2025 International Conference on Data Science, Agents, and Artificial Intelligence (ICDSAAI 2025)
28-29 March 2025, Chennai, India

2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI) | 979-8-3315-3755-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICDSAAI65575.2025.11011776

Machine Learning-powered Job Offer Verification

Vigneshwaran N[*]
Department of IT,
R.M.K. Engineering College.
RSM Nagar, Kavaraipettai,
Chennai-601206
vigneshwaran14n@gmail.com

A. Vijayaraj
Associate professor,
Department of IT,
R.M.K. Engineering College,
RSM Nagar, Kavaraipettai,
Chennai-601206,
satturvijay@gmail.com

V.P. Murugan,
Associate professor,
Department of Mathematics,
Panimalar Engineering College,
Varadharajapuram, poonamallee
Chennai- 600 123.
vpmurugan07@gmail.com

Jebakumar
Associate professor,
Department of computing technologies,
SRM Institute of science and technology,
Kattankulathur, Chennai.
jebakumar@srmist.edu.in

N. Mageshkumar
Assistant Professor,
Department of Computer science and Technology,
Madanapalle Institute of Technology & Science
Andhra Pradesh -517325.
mageshkumarn@mits.ac.in

R.Megavannan,
Faculty of Management,
SRM Institute of Science and Technology,
Kattankulathur.
megavanr@srmist.edu.in

***Abstract* - In the modern world, technological developments and inventions have given us the ability to handle practically each aspect of our lives, including keeping track of money, education, job appearances, and security. But this reliance on technology has also made it easier for con artists to defraud people and gain quick money. Fake work notices are a specific approach that has recently become more common. Unsuspecting people submit applications for these fake job openings, handing over their personal information and paying the application fees to the con artists as though they were going to fall for the fraud and waste their hard-earned money. In this paper, the Natural Language Processing (NLP) approach is used to identify bogus job postings. We collected relevant information from job advertisements, after that we applied the random forest classifier to train and evaluate our model. The results we obtained show that our method can recognize fake job advertisements with a 97% accuracy rate.**

Keywords— Fake job, postings, Random Forest algorithm, Natural Language Processing.

I.INTRODUCTION

An increase in the quantity of fraudulent job listings has taken place in recent years as a result of the expansion of online job portals and the easy availability of job postings. These job postings regularly attract job seekers with attractive career opportunities, but in fact they serve just as a way for fraudsters to swindle unsuspecting victims out of their personal information or money. Therefore, it is essential to address the problem of identifying such fake job advertising. Computerized job posting fraud detection using natural language processing (NLP) methods and machine learning algorithms is a solution to this issue. In this research, we present a novel approach for detecting fake jobs in the EMSCAD dataset by applying the Random Forest algorithm. The EMSCAD dataset is a freely available dataset that contains both real and fake job listings that were scraped from various online employment portals. This dataset was used to develop and evaluate our model. The approach we propose includes pre-processing the job advertisements to extract relevant elements, such as keywords, job descriptions, and company data, using NLP approaches. Based on these collected features, we next use the Random Forest algorithm to categorize job listings as either genuine or fake. Accuracy, precision, recall, and the F1-score were some of the performance indicators used to assess the effectiveness of our suggested approach. Our experimental results show that our approach to fake job detection achieves high accuracy and beats

979-8-3315-3755-5/25/$31.00 ©2025 IEEE

existing state-of-the-art approaches. In conclusion, using NLP methods and machine learning algorithms, our suggested approach has the potential to automatically identify fake job advertisements. This has significant implications for the job market since it can protect job seekers against fraud and raise the standard of job ads on job boards on the internet.

II. RELATED WORK

In the realm of online deception and fraudulent activities, researchers have diligently explored various facets to understand and combat misinformation. Simmons [2] conducted an insightful investigation into the market incentives driving fraud, particularly focusing on the intricate relationship between reach and frequency due to the lack of attention [1]. This exploration provided valuable insights into the motivations behind deceptive practices in the digital landscape building on the understanding of information dissemination dynamics, Vosoughi, Roy, and Aral [3] conducted a seminal study on the spread of true and false news online. Their research not only delineated the patterns of information propagation but also shed light on the factors influencing the virality of both accurate and deceptive content. Cyr, Head, Lim, and Stibe [4] delved into the realm of online persuasion through website design, employing the Elaboration Likelihood Model to understand how the design of online platforms can influence user perceptions and interactions. This work contributes valuable perspectives on the persuasive elements inherent in digital interfaces. Examining the nuanced meanings conveyed through one-click interactions in social media, Hayes, Carr, and Wohn [5] offered insights into paralinguistic digital affordances. Understanding the varied interpretations of user actions in the digital space is crucial for discerning genuine interactions from potentially deceptive or misleading behaviours. Individual susceptibility to online influence became a focal point in the study by Williams, Beardmore, and Joinson [6], highlighting the diverse ways individuals respond to persuasive techniques and deceptive content. Cook, Lewandowsky, and Ecker [7] advanced the field by proposing inoculation strategies to neutralize misinformation, emphasizing the exposure of misleading argumentation techniques. In the realm of text classification, Zhang, Yoshida, and Tang [8] utilized support vector machines based on multi-word features, showcasing the relevance of advanced algorithms in discerning patterns within textual data. Meanwhile, Chen, Huang, Tian, and Qu [9] explored feature selection for text classification using Naïve Bayes, contributing to the ongoing discourse on effective classification methodologies. The optimization of bag-of-words models for image classification became a focus in the work of Wang and Huang [10], providing cross-disciplinary insights into the application of text-based models in image-related tasks. These studies collectively underscore the importance of robust text classification methods in diverse applications. In the specific context of online recruitment fraud, Vidros, Kolias, Kambourakis, and Akoglu [11] proposed an automatic detection approach, laying the groundwork for subsequent investigations into identifying fraudulent job postings. Ahmed, Traore, and Saad [12] and [13] applied n-gram analysis and machine learning techniques to detect fake news and opinion spams, showcasing the versatility of computational approaches in identifying deceptive content. Furthering the exploration into the realm of fake job recruitment, Dutta and Bandyopadhyay [14] undertook a comprehensive examination using a machine learning approach. Shibly, Sharma, and Naleer [15] contributed to this area by comparing the performance of two class boosted decision tree and two class decision forest algorithms in predicting fake job postings. Moving towards the integration of deep learning, Anita, Nagarajan, Sairam, Ganesh, and Deepakkumar presented a study on fake job detection and analysis, leveraging machine learning and deep learning algorithms.

III. METHODOLOGY

A. Examine Spam Detection

In the modern digital era, identifying spam is a critical component of email communication management. Unwelcome, unwelcome, and irrelevant email messages that are delivered in mass to many recipients are referred to as spam. By contaminating computers via malware or phishing scams, spending time sifting through pointless communications, and decreasing productivity in businesses, spam messages can be harmful. Therefore, spam detection is essential for email management to stop such attacks and boost the effectiveness of email communication. The process of spam detection involves identifying and removing spam communications using a variety of methods and algorithms. Blacklist databases, which kept a list of well-known spammers and spam domains, were one of the first techniques for spam detection. This strategy, however, fell short because spammers routinely alter their e-mail addresses and domains to evade detection. Because of this, contemporary spam detection systems employ increasingly sophisticated and cutting-edge Unwanted bulk emails, often known as spam emails, are frequently sent to users and can result in storage problems and an increase in bandwidth usage. Neural network-based filters for spam are used by email service providers including Google Mail, Yahoo Mail.

B. Detection for Fake News

Fake news on social media can frequently be identified from three different angles: how it is generated, how it spreads, and how engaged an individual person is with it. Echo chambers and rogue user accounts are used to distribute fake news. Through the acquisition of data relevant to the news's social context and content, machine learning algorithms are used to identify fake news. Videos has put up a method for detecting job fraud, but their method has only been evaluated on balanced datasets; it is uncertain how well it works on imbalanced datasets. To make sure that prediction models are trustworthy, it is crucial to test them using an unbalanced dataset. An ensemble-based model called the ORF Detector is used to identify online fraud, but it only works well on balanced datasets and is not as accurate on imbalanced datasets. The model employs three baseline classifiers—J48, logistic regression, and random forest—to apply an average number of cast votes, a majority of votes, and the maximum number of cast votes. Various methods and algorithms are used to identify and categorise stories as either genuine or bogus in the detection of bogus information. Content-based analysis is one of the most widely used methods for spotting fake news. In content-based analysis, the language and structure of news stories are examined in order to spot trends that are indicative of fake news. For instance, false news pieces frequently use sensational language, make bold assertions, and have questionable sources. In order to identify the traits of fake news and categorise new articles as either genuine or false, machine learning algorithms like Naive Bayes and Support Vector Machines could be trained on a huge dataset of labelled news articles. Social network analysis is another method for spotting bogus news. To identify the propagation of false information, social network analysis looks at user behaviour and social interactions methods. Content-based filtering is the method used most frequently to identify spam messages. With this approach, the email message's content is examined for signs of spam. To assess the possibility that a communication is spam, content-based filtering employs a number of techniques, including analysis of on social media platforms. For instance, false news items frequently contain a high number of shares, likes, & comments, which can show that they are disseminated quickly via social media networks. Algorithms can spot false news stories and take action to stop their dissemination by examining trends in social interactions and behaviour. Another method for spotting bogus news is fact-checking. Examining and evaluating assertions made in news items in order to determine their veracity and correctness is known as fact-checking. This method can take a lot of time, but it has an extremely high degree of accuracy in spotting false information. Journalists or automatic fact-checking systems that make use of machine learning and natural language processing algorithms can both assess the accuracy of news stories. The use of the technology known as blockchain is one last method for identifying bogus news. A distributed ledger system called blockchain technology can safely and openly record and verify transactions. News articles can be authenticated and traced from the source through the reader by utilising blockchain technology. It is challenging to edit or alter news stories using this strategy since it offers an elevated level of security and transparency.

IV. SINGLE CLASSIFIER BASED PREDICTION (MODELS IMPLEMENTED)

The classifiers that have been learned are utilized to predict unknown test cases when identifying fraudulent job postings.

A. Naive Bayes

Since the number of parameters required for naive Bayes classifiers is inversely related to the total number of variables in a learning problem, they are particularly scalable. In contrast to other classifiers that necessitate costly iterative approximation, naive Bayes classifiers may be trained to maximise likelihood in linear time by simply assessing a closed-form expression. By using class labels chosen from a limited collection, these classifiers are used to build models that categorise issue occurrences represented as vectors of feature values. To evaluate the accuracy of this classifier, it is crucial to determine the class's loss of information as a result of the premise of independence rather than focusing on feature dependencies. The reason why Naive Bayes is named "naive" is because it strongly presumes that the characteristics used to create the classification are independent. As a result, the algorithm makes the assumption that the characteristics are independent of one another, which is frequently false in practical situations. Naive Bayes, despite this drawback, is frequently effective in real-world applications, particularly in language processing tasks like sentiment evaluation, identifying spam, and text classification. One of the primary benefits of Naive Bayes is its ease as well as effectiveness. It requires a small quantity of training information and can be trained rapidly, making it appropriate for real-time uses. The algorithm is also strong to insignificant features, which indicates that it may still make precise predictions even if a few of the features are not beneficial for the classification task. Gaussian Naive Bayes algorithm, Multinomial Naive Bayes, & Bernoulli Naive Bayes are the three primary varieties of Naive Bayes algorithms. For continuous data, Gaussian Naive Bayes is utilised, however for discrete data, Multinomial & Bernoulli Naive Bayes are utilised. Multinomial Naive Bayes, which associates each feature with a word in the text and uses the average frequency of the specified word as the feature's value, is a popular method for classifying texts. Regardless of its benefits, Naive Bayes has a few restrictions. As said before, the independence assumption might not hold in practical applications, which may lead to minimised accuracy. In addition, Naive Bayes might struggle with rare or unnoticed features, as it presumes that each feature has been seen in training data. This limitation may be discussed via techniques like Laplace's smoothing and using a larger training dataset.

B. Support Vector Machine

Models for supervised learning called support vector machines (SVMs) may handle problems like classification and regression. They are useful in a variety of situations because they are capable of handling both nonlinear and linear issues. Drawing a distinction among categories in a classification task is the basic premise of SVMs. Maximizing the separation among endpoints on either end of the chosen line is the goal of this line. This method is helpful because, once separation is complete, the model can accurately forecast the desired classes (labels) for future examples. Both linearly separable & non-linearly separable data can be used with SVM successfully. A linear SVM may be used to locate a linear hyperplane which divides data into distinct classes when the data is linearly separable. However, the data cannot be separated linearly in many real-world applications. In these circumstances, a non-linear Support Vector Machine (SVM) can be applied by utilising a kernel function to translate the input information to a greater-dimensional feature space. In order for the algorithm to discover a hyperplane that categorises the data, this puts the data into a space where it's more probable to be linearly separable. The capacity of SVM to handle high-dimensional information as well as outliers is one of its key advantages. It handles data that is not typically dispersed and performs well with data with a lot of properties. Due to its use of the margin to choose the ideal hyperplane and prevent overfitting to the data used for training, SVM is additionally less prone to overfitting. SVM does, however, have some restrictions. With large datasets, it may become computationally expensive and memory-intensive to train. Additionally, the algorithm's performance may be sensitive to which kernel function is selected. SVM might also struggle with overlapping or noisy data.

C. Logistic Regression

A classification procedure makes use of a categorical answer variable. For instance, the variable that responds will have the two values pass and fail when estimating if someone is going to succeed or fail a test based on the amount of time they spend studying. The binomial logistic regression approach is employed when a response parameter has several values, such as 0 and 1, either positive or negative, or either true or false. In cases where the answer variables contain three or more potential values, multinomial logistic regression is applied. The fundamental goal of the logistic regression technique is to identify the line with the best fit between the two classes. Logistic regression, as opposed to linear regression, which forecasts continuous values, forecasts the likelihood that an input will fall into a given class. A number between 0 and 1 that represents the likelihood that the input belongs to the positive category and is the result of a logistic regression. Modelling the link between the input factors and the likelihood of the outcome variable is how logistic regression functions. This is accomplished by utilising a function called a sigmoid, which converts input variables into probabilities for output. Because of its S-shaped curve, the sigmoid function is able to represent non-linear interactions among the variables that are input and output. The logistic regression model's parameters are calculated via maximum likelihood estimation. Finding the values of the parameters that maximise the likelihood of what was observed given the model is necessary to accomplish this. This is accomplished by minimising the price function, which represents the data's negative logarithmic likelihood. A few benefits of logistic regression include its readability and simplicity. The mathematical coefficients of the method, which show the impact of each of the input variables on the probability of the output, are simple to comprehend and analyse. Additionally, computationally effective and able to handle sizable datasets is logistic regression.

D. BASED ON THE ENSEMBLE APPROACH CLASSIFIERS (RANDOM FOREST)

The ensemble technique can be used to combine several machine learning algorithms to increase a system's overall accuracy. Random Forest (RF), an ensemble learning technique that may effectively tackle classification issues, uses regression. The random forest functions as a classifier that blends numerous decision trees on different subsets of the data in order to increase the expected accuracy of the dataset. Several decision tree classifiers can be combined using the meta-estimator known as a random forest. Averaging is used to increase prediction accuracy and manage over-fitting on various sub-samples of the dataset. The maximum amount of samples argument must be true in order to confine the sub-sample size; alternatively, each tree is built using the entire dataset. The Random Forest classifier includes a variety of tree-based classifiers, which are applied to different dataset subsamples. These classifiers decide which class best matches the input by voting for it, resulting in an ensemble-based prediction.

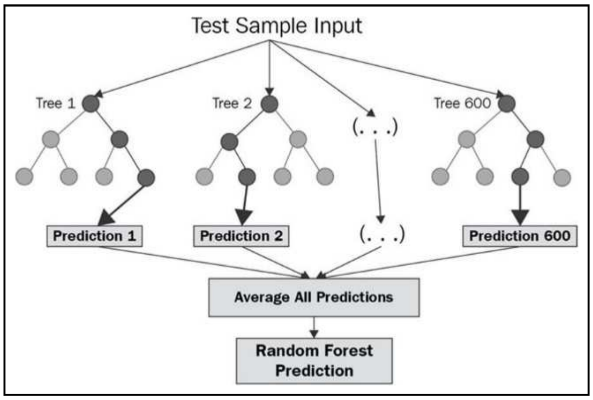
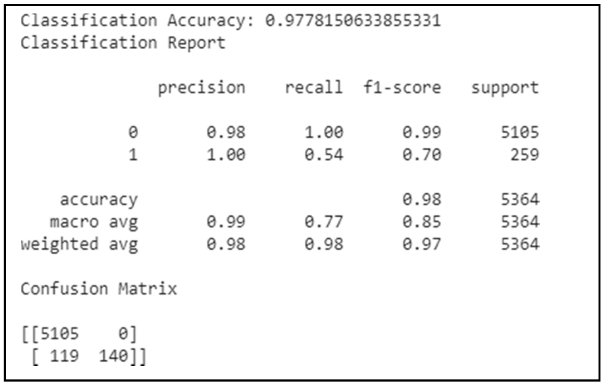


Fig. 1. concept of the Random Forest classifier

E. Metrics for Performance Evaluation

When making classification predictions, there are 4 potential results: "true positive (TP), true negative, false positive, & false (False Negative). To evaluate the success of the system for detecting fraudulent jobs, four indicators were employed. These indicators are:

1. Accuracy: A measure of accuracy is the proportion of accurate forecasts generated by the model. This is calculated by divided (TP + TN) by (TP + FP + FN + TN). When the classes are balanced, or when there are about the same amount of both positive and negative instances, this statistic is useful.
2. Recall: The number of percent of positive events that the model properly detected is expressed as a metric called recall. The equation is TP / (TP + FN). This statistic is helpful when the objective is to find all instances of positivity, even if doing so results in some false positives.
3. Precision: The precision statistic counts the proportion of positive instances the model accurately detected out of all the positive instances it predicted. The equation is TP / (TP + FP). When the objective is to prevent negative results, even if doing so involves ignoring some positive cases, this statistic is helpful.
4. F1 score: The average harmonic of recall and precision is the F1 score. It is a strategy for striking a balance between recall and precision when assessing a classifier's performance. while 2 * (precision * recall) / (precision + recall) is calculated.

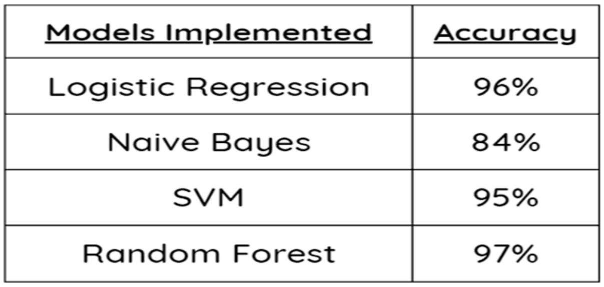


```
Classification Accuracy: 0.9778150633855331
Classification Report

              precision    recall  f1-score   support

           0       0.98      1.00      0.99      5105
           1       1.00      0.54      0.70       259

    accuracy                           0.98      5364
   macro avg       0.99      0.77      0.85      5364
weighted avg       0.98      0.98      0.97      5364

Confusion Matrix

[[5105    0]
 [ 119  140]]
```

Fig. 2. Classification Report

V. EXPERIMENTAL RESULTS

When it comes to fraud detection, a low sensitivity (failure to detect fraudulent job postings) can be dangerous for job-seekers. On the other hand, a low specificity (labelling legitimate job postings as fraudulent) may only lead to additional review by a human since real job postings are usually easy to identify. However, the real problem lies in fraudulent job postings that appear legitimate and can deceive people. Fig 3 proposed model's performance is compared to that of other models that are already in use in Fig 3, which shows that our approach obtains superior accuracy, recall, and precision scores, showing improved fraud detection capabilities.



Models Implemented	Accuracy
Logistic Regression	96%
Naive Bayes	84%
SVM	95%
Random Forest	97%

Fig. 3. Results

VI. CONCLUSION

You will only ever get sincere job offers from us; we promise. Our research offers numerous machine learning techniques to identify fake job listings. We provide answers to this issue and demonstrate the effectiveness of several classifiers at identifying employment fraud using a supervised method. According to the experimental results, the Random Forest is an extremely strong classifier that outperforms its rivals in terms of accuracy in classification. Because it can handle large datasets, handle data that is missing, and take into account complex relationships between variables, the Random Forest algorithm was chosen. The system correctly classified job posts as either legitimate or false by integrating numerous decision trees, using bootstrapping and selecting features approaches, and combining feature selection with bootstrapping. The Random Forest algorithm's excellent accuracy rate in this investigation demonstrates its potential for broad use in the labour market. Early suspicion of suspicious job advertising can assist protect job applicants from falling for scams and ultimately lessen the harm that false job postings cause to both people and organisations. Our suggested solution outperformed the competition with a 97 percent accuracy rate.

REFERENCES

[1]. Becker R (2017) Your short attention span could help fake news spread. overload-twitter-facebook-social-media

[2]. Simmons G (2017) Market incentives that drive fraud: the truth behind reach vs. frequency.

[3]. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151

[4]. A. Vijayaraj , V.P. Murugan, R.Megavannan, V. R.Thejeshwar, Hamelda Lourdus Mary A, Dhiya Shree. S , "Echo Trade: Transforming Waste Into Wealth Through Sustainable Actions", Proceedings of the 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI-2024) IEEE Xplore Part Number: CFP24VL6-ART; ISBN: 979-8-3503-8960-9

[5]. Hayes RA, Carr CT, Wohn DY (2016) One click, many meanings: interpreting paralinguistic digital affordances in social media. J Broadcast Electron Media 60(1):171–187

[6]. Williams EJ, Beardmore A, Joinson AN (2017) Individual differences in susceptibility to online influence: a theoretical review. Comput Hum Behav 72:412–421

[7]. Cook J, Lewandowsky S, Ecker UK (2017) Neutralizing misinformation through inoculation: exposing misleading argumentation techniques reduces their influence. PLoS One 12(5):e0175799

[8]. Zhang W, Yoshida T, Tang X (2008) Text classification based on multi-word with support vector machine. Knowl Based Syst 21(8):879–886

[9]. Chen J, Huang H, Tian S, Qu Y (2009) Feature selection for text classification with Naïve Bayes. Expert Syst Appl 36(3):5432–5435

[10]. Wang C, Huang K (2015) How to use bag-of-words model better for image classification. Image Vis Comput 38:65–74

[11]. Vidros S, Kolias C, Kambourakis G, Akoglu L (2017) Automatic detection of online recruitment frauds: characteristics, methods, and a public dataset. Future Internet 9(1):6

[12]. V Praveena, A Vijayaraj, P Chinnasamy, Ihsan Ali, Roobaea Alroobaea, Saleh Yahya Alyahyan, Muhammad Ahsan Raza , "Optimal deep reinforcement learning for intrusion detection in UAVs" , Computers, Materials & Continua, Volume 70 issue 2, Pp 2639-2653

[13]. Ahmed H, Traore I, Saad S (2018) Detecting opinion spams and fake news using text classification. Secur Priv 1(1):e9

[14]. Dutta S, Bandyopadhyay SK (2020) Fake job recruitment detection using machine learning approach. Int J Eng Trends Technol 68.4(2020):48–53

[15]. Shibly F, Sharma U, Naleer H (2021) Performance comparison of two class boosted decision tree snd two class decision forest algorithms in predicting fake job postings. Ann Rom Soc Cell Biol 25(4):2462–2472