Job portal end detection of fake job posting using machine learning

Prem Anand*, Dr. Vishnu Sharma#

M.Tech student , Jaipur National University,Jaipur

premanand@gmail.com

#Associate Professor, Jaipur National University

vishnu@jnujaipur.ac.in

Abstract — **This research paper contributes to prevent fake job postings over job portals. This study will protect job seekers from unauthorized access of personal information and loss of money with fraud job posting. In this paper we have worked upon a steadfast model a model which can detect the fake job posting on the job portals. I can be integrated with job portal and it is aimed to develop noble online recruitment environments. This research presents a reliable method to classify a posted job as fraudulent or in fair category. Based on the category posted job have been identified, it is decided that this post will be shown on the portal or not. Online fake job posting which creates too many consequences for the job seekers and the job portal may be clogged with this model presented here.. This detection is done while fake job postings takes place. A open dataset available with kaggle.com is employed to apply the model. Pre-processing is done with python before the feature selection process and classification process. The results obtained have shown the accuracy of 97.50% for identification of fraudulent job. Further, the findings presented the main features and significant factors in detection purpose include having a company registration no(CIN No.), official mail id and security deposit.**

Key words – KDD(Knowledge Discovery from Data),DM(Data Mining)

I. Introduction

As we know in these day study is aimed to get a good job. Job may be achieved with campus placements during academics or by appearing in interviews personally. Another way which is most popular is to apply for the job through web portal. Here candidate is required to post CV on the portal or by filling all the required details in the prescribed format. This type of portals are known as job portals. Basically these portals fills the gap between employers and employee. Employers seek for good qualified, experienced, skilled employees and job seekers want to have a suitable post with high salary and other facilities In this paper a machine learning model is developed using which fake job postings may be detected and prevent deceiving of decent job seekers.

II. Background

2.1. Functions of Job portal

1. Employer Job posting on job portal
2. Candidate verification and CV uploading process.
3. Employer search for suitable candidates based on certain keywords from available profiles.
4. Employee search for suitable job based on certain keywords specified with Job portal repository.
5. Suitable jobs and employees are retrieved from the repository and shared with them.
6. Formal processes takes place like telephonic and Email communication.
5.Employer gets appropriate candidate as per their choice and employee get pertinent jobs.

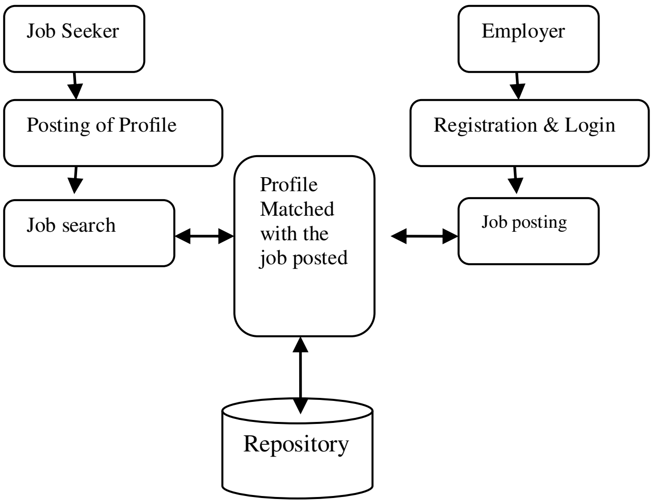


Figure 1 Architecture of Job portal

Online sources enables job seekers to enhancement employment by permitting them to develop peculiar relations with a globally spread people capable of accessing unrestricted content[1]. As per Torok[2] cyber tools are most effective at early stages of a imminent member's unexpected action with extremist thinking—the employment and militancy phase [2]. The "data explosion" and "workload" as contributing factors to analyst and agent oversights was cited Nov 7, 2009 cited [1]. Near about 20,000 electronic documents. Organization involve in terrorist activities involves in redirected communications and advertisement, hiring employees on social media like Face book, LinkedIn and on religious forums [5]. Having workload demands in consideration it was undoubtedly a challenge for the two reviewers who were allocated to the case at that time [3].

Though, there are too many risks augmented by apparent scams and frauds accompany with the ample interest and accepting embedded software like this [6].

2.1. Online Recruitment System

The online Job portals make use of web-based applications which utilizes public internet or intranet for recruitment of employees [7]. These portals brings many benefits to the organization as in such a short span and minimum efforts recruiter can get a suitable or list of suitable candidates. Manual process for the same may take weeks of time. Automatic scrutiny is done of the ineligible candidates. Role of job portals enhances more when a project require certain experience and work profile. On the other hand employees also gets lot many advantages with this. Being in one job he/she may apply for the dream job with just simple registration and posting of CV. Applicant gets regular notification for jobs for which his profile is found suitable. On a single click he may approach to his dream job. Besides, being a affordable and productive, it saves efforts also. The important mechanism of online recruitment also involves keep an eye on the candidate's status, recruiter's website, online recruitment portals, online testing, and social networking portals as well [8]. E-employment serves timeliness, effectiveness, comfort ability, ease, and competency as well [9]. The study has enlightened an assortment of imperative advantages of e-employment such as shortened time, reduction in applied Cost, approaching massive groups (recruiters and jobseekers), Functions filtration process, and Brand image development. Two diverse ways for online employment by (Prasad & Kapoor,2016) which take in recruiter's registration, Job posting with specified profile demanded on requitement portals and supplemented by an online employment page on the company's portal[10].

2.2 Data Mining Tools

Extraction of qualified information from digital text received from diverse sources is termed as Text mining or text classification [11]. Diverse popular Data classification techniques like regression and clustering utilizes labelled data only which is termed as supervised learning method, and involves training and testing phase that make classification od data into a range of classes depending on assigned attributes which are obtained from available dataset [12]. Here are a diverse range of data mining tools are available in these days in which plenty of them are open source tools. Most popular tools are weka, rapid,orange and orange. Most of them provide a user friendly interfaces where data may be retrieved and several operators are utilized to pre-processing of data to model building and result analysis. Rapid Miner is an open source tool that make use of the client-server architecture model and deal with diverse file format like csv,xls,sql etc.[13]. Weka tool is a efficient open source tool for data mining and machine learning tool. it was lunched by Waikato University [14] and offers a vast range of pre-processing data, data modelling algorithms and performance analysis tools.[13]. Orange is also an efficient and easy open source tool for data mining and machine learning it was developed by the University of Ljubljanato, it caters a broad range of drag and drop operators in toolbox with suitable connection methods.[13]. Data mining make use of different methods for identifying patterns(classification), development of prediction model and performance analysis tasks. The most popular method implemented for data mining is Classification and it is identifying the fundamental rules to segregate items into predetermined classes, which comes under predicting task [15].

2.2. Knowledge Discovery from Data (KDD)

KDD can be termed as a discovering and drawing out important, authentic, valuable, unique, and intelligible association or model in the data[16].

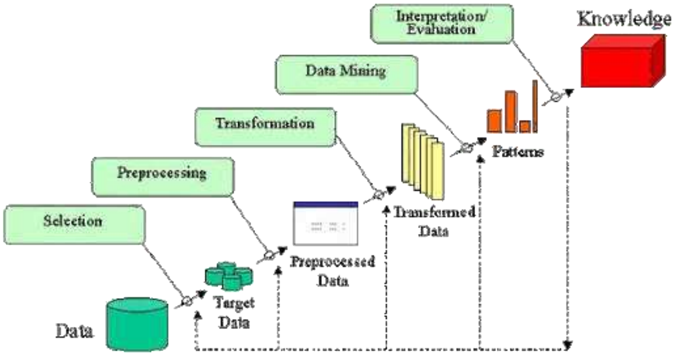


Figure 2 KDD [16]

Figure 2 Exhibits the model of KDD which depends on various sequence of executed processes. Triumph of Each step is dependent on the output of prior one It assert the repetitive nature of the KDD and diverse feedback iterations to point out the result of revision activities [17].

2.3 Data Mining Techniques Related Works

In 2016 Vidros et al. uncovered the frauds which were faced by so called jobseekers with Online employment portals. With the study they determined that ORF is a field where further process may be done .These frauds are categorised into 3 major classes in the online employment processes include

1. Posting of Fake Job Advertisement,
2. Financial scam with fake job advertisements which are made available on online recruitment boards portals, and
3. Presumed business houses announcing fake vacancy on the online protal [18].

2.4. Dataset discovery for fake job identification model

Several official portals provides dataset for diverse problems exist in universe, which may be used by researcher to work upon and serve society by providing solution of existing problem. Dataset related to problem with sufficient features was obtained from kaggle.com.

2.5. The problem of fake job posting

These portals fills the gap between employers and employee. Employers seek for good qualified, experienced, skilled employees and job seekers want to have a suitable post with high salary and other facilities. But it is not so trouble-free as it seems. Here may be fraud of differet In these days like other fields scam has been entered in this gracious field also. Few scammers used to post fake jobs which does not exist. These type of job scams not only create problems to job seekers but also harm to the prestige of job websites and reputed company posts too as they also to be seems to be scams. Job portals hold repository of CV along with verified recruiters. job seekers are worried about getting scammed while job searching now a days. Most of the scams takes place with fake job postings. In a fake job scam, a so called scammer post a job, but the in reality job doesn't exist.

The swindler exploit the job posting to interact with the job seekers and obtain their personal information,like their Aaadhar, PAN card No., credit card no, pin no. The recouped information is then exploited to retrieve job seekers Credit or bank account details to perform some fraud.

Most of the time Fake job scam is done to get job seekers to send money from their bank, or wire money with Western Union money transfer or in seldom conditions scammer ask to transfer money in their bank accounts.

In these days fake job posting is spoiling this gracious field. These are detrimental for job seekers, job portal prestige and employers as well. These fraud job may be stopped if these type of job posting are validated at the time of posting itself. This research is aimed to discover this solution.

In one opinion we can gone through website of particular job posting agency. The rational for keeping the website age information was the tendency of the fraudsters to create fake websites just before posting a fake job advertisement. Nowadays, the ease of creating websites in a few clicks has increased the number of such fake websites where the company does not actually exist, but the website does. Several news reports and government agencies list these sorts of discrepancy [20,21]

III The Proposed Online Recruitment Fraud Detection Model

Proposed Model can be conducted in following steps

1. **Dataset download**
2. **Data cleansing and pre-processing phase**
3. **Set of Feature selection**
4. **Data Splitting(Train and Test set)**
5. **Prediction model**
6. **Performance Analysis**

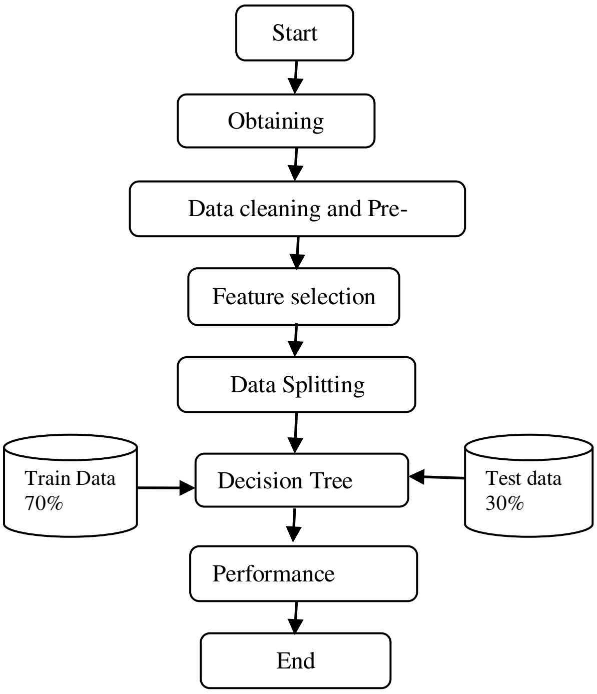


Figure-3 Flow Chart of the Fake job prediction model

3.1 Data set download

Data set of job postings (2019) was obtained from kaggle.com . Data set was received in CSV file having 17881 data rows. This data was containing sufficient fields along the field which disclose the actual status of the jobs whether the job was fraudulent or genuine. Received dataset was requiring data cleaning and pre-processing before being applied to the classifier.

3.2 Data Cleansing and pre-processing

The major aim of data cleaning is to get rid of any data that is erroneous, incomplete, improperly formatted, repeated, or even inappropriate to the purpose of the task.

This is generally accomplished by replacing the values(like replacing incomplete information with word NA), modifying(using a single value or character/word for a range), or in certain conditions erase data that is not appropriate in the task. Hence, it is essential to have a efficient cleaning methods so that the decisions which are being made in determining of fake job are the best possible. However, it is very important to determine the suitable features or variables selection technique to be used to prediction.

Data cleaning involve different techniques based on the problem and the data type. Obtained dataset having 21 columns and lot of missing data and garbage values in many rows was very terrible and error prone. So data cleaning was a mandatory step. Python is used to clean the obtained data. where Different methods can be applied with each has its own trade-offs. Overall, incorrect data is either removed, corrected, or imputed

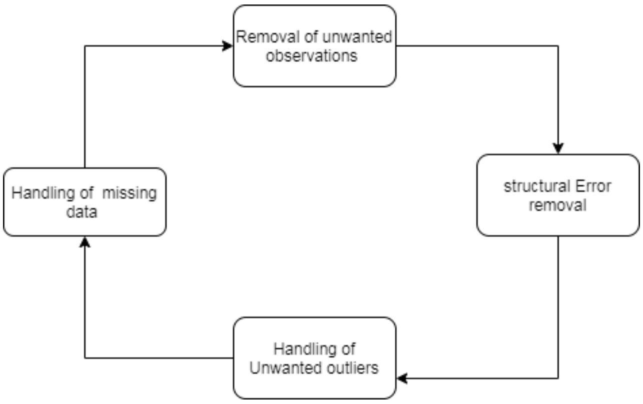

Removal of unwanted observations
structural Error removal
Handling of Unwanted outliers
Handling of missing data

Figure-3.1 Data Cleaning and pre-processing

3.3 Feature Selection :

In most of the cases dataset contains many columns in which only few of them may contribute in the prediction model. Selecting appropriate input variables from the dataset is known as feature selection process. It not only reduces number of input but also enhances the efficiency and reduces the time and efforts in processing the records. So its very important to have feature selection process when developing a predictive model. It is advantageous to shrink the number of input variables it reduces the cost of computation of said modelling and, in some conditions, to get better the performance of the prediction model. One of the most important aspect we need to consider while feature selection is assessing the association between each input and the goal variable with the use of statistics and selection of the inputs which are found to be in strong relationship with the goal variable. This technique is determined as fast and effective, though the selection of statistical measures relies on type of the of the input variables and output variables both.

On the basis of the method to collaborate the feature subsets search along with the development of classification model, feature selection techniques are classified in three categories: filter, wrapper, and embedded techniques[22].As per the data set Wrapper method was found to be best feature selection.

Features selected for the used dataset were job posting company's CIN Number, official Mail id, Security deposit. Max of the fake job posters donot have government CIN number, they do not have official mail id(portal domain), Security is demanded and security amount is very high. These feature are the base of our prediction model.

Here is the truth table based on our selected CIN features

Verified CIN	Fraud	Prediction
N	1	Correct
Y	0	Correct
Y	1	Exception
N	0	Exception

3.4 Decision tree Model for prediction

Four features CIN Number, official Mail id, Security deposit were selected for developing model for prediction of Fraudulent and fair job posting.

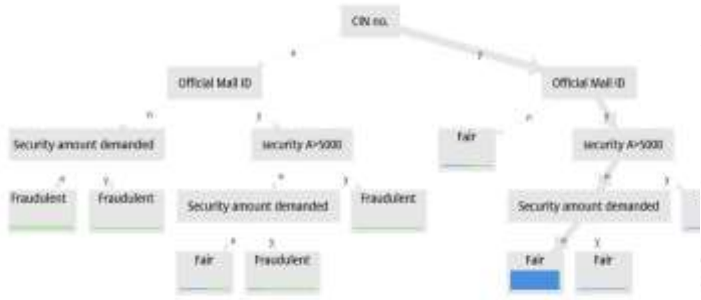
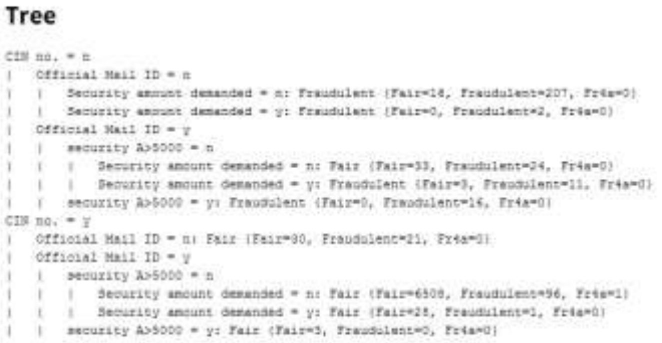


Figure 3.4 Decision tree **for job portal**

Decision tree based algorithm is used to classify jobs into Fake or genuine.



Tree

```
CIN no. = n
|   Official Mail ID = n
|   |   Security amount demanded = n: Fraudulent {Fair=16, Fraudulent=207, Fr4a=0}
|   |   Security amount demanded = y: Fraudulent {Fair=0, Fraudulent=2, Fr4a=0}
|   Official Mail ID = y
|   |   security A>5000 = n
|   |   |   Security amount demanded = n: Fair {Fair=33, Fraudulent=24, Fr4a=0}
|   |   |   Security amount demanded = y: Fraudulent {Fair=3, Fraudulent=11, Fr4a=0}
|   |   security A>5000 = y: Fraudulent {Fair=0, Fraudulent=16, Fr4a=0}
CIN no. = y
|   Official Mail ID = n: Fair {Fair=90, Fraudulent=21, Fr4a=0}
|   Official Mail ID = y
|   |   security A>5000 = n
|   |   |   Security amount demanded = n: Fair {Fair=6508, Fraudulent=96, Fr4a=1}
|   |   |   Security amount demanded = y: Fair {Fair=28, Fraudulent=1, Fr4a=0}
|   |   security A>5000 = y: Fair {Fair=3, Fraudulent=0, Fr4a=0}
```

Figure 3.5 Description of Decision tree

3.5. Implementation of proposed model

For implementation of the model Repidminer tool is used due to its user friendliness and accurate results. Repidminer studio 9.6 is used for this implementation.

Design of the proposed model

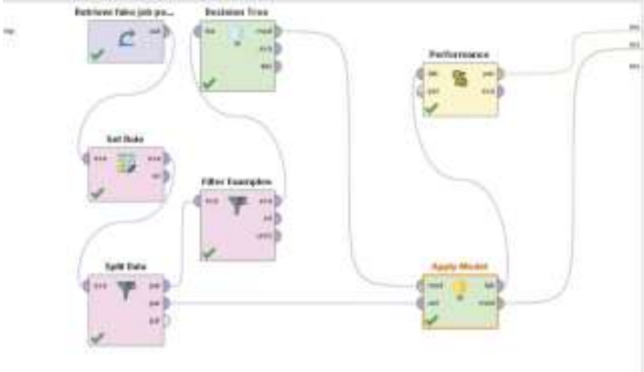


Figure 3.6 Design of model

Step 1 : Retrieve data : Importing of data on the process board

Step 2: Set Role: Column Fraudulent role is determined.

Step3 : Split Data: The Split Data operator takes retrieved data set as an input and outputs the splitted data as per the parameters specified. In our research pre-processed dataset is split into two one is training data which is 70% of the entire dataset and 30% as test set. The number of subsets (or partitions) and the relative size of each partition are specified through the partitions parameter. 1st split data is used to train the model for accurate prediction ad 2nd split is utilized to test whether the prediction model is accurate as target variable fraudulent is already specified.

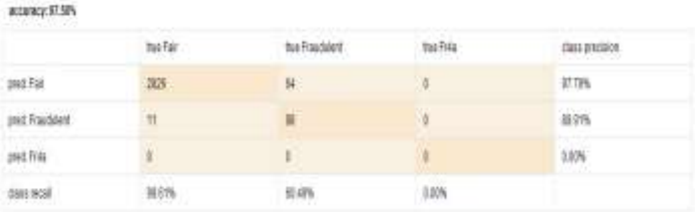
Step: 4 Decision tree model is called where on 1 port 70% training data is implemented

Step: 5 Apply model : Decision model output is sent to apply model and unlabelled data(test data is obtained from split process. Apply model outputs labelled data.

Step:6 Performance classification : This process get input from apply mode in the form of labelled data and outputs on the port. Performance is percentage is set as output.

Figure 3.7 Decision Tree Simulator results

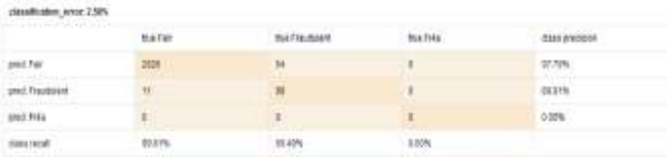
Performance Evaluation :



accuracy: 97.50%

	true Fair	true Fraudulent	true Fr4a	class precision
pred. Fair	2026	64	0	97.79%
pred. Fraudulent	11	98	0	89.91%
pred. Fr4a	0	0	0	0.00%
class recall	99.51%	60.49%	0.00%	

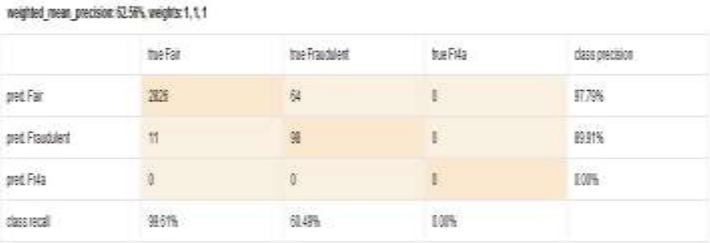
Figure 3.7 Performance-Accuracy of model



classification_error: 2.50%

	true Fair	true Fraudulent	true Fr4a	class precision
pred. Fair	2026	64	0	97.79%
pred. Fraudulent	11	98	0	89.91%
pred. Fr4a	0	0	0	0.00%
class recall	99.51%	60.49%	0.00%	

Figure 3.8 Performance-classification Error of model



weighted_mean_precision: 62.56%, weights: 1, 1, 1

	true Fair	true Fraudulent	true Fr4a	class precision
pred. Fair	2026	64	0	97.79%
pred. Fraudulent	11	98	0	89.91%
pred. Fr4a	0	0	0	0.00%
class recall	99.51%	60.49%	0.00%	

Figure 3.9 Performance-Weight means precision of Model

DECISION TREE LIFT CHART

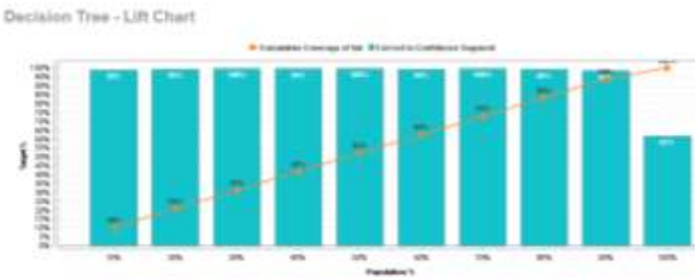


Figure 3.10 Decision tree lift chart

V. Conclusions:

This research has obtained the accuracy of 97.50% for predicting the fraudulent jobs. This model will be effective for the job portals to solve the problem of fake job posting. In very first stage of posting of the job, job will be processed with the model and this process will decide whether to post or deny.

References

[1] Scanlon, J.R. and Gerber, M.S. (2014) Automatic Detection of Cyber-Recruitment by Violent Extremists. Security Informatics , 3, 5. https://doi.org/10.1186/s13388-014-0005-5

[2] R Torok, "Make A Bomb In Your Mums Kitchen": Cyber Recruiting And Socialisation of 'White Moors' and Home Grown Jihadists, in Australian

[3] WH Webster, DE Winter, L Adrian, J Steel, WM Baker, RJ Bruemmer, KL Wainstein, Final report of the William H.Webster Commission on the Federal Bureau of Investigation, counterterrorism intelligence, and the events at Fort Hood, Texas on November 5, 2009. Technical report, Federal Bureau of Investigation (2012)

[4] Artificial Intelligence Laboratory, University Of Arizona, Dark Web Forum Portal: Ansar AlJihad Network English Website (2014). http://cri-portal. dyndns.org.

[5]LA Overbey, G McKoy, J Gordon, S McKitrick, Automated sensing and social network analysis in virtual worlds, in Intelligence and Security Informatics (ISI) (IEEE Vancouver, BC, Canada, 2010), pp. 179–184

[6] Vidros, S., Kolias, C., Kambourakis, G. and Akoglu, L. (2017) Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset. Future Internet , 9, 6. https://doi.org/10.3390/fi9010006

[7]Armstrong, A. (2006) Handbook of Human Resource Management Practice. 10th Edition, Kogan Page Limited, London.

[8] Hada, B. and Gairola, S. (2015) Opportunities and Challenges of E-Recruitment. Journal of Management Engineering and Information Technology , 2, 1-4.

[9] Kaur, P. (2015) E-Recruitment: A Conceptual Study. International Journal of Applied Research , 1, 78-82.

[10] Prasad, L. and Kapoor, P. (2016) Topic: E-Recruitment Strategies. International Journal of Business Quantitative Economic and Applied Management Research , 2,80-95.

[11] Sinoara, R., Antunes, J. and Rezende, S. (2017) Text Mining and Semantics: A Systematic Mapping Study. Journal of the Brazilian Computer Society , 23, 9.https://doi.org/10.1186/s13173-017-0058-7

[12] Diwathe, D. and Dongare, S. (2017) Classification Model Using Optimization Technique: A Review. International Journal of Computer Science and Network, 6,42-48.

[13] Kukasvadiya, M. and Divecha, N. (2017) Analysis of Data Using Data Mining Tool Orange. International Journal of Engineering Development and Research, 5, 836-1840.

[14] Rehman, N. (2017) Data Mining Techniques Methods Algorithms and Tools. International Journal of Computer Science and Mobile Computing , 6, 227-231.

[15] Jyoth, P., Siva Ranjani, R., Mishra, T. and Mishra, S.R. (2017) 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.

[16] Panov, P., Soldatova, L. and D žeroski, S. (2013) OntoDM-KDD: Ontology for Representing the Knowledge Discovery Process. 16th International Conference on Discovery Science , Singapore, 6-9 October 2013, 126-140. https://doi.org/10.1007/978-3-642-40897-7_9

[17] Cios, K.J., Pedrycz, W., Swiniarski, R.W. and Kurgan, L.A. (2007) Data Mining: A Knowledge Discovery Approach. Springer, New York.

[18]Vidros, S., Kolias, C. and Kambourakis, G. (2016) Feature: Online Recruitment Services: Another Playground for Fraudsters. Computer Fraud & Security, 2016, 8-13.https://doi.org/10.1016/S1361-3723(16)30025-2

[19] Mahbub, Syed & Pardede, Eric. (2018). "Using Contextual Features for Online Recruitment Fraud Detection" The 27th International Conference on Information Systems Development, At Lund, Sweden Aug 2018.

[20] Scam employment website targets jobless miners – ABC News (2016),http://www.abc.net.au/news/2016-11-30/jobless-miners-targeted-in-online-scam/8079304. Accessed: April 13, 2018

[21]. WAScamNet: Mining Recuirtment Scan – Government of Western Australia (2012),http://www.scamnet.wa.gov.au/scamnet/Scam_types-Jobs__Investment-Jobs__Employment-Mining_recruitment_scam.htm. Accessed: April 13, 2018

[22] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.View at: Publisher Site | Google Scholar