Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

# Real and Fake Job Posting Using Machine Learning Technique

J Refonaa [1], Keerthigha M [2], Arthi Nandhini [3], A Viji Amutha Mary [4], S L Jany Shabu [5]
[1]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
refonna.cse@sathyabama.ac.in
[2]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
[3]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
[4]Department of Information Technology,
Sathyabama Institute of Science and Technology, Chennai, India
vijiamuthamary.cse@sathyabama.ac.in
[5]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
janyshabu.cse@sathyabama.ac.in

***Article Info***
***Page Number: 562 - 570***
***Publication Issue:***
***Vol 71 No. 3s2 (2022)***

***Article History***
***Article Received****: 28 April 2022*
***Revised****: 15 May 2022*
***Accepted****: 20 June 2022*
***Publication****: 21 July 2022*

***Abstract***
Everything is being online now it allows people to reduce their manual efforts all of the job postings are posted online now so it gives the company a wide range of area to gather the talented candidates and for the people who are searching they can also know the information most companies can directly post the job. All job postings which are posted are not true there are fraudulent job postings also. So try to classify the fraudulent posting from real. The aim is to find the fraudulent job posting using Machine learning based concepts with correct accuracy. The analysis of dataset is done by Supervised ML to predict details like, variable finding, variate analysis, missed values treating and analysing data validation, cleaning data and visualizating the data would be carried on total dataset given.

**Keywords**: - Fake Job, Online Recruitment, Machine Learning, Accuracy, Random forest algorithm.

## I. INTRODUCTION

The data driven will follow and store the represented data from a similar articles for fraud job detection. But these will still have many drawbacks: Because of the difficulties in gathering fraud jobs, existing dataset's capacity are low, and there are unverified jobs which misses data in similar articles which makes to identify the credibility of methods tough.

Especially, difference between the fraud and true jobs are not limited to whether there is true feature in their similar article, but they include unseen difference at the linguistics level, such as the perspective of emotional expressions, writing flow. The method existing are hard to totally catch these difference. The proposed model has developed a machine learning model to find and split the original or fraud job posting to erradicate this method is to do a machine learning approach of GUI's

user interface. The dataset is first preprocessed and the dataset is used to do the needful rows are examine to view the dependent and independent variable and then various machine learnings techniques will be used to get the patterns and to get result with maximum accuries.

These reports are to the investigate of applicability of ml techniques for job posting classification. Atlast, it highspot few observation on the future findings, challenges, and their needs. The main idea is to do a model for original or fraud job Prediction finding result of matched accuracy by comparing them with supervised techniques.

## II. LITERATURE REVIEW

It is a form of data which tends to check the relevant points prevailing knowledge method wise approach to a significant theme. It is a another source of data and says published details in a particular topic and sometimes details in a specific topic in a span of time. Its main aim is to make the user to be with the recent knowledge with recent literature on data like heareafter the research which will use on places, follows presentation is a small collection files.

They provide an explanation of source / bind explanation and finds creative development in field, moreover it includes debate. On basis of the criteria, the review might estimate the files, tell user the most suitable / appropriate in these. It is a another source of data and says published details in a particular topic and sometimes details in a specific topic in a span of time.

### *REVIEW OF LITERATURE SURVEY*

*Title:* Predicting of Jobs Failures in Clouds Basis on Online Extreme Machine:

*Author:* CHUHONG LIU1, 2, JINGJNG HAN2, YNLEI SHANG1, CHANCHANG LIU1, BOJCHENG1, AND JNLIANG CHEN1

*Year:* 2017

Early finding of job fails and basic disposal on previous can constantly increase the competence of resource intake on large number data centers. The prevailing machine learning- methods collectively get the doing, working pattern, that cannot be taken for online guesses in practical data in which data arrive orderly. To eradicate these issues, a new model based on Online Extreme Learning Machine is presented in the paper to say and predict the online jobs terminations.

*Title:* Fraud Jobs Recruitment Finder Using MachineLearning Approach

*Author:* Samr Bandypadhyay, Shwni Duta

*Year:* 2020

To neglect fake post for jobs in the media, a tool which is automated using ML basis of methods are presented here. Various types are deployed for validating fake post for internet and final result for these are used in finding accurate scan finding model. This makes us easy for finding fraud work posting in a large value post collection. There are two main type of classifiers. However, experimental finding says that ensemmble classifier are the accurate division to find scam on the single only classifier.

*Title:* Classification of job posting

*Author:* Ibrahm M. Nassr1 and Amjd H. Alzaann2

*Year:* 2020

In this, we found various ML classifiers which includes the following: Naive Bayes, Random Forest, Decision Tree algorithm, Support Vector Machine, KNN of data classified statement. The data that are been using here has original and fraud job posting. Applied TFIDF for creature extract and processed all the data. Hereafter, Evaluated all the data after implementing the classifiers. Evaluated metrics used are precisionate, recalling, f-measured, and accurate. For every classifier, all result are done and compared with each others.

*Title:* Online Recruitment fraud automatic detection

*Author:* Sokrats Vdros 1, Cnstantinos Kolias 2

*Year:* 2017

The critical process of hiring has relatively recently been ported to the cloud. Especially, the automated systems that are responsible for finishing the recruitment process of new workers in an internet mode. The internet mode has a goal to make the hiring process more reliable and easy and accurate. But, the internet exposure of these business programs has proposed new procedures of failure which might cause some privacy losses for users and reduce organization reputation. So far, the employment scam is the most common case of Online Recruitment Fraudulent. Unlike these online fraudulent problem the way of pullying of Online recruitment fraudulent has not yet git any good attention, remaining mostly unexplored till now. At the same time, it lends and evaluates to ourself about the datasets that are available of 17,888 jobs adverts got from real life system.

*Title:* Enhanced RSA Algorithm using Fraud Modulusand Fraud Public Key Exponent

*Author:* Raghunanhan K R, Ganesh Aithal, Suendra Shetty,Rakshih N

*Year:* 2018

In data communication, cryptography Public key cryptography has a most important role. There are two different keys in Public key to encrypt data and decrypt data The Public key cryptography has a efficient algorithm called RSA .Efficiency of this is mostly dependent on how the public key components is shared that is modulus n and public key exponent e In this an enhanced RSA algorithm is presented to make high the complexity of the public keys paper have used unreal public key exponent r instead of e and Y instead of n. This Paper also uses standard metrics and gives comparative analysis of the proposed work.

## III PROPOSEDWORK

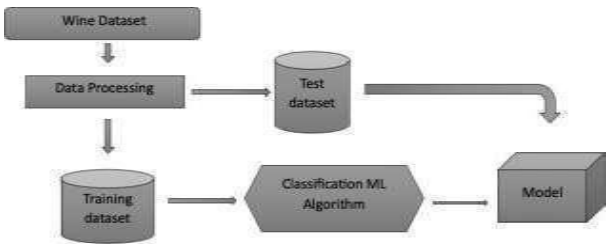



Fig 1. Architecture of Proposed model

To find whether a job post is fraud or not is the main target of this studies. By identifying and removing all these fraudulent job posts will make the job finders to concentrate on the original and real jobs .In this project a dataset is derived from Kaggle and it gives information about the jobs posted whether they belong to original ornot original. The dataset has the theme as depicted.

| | |
|---|---|
| job_id | int64 |
| title | object |
| location | object |
| department | object |
| salary_range | object |
| company_profile | object |
| description | object |
| requirements | object |
| benefits | object |
| telecommuting | int64 |
| has_company_logo | int64 |
| has_questions | int64 |
| employment_type | object |
| required_experience | object |
| required_education | object |
| industry | object |
| function | object |
| fraudulent | int64 |

Fig. 2. Schema of dataset

There are 17,888 job posts in this dataset above. The overall performance of this proposed methods are tested using this dataset. Some preprocessing algorithms are being used in this dataset before using these data to any classifiers. Pre- processing techniques include many algorithmic techniques. This makes the dataset to be transformed to a encoder to make them a vector. Several classifiers are used in these vectors. The following Fig. 3 says a description about the working criteria of a classifierfor prediction.

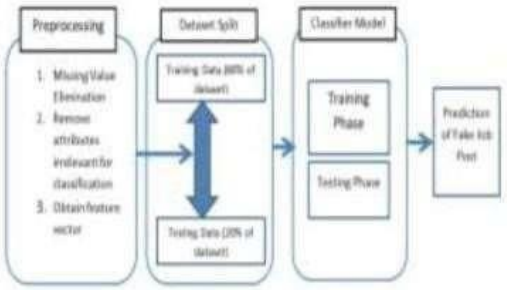



Fig 3.Detailed description for working of Classifiers

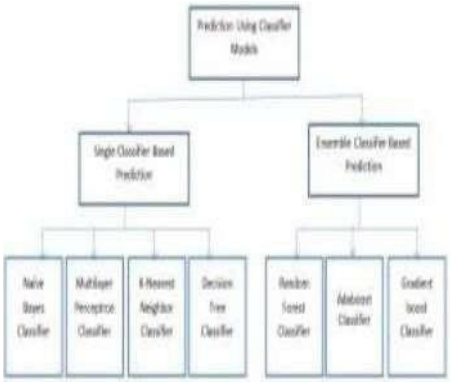


Fig 4. Classification models used in this framework

As detailed in Fig. 4, a couple of classifiers are employed such as NB Classifier, Decision Tree Classifier, KNN Classifier, and Random Tree for classifying job post as fake. It is to be stated that the fake posts of the dataset is the main target of classification proposed. 85% of dataset is trained

and again 15% of the total dataset is used for prediction process. The metrics like accuracy, F measurement, and CK score is used in finding and predicting classifiers. Atlast the classifier which has the best performance with all the metrics is given as the best candidate.

### *A. Implementing Classifiers*

Classifiers here are being prepared with accurate parameter. The parameters that are default are not enough for increasing the performance of the models. The Best one for finding and eradicating the fraud jobs from job finders are made by adjusting the parameters which infact increases the reliability of the model.

MLP is used as a form of five concealed layers which include: 128,64,32,16,8 sizes respectively. Considering all the metrics The K- NN classifier predict a result , k as 5 . On another fact, ensemble classifiers, like RF, Adaa-Boost, GB classifiers are terminated in which five hundred number result are built. Training data are allocated into this after proposing all the classification. Dataset is used in finding the result. Based on the result value and the original value the performance and activity are evaluated after the prediction is completed using the original and the original values.

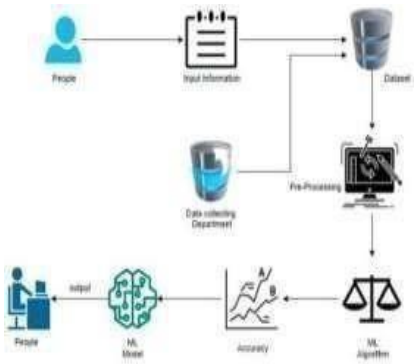
### *B. Performance Evaluation Metrics*

It is necessary to have some metrics in order to evaluate some skills of a model. For evaluating model accuracy is not enough accomplishment because it doesn't foretell correctcases.

This creates a problem if a fraudulent post is considered as a genuine post. So, it is important that the probabilities of the desirable and the non-desiraible outputs is taken into account. To calculate this exactness it is required to be taken into account. Exactness tells us if the probability of the number of correct results over the number of correct obtained results. Reccall gives number of correct result/ number of sample. F1-result is a framework is calculated as accuracy and reccall.CK result is taken into account to calculate grade. Here the grade is a measurement which that gives the agreement for problem.

MS Error is used for the measurement of the various values between the predicted and the obtained result of test specimen. The values of all the MSE and precision, the F measure, CP score gives the better model.

WORKING



This proposed method is built a machine learning model to classify the real or fake job posting, a ML method is used in this model. The dataset is first preprocessed and the columns are analyzed to see the dependent and independent variable and various ML algorithms is implemented for extracting the desirable pattern to get the result with higher precision Many data file from various

location is merged to get a common data file, and various ML algorithm will be implemented to obtain the desirable patterns to get the result with higher precision. Next, it loads, cleans and trims the data source for estimation. It must be taken into account that the steps must be followed carefully to clean .Data file gathered to predict data to Train and Test the file. Usually, 7:4 ratio is deployed to separate the Training Test file. Data structure that were formed with ML techniques is made to fall in line training file in accordance with the result precision and prediction is carried out at ease.

## IV. Result

The machine learning approach is developed in this model to predict the real or fake job posting. To view the dependent and independent variable it is preprocessed and examined, and to get the results various machine learning algorithms are used. Dataset are combined to form a common dataset and these would be used to get the pattern and result with higher accuracy. The data will be loaded, cleaned and trimmed for result.

## V. Conclusion and Future Work

This project will be helpful to avoid the fake job notification to the job seekers to avoid any scam of any sort. This model is built starting from the process of dataset cleaning and preprocessing. This model will be used to find the fraudulent occurring during online hiring process. The Future work includes fraudulent job notification to connect with cloud. To amend work in AI domain.

## References

[1] I. Rish, “An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier”, no. 11, January 2001, pp. 41–46, 2014.

[2] D. E. Walters, “Bayes’s Theorem and the Analysis of Binomial Random Variables”, Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

[3] Murtagh, “Multilayer perceptrons for classification and regression”, Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[4] P. Cunningham and S. J. Delany, “K -Nearest Neighbour Classifiers”, Mult. Classif. Syst., pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)000996.

[5] H. Sharma and S. Kumar, “A Survey on Decision Tree Algorithms of Classification in Data Mining”, Int.J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.

[6] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, O. Adetunmbi, and O. E. Ajibuwa,

"Machine learning for email spam filtering: review, approaches and open research problems", Heliyon, vol. 5, no. 6, 2019, doi:10.1016/j.heliyon.2019.e01802.

[7] L. Breiman, "ST4_Method_Random_Forest", Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi:10.1017/CBO9781107415324.004.

[8] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, "Bagging classifiers for fighting poisoning attacks in adversarial classification tasks", Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37.

[9] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial", Front. Neurorobot., vol. 7, DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[10] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, "Spam review detection techniques: A systematic literature review", Appl. Sci., vol. 9, no. 5, pp. 1–26, 2019, doi: 10.3390/app9050987.

[11] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "FakeNews Detection on Social Media", ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.

[12] Shivam Bansal (2020, February). [Real or Fake]Fake Job Posting Prediction, Version 1. Retrieved March 29, 2020 from https:// www.kaggle.com/shivamb/real-or-fake-fake- job posting-prediction

[13] 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.

[14] I. Rish, "An Empirical Study of the Naïve Bayes Classifier An empirical study of the naïve Bayes classifier", no. January 2001, pp. 41– 46, 2014.

[15] D. E. Walters, "Bayes's Theorem and the Analysis of Binomial Random Variables", Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

[16] Murtagh, "Multilayer perceptrons for classification and regression", Neurocomputing, vol.2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[17] P. Cunningham and S. J. Delany, "K-Nearest Neighbour Classifiers", Mult. Classif. Syst., May, pp.1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[18] H. Sharma and S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining", Int. J. Sci.Res., vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.

[19] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems", Heliyon,vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[20] L. Breiman, "ST4_Method_Random_Forest", Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[21] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, "Bagging classifiers for fighting poisoning attacks in adversarial classification tasks", Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350– 359, 2011, doi:10.1007/978-3-642-21557-5_37.

[22] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial", Front. Neurorobot., vol. 7, no.DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[23] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, "Spam review detection techniques: A systematic literature review", Appl. Sci., vol. 9, no. 5, pp. 1–26, 2019, doi: 10.3390/app9050987.

[24] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media”, ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.

[25] Shivam Bansal (2020, February). [Real or Fake] Fake Job Posting Prediction, Version 1. Retrieved March 29, 2020 from https://www.kaggle.com/shivamb/real- or-fake-fake-job posting-prediction

[26] 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

[27] D. E. Walters, “Bayes’s Theorem and the Analysis of Binomial Random Variables”, Biometrical J., vol. 30, no. 7, pp. 817–825, 1988,doi: 10.1002/bimj.4710300710.

[28] Murtagh, “Multilayer perceptrons for classification and regression”, Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[29] 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.

[30] H. Sharma and S. Kumar, “A Survey on Decision Tree Algorithms of Classification in Data Mining”, Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016,doi: 10.21275/v5i4.nov162954.

[31] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems”, Heliyon,vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[32] L. Breiman, “ST4_Method_Random_Forest”, Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[33] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, “Bagging classifiers for fighting poisoning attacks in adversarial classification tasks”, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37.

[34] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial”, Front. Neurorobot., vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[35] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, “Spam review detection techniques: A systematic literature review”, Appl. Sci., vol. 9, no. 5, pp. 1–26, 2019, doi: 10.3390/app9050987.

[36] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media”, ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.

[37] Kanyadara Saakshara, Kandula Pranathi, R.M. Gomathi, A. Sivasangari, P. Ajitha, T. Anandhi, "Speaker Recognition System using Gaussian Mixture Model", 2020 International Conference on Communication and Signal Processing (ICCSP), pp.1041-1044, July 28 - 30, 2020.

[38] R. M. Gomathi, P. Ajitha, G. H. S. Krishna and I. H. Pranay, "Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862048.

[39] Subhashini R , Milani V, "IMPLEMENTING GEOGRAPHICAL INFORMATION SYSTEM TO PROVIDE EVIDENT SUPPORRT FOR CRIME ANALYSIS", Procedia Computer Science, 2015, 48(C), pp. 537–540

[40] Harish P, Subhashini R, Priya K, "Intruder detection by extracting semantic content from surveillance videos", Proceeding of the IEEE International Conference on Green Computing, Communication and Electrical Engineering, ICGCCEE 2014, 2014, 6922469

[41] Sivasangari, A., Krishna Reddy, B.J., Kiran, A., Ajitha, P.(2020), " Diagnosis of liver disease using machine learning models", ISMAC 2020, 2020, pp. 627–630, 9243375

[42]

[43] Sivasangari, A., Nivetha, S., Pavithra,, Ajitha, P., Gomathi, R.M. (2020)," Indian Traffic Sign Board Recognition and Driver Alert System Using CNN", 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020, 2020, 9315260

[44] Ajitha, P., Lavanya Chowdary, J., Joshika, K., Sivasangari, A., Gomathi, R.M., "Third Vision for Women Using Deep Learning Techniques", 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020, 2020, 9315196

[45] Ajitha, P.Sivasangari, A.Gomathi, R.M.Indira, K."Prediction of customer plan using churn analysis for telecom industry",Recent Advances in Computer Science and Communications,Volume 13, Issue 5, 2020, Pages 926-929.

[46] Gowri, S. and Divya, G., 2015, February. Automation of garden tools monitored using mobile application. In International Confernce on Innovation Information in Computing Technologies (pp. 1-6). IEEE.

[47] Gowri, S., and J. Jabez. "Novel Methodology of Data Management in Ad Hoc Network Formulated Using Nanosensors for Detection of Industrial Pollutants." In International Conference on Computational Intelligence, Communications, and Business Analytics, pp. 206-216. Springer, Singapore, 2017.