Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **622:627***

Identifying Real and Fake Job Posting using Machine Learning

Sherina Sara Jaison[1*], Dr Mallikarjuna Kodabagi[2]

[1,2]School of Computing and Information Technology, Reva University
[1]r19mds10@Cit.Reva.Edu.In
[2]mallikarjun.mk@reva.edu.in

**Corresponding author's E-mail: r19mds10@Cit.Reva.Edu.In*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 29 Nov 2023

Abstract

According to researches, there are around 188 million unemployed people around the globe. We may find many job vacancies on job portals and across the internet to help the job seekers. India alone has more than a hundred job portals. One major issue people face here is that the job seekers are not sure if the employer is real or fake. Most of these portals do not have a system that could check if the employer, posting a job is real or fake. Scammers are making use of this opportunity to post fake job offers which might look genuine to the job seekers applying for it. This way the poor job seekers might lose a large amount of money and time. A best possible solution for this problem would be that the job portal itself being able to identify if the job being posted is real or fake. This paper suggests using a machine learning model to achieve this goal. The idea here is to use natural language processing to understand and analyze the job posting and then making use of a machine learning model to predict if the job posting is real or fake. The first step is to import a dataset which has real life real and fake job posting. In this project, Employment Scam Aegean Dataset provided by University of Aegean Laboratory of Information and Communication system Security is being used. This dataset contains 18000 samples containing real life job postings. Various text cleaning techniques like lemmatization, stop words removal and special characters and punctuation removal is done on the data. Once the text data is processed, various algorithms like Random Forest, Linear SVC, Gradient Boosting Classifier, Gaussian naïve Bayes classifier and XGB classifier is used to test the performance of the model. The best two algorithms with respect to the percentage of accuracy with which the models could classify real and fake job posting was taken into consideration. Random Forest and Linear SVC could give accuracy close to 98%. Both of these algorithms were tuned using GridSearchCV , a library function which is a part of sklearn's model selection package. After tuning, the performance of both these algorithms increased and Linear SVC gave a better accuracy score of 99%. Hence Linear SVC is being used in this project for predicting real and fake job posting on a job portal..

Keywords: *Random Forest, Linear SVC, Natural Language Processing, Machine Learning*

1. Introduction

Need for a secure job is one of the most important aspects that is required for an individual. The spammers are making use of this opportunity to post fake job postings and grab money from the job seekers. Every day five out of ten people fall prey to this trap. The main reason for this is the lack of background verification and checks made by the job portals with respect to the authenticity of the employer. The basic procedure to register as an employer for most of the job portals is quite simple. The Employer has to just enter the email-ID and the company details in order to register to the job portal. This way it is easy for the scammers to fake his/her identity and register as an employer and trap the job seekers promising them jobs of high salaries. Through this project, we have implemented natural language processing to process the text data of 18,000 real life fake and real job posting and finally use a machine learning model to classify if a job is real or fake based on the text data entered by the employer. On testing the accuracy of prediction with respect to different algorithms, we found that Linear SVC shows the best percentage of accuracy which is around 99%. Such an algorithm can be implemented on a job portal in order to identify if a job is real or fake when the employer posts it.

Literature Survey

Distinctive papers were referred towards building this model. In the papers referred, the precision tiers of various data sets were computed using different algorithms under machine learning.

One of the research projects uses MLP classifier as an assortment of 5 hidden layers of size 128, 64, 32, 16 and 8 individually [1]. The K-NN classifier gives a promising outcome for the worth k=5 considering all the assessing metric. Then again, gathering classifiers, for example, Random Forest, AdaBoost and Gradient Boost classifiers are fabricated dependent on 500 quantities of assessors on which the boosting is ended. Subsequent to building these order models, information are fitted into it. Later the testing dataset are utilized for expectation reason. After the expectation is done, execution of the classifiers are assessed dependent on the anticipated worth and the genuine worth. The accuracy that this model could achieve was 98%

A mechanized instrument utilizing AI based order methods is proposed in another research. Various classifiers are utilized for checking fake job posts in the web and the accuracy of those classifiers are thought about for distinguishing the best model. The paper suggests distinguishing counterfeit occupation posts from a gigantic number of posts. Two significant kinds of classifiers, for example, single classifier and gathering classifiers are considered for false job posts recognition.

Another research makes use of different algorithms to classify Fake and real news posted on social media [5]. This paper also deals with the technique of processing textual data. The paper sums up the normal precision of all calculations over the 4 datasets. By and large, the best performing calculation is Random Forest classifier (choice trees) (precision 94%), though the most exceedingly terrible performing calculation is Wang-Bi-LSTM (exactness 64.25%). Singular learners' precision is 77.6% while the exactness of outfit learner is 92.25%.

Another research audits different Machine learning approaches in recognition of phony and created news [6]. The constraint of such and approaches and ad lib via executing profound learning is likewise checked on.

Each one of the papers alluded had a few downsides. So as to conquer the disadvantages of the papers alluded, and furthermore to improve the precision levels we have built up a model utilizing Natural Language Processing and Machine Learning. Using Linear SVC algorithm could give commendable level of accuracy. This algorithm could meet the expectations to a great extent. This calculation demonstrates to have preferred execution over all the strategies utilized in the papers referred.

Proposed System

Knowing the importance of a job vacancy, scammers try to post fake job offers with the intention to cheat job seekers. This could deteriorate the dignity of a job portal as well. Therefore, the best solution that a job portal can implement is to use some kind of an algorithm to detect if a job posting being posted is real or fake. This project makes use of Natural Language Processing to process the textual data and then trains a Machine Learning model to classify is a job being posted is real or fake.

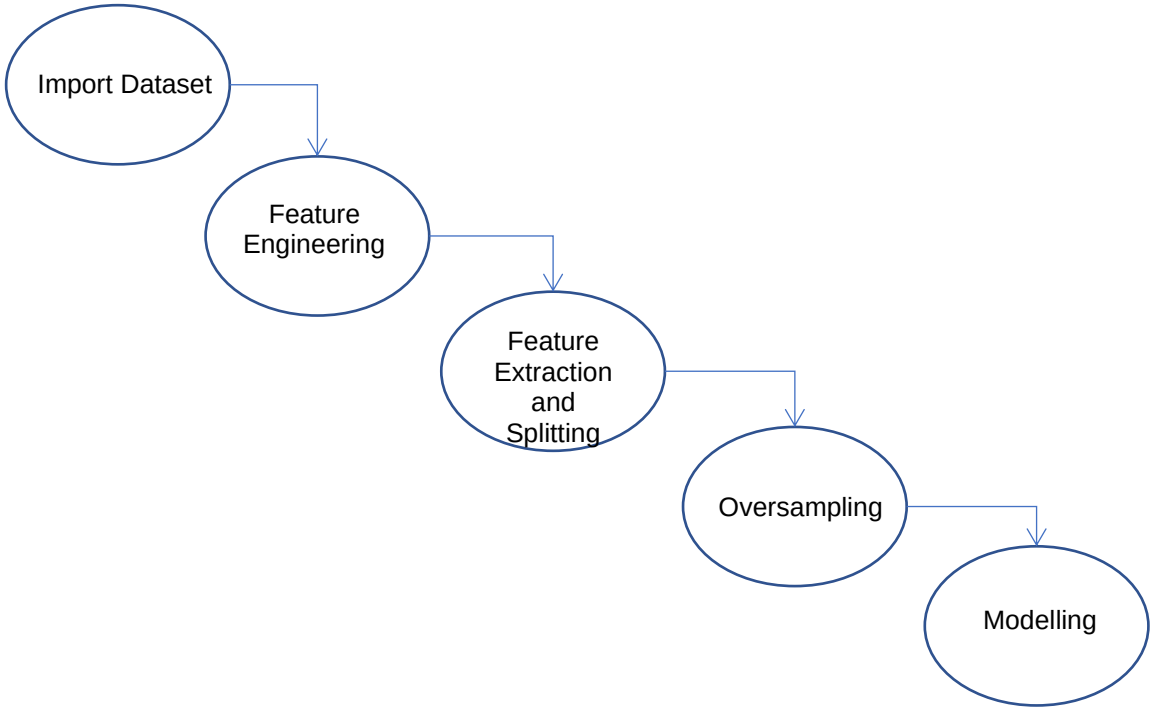

Fig 1: Workflow of the Proposed System

A. Importing the Dataset

The dataset being use for this project is Employment Scam Aegean Dataset, which consists of 18000 records consisting real life job posting. The dataset includes both real and fake job postings. Once the dataset is imported, the data set is analysed and the following findings are made:

- The job id is a unique identifier for a specific job opening
- Some of the salary ranges are written in ranges while others are in values and some others are blank. There are, however a wide variety of salary ranges.
- When it comes to the company profile and description, all the data is textual data.
- These are several missing values in different fields like “required education”, “required experience”.

The dataset is imbalanced; the ratio of real and fake job offers on the dataset is 20:1.

B. Exploratory Data Analysis

The percentage of fake job posting with respect to each variable is observed. The results are plotted using Matplotlib and Seaborn. The dataset is visualized using these libraries and the following findings are made.

- A job is allowing work from home is more probable to be fake.
- A job without a company logo is has a greater probability to be fake.
- Considering the different employment type in the dataset, the part time employment type has greater probability of being fake.
- Another observation is that the high school education category under required education has highest number of fake job postings

The salary range having exceptionally high range is found to have more fake job openings.

C. Feature Engineering

Feature engineering is a process by which each field in the dataset is processed and formatted. All the category columns in the dataset will be made into one textual data. This text is then processed using various text cleaning methodologies. At the end the required features are extracted using Tfldf Vectorizer. Firstly, all the numerical data in the dataset is converted into text. The next step is to merge all the text into a single column. There are various text cleaning methods used on this data such as:

1. **Converting to lower case**: A capital letter is used either at the beginning of a sentence or for a noun. So, to increase the predictive accuracy of the model, all the data can be converted to lower case.
2. **Removal of Stop words and lemmatization**; There are words which does not contribute much meaning to a sentence such as “as”, ”The” etc. These words are called stop words which can be removed. Lemmatization is converting each word in a sentence into its base form. This could improve the accuracy of the model.
3. **Special Character removal and Punctuation Removal**: Punctuations and special characters can add noise to the data that is being processed. The best practice is removing these words before the data is processed.

D. Feature Extraction and Splitting

For the machine learning model to understand the data, the data is converted into a matrix form. To convert it into a matrix, TfldfVectorizer is being used. TfldfVectorizer calculates the importance of each word in a sentence and converts it into a matrix for the machine learning model to process. Once this is done, the data is split into training and testing set. 80% of the data will be considered for training and the remaining 20% will be considered for testing

E. Oversampling and Modelling

Since the dataset is imbalanced. Oversampling is being done. ADASYN method, expanded as Adaptive Synthetic Sampling Method is the oversampling method that is being used. ADASYN calculated the ratio of minority to majority features. The next step is to calculate the total count of data to be generated. The K-Nearest Neighbour of every individual sample is considered. The amount of examples to be generated for each neighbour is calculated. This way the imbalance issue will be resolved.

The data is trained using different algorithms like Random Forest, Linear SVC, Gradient Boosting Classifier and XGB classifier without tuning the hyper-parameter. Random Forest exhibited greater accuracy along with greater speed. Although Linear SVC performed slightly better than Random Forest, both Linear SVC and Random Forest are considered for tuning.

3. Results and Discussion

The Employment Scam dataset that was used for this project has a sum of 18000 records. The dataset was split into two considering 80% for training and the remaining 20% for testing. The table below shows the accuracy of each of the algorithms used in the project before considering the hyper-tuning parameter.

TABLE I Table showing accuracy results with respect to each algorithm

Algorithm	Training Accuracy	Testing Accuracy
Random Forest	100%	98.3%
Linear SVC	100%	98.8%
Gradient Boost	98.1%	95.4%
XG Boost	98.1%	95.8%

From Table 1 we see Random Forest that works by constructing different decision trees gave an accuracy of 100% while training, but the results dropped to 98% while testing which is also a good range of accuracy at the testing level. Linear SVC which works by finding a hyper plane that best fits the data gave an accuracy of 100% while training, but the accuracy dropped to 98.8% while testing which is also considered satisfactory. Gradient Boost algorithm gave an accuracy of 98.1% while training but the accuracy dropped to 95% while testing. XG Boost gave 98.1% accuracy while training, but the accuracy dropped to 95.8% while testing.Considering these results, the best algorithm was considered for tuning.

A. Tuning using GridSearchCV

GridSearchCV is used to loop through different hyper parameters and fit the model onto the training set. On concluding this process, we are able to select the best parameter from the hyper parameters that are listed. We pass predefined values for hyper parameters to the GridSearchCV work. We do this by

characterizing a word reference in which we notice a specific hyperparameter alongside the qualities it can take. Here is an illustration of it. Note that the remainder of the hyper parameters will be set to their default esteems. GridSearchCV attempts every one of the blends of the qualities passed in the word reference and assesses the model for every mix utilizing the Cross-Validation strategy. Thus in the wake of utilizing this capacity we get precision/loss for each blend of hyper parameters and we can pick the one with the best exhibition.

TABLE II Accuracy post tuning SVC Algorithm

Algorithm	Testing Accuracy Before Tuning	Testing Accuracy After tuning
Linear SVC	98.8%	99.7%

On tuning Linear SVC, the accuracy of Linear SVC increased as shown in Table II. Hence Linear SVC is considered to have a better accuracy score than the existing systems and linear SVC is used to classify if a job being posted is real or fake.

To demonstrate the implementation of this model, a UI is developed using python tkinter module. This UI is similar to a job portal page when an employee can register his/her company and recruitment details. Once the employee registers, the algorithm identifies is the job being posted by the employer is real or fake purely based on the textual data that the employer has entered. Once this is identified, the portal blocks the employer from posting the job. This could help save many job seekers who fall prey to fake job offers.

4. Conclusion

This project has been implemented using Linear SVC which classifies a real and fake job posting with commendable level of accuracy. Different algorithms were used to test the performance of the model and the best two algorithms such as Random Forest and Linear SVC were considered for tuning. On tuning the accuracy of both the models increased. However, Linear SVC performed better in terms of speed and accuracy. Linear SVC could give an accuracy close to 99% which is the best accuracy score when it comes to researches conducted in the field of fake and real job posting detection. To demonstrate the implementation of this model, a UI which replicates a job portal employee registration page is made. When the employer posts the job, NLP is used to process the textual data that the employer has entered and TfldfVectorizer is used to convert this textual data to a matrix form which is fed into the machine learning model which classifies is the job being posted is real or fake which commendable level of accuracy. This way the job portal can itself block a scammer from posting a fake job opening. This can help to retain the reputation of a job portal and also reduce the complexity of the job seekers.

Acknowledgement

Any given task achieved is never the result of efforts of a single individual. There arealways a bunch of people who play an instrumental role in leading a task to its completion. Our joy at having successfully finished our project work would be incomplete without thanking everyone who helped us out along the way. We would like to express our sense of gratitude to REVA University for providing us the means of attaining our most cherished goal.

We would like to thank our Hon'ble Chancellor, Dr. P. Shyama Raju, Hon'ble Vice-Chancellor, Dr. S. Y. Kulkarni for their immense support towards students to showcase innovative ideas. A big thanks goes out to Director, Dr. Sunilkumar S. Manvi for providing a highly conducive environment and encouraging the growth and creativity of each and every student, Prof.Mallikarjun Kodabagi for continuously supporting and guiding in every endeavour and for taking a keen and active interest in the progress of every phase of this Project.

References:

[1] Samir Bandyopadhyay, Shwani Dutta"Fake Job Detection using Machine Learning Approach",published on Research Gate on April 2020

[2] Kafaru Simmie, "Machine Learning for fake job detection", published on Medium data driven investor journal, June 2020

[3] Jawed Zhang1, Bowen Dong2, Philip S. Yu2 "FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network" IFM Lab, Department of Computer Science, Florida State University, FL, USa

[4] Sokratis Vidros , Constantinos Kolias , Georgios Kambourakis and Leman Akoglu ,"Automatic Detection of Online Recruitment Frauds": Characteristics, Methods, and a Public Dataset (2019)

[5]"Fake News Deection using ensemble methods"by Iftihar Ahmed, Suhail Yousaif, Muhammed Ovasis.,Hindawayi Journals, 17 October 2020
[6] Syed Ishfaq Manzoor, Jimmy Sangla, Nikitha "Fake News Deection using machine learning approaches,", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)
[7] Armstrong, A. (2006) Handbook of "Human Resource Management Practice. 10th Edition, Kogan Page Limited," London.
[8] Hada, B. and Gairola, S, "Opportunities and Challenges of E-Recruitment. Journal of Management Engineering and Information Technology", (2015)
[9] Kaur, P, "E-Recruitment: A Conceptual Study. International Journal of Applied Research", 1, 78-82. . (2015)
[10] Prasad, L. and Kapoor, P, "Topic: E-Recruitment Strategies. International Journal of Business Quantitative Economic and Applied Management Research", 2, 80-95. . (2016)
[11] Panov, P., Soldatova, L. and Dzeroski, S, "OntoDM-KDD: Ontology for Representing the Knowledge Discovery Process". 16th International Conference on Discovery Science, Singapore, 6-9 October 2013, 126-140 . (2013)
[12] Cios, K.J., Pedrycz, W., Swiniarski, R.W. and Kurgan, L.A, "Data Mining: A Knowledge Discovery Approach. Springer", New York. (2007)
[13] Hussain, S. "Survey on Current Trends and Techniques of Data Mining Research." London Journal of Research in Computer Science and Technology, 17, 7-15. (2017)
[14] Sinoara, R., Antunes, J. and Rezende, S,Text Mining and Semantics: A Systematic Mapping Study. Journal of the Brazilian Computer Society, . (2017)
[15] Diwathe, D. and Dongare, S. "Classification Model Using Optimization Technique: A Review". International Journal of Computer Science and Network, 6, 42-48. (2017)
[16] Singh, G. and Singh, A. "A Review Paper: Using Data Mining Clustering Technique to Predict Criminal Behavior." International Journal of Computer Science and Mobile Computing, 6, 160-167. (2017)
[17] Witten, I. and Frank, E, "Data Mining Practical Machine Tools and Techniques. Morgan Kaufmann Elsevier", San Francisco. . (2005)
[18] Kukasvadiya, M. and Divecha, N. "Analysis of Data Using Data Mining Tool Orange". International Journal of Engineering Development and Research, 5, 836-1840. (2017)
[19] Rehman, N. Data Mining Techniques Methods Algorithms and Tools. International Journal of Computer Science and Mobile Computing, 6, 227-231. (2017)
[20] Jyoth, P., Siva Ranjani, R., Mishra, T. and Mishra, S.R. "A Study of Classification Techniques of Data Mining Techniques in Health Related Research." International Journal of Innovative Research in Computer and Communication Engineering, 5, 13779-137876. (2017)
[21] Vidros, S.,Kolias, C. and Kambourakis, G."Feature: Online Recruitment Services: Another Playground for Fraudsters. Computer Fraud & Security", 2016, 8-13.
[22] Yasin, A. Abuhasan, A. "An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security & Its Applications", (2016)
[23] Al-garadi, M.A., Varathan, K.D. and Ravana, S.D, Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network. Computers in Human Behavior . (2016)
[24] Sharaff, Nagwani, N.K. and Swami, K. "Impact of Feature Selection Technique on Email Classification. International Journal of Knowledge Engineering", (2015)
[25] Sornsuwit, P. and Jaiyen, S. "Intrusion Detection Model Based on Ensemble Learning for U2r and R2l Attacks". In: 7th International Conference Information Technology and Electrical Engineering, IEEE, Chiang Mai. ((2015)
[26] Gaikwad, D.P. and Thool, R.C."Intrusion Detection System Using Bagging Ensemble Method of Machine Learning." Computing Communication Control and Automation, Pune, 26-27 February 2015, 291-295.
[27] Zuhaira, Selmat, A. and Salleh, M,The Effect of Feature Selection on Phish Website Detection: An Empirical Study on Robust Feature Subset Selection for Effective Classification. International Journal of Advanced Computer Science & Applications. . (2015)
[28] Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of Fraudulent Emails by Employing Advanced Feature Abundance. EgyptianInformaticsJournal, 15, 169-174.
[29] Shrivas, K. and Dewangan, A.K, "An Ensemble Model for Classification of Attacks with Feature Selection Based on KDD99 and NSL-KDD Data Set. International Journal of Computer Applications. (2014)