Revolutionary Advances in Computing and Electronics: An International Journal
Volume 01, Issue 01, pp.41--51, Jul-Sep 2025

Received: 17/06/2025, Accepted: 13/07/2025

Review Article

The Growing Threat of Fake Job Postings: A Review of Machine Learning and NLP-Based Detection Approaches

Shashikant Kumar, Sadiya Yasmeen, Prabhat Kumar

Computer Science and Engineering, NIT Patna, India

shashikantk.pg24.cs@nitp.ac.in, sadiyay.pg23.cs@nitp.ac.in, prabhat@nitp.ac.in

ABSTRACT

This review paper discusses the Fake Job Post Detection, spotlighting machine-learning, deep-learning, and natural-language-processing tools built to shield young job seekers from con artists and keep the hiring scene honest. To ground the discussion, an in-depth survey of past studies maps the techniques and trends that now dominate the field. The main contribution is a side-by-side comparison of existing systems, spelling out where each shines and where it falls short. The paper goes on to summarize the key hurdles and opportunities investigators still face. A pressing call for larger, cleaner datasets, closer industry partnerships, and fraud alerts that fire in real time runs through the review. By weaving these threads together, the article offers a fresh perspective on what is known today and sketches concrete directions for future studies and policy action in the fast-changing area of fake-job-ad detection.

Keywords: *Fake Job, Threat, Detection, Cyber Fraud, Machine Learning, Natural Language Processing, Deep Learning.*

1. Introduction

A Job Post is the medium through which an employer interacts with a job seeker to fulfill their requirements and vice versa. In earlier times, the traditional approaches, such as newspaper posting, were used for recruitment. Still, nowadays, with the evolution of the web, this process has seen a drastic change, shifting from the traditional methods to online job posting on different platforms like Naukri, Indeed, LinkedIn, etc. With the pros like cost-effectiveness, time saving, and huge reach, the online job market has also come with cons like online recruitment fraud (ORF) through fake job postings. Job Postings, which are fabricated advertisements crafted to mislead job seekers by presenting fraudulent or non-existent employment opportunities, are classified as fake job posts. These posts aim to steal personal, sensitive information and cause financial exploitation. As the spread of the job market is increasing, online recruitment fraud is also growing in parallel because it is tough to verify every job post manually, whether fake or genuine. Here, technology plays its role, and it can be used to detect fake job posts. Fake job-post detection uses computer methods, mostly machine learning, deep learning, and natural-language processing, to spot phoney adverts hiding on job boards. Because anyone can publish a listing with a few clicks, scammers now flood these sites with fake openings hoping to trick desperate job seekers. Behind each bogus post is an attempt to collect money, passwords, or other private data from young people anxious about unemployment. Spotting such postings matters not only for protecting candidates; it also keeps the wider online hiring system credible.

The fake job posts lead to psychological effects, financial losses, identity theft, and wasted time. Ranging from a small-scale depression to suicide cases, theft of personal information to sensitive information like credit card details, the impact of fake job posts can be huge. For organizations and online job platforms, the undetected scams can damage their reputation and result in a loss of confidence among users. In severe cases, platforms may face legal problems or user boycotts due to repeated scams.

*Corresponding author: Shashikant Kumar, Computer Science and Engineering, NIT Patna, India. (shashikantk.pg24.cs@nitp.ac.in)

As hiring moves almost entirely online, the sheer flood of job ads now appearing makes it nearly impossible for people to check each one by hand. That reality pushes companies to build automated tools that can sift through the listings and flag any that look suspicious. While many methods could power such tools, this paper zooms in on machine-learning, deep-learning, and natural-language-processing approaches to get the job done.

2. Literature Review

Due to the rising volume of online job scams, research into spotting fake ads has received fresh interest. Vidros et al. (2017) [1] were the first to share a varied collection of 17,880 listings, including nearly 17,000 real postings and roughly 900 that proved fraudulent. Even so, the group acknowledged the set was heavily skewed, a clear weakness for training. They represented each ad with standard bag-of-words features and then tested their models, Random Forest, Logistic Regression, Naïve Bayes, and Decision Trees, to spot the counterfeits automatically.

Ahmed and colleagues [2] proposed a system to spot fake news online. Their approach blended machine-learning tricks with standard n-gram analysis. The team tested six widely used classifiers alongside two feature methods to see which worked best. Results showed that pairing the Term Frequency-Inverse Document Frequency scheme with a Linear SVM delivered the highest score. Under this setup, the model hit an accuracy rate of roughly 92%.

Shawni Dutta and Samir B. [3] proposed an ML-based classification method as an automated tool for the detection of fake job postings. Different ML classifiers are used to identify fake job posts on the internet, and the outcomes of those classifiers are compared to find the best phoney job posting detection model. The selected model helps find fake job postings from many online job postings. Both single-classifier and ensemble-based classifiers are used to detect fake job postings. The ensemble-based classifiers showed better results in the fake job post detection tasks experiments.

Marcel Naudé [4] discussed making different categories of fake job postings. This paper also tried to find which features are more relevant in identifying fake job postings. In this research, researchers have proposed and validated a Machine Learning based method for detecting identity theft, corporate identity theft, and multi-level marketing types of fake job postings among the advertisements. Researchers have used four features: empirical rule set-based features, bag-of-words models, most recent state-of-the-art word embeddings, and transformer models for different machine learning models. The models were tested on the publicly available dataset consisting of job descriptions. The experiment's outcome indicated that the word embeddings and transformer-based features consistently performed better than the handcrafted rule-set-based features. Finally, a Gradient Boosting classifier and a combination of empirical rule-set-based features, parts-of-speech tags, and bag-of-words vectors gave the F1-score of 0.88, which was best among all the experiments.

Shibly [5] used Microsoft Azure Machine Learning Studio to investigate the use of the proposed model. The author did a comparative study on the performance of a two-class boosted decision tree and two-class decision forest algorithms. They used recall, precision, F1 score, and accuracy to compare the two algorithms. The experiment results have shown that the two-class forest decision algorithm has performed better for detecting fake job posts than the other algorithm. Thus, a two-class decision forest algorithm will better detect Fake job postings. Therefore, a two-class decision forest algorithm can be used to find and identify Fake job posts.

In a recent study, Aashir Amaar [6] presented a system that combines supervised machine learning with natural language processing to spot fake job ads on online boards. For features, he relied on the classic Bag-of-Words approach alongside Term Frequency-Inverse Document Frequency. He then pitted six popular classifiers against each representation to see how well each learned the task. A key hurdle,

however, lay in the skewed labels: most posts were genuine, leaving the fake samples largely ignored by the networks. To bridge that gap, Aashir applied ADASYN, a technique that crafts realistic synthetic observations of the minority class. He ran two rounds of testing- one on the raw imbalance and the other on the boosted set- and then compared the results side by side. With ADASYN and TF-IDF powering the features, the top model soared to the best accuracy in every metric. He also benchmarked his pipeline against modern deep networks and other up-sampling strategies, proving that simplicity can still outperform the latest black boxes in this domain.

Afzal H. [7] mixed principal component analysis (PCA) with Chi-square to identify the most important features. He then used the synthetic minority over-sampling technique (SMOTE) to see how fixing class imbalance changed results. The author also pitted his set-up against leading models to highlight any gains. Tests revealed that pairing SMOTE with Chi-square selection delivered the sharpest accuracy on his system.

Khushboo Taneja [8] leaned on transfer learning to build a Fraud-Bert framework based on bidirectional transformers. She ran her model on the uneven EMSCAD Fake Job Post corpus to simulate real noise. A side-by-side benchmark showed her approach outperformed older methods across the board. It finally claimed the top spot with an F1 score of.93 and solid overall accuracy, proving its strength against skewed data.

A recent study by Dinh-Hong [9] introduces a fresh method that leans on deep-learning-based natural-language processing to spot phony job ads more quickly. For the first step, the authors apply Word2Vec to pull dense vector representations straight from the text. They then marry those text vectors with relevant metadata, creating a blended dataset that further sharpens the detection task. Together, these tweaks push the model's accuracy by a noticeable margin. Finally, the team runs head-to-head tests against leading techniques and shows their system winning in the ongoing battle against fake postings.

3. Commonly used Dataset

The researchers are using one publicly available dataset on Kaggle, which was first published by Vidros et al. (2017) [1]. This dataset is known as the Employment Scam Aegean Dataset (EMSCAD).

The Employment Scam Aegean Dataset (EMSCAD) is a publicly accessible collection featuring 17,880 job advertisements. Its purpose is to offer researchers a transparent view of the Employment Scam issue. Records in EMSCAD have been manually labelled and categorised into two groups: 17,014 legitimate job ads and 866 fraudulent ones, all published from 2012 to 2014.

S. No.	Column Name	Non-Null Count	Data Type
1	job_id	17880	int64
2	title	17880	object
3	location	17534	object
4	department	6333	object
5	salary_range	2862	object
6	company_profile	14572	object
7	description	17879	object
8	requirements	15184	object
9	benefits	10668	object
10	telecommuting	17880	int64
11	has_company_logo	17880	int64
12	has_questions	17880	int64
13	employment_type	14409	object

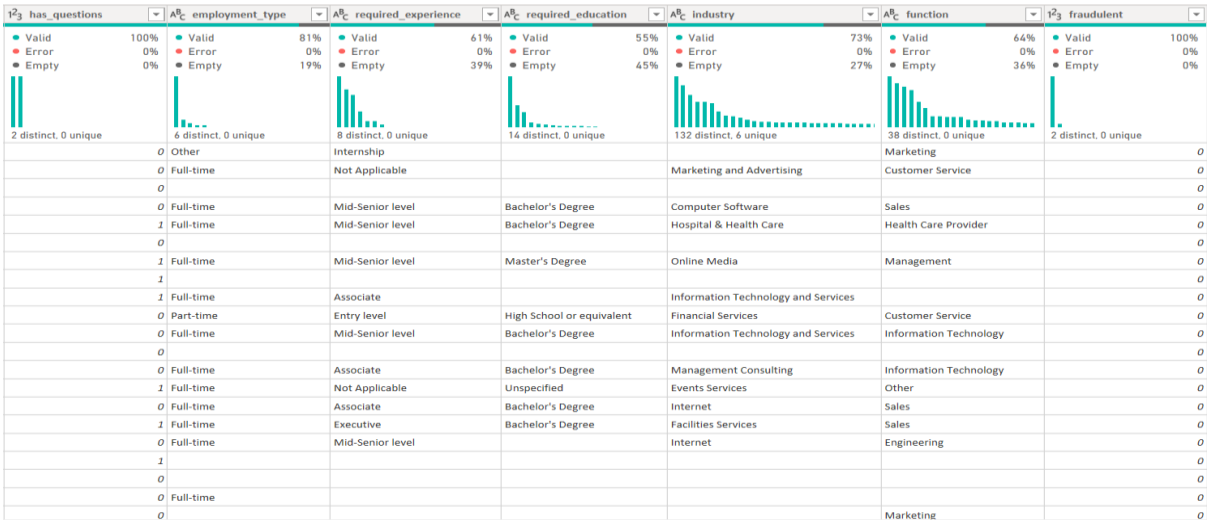
14	required_experience	10830	object
15	required_education	9775	object
16	industry	12977	object
17	function	11425	object
18	fraudulent	17880	int64

Table 1: Description of the dataset

Figure 1: Sample of dataset with column distribution and column quality

	job_id	title	location	department	salary_range	company_profile
Valid	100%	100%	98%	35%	16%	81%
Error	0%	0%	0%	0%	0%	0%
Empty	0%	0%	2%	65%	84%	19%
	17880 distinct, 17880 unique	11231 distinct, 9501 unique	3106 distinct, 1614 unique	1338 distinct, 815 unique	875 distinct, 502 unique	1710 distinct, 537 unique
1		Marketing Intern	US, NY, New York	Marketing		We're Food52, and we've created a groundbreaking and award-winnin...
2		Customer Service - Cloud Video Production	NZ, , Auckland	Success		90 Seconds, the worlds Cloud Video Production Service.90 Seconds is t...
3		Commissioning Machinery Assistant (CMA)	US, IA, Wever			Valor Services provides Workforce Solutions that meet the needs of co...
4		Account Executive - Washington DC	US, DC, Washington	Sales		Our passion for improving quality of life through geography is at the h...
5		Bill Review Manager	US, FL, Fort Worth			SpotSource Solutions LLC is a Global Human Capital Management Cons...
6		Accounting Clerk	US, MD,			
7		Head of Content (m/f)	DE, BE, Berlin	ANDROIDPIT	20000-28000	Founded in 2009, the Fonpit AG rose with its international web portal ...
8		Lead Guest Service Specialist	US, CA, San Francisco			Airenvy's mission is to provide lucrative yet hassle free full service sho...
9		HP BSM SME	US, FL, Pensacola			Solutions3 is a woman-owned small business whose focus is IT Service ...
10		Customer Service Associate - Part Time	US, AZ, Phoenix			Novitex Enterprise Solutions, formerly Pitney Bowes Management Ser...
11		ASP.net Developer Job opportunity at United States,New Jersey	US, NJ, Jersey City		100000-120000	
12		Talent Sourcer (6 months fixed-term contract)	GB, LND, London	HR		Want to build a 21st century financial service?We're convinced that th...
13		Applications Developer, Digital	US, CT, Stamford			Novitex Enterprise Solutions, formerly Pitney Bowes Management Ser...
14		Installers	US, FL, Orlando			Growing event production company providing staging, scenic, and dra...
15		Account Executive - Sydney	AU, NSW, Sydney	Sales		Adthena is the UK's leading competitive intelligence service for Google...
16		VP of Sales - Vault Dragon	SG, 01, Singapore	Sales	120000-150000	Jungle Ventures is the leading Singapore based, entrepreneur backed, ...
17		Hands-On QA Leader	IL, , Tel Aviv, Israel	R&D		At HoneyBook we're re-imagining the events industry and building a p...
18		Southend-on-Sea Traineeships Under NAS 16-18 Year Olds Only	GB, SOS, Southend-on-Sea			Established on the principles that full time education is not for everyo...
19		Visual Designer	US, NY, New York			Kettle is an independent digital agency based in New York City and the...
20		Process Controls Engineer - DCS PLC MS Office - PA	US, PA, USA Northeast			We Provide Full Time Permanent Positions for many medium to large ...
21		Marketing Assistant	US, TX, Austin			IntelliBright was created to leverage enterprise level online business p...

	description	requirements	benefits	telecommuting	has_company_logo
Valid	100%	85%	60%	100%	100%
Error	0%	0%	0%	0%	0%
Empty	0%	15%	40%	0%	0%
	14802 distinct, 13612 unique	11970 distinct, 10911 unique	6208 distinct, 4985 unique	2 distinct, 0 unique	2 distinct, 0 unique
	Food52, a fast-growing, James Beard Award-winning online food com...	Experience with content management systems a major plus (any blogg...		0	1
	Organised - Focused - Vibrant - Awesome!Do you have a passion for cu...	What we expect from you:Your key responsibility will be to communic...	What you will get from usThrough being part of the 90 Seconds team ...	0	1
	Our client, located in Houston, is actively seeking an experienced Com...	Implement pre-commissioning and commissioning procedures for rota...		0	1
	THE COMPANY: ESRI – Environmental Systems Research InstituteOur p...	EDUCATION: Bachelor's or Master's in GIS, business administration, or...	Our culture is anything but corporate—we have a collaborative, creati...	0	1
	JOB TITLE: Itemization Review ManagerLOCATION: Fort Worth, TX ...	QUALIFICATIONS:RN license in the State of TexasDiploma or Bachelors...	Full Benefits Offered	0	1
	Job OverviewApex is an environmental consulting firm that offers stab...			0	0
	Your Responsibilities: Manage the English-speaking editorial team and ...	Your Know-How: ...	Your Benefits: Being part of a fast-growing company in a booming indu...	0	1
	Who is Airenvy?Hey there! We are seasoned entrepreneurs in the hea...	Experience with CRM software, live chat, and phones, including one ye...	Competitive Pay. You'll be able to eat steak everyday if you choose to. ...	0	1
	Implementation/Configuration/Testing/Training on:HP Service Health ...	MUST BE A US CITIZEN.An active TS/SCI clearance will be required.Add...		0	1
	The Customer Service Associate will be based in Phoenix, AZ. The right...	Minimum Requirements:Minimum of 6 months customer service relat...		0	1
	Position : #URL_86fd830a95a64e2b30ceed829e63fd384c289e4f01e3c...	Position : #URL_86fd830a95a64e2b30ceed829e63fd384c289e4f01e3c...	Benefits - FullBonus Eligible - YesInterview Travel Reimbursed – Yes	0	0
	TransferWise is the clever new way to move money between countrie...	We're looking for someone who:Proven track record in sourcing acros...	You will join one of Europe's most hotly tipped startups with plenty of ...	0	1
	The Applications Developer, Digital will develop or participate in the d...	Requirements:4 – 5 years' experience in developing and deploying we...		0	1
	Event Industry Installers Needed!! (Orlando, FL - near Florida Mall)Do ...	Valid driver's license,Somewhat Clean driving recordIf you can drive a ...		0	1
	Are you interested in a satisfying and financially rewarding role in a hig...	You'll need to be smart and passionate and have 2 years experience se...	In return we'll pay you well, give you some ownership in the company ...	0	1
	About Vault Dragon Vault Dragon is Dropbox for your physical stuff - a ...	Key Superpowers3-5 years of high-pressure sales experience, but if yo...	Basic: SGD 120,000Equity negotiable for a rock starGround floor oppor...	0	1
	We are looking for a Hands-On QA Leader for our talented R&D t...	Previous experience in client & server testingExperience in Leadin...		0	1
	Government funding is only available for 16-18 year olds.We have 10 v...	16-18 year olds only due to government funding.Career prospects	Career prospects.	0	1
	Kettle is hiring a Visual Designer!Job Location: New York, NYKettle is a ...			0	1
	Experienced Process Controls Engineer is required having responsibilit...	Must have 5 or more years of experience with DCS programming, trou...		0	0
	IntelliBright is growing fast and is looking for a Marketing Assistant to j...	Job RequirementsAssist in creating client online marketing campaigns...		0	1



has_questions	employment_type	required_experience	required_education	industry	function	fraudulent
Valid 100% Error 0% Empty 0%	Valid 81% Error 0% Empty 19%	Valid 61% Error 0% Empty 39%	Valid 55% Error 0% Empty 45%	Valid 73% Error 0% Empty 27%	Valid 64% Error 0% Empty 36%	Valid 100% Error 0% Empty 0%
2 distinct, 0 unique	6 distinct, 0 unique	8 distinct, 0 unique	14 distinct, 0 unique	132 distinct, 6 unique	38 distinct, 0 unique	2 distinct, 0 unique
0	Other	Internship			Marketing	0
0	Full-time	Not Applicable		Marketing and Advertising	Customer Service	0
0						0
0	Full-time	Mid-Senior level	Bachelor's Degree	Computer Software	Sales	0
1	Full-time	Mid-Senior level	Bachelor's Degree	Hospital & Health Care	Health Care Provider	0
0						0
1	Full-time	Mid-Senior level	Master's Degree	Online Media	Management	0
1						0
1	Full-time	Associate		Information Technology and Services		0
0	Part-time	Entry level	High School or equivalent	Financial Services	Customer Service	0
0	Full-time	Mid-Senior level	Bachelor's Degree	Information Technology and Services	Information Technology	0
0						0
0	Full-time	Associate	Bachelor's Degree	Management Consulting	Information Technology	0
1	Full-time	Not Applicable	Unspecified	Events Services	Other	0
0	Full-time	Associate	Bachelor's Degree	Internet	Sales	0
1	Full-time	Executive	Bachelor's Degree	Facilities Services	Sales	0
0	Full-time	Mid-Senior level		Internet	Engineering	0
1						0
0						0
0	Full-time					0
0					Marketing	0

Features	**Description**
Job_id	The unique job ID for each job post is used to identify each job posting.
Title	The title of the job posting entry.
Location	The location of the job posting.
Department	Department in which the job is posted.
Salary_range	Salary offer range (e.g., $50,000-$60,000)
Company_profile	Information about the company.
Description	Detailed information about the job post.
Requirements	Requirements necessary for the job.
Benefits	Benefits offered in the job.
Telecommuting	True/False for telecommuting positions.
Has_company_logo	Job posting Company have a logo or not?
Has_questions	Had questions during the job application
Employment_type	Permanent or contract based.
Required_experience	Experience required for the job
Required_education	Education required for the job
Industry	The job post is from which type of industry?
Function	Role of the Employee

Table 2: Attribute details of the used dataset

4. Data Pre-processing

In Machine Learning and Deep Learning based approaches, the initial step required is data preprocessing, which greatly impacts the ML and DL models' performance. Different techniques were used for data preprocessing. The dataset contains numerical and textual data that must be preprocessed separately. The numerical attributes like telecommuting, has_company_logo, and has_questions have only two values, 0 and 1, so they didn't require any preprocessing and can be directly fed into an ML or DL model. Category columns like department, employment_type, etc., must first be converted into numerical attributes using One Hot Encoding (OHE) or label encoding. Ordinal Encoding can be used for Features like required_experience and required_education. In most works done till now, one-hot encoding was preferred. The textual columns like company_profile, description, requirements, and benefits must first be converted into numerical representations (Feature Vectors). Many methods are available for this purpose, like Bag of Words (BOW), Term Frequency-Inverse Document Frequency

(TF-IDF), Word2Vec embedding, Glove embedding, BERT/Sentence Transformers, and One Hot Encoding. The initial works focused on the BOW and TF-IDF technique [4] for the numerical representations, but in recent times, Word2Vec [9] has been most widely used in different works. BERT [8] has also been tested in some tasks. A comparative table is provided in Table 5.1 for selecting the best technique depending on the Requirements. Before applying the word embedding techniques, the text needs to be pre-processed a little bit, like lowercasing, removing punctuation, tokenization, stopword removal, and Lemmatization.

Technique	Definition	Typical Use Cases
Bag of Words	Represents the text as a vector of word counts, ignoring grammar and word order.	Simple text classification (e.g., spam detection), fast prototyping.
TF-IDF	Extends BOW by weighing terms based on the frequency and inverse document frequency.	Document classification and information retrieval where term importance matters.
One-Hot Encoding (OHE)	Represents each word as a binary vector with only one high (1) value per word.	Very simple models, vocab size is small; token identification tasks.
Word2Vec	Learns dense vector representations of words based on surrounding context.	Capturing word similarity and relationships, semantic search.
Glove	Like Word2Vec but uses global word co-occurrence statistics for vectors.	Word-level semantic tasks work well with fixed vocabulary.
Bert	Contextual language model that generates dynamic word embeddings based on context.	Complex NLP tasks: sentiment analysis, question answering, NER, etc.

Table 3: Comparison of Common Text Representation Techniques in NLP

We can also apply feature engineering to select or remove some columns depending on the correlation with the target column. If the target has very low dependency on a particular feature, then we can drop that feature from our dataset. The dependency can be calculated using the correlation matrix or a chi-squared test. The less correlated features will not play any significant role in model prediction. In almost all the works related to fake job post detection, the feature engineering technique is used, where some columns that are not correlated with the target are dropped. After all these pre-processing steps, the data is ready for the ML or DL models.

5. Machine Learning Approaches in Fake Job Post Detection

Different approaches were used for fake job post detection based on Machine learning and Deep Learning techniques. Most of them have also utilized natural language processing concepts in their studies. After applying the pre-processing steps and feature engineering mentioned above, researchers have tried to detect fake job posts using different ML models like logistic regression, Naïve Bayes, Decision Tree, and SVM. Many researchers have also explored the ensemble methods, using Random Forest and XGB classifiers. Among the ML techniques, the researchers found that Logistic Regression performs best. Still, the ensemble-based techniques have outperformed all the other models, delivering the best performance among all ML models.

Ensemble methods are a way of improving the performance of machine learning models by combining several models. Instead of relying on just one model to make predictions, ensemble methods combine the outputs of multiple models to make a final decision. The idea is that while one model might make mistakes, a group of models working together can often get things right more consistently.

Types of ensemble methods:

- Bagging: This method creates multiple model versions by training each one on a different random sample of the data. Then, their predictions are averaged (for numerical tasks) or voted on (for classification). A common example is the Random Forest, which combines many decision trees.
- Boosting: In boosting, models are trained one after another, and each new model tries to fix the errors made by the ones before it. Over time, the system gets better at handling tough cases. Well-known boosting methods include AdaBoost and XGBoost.
- Stacking: This method combines completely different models (like a decision tree, a logistic regression, and a neural network) and then uses another model — called a meta-model — to learn the best way to combine their predictions.

People mainly use ensemble methods because they often perform better than a single model. They can reduce overfitting, improve accuracy, and make predictions more reliable, especially on complex tasks like detecting fake job postings, spam filtering, or predicting stock prices.

6. Deep Learning Approaches in Fake Job Post Detection

The fake job posts look very similar to the original ones. Here, the ML models cannot learn those more complex patterns for distinction between genuine and fake job posts, so the Deep Learning models are used for this purpose. Also, ML models do not capture the contextual information of the textual data, but DL models like LSTM, GRU, and Transformers can do that. DL models often require more resources than ML models, but usually perform well. So, considering the risk of ORF DL models, they are preferred to achieve the task. The researchers [9] have used LSTM, BiLSTM, and Transformer [8] models to classify fake job posts. These models are used to deal with sequential data, where models can capture the context of the sequential data, which helps the models perform better than ML models. Researchers have also tried using the CNN and BERT for feature extraction and feeding the output to an ML model for classification. This technique was also performing well, but the deep learning models like BiLSTM and transformers performed best among all the methods.

7. Performance Evaluation

Different evaluation metrics are used to evaluate the models' performance, such as accuracy, precision, recall, F1 score, ROC-AUC, and confusion matrix. Almost all [6] [8] researchers have used all these evaluation metrics to evaluate the model's performance. Dinh [9] has also used metrics like sensitivity, specificity, and other measures not mentioned above. Naude [4] has used Matthews' corr. Coeff other than accuracy, precision, recall, and F1 score for evaluating the model [10].

Accuracy: The percentage of correct predictions out of all predictions.

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN)$$

Precision: out of all the predicted positive cases, how many are positive

$$\text{Precision}=TP/(TP+FP)$$

Recall: out of all actual positive cases, how many did the model correctly identify

$$\text{Recall}=TP/(TP+FN)$$

F1 Score: The harmonic mean of precision and recall.

$$\text{F1 score}=2*(\text{precision}*\text{recall})/(\text{precision}+\text{recall})$$

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): Measures how well the model can distinguish between classes.

Confusion Matrix: A Matrix showing correct and incorrect predictions by category.

Matthews Correlation Coefficient (MCC): A balanced measure that considers TP, TN, FP, and FN. Good for imbalanced data. Value ranges from -1 (worst) to 1 (perfect); 0 means random guessing.

Metric	Best For	Limitation
Accuracy	Balanced datasets	Misleading with imbalanced data
Precision	When false positives are costly (e.g., spam)	Ignores false negatives
Recall	When false negatives are costly (e.g., medical)	Ignores false positives
F1 score	Balancing precision and recall	Lower when one value (precision or recall) is low
ROC-AUC	Binary classification, model discrimination	Not intuitive for multi-class problems
Confusion matrix	Understanding prediction breakdown	Does not summarize performance in a single number
MCC (Matthews Correlation Coefficient)	Imbalanced datasets, binary classification	More complex to interpret

Table 4: Usability and limitations of different metrics

8. Challenges in Fake Job Post Detection

Researchers are trying very hard to deal with the ORF through Fake Job Post Detection, but they face some critical challenges in achieving the goal. Firstly, there is a lack of datasets available, and the available ones have also not been updated. The technique of Fraud Job Post may have evolved, but the dataset does not contain the updated entries, making ML or DL models inefficient for detecting Fake Job Posts.

The available dataset is highly imbalanced, with 17014 records of genuine job posts and 866 records of fake job posts, causing the ML and DL models to give a biased classification. Different techniques for handling the imbalanced data are being used, but the best case will be to get balanced data so that the models can give the best and accurate performance.

The dataset also lacks features like Unique Business Identifier (UBI), Companies posting frequency, etc, which can improve the model performance if added.

Job posts in different regions of the world are posted in various languages, which causes difficulty in identifying fake job posts. The models trained on an English dataset will not perform well on job posts in other languages.

9. Opportunities in Fake Job Post Detection

There is strong potential for the practical implementation of fake job detection systems in popular recruitment platforms like LinkedIn or Glassdoor. These platforms can integrate machine learning models to automatically scan and flag suspicious job listings before they go live. This protects users and improves the platform's reputation, ensuring a safer and more reliable environment for job seekers and employers.

There is a clear opening for applying fake-job-detection tools on major recruiting sites such as LinkedIn, Indeed, and Glassdoor. By weaving machine-learning models into their workflows, these sites could

quietly review and flag questionable listings before they even appear to the public. Doing so would shield users, boost the platform's credibility, and create a smoother, more trustworthy space for job seekers and employers.

When trained on patterns gathered over time, fake-job detectors can serve as early-warning panels that spot repeat offenders. Should one employer account reappear with dodgy or flagged ads, the tool signals moderators or quietly blocks further posts. That heads off large-scale schemes and helps keep job sites tidy and trustworthy.

A lively feedback loop lets users flag odd listings and feeds fresh examples to the model, allowing it to learn on the fly. Those crowd-sourced labels are later used to retrain or tweak the engine, sharpening its eye for new tricks. The human-in-the-loop setup keeps the system relevant and responsive.

10. Future Scope

In this fast-growing job industry, fake job post methods have also evolved. There is a need for a robust system for fake job post detection to integrate it with the job posting platforms to flag the post as fake or real in real time. This can warn users of fraudulent job posts and help the platforms manually verify the authenticity of posts that are flagged as fake. Once verified, the job posts can be added to the database with the class label for the Models to keep learning the evolving patterns of fake job posts to maintain the model's efficiency with future job posts.

Enhanced dataset collection to improve model training and evaluation requires more comprehensive, multilingual, and balanced datasets with labels. Some features, like the business license number, which is unique for each company, and the company posting frequency, need to be integrated to better model performance in detecting fake job posts. There is a huge requirement for a multilingual fake job post detection system, as people in different areas prefer to use other languages for job posts. With the data added to the currently available data, we can analyze different ML and Dl models' performance to investigate any improvements concerning the Present system.

There is a major issue of class imbalance in the present dataset. We can try using different techniques to solve the class imbalance problem and investigate the effects of class imbalance on the model's performance.

11. Conclusions

This review paper presents a comprehensive study of the landscape of fake job post detection. It examined the traditional ML Techniques, ensemble-based techniques, Deep Learning techniques, and transformer-based techniques, highlighting how each method contributes to tackling the fake job post detection Challenge. This paper also explores the dataset, pre-processing techniques, evaluation metrics, challenges in the domain, and opportunities.

The detection of fake job postings is now a social necessity as the rise in job platforms has increased the chances of scams with job seekers. This has increased the necessity of the development and deployment of a robust detection system. The efficient and scalable systems can help the platforms maintain credibility and protect users from harm.

As the tactics used by scammers continue to evolve, the detection techniques need to develop in parallel for fake job post detection. The future work should focus on developing real-time and generalized models that can be deployed. There is a requirement for collaboration between researchers, industry, and policymakers of the country to advance in this area of study.

Funding source

No funding was received for this study.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] 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, doi: https://doi.org/10.3390/fi9010006.

[2] H. Ahmed, I. Traore, S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Technique," In: Traore, I., Woungang, I., Awad, A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science (), vol 10618. Springer, Cham. https://doi.org/10.1007/978-3-319-69155-8_9

[3] S. Dutta and S. K. Bandyopadhyay, "Fake Job Recruitment Detection Using Machine Learning Approach," International Journal of Engineering Trends and Technology, vol. 68, no. 4, pp. 48–53, Apr. 2020, doi: https://doi.org/10.14445/22315381/ijett-v68i4p209s.

[4] M. Naudé, K. J. Adebayo, and R. Nanda, "A machine learning approach to detecting fraudulent job types," AI & SOCIETY, May 2022, doi: https://doi.org/10.1007/s00146-022-01469-0.

[5] F. H. A. Shibly, U. Sharma, and H. M. M. Naleer, "Performance comparison of two class boosted decision tree snd two class decision forest algorithms in predicting fake job postings," Annals of the Romanian Society for Cell Biology, vol. 25, no. 4, pp. 2462–2472, Apr. 2021.

[6] A. Amaar, W. Aljedaani, F. Rustam, S. Ullah, V. Rupapara, and S. Ludi, "Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches," Neural Processing Letters, vol. 54, Jan. 2022, doi: https://doi.org/10.1007/s11063-021-10727-z.

[7] H. Afzal, Furqan Rustam, Wajdi Aljedaani, Muhammad Abubakar Siddique, S. Ullah, and I. Ashraf, "Identifying fake job posting using selective features and resampling techniques," Multimedia tools and applications, vol. 83, no. 6, pp. 15591–15615, Jul. 2023, doi: https://doi.org/10.1007/s11042-023-15173-8.

[8] K. Taneja, J. Vashishtha, and S. Ratnoo, "Fraud-BERT: transformer based context aware online recruitment fraud detection," Discover Computing, vol. 28, no. 1, Feb. 2025, doi: https://doi.org/10.1007/s10791-025-09502-8.

[9] D.-H. Vu, K. Nguyen, K. T. Tran, B. Vo, and T. Le, "Improving fake job description detection using deep learning-based NLP techniques," Journal of Information and Telecommunication, pp. 1–13, Aug. 2024, doi: https://doi.org/10.1080/24751839.2024.2387380.

[10] M. Z. Naser and A. H. Alavi, "Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences," Architecture, Structures and Construction, Nov. 2021, doi: https://doi.org/10.1007/s44150-021-00015-8.