Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Detection of Fake Job Advertisements using Machine Learning algorithms

E. Baraneetharan

Associate Professor & Head, Department of EEE, Surya Engineering College, Erode, India

E-mail: hodeee@surya.ac.in

Abstract

Most companies nowadays use digital platforms to host conferences, job interviews, and other business events. The unexpected increase in the need for internet platforms has resulted in a rapid rise of fraud advertising. The agencies as well as fraudsters recruit the job seekers using a variety of techniques, including sources from online job-providing websites. By applying Machine Learning algorithms, researchers aim to decrease the number of such fraudulent and fake attempts. In this article, classifiers such as K-Nearest Neighbour, Support Vector Machine, and Extreme Gradient Boosting algorithms are implemented for fake advertisement prediction. The performances of the machine learning algorithms are evaluated using metrics such as accuracy, F1 measures, precision and recall.

Keywords: Job interviews, fraudulent advertisements, machine learning algorithm, KNN, SVM, XGboost, performance metrics

1. Introduction

Employers post job vacancies on job portals, and candidates submit their applications online. Both the employers and the job seekers have found this online hiring procedure to be more time-efficient and less stressful. These procedures have certain disadvantages as well. Online recruitment-related crimes have recently been the subject of recorded instances. Malicious advertising, like email spam and false reviews, are unavoidably common on Internet as digital advertising becomes a significant channel for marketing. One of the severe challenges recently addressed in the area of online recruitment frauds is employment scam. A reputable company may be the target of fraudulent job adverts in order to affect their reputation. This scam misleads many people, who end up losing a lot of money. A machine learning methodology makes use of numerous classification algorithms in finding fake advertisements. The objective is to train a model that will provide accurate results when

DOI: https://doi.org/10.36548/jaicn.2022.3.006
Received: 21.08.2022, received in revised form: 19.09.2022, accepted: 05.10.2022, published: 14.10.2022
© 2022 Inventive Research Organization. This is an open access article under the Creative Commons Attribution-NonCommercial International (CC BY-NC 4.0) License

evaluating the authenticity of job postings and eliminating its related problems. In this research, many features are considered to train and test the Machine Learning (ML) algorithm such as K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGboost) in detecting fake advertisements.

This research article has been organized as follows. Introduction has been given in the first section. The related research works are provided in the second section, which emphasizes the significance of the study and explains the different methods employed in fake job advertisement detection. The proposed system is discussed in the third section, which demonstrates fraud detection using ML algorithms. The findings are summarized in the fourth section, and the conclusion is presented in the fifth section.

2. Related Works

Mehboob. A et al., [1] proposed a model for detecting fraud recruiting using gradient boosting algorithm, and it included a number of significant organisational, job-description, and kind of compensation features. The hybrid composition of 13 features performed better than three state-of-the-art baselines and displayed an accuracy of 97.94%. The analysis also discovered that "salary range," "company profile," "organization type," "required education," and "holding many jobs" are the best indicators. The conclusions highlighted a variety of study consequences and offered fresh information for identifying online employment fraud.

Amaar. A et al., [2] proposed an approach for identifying false job postings from job portals using supervised machine learning and natural language processing. The features were extracted from the data using the Term Frequency-Inverse Document Frequency (TF-IDF) as well as Bag-of-Words approaches (BoW). Through experimental analysis, TF-IDF was used for feature extraction and ADASYN for oversampling to obtain 99.9% accuracy.

Keerthana. B et al., [3] proposed a model for prediction of fraudulent job adverts. To increase the effectiveness of the model, the authors implemented feature engineering approaches such as one-hot encode, TFIDF Vectorizer, and count Vectorizer. They used classification techniques to forecast fraudulent job postings. The neural network, MLP Classifier, with the “Adam” model provided accurate result i.e., 71%, out of the full set of methods.

To determine whether a job posting is genuine or fake, S. U. Habiba et al., [4] suggested several methods of data mining and classification algorithms such as KNN, SVM,

Decision Tree, Naive Bayes classifier, Multilayer Perceptron, Random Forest classifier and Deep Neural Network (DNN). 18000 samples from Employment Scam Aegean Dataset (EMSCAD) were used in the article. For this classification challenge, a DNN classifier with three layers was used. A fake job advertisement can be predicted with a classification accuracy of about 98% by the trained classifier using DNN.

Chiraratanasopha. B et al ., [5] proposed a set of attributes created to replicate the actions of fraudsters who give false information, to identify fake job advertisements. Missing details, hyperbole, and credibility were the characteristics. The features were intended to be represented as a category and a readability score that could be generated automatically. A classification method for fraudulent job identification was trained using data from the EMSCAD dataset. To classify fraudulent job adverts, the proposed approach achieved 97.64% accuracy, 97% precision and 99 % recall score for its best model.

Ghosh. G et al., [6] analysed 4000 sample recruitment offers from various employment sites, of which 301 were false. To determine which model works best, numerous popular and contemporary categorization models were analysed. The techniques that have been employed include AdaBoost, Logistic Regression, Decision Tree Classifier, Voting Classifier, Random Forest (RF), LightGBM, and Gradient Boosting. According to data, the following prediction models have the highest accuracy: Logistic Regression (94.67%), AdaBoost (95%) Decision Tree Classifier (95%) Random Forest (95%) Voting Classifier (95.34%) LightGBM (95.17%), and Gradient Boosting (95.17%).

I. M. Nasser et al., [7] suggested an Artificial Neural Network-based model to identify fraudulent job postings. The model was trained and tested using the open source dataset (EMSCAD) and the appropriate text pre-processing methods. The model achieved 91.84% precision, 96.02% recall, and 93.88% f-measure.

Bandyopadhyay et al., [8] identified fake job advertisements among a vast number of postings. For the purpose of identifying fake job postings, two main categories of classifiers— single classifiers and ensemble classifiers, were taken into consideration. However, experimental findings showed that ensemble classifiers are superior to single classifiers in their ability to detect fraud. According to results, the Random Forest classifier outperformed comparable classification tools. The accuracy of the suggested strategy, which was substantially greater than the ones used previously, was 98.27%.

To identify whether a job posting is fraudulent or authentic, Devi. A P et al., [9] suggested a system that essentially used an Artificial Neural Network (ANN) classification method based on the Multinomial Naive Bayes algorithm. By using the dataset in a double-blind study and taking into account the various styles of posting jobs on professional websites as well as other websites, the model was trained to be as effective as possible. The suggested system has a User Interface (UI) frontend where the user types the URL into search box and clicks enter. The system on the backend, gathers data from the Web and feeds it into the ANN model, which utilises this algorithm to assess the reliability of the data.

Anita et al., [10] detected fake job postings using deep learning (Bi-LSTM) and machine learning (KNN, RF, LR) algorithmic techniques. When the authors compared the classification algorithms, they discovered that Bi-LSTM provided the most accurate result for spotting fake jobs.

3. Proposed Work

The proposed basic model is shown in Figure 1 as a staged methodology to address the categorization problem.

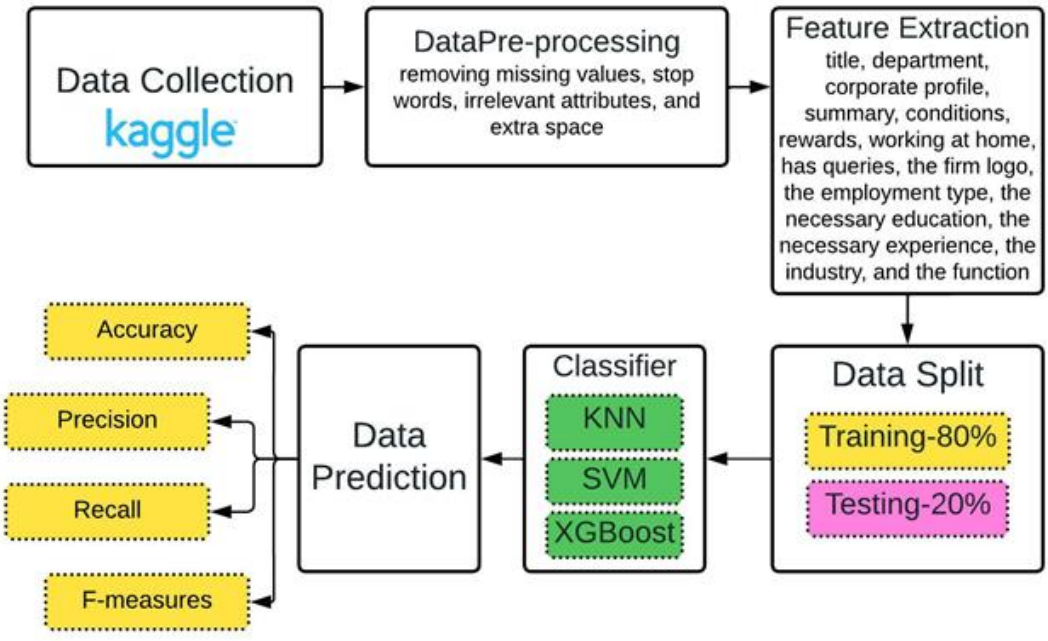

Figure 1. Proposed block diagram

The initial stage involves the gathering of information from Kaggle website. The gathered data are subjected to pre-processing stage, where the text data are processed to be cleared for further examination. After the pre-processing, the data are ready for upcoming

analysis, and the feature engineering techniques are applied to the training and testing processes to choose key features from the text. The dataset is divided into two different groups, where one is for testing and another for training after applying the features extraction technique. 80% data for training and 20% data for testing is the split ratio that has been implemented in the proposed method.

3.1 Data Collection

The proposed dataset contains 17,880 records of advertisements, 17,014 of which are authentic and 866 of which are fake job postings. The dataset is sourced from kaggle labelled as "Real / Fake Job Posting Prediction". Each advertisement is described in terms of the focussed attributes. The data then undergoes various pre-processing methodologies before it can be utilised as a source to any classification technique. The dataset needs to be pre-processed because it contains a lot of missing data and anomalies.

The below figure lists the attributes in evaluating the job prediction as fraud or authenticated.

```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   job_id               17880 non-null  int64
 1   title                17880 non-null  object
 2   location             17534 non-null  object
 3   department           6333 non-null   object
 4   salary_range         2868 non-null   object
 5   company_profile      14572 non-null  object
 6   description          17879 non-null  object
 7   requirements         15185 non-null  object
 8   benefits             10670 non-null  object
 9   telecommuting        17880 non-null  int64
 10  has_company_logo     17880 non-null  int64
 11  has_questions        17880 non-null  int64
 12  employment_type      14409 non-null  object
 13  required_experience  10830 non-null  object
 14  required_education   9775 non-null   object
 15  industry             12977 non-null  object
 16  function             11425 non-null  object
 17  fraudulent           17880 non-null  int64
dtypes: int64(5), object(13)
memory usage: 2.5+ MB
```

Figure 2. Attributes of the proposed dataset in the detection of fake job advertisement [11]

Since there are fewer fake job postings, this could be problematic because the algorithm might train to classify every job posting as authentic. Figure 3 shows the different

modes of job vacancy in the dataset. It is recognized that the majority of real and fraudulent jobs is of full-time employment.

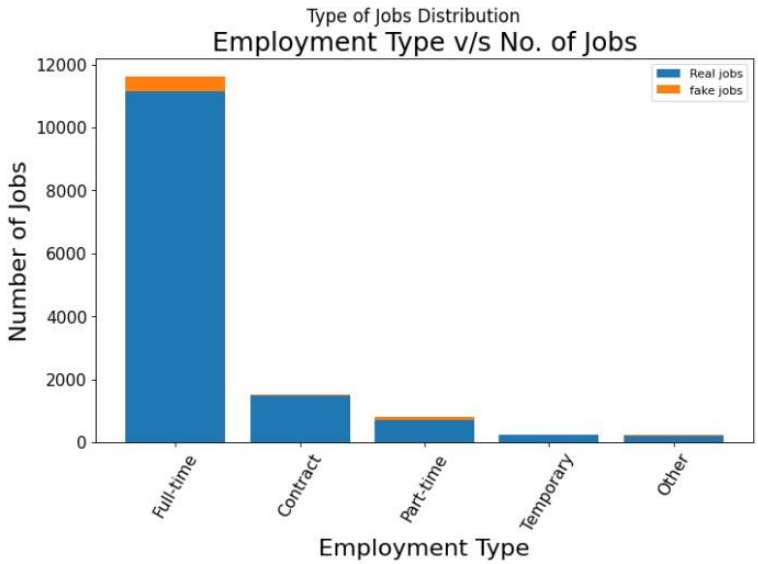

Figure 3. Mode of employment Vs No. of jobs [11]

3.2 Data Pre-processing

In this research, machine learning models are trained using text data. The significant volume of unclear content in the text sample is problematic for machine learning models that in turn affect the model performance of architecture. So, pre-processing of the proposed dataset is carried out to achieve the performance of the system model and to fit any machine learning algorithm. Pre-processing methods include removing missing values, stop words, irrelevant attributes, and extra space. The dataset is now perfect for categorical encoding in order to produce a feature vector.

3.3 Feature Extraction

The goal of the feature extraction technique is to extract useful set of features for the machine learning classification algorithms from the dataset. The number of redundant data in the dataset is decreased with the aid of feature extraction. In the end, the data reduction speeds up the learning and generalisation phases of the machine learning model while enabling the model to be built with less processing. This in turn reduces the training time taken by the algorithm. Considered features of the proposed dataset are title, department, corporate profile, summary, conditions, rewards, working at home, has queries, the firm logo, the employment type, the necessary education, the necessary experience, the industry, and the function.

3.4 Implementation of the classifier model

In this proposal, three supervised machine learning models are applied to the process of predicting future occurrences of fraudulent job postings. Classifiers are trained in this framework using the proper parameters. The default parameters for these models might not be enough to maximise their performance. The accuracy of this model, that may be considered as the most effective one for both identifying and separating the fraudulent job postings from the job seekers, are improved by adjusting these parameters.

3.4.1 KNN (K-nearest neighbour)

KNN is a very straightforward model that is used in machine learning to analyse regression and classification. The information is referred as classes within the nearest neighbours in the technique because it makes use of the available data and organizes the new data according to a distance function. The k-nearest neighbour model in this experiment provides good results when k value is equal to five (k=5). In other words, it looks up the five closest neighbours and decides based on the majority or closest distance [12].

3.4.2 SVM (Support vector machine)

Another ML model used in classifications and regression methods is the SVM. SVM regression is built on mathematical notation and begins with the non-parametric approach. Here, kernel transformation enables the entry of desired data regression problem determination using linear functions [13].

Table 1. Parameters and its specifications of the model

Models	Parameters	Specifications
K-Nearest Neighbour (KNN)	n_neighbours Weight	5 Uniform distribution
Support Vector Machine (SVM)	Cluster size Kernel	2.0 Linear
XGboost algorithm	Booster Verbosity max_depth n_estimators subsample number of folds	Gbtree Default=1 6 100 1 10

3.4.3 XGboost (Extreme Gradient Boosting)

The classification problem of online recruiting fraud detection is addressed using the XGBoost ML approach. The binary predictive algorithm is what comes out of this process. The XGBoost model is used for two reasons: (1) quick execution, and (2) good model performance. In general, XGBoost is quick enough in comparison to other gradient boosting implementations. It predominates in tabulated or structured datasets for tasks involving classification and regression. Gradient boosting, on the other hand, uses new models to anticipate the mistakes of previous models, which are then combined to provide the final prediction.

4. Results and Discussion

This section presents various experimental types that were executed to test the efficiency of the suggested model. The research involves selecting a meaningful subset of features, analysing fraud prediction, analysing features individually, and analysing attributes for fake advertisements. The output from each classifier is examined in order to determine which model is capable of outperforming others in terms of accuracy. A confusion matrix is created to assess the performance of the classification models if they perform equally well.

4.1 Evaluation metrics

The performances of the classification model are evaluated using the four metrics [15].

True Negative (TN) - Accurately categorise the false label

False Positive (FP) - The false class label is misclassified.

False Negative (FN) - The true class label is misclassified.

True Positive (TP) - Accurately categorise the true label

Accuracy: Accuracy $= (TP + TN) / (TN + TP + FN + FP)$

Precision $= TP/ (TP + FP)$

Recall $= TP/ (TP + FN)$

F-Measure $= (2 * TP) /(2 * TP + FP + FN)$

Table 2. Performance metrics of ML classifiers

Model/Metrics	Accuracy	Precision	Recall	F-measures
KNN	91.02	88.02	84.36	87.95
SVM	87.4	82.54	76.32	79.57
XGboost	98.53	97.23	97.52	98.54

The XGboost model performs efficiently than the other two traditional machine learning classification models for the subject features. According to experimental findings, XGboost classifiers outperform the other models as shown in Table 2 in terms of output quality. This classifier has 98.53% accuracy, precision of 97.23%, F1-score of 98.54, and recall of 97.52%. The optimal model for fake job advertisement detection system is therefore the XG boost classifier.

The below graph enhances the visual ability in analysing the performance of the models based on the evaluation metrics.

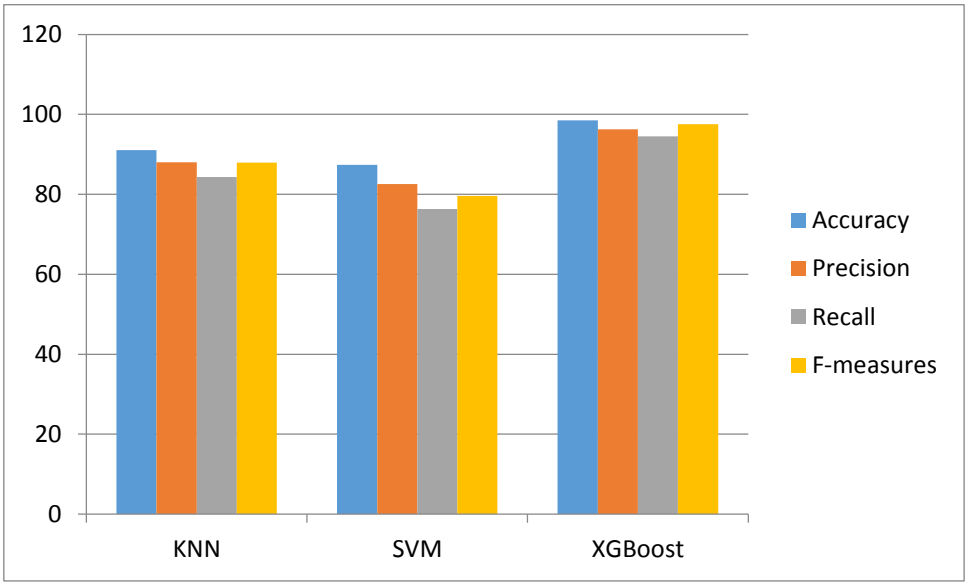

Figure 4. Performance metrics of the ML classifiers

5. Conclusion

Detecting job advertisement scams will lead job seekers to get genuine offers from companies. In this study, three machine learning algorithms are suggested as counters to fake job advertisement detection. For the purpose of scam detection, supervised learning classifiers are employed. This paper has employed three classifiers such as KNN,

SVM and XGboost algorithm with a dataset of 17,880 samples of 18 attributes. The proposed XGboost model achieves an accuracy of 98.53% which is comparatively higher than that of other classifiers, with minimum training time. In future, the model efficiency can be further increased by enhancing the features and evaluating the same with various datasets.

References

[1] Mehboob, A., Malik, M.S.I. Smart Fraud Detection Framework for Job Recruitments. Arab J Sci Eng 46, 3067–3078 (2021). https://doi.org/10.1007/s13369-020-04998-2

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

[3] Keerthana, B., Reddy, A.R., Tiwari, A. (2021). Accurate Prediction of Fake Job Offers Using Machine Learning. In: Bhattacharyya, D., Thirupathi Rao, N. (eds) Machine Intelligence and Soft Computing. Advances in Intelligent Systems and Computing, vol 1280. Springer, Singapore. https://doi.org/10.1007/978-981-15-9516-5_9

[4] S. U. Habiba, M. K. Islam and F. Tasnim, "A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques," *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 2021, pp. 543-546, doi: 10.1109/ICREST51555.2021.9331230.

[5] Chiraratanasopha, B., & Chay-intr, T. (2022). Detecting Fraud Job Recruitment Using Features Reflecting from Real-world Knowledge of Fraud. CURRENT APPLIED SCIENCE AND TECHNOLOGY, 12-pages.

[6] Ghosh, G., Tabassum, H., Atika, A., & Kutubuddi, Z. (2021). Detecting online recruitment fraud by using machine learning (Doctoral dissertation, Brac University).

[7] I. M. Nasser, A. H. Alzaanin and A. Y. Maghari, "Online Recruitment Fraud Detection using ANN," *2021 Palestinian International Conference on Information and Communication Technology (PICICT)*, 2021, pp. 13-17, doi: 10.1109/PICICT53635.2021.00015.

[8] Bandyopadhyay, Samir & Dutta, Shawni. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. International Journal of Engineering Trends and Technology. 68. 10.14445/22315381/IJETT-V68I4P209S.

[9] https://iarjset.com/wp-content/uploads/2021/08/IARJSET.2021.8857.pdf

[10] Anita C, Nagarajan P, Sairam GA, Ganesh P, Deepakkumar G (2021) Fake job detection and analysis using machine learning and deep learning algorithms. Revista Geintec-Gestao Inovacao e Tecnologias 11(2):642–650

[11] https://rishabh20118.medium.com/fake-job-posting-detection-and-getting-useful-job-posting-insights-from-the-dataset-e8edf1870831

[12] P. Cunningham and S. J. Delany, ―K -Nearest Neighbour Classifiers,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

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

[14] Le, H.; Pham, Q.; Sahoo, D.; Hoi, S.C.: URLnet: learning a URL representation with deep learning for malicious URL detection. arXiv preprint arXiv:1802.03162 (2018)

[15] H. M and S. M.N, ―A Review on Evaluation Metrics for Data Classification Evaluations,‖ Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

Author's biography

E. Baraneetharan is an Associate Professor and Head in the Department of EEE at Surya Engineering College, Erode, India. His area of research includes power electronics, electromagnetics, electric drives, IoT and data mining.