning models, namely Bidirectional Encoder Representations from Transformers and Robustly Optimised BERT-Pretraining Approach (RoBERTa), to accurately detect fake job postings. Traditional machine learning and deep learning algorithms have been used in recent studies to detect fake job postings. By combining job ads from three distinct sources, a unique dataset of fraudulent job advertisements is suggested in this study. The effectiveness of current algorithms to identify fake jobs is hampered by the outdated and restricted benchmark datasets, which are based on knowledge of particular job advertisements. We thus update it with the most recent job openings. The class imbalance issue in identifying phoney employment is brought to light by exploratory data analysis (EDA), which causes the model to behave aggressively against the minority class. The work at hand employs 10 of the best Synthetic Minority Oversampling Technique (SMOTE) variations in order to address this issue. Analysis and comparison are done between the models' performances balanced by each SMOTE version. Every strategy that is used is carried out in a competitive manner. However, with a balanced accuracy and recall of almost 90%, BERT+SMOBD SMOTE produced the best results.

I. INTRODUCTION

The internet has fundamentally changed our lives in a variety of ways in this era of sophisticated technology. Nowadays, doing any task the old-fashioned manner has been replaced by the internet. As a result, employment and job searching have also moved online. Productivity, ease of use, and effectiveness are the advantages of an online recruiting system, sometimes known as e-recruitment [1]. To offer job openings to prospective employees, the majority of companies choose online recruiting platforms [2]. Through employment portals, companies post job openings and provide job descriptions, including prerequisites, compensation packages, offers, and facilities to be given. Job searchers go to several online job boards, look for openings that fit their interests, and apply for positions that fit. After that, the business reviews resumes to make sure they meet its needs. After completing further procedures, such as interviewing and choosing possible applicants, the post is closed. During the worldwide COVID-19 epidemic, the tendency of posting job ads online was exaggerated. The World Economic Outlook Report states that the International Monetary Fund (IMF) calculated that at the height of the COVID-19 epidemic in 2020, the jobless rate rose to 13%. In 2018 and 2019, these figures were only 3.9% and 7.3%, respectively. Many businesses made the decision to advertise job positions online during the epidemic in order to accommodate job seekers [3]. However, when a resource is made available

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

to the general population, it also gives internet scammers the opportunity to exploit their gloom.

One of the major issues in the field of online recruitment fraud (ORF) is an employment scam. Online recruiting systems are advantageous to both recruiters and job searchers, but if they are not used properly, they may also be harmful to both parties. For job searchers, it is unlucky since they risk losing their money, privacy, or even their existing position. Furthermore, by damaging reputable organisations' reputations in the employment market, scammers indirectly undermine their trustworthiness [4]. Sophisticated techniques are being used by the scammers to trick consumers, making it very difficult for them to tell the difference between genuine and fraudulent job postings. About 52% of the candidates did not know about ORFs, while the remaining candidates knew just a little bit about them, according to a Flex Jobs study [5]. More than 67% of individuals are now interested in searching for a job online, according to a recent poll conducted by Action Fraud [6]. However, they must be mindful of the rise in employment fraud.

To find ORF, several investigations were carried out. The authors of [7] and [8] classified job posts as either fraudulent or non-fraudulent using conventional machine learning techniques. To increase classification accuracy, the works [9] and [10] used ensemble-based machine learning approaches. The authors in [11] used an Artificial Neural Network (ANN) based model for classification after first performing downsampling to address the imbalance issue. After oversampling the data, the authors in [12] used Random Forest (RF) to increase accuracy after extracting features using TF-IDF. In order to test context-based behavioural traits on traditional machine learning algorithms and get predictions, researchers in [13] constructed their own dataset. Examining the underlying paper reveals that several machine learning techniques have been used to ORF identification. In contrast to traditional machine learning algorithms, transformer-based deep learning techniques are becoming more popular these days. Nevertheless, advanced deep-learning approaches for ORF detection have not yet been fully investigated to address this issue. Therefore, by using transformer-based deep learning models, this study attempts to identify ORF and analyse and warn individuals about the increasingly expanding job frauds. As a result, consumers would no longer be duped by employment frauds. Therefore, folks who are squandering their time and money on such fraudulent operations might be more cautious by spotting ORF.

In this study, we introduced a new dataset of fictitious job advertisements that were classified as "fraudulent" for fictitious job advertisements and "non-fraudulent" for genuine ones. Three distinct sources of job ads are combined to provide the suggested data. To expand the dataset with the most recent job posts, we utilise "Fake Job Postings" as the core dataset and include publicly accessible job ads from Pakistan and the US. We took this action as the benchmark datasets now in use are out-of-date and constrained by the knowledge of particular job ads, which reduces the effectiveness of current algorithms to identify fake employment. The dataset was prepared, and then it was subjected to exploratory data analysis, or EDA. The dataset's unbalanced class distribution was discovered by EDA. The ratio of samples in the minority class to those in the majority class is known as the imbalance class distribution [14]. For regular courses, it may result in high prediction accuracy; for rare classes, it might result in poor predictive accuracy. Anomaly detection [15], facial recognition [16], medical diagnosis [17], text classification [18], and many other real-world areas are affected by the class imbalance issue. SMOTE [19], an oversampling method, became widely used. In order to address

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

class imbalance issues in a variety of fields, researchers have lately used over 85 distinct SMOTE variations that have been described in the literature.

Investigating Online Recruitment Fraud (ORF) and resolving any potential system implementation problems are the goals of this study. The following are some of this study's noteworthy contributions:

• A unique dataset is created by combining job ads from three distinct sources.

• Exploratory Data Analysis (EDA) reveals a severely unbalanced class distribution in the dataset that was gathered. To balance the class distribution ratio, ten of the best SMOTE variations are used.

• To determine if a job ad is fake or not, transformer-based deep learning models are applied to the dataset.

• Both balanced and unbalanced datasets are used for comparative examination of implemented models.

The following parts make up the remainder of the paper: A thorough summary of previous research on the underlying topic is included in Section II. Features of the dataset, suggested approach, and framework of applied models are presented in Section III. In Section IV, the essential findings of the research are discussed and the experimental results are shown. Section V wraps up the research that was given, outlining its shortcomings and offering some suggestions for the future.

II. LITERATURE SURVEY

"E-Recruitment: A Conceptual Analysis,"

Kaur, P.

E-Recruitment, often known as online recruitment, is the process of hiring people online that enables businesses to reach a wide workforce and quickly locate qualified candidates using web-based resources and technology. S.L. Lakshmi (2013). Following COVID-19, a lot of businesses are simplifying their hiring and selection procedures by utilising technology such as video conferencing, chatbots, mobile applications, the internet, and computer-based tests, among other things. This allows candidates to be matched with open positions. Recruiters and businesses are increasingly embracing online social networking to draw in and evaluate applicants as part of the employment process, according to a number of studies. Nickolas Ollington et al. (2013). The report demonstrates how e-recruitment, aided by technology, has improved organisations' ability to save critical time and money while also making their hiring process easier and more efficient. The research aims to comprehend the practices, advantages, and difficulties of e-recruitment.

"Detection and analysis of fake jobs using deep learning and machine learning algorithms,"

G. A. Sairam, P. Ganesh, G. Deepakkumar, P. Nagarajan, and C. S. Anita

The number of online jobs offered on numerous employment sites has significantly increased due to the pandemic crisis. However, some of the jobs that are advertised online are phoney, which makes it possible for important and sensitive information to be stolen. Therefore, by using sophisticated deep learning and machine learning classification algorithms, these phoney jobs may be accurately identified and categorised from a pool of job postings of both phoney and legitimate employment. In order to identify and distinguish between phoney and legitimate employment, this article use machine learning and deep learning techniques. In order to ensure that the classification method used is very exact and precise, this research also suggests data cleaning and analysis. It should be mentioned that the data cleaning phase is a crucial part of any machine learning project as it really affects how accurate the deep learning and machine learning algorithms are. Therefore, this study places a lot of emphasis on the data cleaning and pre-processing steps. It is possible to classify and identify phoney jobs with great precision and

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

accuracy. Therefore, in order to improve accuracy, machine learning and deep learning algorithms must be used on cleaned and pre-processed data. In order to get more accuracy, deep learning neural networks are also used. In order to choose the classification method with the best accuracy and precision, all of these models are finally compared to one another.

"Predicting fake e-job postings using advanced machine learning techniques,"

F. Younas, F. Akhtar, S. Ubaid, and A. Raza,

Even on trustworthy job-posting websites, there are plenty of job advertisements that never seem to be fraudulent. But after the selection, the so-called recruiters start looking for bank account details and money. Many applicants fall into these traps, losing both their current employment and a significant amount of money. Determining whether a job posting posted on the website is authentic or fraudulent is therefore preferable. It is very difficult, if not impossible, to recognise manually! To get rid of fake job advertisements on the internet, an automated online solution (website) based on machine learning-based classification and algorithms is offered. Among the many job advertisements on the internet, it helps identify fraudulent ones.

According to a survey, more millennials than elderly fall prey to employment scams.

Howington, J.

Among the most prevalent frauds in America are mass marketing scams, which also include postal scams. According to an increasing amount of research, age-related impairments in cognitive functioning and social isolation put older persons at higher risk of victimisation and may increase their likelihood of being victims again. Identifying patterns of victimisation related to age, scam type, seasonality, and region, as well as the prevalence of recurrent mass marketing fraud (revictimization) among older persons, is the goal of this research. We make advantage of the United States Postal Inspection Service's (USPIS) twenty years of non-public administrative data. These datasets, which include over 2 million distinct U.S. victims and their transactions with four distinct fraud organisations, were taken during law enforcement investigations into mass mailing scam organisations. Name, address, and a change of address file were used to match victims across databases. We discover that in psychic frauds, revictimization rates rise with age. The average age of the 10,000 victims who replied the most (82–562) was 78 years old, and their total damages were $4,700 per person. Other noteworthy patterns about lottery and sweepstakes frauds surfaced. In contrast to earlier research on fraud victimisation, conclusions on victim characteristics are derived from real-world fraud encounters rather than from surveys or fictitious situations in which victims are required to self-report fraud. The results provide important information on elder victims and the trends of chronic victimisation that is pertinent to policy.

III. SYSTEM ANALYSIS & DESIGN

EXISTING SYSTEM

The first dataset, "Employment Scam Aegean Dataset" (EMSCAD), was formally provided by Vidros et al. [7] in order to identify fraudulent job advertisements. They then used conventional machine learning classifiers on the dataset to identify ORF. They conducted two different kinds of experiments and contrasted the outcomes. Naive Bayes (NB), Zero Rule (ZeroR), One Rule (OneR), Logistic Regression (LR), J48, and Random Forest (RF) are the six classifiers used in the first experiment. With the greatest accuracy of 91.4%, RF is the best classifier in this trial. The empirical ruleset model is used for the second experiment. The accuracy of the empirical ruleset modelling was 90.6% according to the LR, J48, and RF classifiers.

The "fake job postings" dataset was also subjected to machine learning methods by Dutta and Bandyopadhyay [8]. As single classifier-based predictions, NB, K-Nearest Neighbour

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

(KNN), Multi-Layer Perceptron (MLP), and Decision Tree (DT) are used. Ensemble classifier-based predictions are made using RF, Gradient Boosting (GB), and Adaptive Boosting (AdaBoost) classifiers. The RF classifier performs better with an accuracy of 98.27% among ensemble classifier-based predictions, while DT earned the best accuracy of 97.2% among single classifier-based predictions. Alghamdi and Alharby released another study to identify ORFs [9]. To identify pertinent characteristics in the dataset, they used Support Vector Machines (SVM). They used an ensemble-based RF classifier for the classification challenge. This study achieved 97.2% accuracy, which is regarded as excellent and sufficient. In order to construct ORF Detector, Lal et al. [10] used three ensemble techniques—Maximum Vote, Majority Vote, and Average Vote—on three baseline classifiers: RF, LR, and J48. Three fundamental categories—contextual, linguistic, and metadata—are used to group the collected characteristics. The suggested ORFDetector has a 95.5% accuracy rate. Nasser et al. [11] attempted to address this issue by downsampling majority class records using an unbalanced dataset.

This article use Artificial Neural Networks (ANN) to identify online recruitment scams. The suggested model achieves 93.64% accuracy. The EMSCAD dataset was subjected to several data mining approaches in a research by Habiba et al. [20]. Both conventional Machine Learning (ML) and Deep Learning (DL) classifiers have been assessed by them. Among ML classifiers, RF performed the best, with the maximum accuracy of 96.5%, while among DL models, Deep Neural Network (DNN) had the highest accuracy of 99%. Lokku et al. [12] utilised the "EMSCAD" dataset in another investigation. Following preprocessing and data cleaning, the TF-IDF was used to extract features. Since the dataset is unbalanced, they worked on balancing it by boosting the minority class's data points. Then, they used the RF classifier to the balanced dataset. By using this method, they achieved 99% accuracy.

In order to eradicate the issue of job frauds, Nindyati and Nugraha [13] also conducted study. The Indonesian Employment Scam Detection Dataset (IESD) is the dataset they produced. In order to determine if online job opening descriptions are fraudulent, they suggested context-based behavioural characteristics. They used six machine learning algorithms—NN, SVM, LR, DT, NB, and KNN—to evaluate their suggested features. Behavioural characteristics are used to achieve 90% accuracy. Alandjani et al. [21] classified and compared job ads using two characteristics based on machine learning models: DT, NB, RF, and KNN. It is evident that KNN produced encouraging findings for this study.

In order to address the problem of class imbalance, David et al. [25] worked on the "SocIal Media And Harassment (SIMAH)" dataset. For comparison, three distinct experimental sets—BERT+LSTM, BERT+FNN, and BERT+SMOTE+LSTM—have been constructed. The BERT+SMOTE+LSTM model outperformed the other models, according to the data. The issue of class imbalance in online transaction fraud (OTF) detection was addressed by Singla et al. [26]. Three datasets—Credit Card, Banksim, and IEE CIS—that contain transactional data are used. With a distinct set of hyperparameters, the DNN architecture with two hidden layers was built up for every dataset, and its performance was noticeably better than that of other baseline techniques.

Disadvantages

• Data complexity: In order to identify Online Recruitment Fraud (ORF), the majority of machine learning models now in use need to be able to properly analyse vast and complicated datasets.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

• Data availability: In order to provide precise predictions, the majority of machine learning models need a lot of data. The accuracy of the model may degrade if data is not accessible in large enough amounts.
• Inaccurate labelling: The accuracy of the machine learning models that are now in use depends on how well the input dataset was used for training. Inaccurate labelling of the data prevents the model from producing reliable predictions.

PROPOSED SYSTEM

We introduced a brand-new dataset of phoney job advertisements that were classified as "fraudulent" for phoney job advertisements and "non-fraudulent" for genuine ones. Three distinct sources of job ads are combined to provide the suggested data. To expand the dataset with the most recent job posts, we utilise "Fake Job Postings" as the core dataset and include publicly accessible job ads from Pakistan and the US.

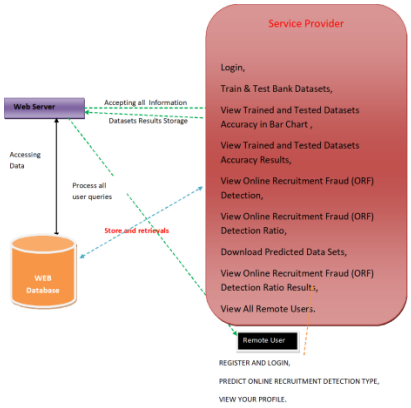
We took this action as the benchmark datasets now in use are out-of-date and constrained by the knowledge of particular job ads, which reduces the effectiveness of current algorithms to identify fake employment. The dataset was prepared, and then it was subjected to exploratory data analysis, or EDA. The dataset's unbalanced class distribution was discovered by EDA. The ratio of samples in the minority class to those in the majority class is known as the imbalance class distribution [14]. For regular courses, it may result in high prediction accuracy; for rare classes, it might result in poor predictive accuracy. Anomaly detection [15], facial recognition [16], medical diagnosis [17], text classification [18], and many other real-world areas are affected by the class imbalance issue. SMOTE [19], an oversampling method, became widely used. In order to address class imbalance issues in a variety of fields, researchers have lately used over 85 distinct SMOTE variations that have been described in the literature.

Advantages

• A unique dataset is created by combining job ads from three distinct sources.
• Exploratory Data Analysis (EDA) reveals a severely unbalanced class distribution in the dataset that was gathered. To balance the class distribution ratio, ten of the best SMOTE variations are used.

To determine if a job posting is fake or not, transformer-based deep learning models are applied to the dataset. Both balanced and unbalanced datasets are used for comparative study of the models that are applied.

SYSTEM ARCHITECTURE



IV. IMPLEMENTATIONS

Modules

Service Provider

The Service Provider must use a working user name and password to log in to this module. Following a successful login, he may do several tasks including training and testing bank datasets, See the Accuracy of Trained and Tested Datasets in a Bar ChartView Accuracy Results for Trained and Tested Datasets, View the Detection Ratio for Online Recruitment Fraud (ORF), Online Recruitment Fraud (ORF) Detection, Get Predicted Data Sets here. View the Results of the Online Recruitment Fraud (ORF) Detection Ratio and View Every Remote User.

View and Authorize Users

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

The administrator may see a list of all registered users in this module. Here, the administrator may see the user's information, like name, email, and address, and they can also grant the user permissions.

Remote User

A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database. Following a successful registration, he must use his password and authorised user name to log in. Following a successful login, the user may do tasks including registering and logging in, predicting the kind of online recruitment detection, and seeing their profile.

ALGORITHMS

K-Nearest Neighbors (KNN)

- A simple but very effective classification technique that uses a similarity metric
- Non-parametric learning and lazy learning don't "learn" until the test example is provided.
- We use the training data to determine the K-nearest neighbours of any fresh data that has to be classified.

Example

- The training dataset comprises the k-closest examples in feature space, which is defined as a space containing categorisation variables (non-metric variables).
- Learning is based on instances, which also facilitates lazy learning because it may take some time for an instance near the input vector for testing or prediction to occur in the training dataset.

Logistic regression Classifiers

The relationship between a collection of independent (explanatory) factors and a categorical dependent variable is examined using logistic regression analysis. When the dependent variable simply has two values, like 0 and 1 or Yes and No, the term logistic regression is used. When the dependent variable contains three or more distinct values, such as married, single, divorced, or widowed, the technique is sometimes referred to as multinomial logistic regression. While the dependent variable's data type differs from multiple regression's, the procedure's practical application is comparable.

When it comes to categorical-response variable analysis, logistic regression and discriminant analysis are competitors. Compared to discriminant analysis, many statisticians believe that logistic regression is more flexible and appropriate for modelling the majority of scenarios. This is due to the fact that, unlike discriminant analysis, logistic regression does not presume that the independent variables are regularly distributed.

Both binary and multinomial logistic regression are calculated by this software for both category and numerical independent variables. Along with the regression equation, it provides information on likelihood, deviance, odds ratios, confidence limits, and quality of fit. It does a thorough residual analysis that includes diagnostic residual plots and reports. In order to find the optimal regression model with the fewest independent variables, it might conduct an independent variable subset selection search. It offers ROC curves and confidence intervals on expected values to assist in identifying the optimal classification cutoff point. By automatically identifying rows that are not utilised throughout the study, it enables you to confirm your findings.

Naïve Bayes

The supervised learning technique known as the "naive bayes approach" is predicated on the straightforward premise that the existence or lack of a certain class characteristic has no bearing on the existence or nonexistence of any other feature. However, it seems sturdy and effective in spite of this. It performs similarly to other methods of

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

guided learning. Numerous explanations have been put forward in the literature. We emphasise a representation bias-based explanation in this lesson. Along with logistic regression, linear discriminant analysis, and linear SVM (support vector machine), the naive bayes classifier is a linear classifier. The technique used to estimate the classifier's parameters (the learning bias) makes a difference.

Although the Naive Bayes classifier is commonly used in research, practitioners who want to get findings that are useful do not utilise it as often. On the one hand, the researchers discovered that it is very simple to build and apply, that estimating its parameters is simple, that learning occurs quickly even on extremely big datasets, and that, when compared to other methods, its accuracy is rather excellent. The end users, however, do not comprehend the value of such a strategy and do not get a model that is simple to read and implement.

As a consequence, we display the learning process's outcomes in a fresh way. Both the deployment and comprehension of the classifier are simplified. We discuss several theoretical facets of the naive bayes classifier in the first section of this lesson. Next, we use Tanagra to apply the method on a dataset. We contrast the outcomes (the model's parameters) with those from other linear techniques including logistic regression, linear discriminant analysis, and linear support vector machines. We see that the outcomes are quite reliable. This helps to explain why the strategy performs well when compared to others. We employ a variety of tools (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b, and RapidMiner 4.6.0) on the same dataset in the second section. Above all, we make an effort to comprehend the outcomes.

Random Forest

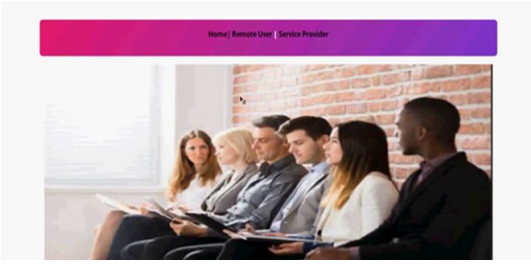
Random forests, also known as random decision forests, are ensemble learning techniques that build a large number of decision trees during training for tasks like regression and classification. The class chosen by the majority of trees is the random forest's output for classification problems. The mean or average forecast of each individual tree is given back for regression tasks. The tendency of decision trees to overfit to their training set is compensated for by random decision forests. Although random forests are less accurate than gradient enhanced trees, they often perform better than choice trees. However, their performance may be impacted by data peculiarities.

Tin Kam Ho[1] developed the first algorithm for random decision forests in 1995 by using the random subspace technique, which in Ho's definition is a means of putting Eugene Kleinberg's "stochastic discrimination" approach to classification into practice.

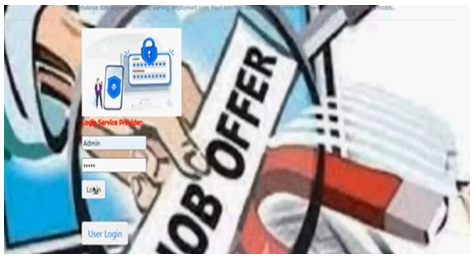
Leo Breiman and Adele Cutler created an algorithm extension and filed for a trademark in 2006 for "Random Forests" (owned by Minitab, Inc. as of 2019). The extension builds a set of decision trees with controlled variance by combining Breiman's "bagging" concept with random feature selection, which was initially proposed by Ho[1] and then separately by Amit and Geman[13].

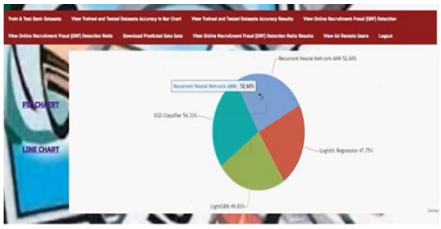
Businesses often employ random forests as "blackbox" models since they need minimal setup and provide accurate forecasts across a variety of inputs.

V. SCREENSHOTS



International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392





VI. CONCLUSION

The issue of ORF detection is carefully examined in this study. A unique dataset of fraudulent job ads was given in this article. Three distinct sources of job ads are combined to provide the suggested data. The class distribution in the gathered dataset was found to be very unbalanced upon doing EDA. The top ten most successful SMOTE variations were applied to the unbalanced data in order to correct this class distribution imbalance. The effect of using SMOTE variations on predictive models was then examined using a type error analysis. To get a more thorough understanding of the trials, transformer-based classification models, BERT and RoBERTa, were used to both the balanced and unbalanced data. The outcomes were then compared. The performance of the applied strategies was compared using a variety of assessment indicators. Only accuracy as an assessment criterion was unable to accurately reflect overall performance due to the class imbalance problem. Since the majority class's strong prediction accuracy may obscure the minority class and result in an incomplete judgement, it may be deceptive. Therefore, improving balanced accuracy and recall as assessment measures was given top priority in this research. Every strategy that was used performed admirably. However, it was found that BERT, when combined with the SMOBD SMOTE approach, performed very well on our data and produced the best results based on the type error and classification findings.

The experiments conducted in this study may provide reputable organisations and job searchers important guidance on how to better comprehend fact-based insights on employment scams and their impact on society. As a result, individuals would no longer fall victim to job frauds. such who were squandering their time and money on such fake operations may now be alert thanks to the differentiation of ORF. When class imbalance issues are not taken into account,

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

traditional fraud detection methods might provide inaccurate results for employers and job seekers alike. This issue must also be resolved in order to get a genuine set of findings. Although we significantly enhanced the system's performance in this study and obtained useful outcomes based on balanced data, there are still a lot of unanswered questions that may be addressed in the future. Only job advertisements that are posted in English are subject to all sets of analysis.

A comparable study of job posts published in languages other than the present dataset might be performed for a more thorough investigation. The dataset may be enhanced by including the most recent job posts, especially as internet recruiting is becoming more and more common. This survey includes job advertisements from all across the globe. For the benefit of the public, a similar study may be carried out especially for job posts in a certain area to determine the percentage of false postings. Furthermore, considering the high frequency of fraudulent activity linked to online employment from home, it might be considered essential to include job posts that relate to remote work possibilities via online platforms in order to create a unique dataset.

To overcome the imbalance in class distribution, a variety of SMOTE variations were used in the current study. Using hybrid oversampling approaches might be taken into consideration to provide even more accurate findings. Future studies should investigate innovative transformer-based hybrid models and explainable AI.

REFERENCES

[1] P. Kaur, ‘‘E-recruitment: A conceptual study,’’ Int. J. Appl. Res., vol. 1, no. 8, pp. 78–82, 2015.

[2] C. S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar,‘‘Fake job detection and analysis using machine learning and deep learning algorithms,’’ Revista Gestão Inovação e Tecnologias, vol. 11, no. 2, pp. 642–650, Jun. 2021.

[3] A. Raza, S. Ubaid, F. Younas, and F. Akhtar, ‘‘Fake e job posting prediction based on advance machine learning approachs,’’ Int. J. Res. Publication Rev., vol. 3, no. 2, pp. 689–695, Feb. 2022.

[4] Online Fraud. Accessed: Jun. 19, 2022. [Online]. Available: https://www.cyber.gov.au/acsc/report

[5] J. Howington, ‘‘Survey: More millennials than seniors victims of job scams,’’ Flexjobs, CO, USA, Sep. 2015. Accessed: Jan. 2024 [Online]. Available: www.flexjobs.com/blog/post/survey-results-millennials-seniors-victims-job-scams

[6] Report Cyber. Accessed: Jun. 25, 2022. [Online]. Available: https://www.actionfraud.police.uk/

[7] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, ‘‘Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset,’’ Future Internet, vol. 9, no. 1, p. 6, Mar. 2017.

[8] S. Dutta and S. K. Bandyopadhyay, ‘‘Fake job recruitment detection using machine learning approach,’’ Int. J. Eng. Trends Technol., vol. 68, no. 4, pp. 48–53, Apr. 2020.

[9] B. Alghamdi and F. Alharby, ‘‘An intelligent model for online recruitment fraud detection,’’ J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.

[10] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, ‘‘ORFDetector: Ensemble learning based online recruitment fraud detection,’’ in Proc. 12th Int. Conf. Contemp. Comput. (IC3), Noida, India, Aug. 2019, pp. 1–5.

[11] I. M. Nasser, A. H. Alzaanin, and A. Y. Maghari, ‘‘Online recruitment fraud detection using ANN,’’ in Proc. Palestinian Int. Conf. Inf. Commun.Technol. (PICICT), Sep. 2021, pp. 13–17.

[12] C. Lokku, ‘‘Classification of genuinity in job posting using machine learning,’’ Int. J. Res.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025
https://doi.org/10.62647/ijitce.2025.v13.i2.pp382-392

Appl. Sci. Eng. Technol., vol. 9, no. 12, pp. 1569–1575, Dec. 2021.
[13] O. Nindyati and I. G. Bagus Baskara Nugraha, ‘‘Detecting scam in online job vacancy using behavioral features extraction,’’ in Proc. Int. Conf. ICT Smart Soc. (ICISS), vol. 7, Bandung, Indonesia, Nov. 2019, pp. 1–4.
[14] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, ‘‘Handling imbalanced datasets: A review,’’ GESTS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1,pp. 25–36, 2006.
[15] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, ‘‘Toward credible evaluation of anomaly-based intrusion-detection methods,’’ IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 5, pp. 516–524, Sep. 2010.
[16] Y.-H. Liu and Y.-T. Chen, ‘‘Total margin based adaptive fuzzy support vector machines for multiview face recognition,’’ in Proc. IEEE Int. Conf.Syst., Man Cybern., Waikoloa, HI, USA, Oct. 2005, pp. 1704–1711.
[17] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, ‘‘Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance,’’ Neural Netw., vol. 21, nos. 2–3, pp. 427–436, Mar. 2008.
[18] Y. Li, G. Sun, and Y. Zhu, ‘‘Data imbalance problem in text classification,’’in Proc. 3rd Int. Symp. Inf. Process., Luxor, Egypt, Oct. 2010, pp. 301–305.
[19] N. V. Chawla, K.W. Bowyer, L. O. Hall, andW. P. Kegelmeyer, ‘‘SMOTE:Synthetic minority over-sampling technique,’’ J. Artif. Intell. Res., vol. 16,pp. 321–357, Jun. 2002.
[20] S. U. Habiba, Md. K. Islam, and F. Tasnim, ‘‘A comparative study on fake job post prediction using different data mining techniques,’’ in Proc.2nd Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST), Dhaka,Bangladesh, Jan. 2021, pp. 543–546.
[21] G. Othman Alandjani, ‘‘Online fake job advertisement recognition and classification using machine learning,’’ 3C TIC, Cuadernos de Desarrollo Aplicados a las TIC, vol. 11, no. 1, pp. 251–267, Jun. 2022.
[22] A. Gosain and S. Sardana, ‘‘Handling class imbalance problem using oversampling techniques: A review,’’ in Proc. Int. Conf. Adv. Comput.,Commun. Informat. (ICACCI), Delhi, India, Sep. 2017, pp. 79–85.
[23] F. Akhbardeh, C. O. Alm, M. Zampieri, and T. Desell, ‘‘Handling extreme class imbalance in technical logbook datasets,’’ in Proc. 59th Annu.Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang.Process., 2021, pp. 4034–4045.
[24] J. Ah-Pine and E.-P. Soriano-Morales, ‘‘A study of synthetic oversampling for Twitter imbalanced sentiment analysis,’’ in Proc. Workshop Interact. Between Data Min. Nat. Lang. Process. (DMNLP), Riva del Garda, Italy,Sep. 2016, pp. 17–24.
[25] J. David, J. Cui, and F. Rahimi, ‘‘Classification of imbalanced dataset using BERT embeddings,’’ Dalhousie Univ., Halifax, Canada, Jan. 2020. Accessed: Jan. 2024. [Online]. Available: https://fatemerhmi.github.io/files/Classification_of_imbalanced_dataset_using_BERT_embedding.