Exposing Fraudulent Job Advertisements using the Power of Machine Learning Models

NABIL HOSSAIN CHOWDHURY* and MITUL ROY TANNY*, BRAC University, Computer Science Department, Bangladesh

Job finding in modern times have become greatly accessible to the general public mainly due to the internet. Both the companies and the job seekers receive this benefit. Companies are easily able to list a job online, which are easily and quickly accessed by job seekers. Unfortunately, such ease of access also allows malicious actors to bait unaware seekers into scams that can end up costing the person lots of money. These fake job advertisements often tend to be attractive and when people apply for these jobs, there is often an application fee. So the job seeker gives the scammer both their money and their personal information. Our objective is to understand if a posted job is fraud or not with help from machine learning algorithms. Models will be trained on previous data of fake and legitimate postings. Several models will be used to indicate the performance and to show which models perform best at determining the legitimacy of an advertisement.

Additional Key Words and Phrases: Job, Advertisement, Fraud, Fraudulent, Machine Learning, NLP, training, model

1 Introduction

As opportunities to work online job portals and remote work increases, so have the people who will try to take advantage of this. Falling for such a scam leads to significant financial and emotional stress to job seekers, not to mention sending data to scammers. Thus, it is important to have a system in place that will help seekers to filter out false advertising and scam, therefore increasing efficiency and protecting the people. While it is possible to manually identify jobs that may be fraudulent, it is not going to be feasible in the actual market with millions of postings. Machine Learning offers an efficient solution to automate this filtration at a large scale.

Fraud and Scams in job advertisements not only harm the users, but also the platforms that host job advertisements. Genuine employees may be driven away by the existence of these false advertisements. So if the platforms can implement such an automated technique to remove the malicious advertisements, users will tend to put more trust on the system.

Fraud Jobs also affect the entire recruiting system as a whole economically. Resources are wasted for both job seekers and businesses, potentially leading to financial losses.

2 Related Work

There are plenty of research on detecting fake job posting with the help of machine learning. It originally emerged in 2016, where it was decided that there is a need for research on this topic and bring attention to the scammers who exploit job portals and steal information for malicious purpose.[4] It was stated that the increase in a more digitized system, Recruitment fraud will be more prevalent.

Subsequently more and more researchers decided to respond to this issue and expand on the previous works, while focusing on using different classifiers and models to optimize results.

Vidros et al.(2016)[8] showed two different groups of fraudulent jobs. One group tries to social engineer sensitive

*Both authors contributed equally to this research.

Authors' Contact Information: Nabil Hossain Chowdhury, nabil.hossain.chowdhury@g.bracu.ac.bd; Mitul Roy Tanny, mitul.roy.tanny@g.bracu.ac.bd, BRAC University, Computer Science Department, Merul Badda, Dhaka, Bangladesh.

information such as social security number from the job seeker or bait them into depositing money. Another group aims to steal personal information such as phone numbers, email, etc. Which might then be sold to other malicious parties that can use these information to make spam calls/emails/texts. A malicious person will try to disguise themselves into a legitimate employer and even might schedule interviews/assessments.

Swetha et al. (2023)[7] have suggested deep neural network to anticipate fraud. They used categorical characteristic of the EMSCAD (Employment Scam Aegean Dataset) dataset and applied training techniques. This method efficiently reduces the number of attributes for training. Naive Bayes classifier, K Nearest Neighbor, decision tree, random forest classifier, support vector machine and neural network models were compared on the dataset.

3 Fraud Detection

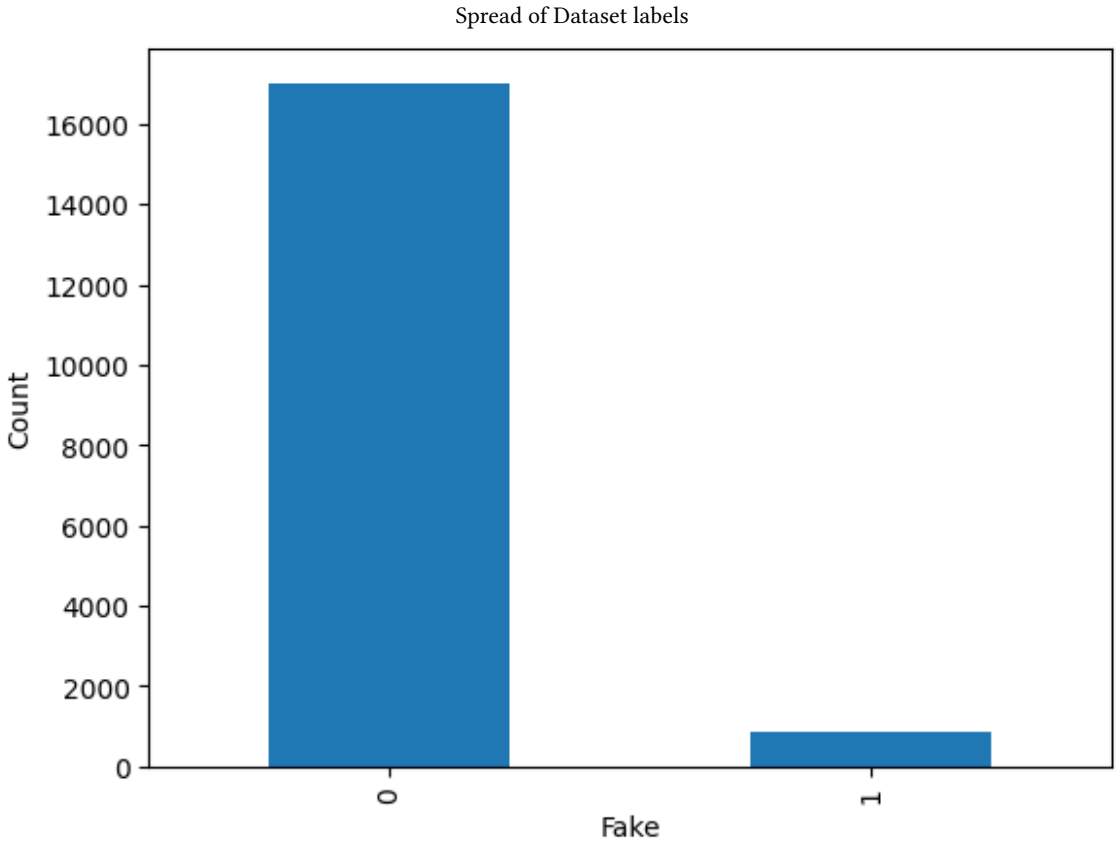
Fraud detection has been extensively studied in domains such as credit card transactions, insurance claims, and email phishing. Ayoyemi et al (2017)[2] developed a model to detect credit card fraud and made comparisons between different techniques. Fraud detection research in other domains also provide insights in detecting fraudulent behaviour with Machine learning models.

4 Machine Learning Models in Fraud Detection

Machine learning models are commonly used in fraud detection across different domains. Different Researchers compared different models such as Logistic Regression against a more modern model like neural networks to detect spams or fraud in financial transaction.

5 Dataset

EMSCAD has been used in many different research to train models to identify fake postings. The data has 18000 samples and 18 features in total. The features are description, requirements, benefits, telecommunication, has company logo, job id, title, location, company profile department, employment type, required experience, industry, function, salary range and required education along with the label, that shows whether a sampleis fraudulent or not. One problem with our dataset is that among the 18000 samples, only 800 of them are fake, so the data heavily skewed. Others have validated the work on the same dataset with some focusing on two main problems: performance improvement of prediction by trying out diferent models/classifiers (Lal et al (2019)[5] and Dutta and Bandhyopadhyay (2020)[3])



5.1 Natural Language Processing

NLP, or Natural Language Processing, has proven to be very effective in detecting different types of fraud where textual data plays a key role, such as fake reviews, phishing and other scams. Fraudulent activities often involve deceptive language and misleading information that are particularly targeting to bait unsuspecting people, which NLP techniques can uncover by analyzing text patterns, word usage, and context. However, we will use and compare some of the traditional machine learning models to see which of them provides more accurate results.

6 Classifiers

We used five different classifier models, which are Logistic Regression, Random Forest, Nearest Centroid, Support Vector Machines and Decision Tree.

6.1 Random Forest

Random Forest is a combination of several decision trees. Every tree is based on a subset of the features, and the final outcome is decided by combining the predictions from all of the trees. It reduces over fitting and improves accuracy, especially on complex datasets.

6.2 Decision Tree

A Decision Tree is a model that shows a tree structure, the nodes showing a feature, the branches representing a decision, and the leaf nodes representing a prediction or outcome. The data is split based on feature values to maximize information gain or reduce entropy.

6.3 Nearest Centroid

In the Nearest Centroid model, the classifier assigns a new sample to the class whose centroid (mean of all points in the class) is closest to the sample. It's a simple classification method based on the straight line distance (Euclidean) between points and centroids.

6.4 Support Vector Machines

SVM finds a hyperplane that separates data points of different classes in a multi-dimensional space. The goal is to create the maximum possible margin between classes, both non-linear and linear data is handled using kernel functions (polynomial, radial basis function, etc.).
The Support Vector Classifier (SVC) is a category of SVM, it similarly creates the hyperplane for dividing and maximizing the distance between the nearest data points from both classes, which are also called support vectors. In cases where data isn't linearly separable, SVC projects the data into higher dimensions, making it easier to separate classes.

6.5 Logistic Regression

Logistic Regression model calculates the probability that a particular sample belonging to a particular class. It uses the logistic function to output probabilities, which then thresholds to predict class labels. It is a linear model for binary classification, though it can be extended to multi-class problems.

7 Methodology

We collected our data from The University of the Aegean's Laboratory of Information and Communication Systems Security, the Employment Scam Aegean Dataset (EMSCAD). This is a huge dataset with about 18,000 job descriptions. The dataset contains several information about the jobs.

7.1 Data Preprocessing

Before we apply any models, the data is processed to transform it into a more machine readable form. In our dataset, there are several empty or N/A values, this does not help us in training, so we first identify if the feature that contains such a large number of N/A values are important for classification. The columns that are deemed not important have been removed, while others that have N/A values have been filled. We used the Simple Imputer module's mean calculation to fill the other features' N/A values. Many of the features have been converted into categorical features and the values have been encoded into numeric values for training the models. Other researchers, such as Adebayo et al.[6] chose to create a new categorical numeric column identity theft, real job, multi-level marketing and corporate identity theft. These are given numeric values. The 4 types define what type of data is being stolen and how.
Our objective is to create classification models that use the textual data to decide if a job advertisement is fraudulent or not. The models identify the key traits of a job advertisement to detect which ones are fraudulent in nature.

After the pre-processing stage, the dataset is divided in two parts, the testing set (20%), and the training set (80%). Now, as we mentioned earlier, our dataset is not balanced and has a much larger count of non-fraudulent jobs, this is a detriment as it can result in over fitting. To make sure none of the features are too dominant in deciding the label, we use scaling. Scaling is a technique that is used for adjusting the range of features in a dataset so that they have comparable magnitudes. It is necessary because many models are sensitive to the data scale, and can give biased results if some of the features dominate the scale. We use Standard Scaler to scale our processed dataset.
Using the scaled data, we train each of the models.

8 Results

We evaluated our models in several metrics to ensure fair comparison.

8.1 Accuracy

Accuracy measures the proportion of accurate predictions (both true negatives and true positives) out of all the predictions. It gives an overall performance of the model but can be misleading when the data is not balanced. Since our dataset has an unbalanced class distribution, this is not a very good metric for our case.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

8.2 Precision

Precision calculates the ratio of accurately predicted positive cases out of total count of predicted positive cases. It displays the number of positive predictions that were actually correct. It is usually important when having false positives is costly.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

8.3 Recall

Recall is the ratio of the positive cases that were properly identified by the model. It shows how well the model captures positive instances. It is crucial when missing positive instances has a high cost.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

8.4 F1-Score

The F1-score is a combination recall and precision into a singular value to provide a balanced measure, especially useful dealing with imbalanced datasets, which is vital for our dataset. Unlike arithmetic mean, harmonic mean allows more weight to lower values. So this score is the harmonic mean calculation between recall and precision. Therefore, F1-score will be high given both Recall and Precision are high.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

8.5 Support Vector Classification

The Classification Report for Support Vector Classifier:

Classification Report:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	3395
1	0.89	0.17	0.29	181
accuracy			0.96	3576
macro avg	0.92	0.59	0.63	3576
weighted avg	0.95	0.96	0.94	3576

Fig. 1. Report for SVC

8.6 Random Forest Classifier

The Classification Report for Random Forest Classifier:

Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	3395
1	0.96	0.67	0.79	181
accuracy			0.98	3576
macro avg	0.97	0.84	0.89	3576
weighted avg	0.98	0.98	0.98	3576

Fig. 2. Report for Random Forest

8.7 Logistic Regression

The Classification Report for Logistic Regression:

Classification Report:

	precision	recall	f1-score	support
0	0.95	1.00	0.98	3395
1	0.91	0.06	0.10	181
accuracy			0.95	3576
macro avg	0.93	0.53	0.54	3576
weighted avg	0.95	0.95	0.93	3576

Fig. 3. Report for Logistic Regression

8.8 Nearest Centroid

The Classification Report for Nearest Centroid:

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.78	0.87	3395
1	0.15	0.70	0.24	181
accuracy			0.78	3576
macro avg	0.56	0.74	0.56	3576
weighted avg	0.94	0.78	0.84	3576

Fig. 4. Report for Nearest Centroid

8.9 Decision Tree

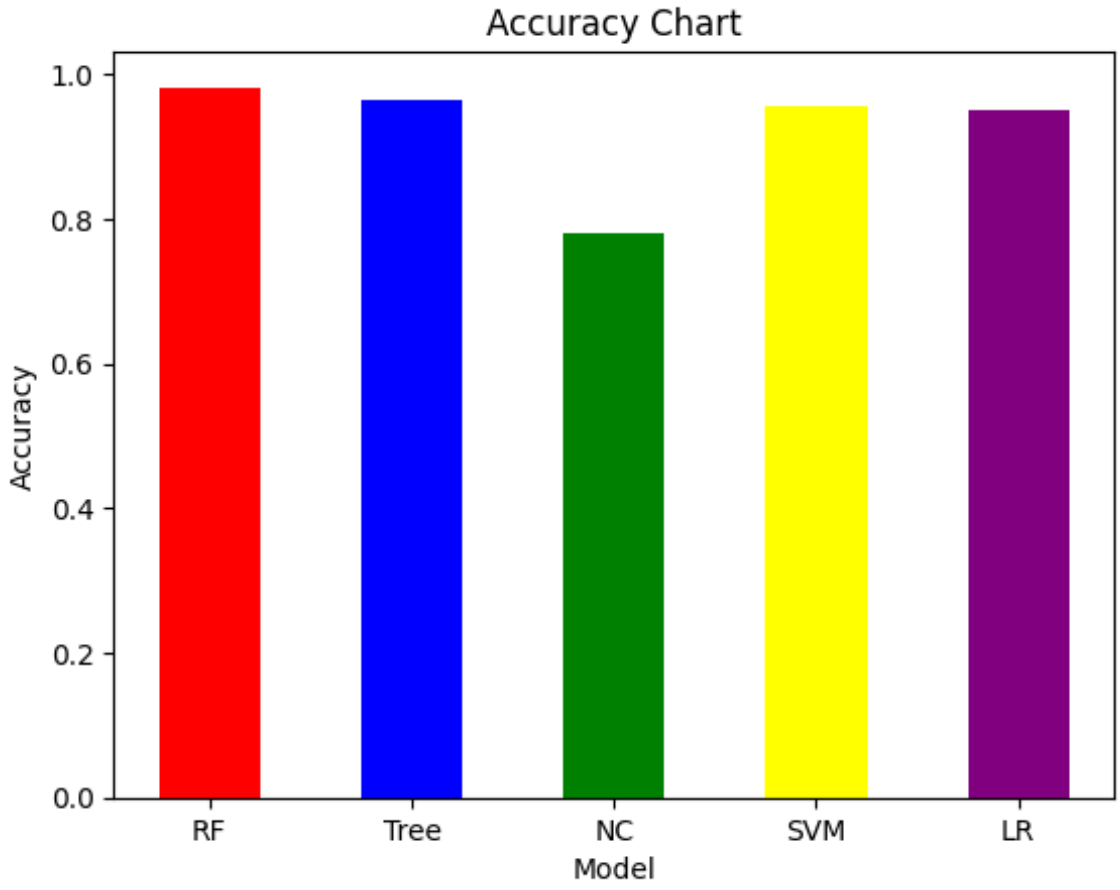
The Classification Report for Decision Tree:

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	3395
1	0.65	0.69	0.66	181
accuracy			0.97	3576
macro avg	0.81	0.83	0.82	3576
weighted avg	0.97	0.97	0.97	3576

Fig. 5. Report for Decision Tree

After making comparison of each model, we created the following chart:



Our Results show that Random Forest Classifier is the best performing model with an accuracy of 98%, likewise Decision Tree (95%), SVM (96%) and Logistic Regression (95%) all performed decent with over 90% accuracy.

9 Conclusion

In our paper, we investigated the use of five different machine learning models for detecting fraudulent job advertisements. We implemented and put the performance of Logistic Regression, Random Forest, Support Vector Classifier, Nearest Centroid, and Decision Tree classifiers side by side for a labeled dataset of job postings and collected the results. Among these, Random Forest Classifier achieved the highest among the models with an accuracy of 98%, showing its effectiveness in identifying deceptive patterns in job advertisements.
Our research contributes to the field of fraud detection by applying machine learning to a relatively less explored area of text-based fraud, job advertisement fraud. The strong performance of Random Forest highlights its potential for practical use in detecting fraudulent job postings on online job posting platforms or portals. Implementing a machine learning model such as these could significantly reduce the number of fraudulent advertisements encountered by job seekers by filtering and removing the fake jobs, improving both the security and trustworthiness of these platforms.

9.1 Limitations

Despite promising results from our code, our study has some limitations. The dataset size, while sufficient for this study, could be expanded to further validate the generalization of our findings. Also, the dataset has a very one sided label, with only about 800 of the 18000 samples being fake. A more balanced dataset would help better train the models. To overcome this limitation, Amaar et al. (2022) [1] used an approach called the adaptive synthetic sampling, this approach balances among the target classes by artificially creating the number of samples for the minor class.
Our study focused exclusively on English-language job postings, and future research should investigate the model's applicability to other languages and regions to better access those that fall victim to fake job advertisements.

9.2 Our Future Objective

In the future, we plan to explore larger, more diverse datasets, as well as models with multiple languages to enhance fraud detection capabilities globally. Additionally, implementing non-textual features, such as user interaction data and external reputation systems, could make the detection system more robust. These enhancements may further boost accuracy and help combat fraud more effectively in the real-world.

10 Acknowledgments

Acknowledgments

We would like to express our gratitude to Annajiat sir and Sabbir sir for their invaluable guidance and support throughout this research.

References

[1] Aljedaani W. Rustam F. et al. Amaar, A. 2022. Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches. *Neural Process Lett 54* (2022), 2219–2247. https://doi.org/10.1007/s11063-021-10727-z

[2] Adetunmbi A. O. Oluwadare S. A. Awoyemi, J. O. 2017. Credit card fraud detection using machine learning techniques: A comparative analysis. *In 2017 international conference on computing networking and informatics (ICCNI)* (2017), 1–9.

[3] Bandyopadhyay S. K. Dutta, S. 2020. Fake job recruitment detection using machine learning approach. *International Journal of Engineering Trends and Technology* (2020), 48–53.

[4] Utekar A. Dhonde A. Karve S. S. Khandagale, P. 2022. Fake Job Detection Using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology* 10, 4 (2022), 1822–1827. https://www.academia.edu/download/99408900/ijraset.2022.pdf

[5] Sardana N Verma A Kaur A Mourya R Lal S, Jiaswal R. 2019. ORFDetector: ensemble learning based online recruitment fraud detection. *2019 Twelfth international conference on contem porary computing (IC3). IEEE* (2019), 1–5.

[6] Adebayo K.J. Nanda R Naudé, M. 2023. A machine learning approach to detecting fraudulent job types. *AI Soc 38* (2023), 1013–1024. https://doi.org/10.1007/s00146-022-01469-0

[7] Sravani K. Swetha, K. 2023. Fake job detection using machine learning approach. *Journal of Engineering Sciences* (2023), 67–74.

[8] Kambourakis G Vidros S, Kolias C. 2016. Online recruitment services: another playground for fraudsters. *Comput Fraud Secur 2016* (2016), 8–13.