International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Fake E Job Posting Prediction Based on Advance Machine Learning Approachs

Ali Raza^{a,}, Saqib Ubaid^b, Faizan Younas^c, Farhan Akhtar^d*

^{a,b,c}Department of Computer Science, Khwaja Fareed University of Engineering & Information Technology, Rahim Yar Khan, 64200, Pakistan.

^dDepartment of Mathematics, Khwaja Fareed University of Engineering & Information Technology, Rahim Yar Khan, 64200, Pakistan.

DOI: <https://doi.org/10.55248/gengpi.2022.3.2.7>

ABSTRACT

There are many jobs adverts on the internet, even on reputable job posting sites, that never appear to be false. However, following the selection, the so-called recruiters begin to seek money and bank information. Many candidates fall into their traps and lose a lot of money as well as their existing job. As a result, it is preferable to determine whether a job posting submitted on the site is genuine or fraudulent. Manually identifying it is extremely difficult, if not impossible! An automated online tool (website) based on machine learning-based categorization and algorithms are presented to eliminate fraudulent job postings on the internet. It aids in the detection of bogus job postings among the vast number of postings on the internet.

Keywords: Fraudulent, Job, Machine learning, Real, Fake, Job advertisement, Classification.

1. Introduction

Employment scams are one of the more important concerns that have recently been addressed in the realm of Online Recruitment fraud. We are living in unprecedented times as a result of the COVID-19 epidemic, which is wreaking havoc on economies around the globe. Unemployment rates are rising daily, with the United States reporting over 26 million people, the most recorded in the country's long history.

In its most recent World Economic Outlook report, the IMF (International Monetary Fund) forecasts unemployment in Pakistan at 13% in 2020, up from 7.3 percent in 2019 and 3.9 percent in 2018. Many organizations now choose to list their job openings online so that job searchers can access them readily and quickly. However, this might be a form of scam perpetrated by fraudsters who promise work to job seekers in exchange for money.

To undermine a reputable company's credibility, fraudulent job adverts might be issued. These fraudulent job post detections pique people's interest in acquiring an automated tool for recognizing bogus jobs and alerting them to individuals to avoid applying for such positions. A machine learning technique is used for this goal, which utilizes numerous classification algorithms for detecting bogus postings. In this scenario, a classification tool detects and warns the user when it detects bogus job postings among a bigger set of job adverts.

According to a Dawn poll of 300 women, sexual harassment, abuse, and discrimination are widespread in Pakistan's workplaces, including universities, and are usually unreported and overlooked by senior management. When asked if women were forced to remain silent about workplace harassment, 61% stated their employers did not persuade them to do so, yet a substantial 35% were urged to be silenced by their colleagues and managers.

To begin addressing the challenge of spotting job advertising frauds (Alharby, 2019), supervised learning algorithms as classification approaches are being examined. A classifier uses training data to translate input variables to target classes. The classifiers covered in the research for distinguishing phony job postings from others are briefly outlined. These classifier-based forecasts may be divided into two types: single classifier-based predictions and ensemble classifier-based predictions.

* Corresponding author.

E-mail address: alirazaatofficial@gmail.com

2. Research contributions

This project will develop and deliver a new online automated tool/website using Python. This new online automated tool/website will display several job posts which will be real not contain any fake posts. The users will be able to apply online for a job that contains real jobs which will save their money cost and time. This new online automated tool/website will contain a huge database of job posts as a record. This online automated tool/website will be developed to enable additional features to be added to it over some time and be easy to maintain. This proposed solution contains a lot of advantages that will be proved fruitful for online job seekers. The proposed solution advantages are Easy to use, Time saving, cost-effective (Saving money), and contains a huge database of job posts.

3. Related work

Several studies have found that review spam detection (B. Biggio, 2011) email spam detection, and fake news identification have gotten a lot of interest in the realm of online fraud detection (Delany, 2007).

People frequently submit reviews of the things they buy on online forums. It might help other buyers decide which things to buy. Spammers can alter reviews to generate profit in this setting, hence approaches for detecting spam reviews are essential. This may be accomplished by extracting characteristics from the reviews using Natural Language Processing (NLP). The characteristics are then subjected to machine learning algorithms. Lexicon-based approaches might be an alternative to machine learning techniques that employ dictionaries or corpora to remove spam reviews.

Unwanted bulk communications, which fall under the category of spam emails, frequently arrive in user mailboxes. This may result in an unavoidable storage issue as well as increased bandwidth usage. To address this issue, Gmail, Yahoo Mail, and Outlook service providers have implemented spam filters based on Neural Networks (E. G. Dada, 2019). Adaptive spam filtering techniques are considered for tackling the problem of email spam detection. These approaches include content-based filtering, case-based filtering, heuristic-based filtering, memory or instance-based filtering.

Malicious user-profiles and echo chamber effects are characteristics of fake news on social media. The basic research of false news identification is based on three perspectives: how fake news is created, how fake news spreads, and how a user is connected to fake news. To detect false news, features linked to news content and social context are retrieved and a machine learning model is applied.

4. Literature review

The undertaking according to the literature research, no current system is still installed or functioning in the same way as this project is. Several research investigations are offered, however, there has never been a system like this before. The literature study gives valuable insight into the field of machine learning. People conduct research using a broad range of methodologies. There are several active studies on machine learning approaches (Knoll, DEC, 2013). This section's research findings show the potential benefits and popularity of machine learning. A survey of the literature allowed us to have a better understanding of the various 19 algorithms. The majority of the research discussed in this chapter has consistently uncovered the primary benefits of depending on machine learning technologies.

Table 1 - Literature review analysis.

Sr no.	Title	Author	Publication	Remark
1	Fake Job Recruitment Detection Using Machine Learning Approach.	Shawni Dutta, Prof. Samir Kumar	4-April 2020.	Naive Bayes, K-nearest Neighbor, Decision Tree Classifier.
2	Machine Learning and Job Posting Classification.	Ibrahim M. Nasser, Amjad H. Alzaanin	9-September 2020.	Multinomial Naive Bayes, Support Vector Machine, Decision Tree, Random Forest Classifier.
3	Smart Fraud Detection Framework for Job Recruitments.	Asad Mehboob, M. S. I. Malik	4-October 2020.	NB, KNN, DT, SVM, Random Forest (RF), XGBoost Classifier.
4	Comparative study on various algorithms for detection of fake job postings.	Dhanamma Jagli, Vishal Saroj Gupta	09-September 2020.	SVM, Logistic Regression, KNN, RF, DT Classifier.

5. Proposed Methodology

The job post dataset is utilized in this research study. To get accurate results from the proposed model dataset preprocessing is applied. The preprocessing involves text noise cleaning, data transformation, and data normalization. The TFIDF is utilized for feature engineering. The dataset splitting is conducted in the next phase. The 70% portion of the dataset is used for the proposed model and 30% is used for model testing and result evaluation. The proposed machine learning approach is fully hyperparameter tuned. The best fit model is deployed behind the fake job website. Now users can check fake jobs by just providing the job post URL.

Fig. 1 - Proposed system methodological architecture.

6. Job Post Dataset

The Employment Scam Aegean Dataset (EMSCAD) is a publicly accessible dataset including 17,880 real-life job advertising that intends to provide the academic community with a comprehensive view of the Employment Scam issue and can serve as a helpful testbed for scientists working in the field (Rish, January 2001). EMSCAD records were manually annotated and divided into two groups. The collection, in particular, comprises 17,014 real and 866 fraudulent job adverts issued between 2012 and 2014. The dataset may be used to train classification algorithms to recognize bogus job descriptions. This step indicates that the dataset is full and ready to train, test, and apply the Machine Learning model.

Table 2 - Dataset Attributes information.

Sr no.	Name	Description
1	Title	The title of the job ad entry.
2	Location	The geographical location of the job ad.
3	Department	Corporate department (e.g., sales).
4	Salary range	Indicative salary range (e.g., \$50,000-\$60,000)
5	Company profile	A brief company description.
6	Description	A detailed description of the job ad.
7	Requirements	Enlisted requirements for the job opening.

8	Benefits	Enlisted offered benefits by the employer.
9	Telecommuting	True for telecommuting positions.
10	Company logo	True if the company logo is present.
11	Questions	True if screening questions are present.
12	Fraudulent	Classification attribute.
13	In balanced	Selected for the balanced dataset
14	Employment type	Full-type, Part-time, Contract, etc.
15	Required experience	Executive, Entry level, Intern, etc.
16	Required education	Doctorate, Master's Degree, Bachelor, etc.
17	Industry	Automotive, IT, Health care, Real estate, etc.
18	Function	Consulting, Engineering, Research, Sales, etc.

6.1. Dataset Correlation Analysis

Correlation analysis is a widely used approach for detecting interesting correlations in data. These connections assist us in determining the significance of traits in relation to the target class to be forecasted. The findings reveal intriguing relationships in the fraudulent characteristic for job post categorization.

Fig. 2–Job Post Dataset Correlation Analysis.

7. Data Preprocessing

Data pre-processing is a data mining approach that entails converting raw data into a usable format. Real-world data is frequently inadequate, inconsistent, and/or deficient in specific behaviors or patterns, and it is rife with inaccuracies (Walters, 1988). Data pre-processing is a tried-and-true way for overcoming such problems. In the actual world, data is usually incomplete: it lacks attribute values, it lacks particular qualities of interest, or it just contains aggregate data. Noisy: including mistakes or outliers Inconsistent: having differences in codes or names.

8. Splitting dataset into Train and Test Set

In any Machine Learning model, we will divide the data into two independent sets: Training Set and Test Set. In this case, it is your algorithm model that will learn from your data to create predictions. In general, we divided the data set into 70:30 or 80:20 ratios. We have taken 70% of the data in train and 30% of the data in the test. However, the Splitting might vary depending on the form and size of the dataset.

9. Proposed Support Vector Machine

The goal of this method is to determine whether or not a job posting is fake. Identifying and deleting bogus job adverts will allow job seekers to focus solely on authentic job postings. The "Support Vector Machine" (SVM) is a supervised machine learning technique that may be used to solve classification and regression problems. It is, however, largely employed in categorization difficulties. In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of features), with the value of each feature being the value of a certain coordinate. Then, we accomplish classification by locating the hyperplane that best distinguishes the two classes (look at the below snapshot) [2]. Individual observation coordinates are used to calculate support vectors. The SVM classifier is a frontier that best distinguishes between the two classes (hyper-plane/line).

$$w^T x + b = 0 \quad (1)$$

$$f(x) = \sum_i a_i y_i (x_i^T x) + b \quad (2)$$

It is simple to create a linear hyperplane between these two classes in the SVM classifier. But another pressing concern is if we need to integrate this functionality manually to have a hyper-plane. No, the SVM method employs a technique known as the kernel trick. The SVM kernel is a function that takes a low-dimensional input space and changes it to a higher-dimensional space, converting a non-separable problem to a separable one. It is especially beneficial when dealing with non-linear separation problems. Simply told, it does some fairly sophisticated data transformations before determining how to divide the data based on the labels or outputs you've specified.

10. Human-Computer Interface (HCI)

The way of communication between a specific user and a computer system, namely the usage of input/output devices and related software. They must be set up in such a way that they promote an efficient and desired interaction between a human and a machine.

Fig. 3 - (a) website with URL search tool; (b) job post results related to a user query.

The development phase is discussed in this section. Describe the development of web pages. In addition, user interface samples from the website are displayed to provide a clear project view. This phase indicates that the project is finished and ready for testing and implementation.

11. Testing of the proposed system

The project is created, implemented, and tested in stages (a bit more is added each time) until the product is completed. It entails both development and upkeep. When a product meets all of its specifications, it is said to be done. First, we pre-processed the dataset before building models on top of it. It consists of model training and testing. We create the website's front end (user interface) using HTML, CSS, and the Flask Framework [5]. After completing that process, we tested the component to see if it works, whether it fits the surroundings, and if it is engaging in user interaction. The model connection was created in the second phase. After building that element, we verified that the dataset is responding to the queries and can be correctly edited. After doing that operation, we tested the final component. All of these phases were constructed one at a time, and each module was added one at a time. In this section, we wrap up all of the project's testing and development work. It is checked in several ways to ensure that there are no errors. Whether or not the system is operational. Test cases also show how they were carried out.

Table 3–Test case analysis.

Parameter	Value
TestCaseID	01
ApplicationName	Fake E Job Posting Prediction
Purpose	Todescribetheapplicationforminwhich the customer Provides a job webpage URL.
Environment	Python, Flask
Pre-Requisite:	Userwillbethecustomer
Strategy:	User will perform the following operation a. Provide Job Post webpage URL. b. Show Results.
Expected Result	Users should enter the main form with their relevant privileges.
Observations	The user has successfully entered the system having defined by the admin

Test Case ID	02
Application Name	Fake E Job Posting Prediction
Purpose	To describe the application form in which customers Check the job post is real or fake
Environment	Python, Flask
Pre-Requisite:	User will be the customer
Strategy:	The user will perform the following operation a. Submit the job post URL. b. check prediction results.
Expected Result	User should see their results successfully, if any incorrect information the error occurs.
Observations	The user has successfully entered the system having defined by the admin

12. Results and Evaluation

While data preparation and training a machine learning model are crucial steps in the machine learning pipeline, measuring the performance of this trained model is as important. The ability of the model to generalize on previously unknown data is what distinguishes adaptive machine learning models from non-adaptive machine learning models. We should be able to increase the overall prediction capability of our model before we push it out for production on unknown data by employing multiple measures for performance evaluation. Without doing a full evaluation of the ML model using various metrics, and relying just on the accuracy, it might lead to a problem when the related model is deployed on unknown data, resulting in bad predictions. Figure 4 depicts the Features analysis.

Fig. 4 - (a) country-wise job post analysis; (b) number of jobs with experience analysis.

The proposed classifier is trained and tested to recognize false job posts in a dataset that includes both fake and authentic posts. It provides us with an average accuracy of 97%. This covers all of the model's precision, recall, and f1-score results.

Table 4–Proposed approach results evaluation.

Category no.	Precision-score	Recall-score	F1-score	Support-score
0	0.97	1.00	0.98	5091
1	0.99	0.32	0.49	270

A confusion matrix is a depiction of the prediction outcomes of any binary testing that is frequently used to explain the performance of the classification model (or "classifier") on a set of test data with known true values. The confusion matrix itself is straightforward to grasp, but the language associated with it may be perplexing.

Fig. 5–Proposed approach confusion matrix analysis.

In this section, we wrap up all of the project's testing and development outcomes. It is tested in many ways to ensure that there are no errors and to determine model accuracy and outcomes.

13. Conclusion

Fake Job Postings on the Internet Prediction will help job searchers receive only real employment offers from firms. Several machine learning methods are offered as countermeasures in this Project to combat Online Fake Job postings. Finally, we conclude that our application will be the finest application in the We worked on those difficulties, fixed them, and applied the changes in our project. In terms of future work, we will undoubtedly work on it to meet the most recent technology and please our applicant. In the future, we will cover additional work eras and add more job categories to our application. That further modification will be performed in response to future requirements. For example, adding Scholarships posts.

Acknowledgements

We gratefully acknowledge the support of our institute KFUEIT and our research supervisors for their support and appreciation. We would like to thank everyone who had contributed to the successful completion of this research.

REFERENCES

- Alharby, B. A. (2019). An Intelligent Model for Online Recruitment Fraud Detection.
- B. Biggio, I. C. (2011). Bagging classifiers for fighting poisoning attacks in adversarial classification tasks.
- Delany, P. C. (2007). K -Nearest Neighbour Classifiers.
- E. G. Dada, J. S. (2019). Machine learning for email spam filtering: review, approaches and open research problems,.
- Knoll, A. N. (DEC, 2013). Gradient boosting machines a tutorial. vol. 7.
- Rish, I. (January 2001). —An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,.
- Walters, D. E. (1988). Bayes's Theorem and the Analysis of Binomial Random Variables.