International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0361
e ISSN: 2584-2854
Volume: 02
Issue: 08 August 2024
Page No: 2499-2504

Real or Fake Job Posting Detection

K. Sridevi[1], G. Likitha[2], P. Chandana[3], Shrutika Shamarthi[4]

[1]Assistant Professor, G Narayanamma Institute of Technology and Science, Hyderabad, India.

[2,3,4]Student, G Narayanamma Institute of Technology and Science, Hyderabad, India.

Emails: *kandulasridevia2@gmail.com[1], golilikithareddy0905@gmail.com[2], punnachanduchandana@gmail.com[3], shrutikashamarthi11@gmail.com[4]*

Abstract

This research presents a machine learning approach to distinguish between legitimate and fraudulent job postings in the recruiting sector. The dataset used, labelled as 'authentic list,' comprises approximately 17,880 entries from Kaggle and includes various attributes such as job title, location, salary range, company profile, job description, industry, and indicators of fraudulent activity in job advertisements. The proposed methodology begins with Exploratory Data Analysis (EDA) to gain insights into the multi-class classification of different features and to identify correlations within the dataset. Data pre-processing techniques, including Natural Language Processing (NLP), are employed to prepare the datasets for training and testing. Several machine learning algorithms such as K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest, Logistic Regression, Naive Bayes, and AdaBoost are used to classify job listings as legitimate or fraudulent. The performance of each classifier is evaluated using qualitative metrics such as accuracy, precision, recall, F1-score, selectivity, and specificity. The results show the effectiveness of the system, achieving an accuracy of 99.20% in classifying job postings using the Random Forest classifier.

Keywords: *Classification; Natural Language Processing; Real or Fake Job.*

1. Introduction

An inventive use of machine learning, the Real or Fake Job Posting Classification System [1] aims to improve the dependability and effectiveness of job advertising platforms. There is a requirement of Natural language processing [2] for creating a reliable system that can automatically differentiate between genuine job advertisements and fraudulent or deceptive ones. In this growing internet world there are many frauds apart from this problem [3&4]. The following process is used in our proposal to detect real or fake job. **Data collection**: Compile a wide range of job advertisements, including instances that are authentic and those that are fake. The machine learning models will be trained and assessed using this dataset as the basis.

Feature extraction: Feature extraction is the process of removing pertinent information from job listings, such as the application process, company profile, salary range, and job description. The machine learning models will be trained using these features. **Labelling:** Add labels to the dataset that specify whether or not each job posting is authentic. The machine learning models will be trained and validated using this labeled dataset. Integration: Provide a system that is integrated and enables users to post job openings for categorization. Users should be able to spot possible scams with the use of the system's real-time feedback on the legitimacy of the job posting.

User Interface: Create an intuitive user interface for the system that classifies job postings so that people can interact with it and comprehend the outcomes with ease. **Scalability and Maintenance:** Take into account scalability and maintenance over the long run by putting in place a system that can manage an increasing number of job posts and

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0361
e ISSN: 2584-2854
Volume: 02
Issue: 08 August 2024
Page No: 2499-2504

adjust to shifts in the labour market.

2. Related Work

Machine learning techniques have been utilized in various studies to detect fraudulent job postings. Aashir Amaar [5] employed six different machine learning models to assess the authenticity of job advertisements. They evaluated the effectiveness of each model by combining Bag-of-Words (BoW) and TF-IDF features. To address the imbalance in the dataset, where the number of genuine and fraudulent job postings differs significantly, they utilized the adaptive synthetic sampling (ADASYN) strategy. This approach artificially increases the number of samples in the minority class, helping to improve classifier performance and mitigate overfitting. Two experiments were conducted, one utilizing a balanced dataset and the other an unbalanced one. Through experimental study, ETC was able to extract features with TF-IDF and over-sample with ADASYN, yielding 99.9% accuracy. Additionally, our suggested strategy is thoroughly compared with alternative resampling methods and the most recent deep learning models in this paper. A model to assess the effectiveness of a rule-based highlights model with a sack-of-words model was given by Sokratis and al. [6]. They discovered that an Irregular Woodland classifier employing the latter model could predict dishonest job ads with 91% accuracy. Cardoso Durier da Silva et al.'s survey [7] mostly used data from social media to investigate if machines can be trained to recognize bogus news. The objectives of this work are to define false news, map the current state of fake news identification, and determine which machine learning techniques are most frequently utilized. They concluded that the most often used method for automatically detecting false news is a combination of conventional machine learning algorithms linked by a neural network. Sultana et al. [8] employed both deep learning (Deep Neural Network) and traditional machine learning methods (SVM, KNN, Naive Bayes, Random Forest, and MLP) to predict fraudulent job postings. They found that the Random Forest Classifier achieved the highest accuracy at 99%, while deep neural networks achieved an average accuracy of 97.7%. Research by [9] also used various machine learning techniques, with the Random Forest Classifier achieving nearly 98% accuracy. Anita [10] utilized machine learning and deep learning methods to classify genuine and fraudulent job postings using a large dataset. Techniques included random forest, KNN classifier, logistic regression, and Bi-Directional LSTM for training neural networks.

3. Method

A thorough method for creating a sophisticated prediction model using machine learning techniques is described in the methodology section for "Real or fake job posting detection."

- Prior to loading the dataset, EDA and data preprocessing are carried out to tidy and arrange the data and obtain insights into multi-class categorization.
- Currently, 20% is used for testing and 80% is for training.
- Next, NLP is used to categorize the terms used in genuine and fictitious job descriptions.
- Machine learning models are evaluated using datasets, and their effectiveness is measured using metrics like recall, accuracy, precision, F1-score, selectivity, specificity, and support.

3.1 Dataset Description

The training and testing datasets for this study were sourced from the Kaggle data repository, focusing specifically on distinguishing between real and fake job postings as shown in Figure 1. The dataset consists of approximately 17,880 job listings, each with various attributes such as job ID, title, department, location, pay range, company profile, job description, requirements, benefits, telecommuting options, company logo presence, presence of questions, employment type, required education, required experience, industry, function, and a binary 'fraudulent' indicator. The 'fraudulent' attribute takes binary values of 0 or 1, where 1 indicates a fraudulent job post and 0 indicates a legitimate job advertisement. This dataset comprises job listings gathered from multiple job posting websites.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0361
e ISSN: 2584-2854
Volume: 02
Issue: 08 August 2024
Page No: 2499-2504

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	job_id	title	location	departmer	salary_ran	company_	descriptio	requireme	benefits	telecomm	has_comp	has_quest	employme	required_e	required_e	industry	function	fraudulent
1	1	Marketing	US, NY, Ne	Marketing		We're Foo	Food52, a	Experience	with conte	0	1	0	Other	Internship			Marketing	0
2	2	Customer	NZ, , Auckl	Success		90 Second	Organised	What we e	What you	0	1	0	Full-time	Not Applicable		Marketing	Customer	0
3	3	Commissic	US, IA, Wever			Valor Servi	Our client,	Implement	pre-comm	0	1	0						0
4	4	Account E	US, DC, Wa	Sales		Our passio	THE COMF	EDUCATIC	Our culture	0	1	0	Full-time	Mid-Senio	Bachelor's	Computer	Sales	0
5	5	Bill Review	US, FL, Fort Worth			SpotSourc	JOB TITLE:	QUALIFICA	Full Benefi	0	1	1	Full-time	Mid-Senio	Bachelor's	Hospital &	Health Car	0
6	6	Accountin	US, MD,				Job Overview	Apex is an	environr	0	0	0						0
7	7	Head of C	DE, BE, Be	ANDROID	20000-280	Founded ir	Your Respo	Your Know	Your Bene	0	1	1	Full-time	Mid-Senio	Master's D	Online Me	Manageme	0
8	8	Lead Gues	US, CA, San Francisco			Airenvyâ€	Who is Air	Experience	Competitiv	0	1	1						0
9	9	HP BSM SN	US, FL, Pensacola			Solutions3	Implement	MUST BE A	US CITIZE	0	1	1	Full-time	Associate		Information	Technolc	0
10	10	Customer	US, AZ, Phoenix			Novitex Er	The Custo	Minimum	Requiremer	0	1	0	Part-time	Entry level	High Schoc	Financial S	Customer	0

Figure 1 Sample of The Data

3.2 Exploratory Data Analysis

To understand the structure, patterns, and relationships between variables, the process involves visually examining and representing the data. The main goal of exploratory data analysis (EDA) is to gain insights that can effectively guide further data processing and analysis. EDA uses techniques such as scatter plots, heatmaps, and histograms to visualize correlations and patterns in the data. Through EDA, it is possible to perform multi-class classification, distinguishing attributes such as industry and employment type.

3.3 Data Preprocessing

An essential initial stage in data analytics and machine learning is preparing the data. It describes the procedure for converting unprocessed data into a format appropriate for modelling and analysis. Preparing and cleaning data so that it can be used for additional analysis is the main objective of data preparation. Several processes may be involved in data preparation, contingent on the type of data and the issue that has to be resolved.

4 Results and Discussion

To address underfitting concerns, the original dataset was modified by augmenting both legitimate and fraudulent job listings from additional sources, increasing the proportion of fraudulent jobs from 5% to 15%. The proportion of legitimate jobs remained unchanged, but the number of fraudulent jobs has climbed from 866 to 3044 as shown in Figures 2 & 3. Of the tasks in this dataset, 15% (15,800) are utilized for testing, while the remaining 20% (3,900 jobs) are used for training. In the subsections that follow, the constructed model's qualitative and quantitative outcomes are discussed.

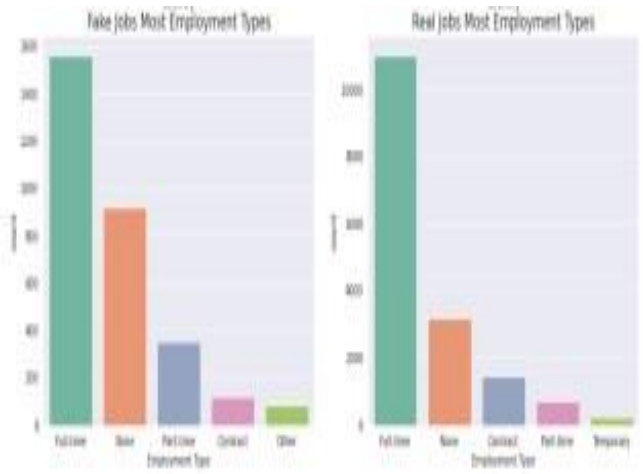

Figure.2. Number of Real and Fake Jobs in Various Employment Types

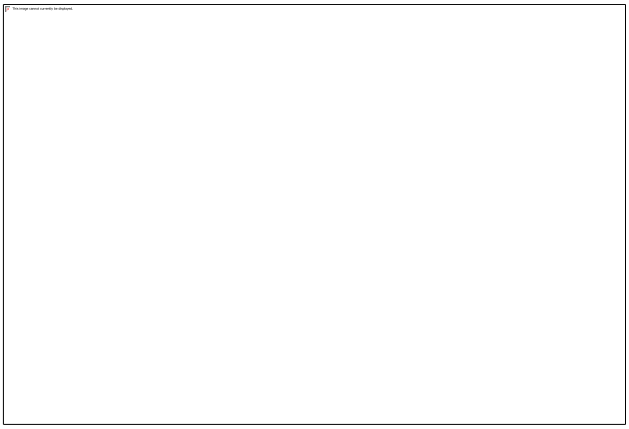

Figure 3 Number of Real and Fake Jobs in Various Industries

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0361
e ISSN: 2584-2854
Volume: 02
Issue: 08 August 2024
Page No: 2499-2504

4.1 Qualitative Results

The quantity of actual and fictitious jobs across a range of industries and occupational categories that was discovered during the dataset's EDA. Real jobs come from the IT sector, while the accounting sector generates the majority of phony positions. It appears that full-time work has the greatest concentration of both legitimate and fraudulent positions.

4.2 Quantitative Results

To address underfitting concerns, the original dataset was modified by augmenting both legitimate and fraudulent job listings from additional sources, increasing the proportion of fraudulent jobs from 5% to 15%. The proportion of legitimate jobs remained unchanged.

4.3 Results for Training and Testing Datasets

Tables 1 and 2 provide concise summaries of the evaluation metrics for the training and testing datasets using various machine learning classifiers in the proposed model. The model is denoted as M in these tables, with metrics including accuracy, precision, recall, F1-score, specificity, selectivity, and represented by abbreviations such as NB (Naive Bayes), LR (Logistic Regression), and RF (Random Forest). In this context, '0' indicates legitimate job postings, while '1' indicates fraudulent job postings.

Table 1 Evaluation Metrics of Machine Learning Models for Training Data

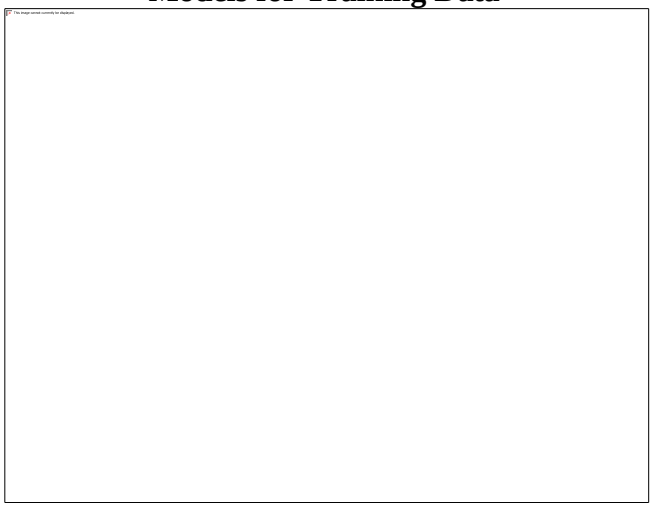

Table 2 Evaluation Metrics of Machine Learning Models for Testing Data

M	A	P		R		F1		SP	SE
		0	1	0	1	0	1		
NB	0.94	0.98	0.75	0.95	0.88	0.96	0.81	0.94	0.88
LR	0.99	0.99	0.95	0.99	0.94	0.99	0.96	0.98	0.92
KNN	0.93	0.99	0.70	0.93	0.96	0.96	0.81	0.92	0.95
	0.87	0.98	0.55	0.87	0.90	0.92	0.69	0.86	0.90
RF	0.99	0.99	0.98	0.99	0.93	0.99	0.96	0.99	0.96
	0.99	0.99	0.98	0.89	0.93	0.99	0.97	0.99	0.93
SVM	0.98	0.99	0.93	0.99	0.97	0.99	0.95	0.98	0.96

4.4 Authentic and Deceptive Words

It is recommended that users proceed with caution while reading job descriptions in order to identify relevant terminology such as "make solid," "cash related terms," "easy to work with," and financial incentives as in Table.3.

Table 3 Real or Fake Words

REAL WORDS		FAKE WORDS	
• Develop • Analyst • Specialist • Director • Office manager	• Project manager • Work environment • Business process • Knowledge • Communication skills etc.,.	• Custom service • Home • Earn • Daily • Want urgent	• Posit earn • Posit work • Base payroll • Cash • Work easy • Benefit include • Make solid • Provide efficiency etc.,.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0361
e ISSN: 2584-2854
Volume: 02
Issue: 08 August 2024
Page No: 2499-2504

4.5 Comparison of the Proposed Model with Existing Models

A detailed comparison of accuracy for each machine learning model is shown in Figure. 4. Our proposed model notably outperformed existing techniques, achieving the highest accuracy of 99%. This improvement is attributed to refining the dataset by adding more fake job listings to address underfitting issues. Additionally, preprocessing techniques were employed to handle incomplete feature sets and missing data. Additionally, the suggested approach expands its capabilities beyond the prediction of actual and phony jobs by providing classification based on employment type.

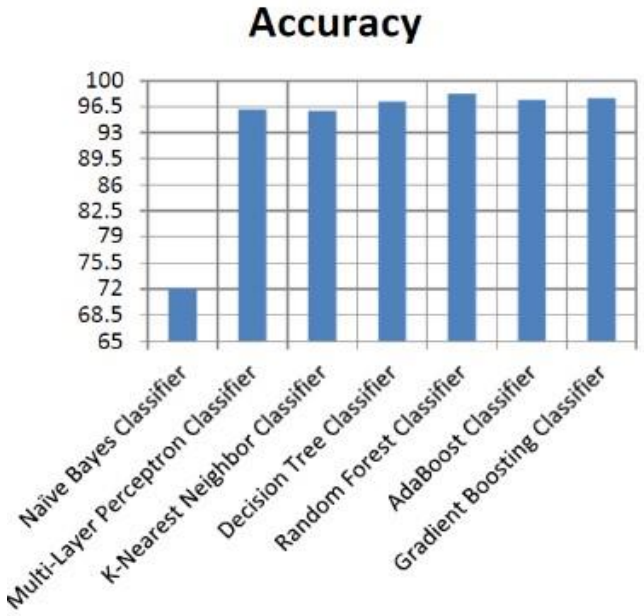

Figure 4 Comparison of Accuracy for All Specified Machine Learning Model

Conclusion

In this study, we present a model that utilizes various machine learning classifiers and Natural Language Processing (NLP) to differentiate between real and fake jobs. The findings highlight the performance of the random forest classifier, achieving a selectivity of 96% and an accuracy of 99.2%. To further enhance classification performance, future research could explore more advanced machine learning models and leverage larger datasets.

References

[1]. E. Baraneetharan, 'Identifying Fake Job Listings with Machine Learning Algorithms,' Journal of Artificial Intelligence, Vol. 4, No. 3, 2022, pp. 200-210.

[2]. Hefu Liu, Qian Huang, Yue Zhao, Cai Chee-Wee Tan, and Kang. 'Natural Language Processing (NLP) in Management Research: An Overview of the Literature,' Journal of Management Analytics, No. 2 (2020), pp. 139–172.

[3]. D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla. 'Machine Learning Methods for Credit Card Fraud Detection,' Proceedings of the 18th International Symposium, 2019.

[4]. Vidros, Sokratis, Georgios Kambourakis, Constantinos Kolias, and Leman Akoglu. 'Automated Detection of Online Recruitment Frauds: Characteristics, Techniques, and a Public Dataset,' Future Internet, 2017, Vol. 9, Issue 1, Article 6..

[5]. Stephanie Ludi, Aashir Amaar, Wajdi Aljedaani, Furqan Rustam, Saleem Ullah, and Vaibhav Rupapara. 'Using Machine Learning and Natural Language Processing to Detect Fake Job Postings,' Letters on Neural Processing M. Schoenberger, "Exploratory data analysis." 27 No. 5 (1979): 563-564 in IEEE Transactions on Acoustics, Speech, and Signal Processing.

[6]. Cardoso Durier da Silva, Ana Cristina Garcia, Fernando, and Rafael Vieira. 'Can Robots Detect Fake News? A Social Media Survey,' 2019.

[7]. Farzana Tasnim, Habiba Sultana Umme, and Md. Khairul Islam. 'Comparative Analysis of Fake Job Post Prediction Using Various Data Mining Techniques,' Proceedings of the 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), IEEE, 2021.

[8]. Using machine learning, Reddy, Y. Venkataramana, B. Sai Neeraj, K. Puneeth

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0361
e ISSN: 2584-2854
Volume: 02
Issue: 08 August 2024
Page No: 2499-2504

Reddy, and P. Bhargav Reddy developed an online fake job advertisement detection application. Engineering Sciences Journal, Volume 14, Issue 03 (2023).

[9]. Anita, C. S., G. Aditya Sairam, P. Ganesh, G. Deepakkumar, and P. Nagarajan. "Fake job detection and analysis using machine learning and deep learning algorithms." 11 no. 2 (2021): 642650 Revista Geintec-Gestao Inovacao e Tecnologias.

[10]. Kumar, Ankit."Self-Attention GRU Networks for Fake Job Classification." Innovative Science and Research Technology International Journal 6, no. 11 (2021).