ORFDetector

ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection

Sangeeta Lal[1], Rishabh Jiaswal[1], Neetu Sardana[1], Ayushi Verma[1], Amanpreet Kaur[1], Rahul Mourya[2]
[1]Jaypee Institute of Information Technology, Noida
[2]Heriot-watt university, Edinburgh
sangeeta@jiit.ac.in, jaiswalrishabh96@gmail.com, neetu.sardana@jiit.ac.in , ayushiv26@gmail.com, amanpreet.kaur@jiit.ac.in, mourya.rahul1981@gmail.com

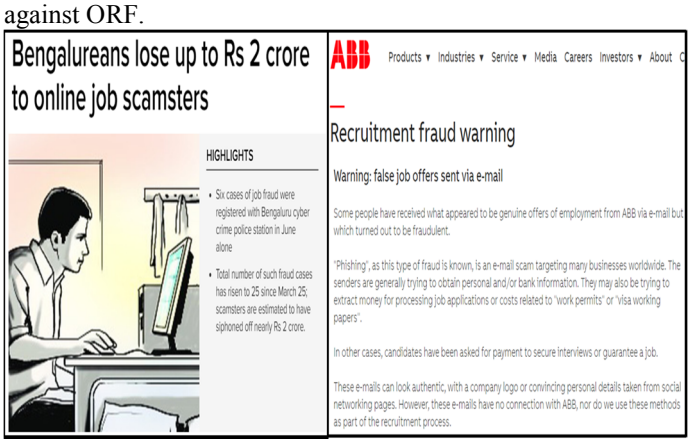
***Abstract*— Online recruitment fraud (ORF) is a new challenge in the cyber security area. In ORF, scammers give job seekers lucrative job offers and in-return steal their money and personal information. In India, scammers have stolen millions of moneys from innocent job seekers. Hence, it is important to find solution to this problem. In this paper, we propose, *ORFDetector*, an ensemble learning based model for ORF detection. We test the proposed model on publicly available dataset of 17,860 annotated jobs. The proposed model is found to be effective and give average f1-score and accuracy of 94% and 95.4, respectively. Additionally, it increases the specificity by 8% as compared to the baseline classifiers.**

Keywords—Online fraud detection, machine leanring, ensemble learning

I. INTRODUCTION

These days' recruitments are mainly done online through online portals such as naukri.com [1], monster.com [2]. Organizations put their job advertisement with desired skills required on these portals. Job seekers or candidates put their resumes and skill details on these portals. Now, companies can scan the profiles of desired candidates and contact the candidates as well as candidates can also apply to the job profiles in which they are interested. After first screening, companies contact the shortlisted candidates for further processing and recruit the suitable candidates. Online recruitment is beneficial for both candidates as well as the companies. In Dec 2016, Naukri.com had a database of about 49.5 million registered users, 11000 resumes were getting added daily [8]. This shows the impact that these online job portals have on users.

The online recruitment is beneficial for both recruiter as well as candidates. However, in the recent years scammers have started this online recruitment industry which has given a new type of fraud, i.e., Online Recruitment Fraud (ORF). In ORF spammers give lucrative job offers to the candidates and steal their money and private information. ORF not only harms the users but it is also problematic for the companies. As, it damages the reputation of companies and leaves a negative impact in the mind of job seekers about the given company. Figure 1 shows some of the news snippets that are showing the damage caused by the ORF problem. Figure 1(a) shows a snipped from news media that job seekers have lost nearly 2 Crore rupees because of ORF [3]. Figure 1 (b) shows a warning sign issued by one of the MNC (ABB corporation) against ORF.



(a) (b)

Fig. 1. News snippets showing the ORF problem. a: the amount of loss Happened because of ORF, b: One of the MNC issuing waning against ORF

ORF detection is an important problem to solve but it has not received much attention from the research community and it is currently a relatively unexplored area. Detection of fraud job offers from a legitimate set of job is a technically challenging problem. The main challenge the class imbalance problem as the number of fraud jobs are relatively less as compared to the legitimate jobs. This makes learning the features of fraud jobs for automated prediction a challenging task.

In this work, we proposed an ensemble-based model *ORFDetector* for ORF detection. We have taken three baseline classifies, J48, Logistic Regression (LR), and random Forest (RF). We applied three ensemble techniques Average Vote (AV), Majority Vote (MV) and Maximum Vote (MXV) on these baseline classifiers to build *ORFDetector* framework. We evaluated the proposed *ORFDetector* model on a publicly available dataset of 17860 jobs. The proposed model is found to be effective and give average f1-score and accuracy of 94% and 95.4, respectively. Additionally, it increases the specificity by 8% as compared to the baseline classifiers.

II. Related Work and Research Contributions

A. Related Work

ORF detection is a relatively new field and there is not much work done in this area. In [9-12], there is large portion important mentions found to teach job seekers on distinguishing deceitful employment opportunity and job boards caution job seekers on the results of employment scams and give detailing reports on any malevolent action. Then there are some indirect methods to solve ORF to little extent such as Email Spam filtering [13] that restricts to send email to user that has some advertising content, anti-phishing techniques [14-15] to identify fake websites, countermeasures against opinion fraud [16] to help identify the posting of deceptive and misleading fake reviews.

To the best of our knowledge only Vidros et al. [5] propose a method to detect the fraud jobs. They identified 22 features from the dataset and propose a machine learning based approach for ORF prediction. However, they worked only with balanced dataset and the performance of prediction algorithms on imbalance dataset set is not known. In real world data is often imbalanced in nature. And sometimes it is difficult to apply a suitable balancing technique. Hence it is important to evaluate the prediction models on imbalanced dataset. In this work, we propose an ensemble learning based method, *ORFDetecteor,* for ORF prediction. We evaluated the performance of the proposed model on the imbalanced dataset.

B. Research Contributions

In context to related work, this work makes the following novel and unique research contributions:

1. We propose an ensemble learning based model, *ORFDetector,* for online fraud detection.
2. We present results of comprehensive analysis of the proposed model on real world dataset.
3. We tested the performance of the proposed model on imbalanced dataset.

III. Background

In this section, we describe the background of the algorithms that we selected for building the ORFDectector system

A. Logistic Regression

The logistic regression (LR) [31] model is a generalization of the linear regression model for binary classification. The LR algorithms compute a score for each data point. If the score is greater than 0.5 the instance is predicted as fraud otherwise it is predicted as legitimate. The score computation is done by assigning weights to each feature and inputting them in equation 1.

$$P(d_i) = \frac{e^{\alpha + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}}{1 + e^{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}} \quad (1)$$

B. J48

The J48 algorithm is the WEKA [6] implementation of the C4.5 algorithm. J48 algorithm first builds a decision tree. This tree is built using information of each attribute. At each step the attribute having maximum information gain is selected and the branches are created based on the unique values of the selected attribute. At the time of prediction, the new data instance traverses the tree from root to leaf, based on its attribute values, to find its class label.

C. Random Forest

Random Forest (RF) [7] is an ensemble-based algorithm that creates several decision trees in the training phase. At the time of prediction, a majority of all the trees is taken. At the time of tree creation, it randomly selects some of the attributes and select one of them for splitting.

D. Ensemble Techniques

Ensemble learning helps in improving the results of machine learning algorithm by combining the predictive power of multiple algorithms. The ensemble techniques are used to combine the prediction of base classifiers in some manner. In this work, we use three ensemble techniques, Average Vote (AV), Majority Vote (MV) and Maximum Vote (MXV). The AV techniques computes the confidence score of base classifiers for each data point. It then calculates average of confidence scores given by each base classifier. It then compares this average score with some threshold value. If the average score is higher than the confidence score the data point is predicted as legitimate otherwise it is predicted as fraud. Similar to this, the Maximum Vote (MXV) ensemble algorithm identifies the maximum score among the confidence scores given by each base classifier and compares it with the threshold value. The Majority Vote (MV) ensemble learning algorithm takes the majority vote of the predictions of these base classifiers to make the prediction. If more than 50% of the classifiers predicts a data point as fraud, the data point is predicted as fraud, otherwise it is predicted as legitimate.

IV. Proposed Framework

In this section, we describe the steps that we used to create the proposed ORFDetector framework.

A. *Instance Collection (Step 1):* We use publically available datatset shared by Vidros et al. [5]. This dataset consists of a total of 17,880 jobs out of which 866 are fraud and 17014 are legitimate.

B. *Feature Extraction (Step 2):* We next extract 21 features identified by Vidros et al. [5]. These features are placed into three categories: *Lingustic, Contextual* and *Metadata. Lingustic* features are related to the type of that the employer has used in the job description. For example, uses of the word 'money' in the job description in the title. *Contextual* features are related to the context of the job. For example, the level of education the employer is looking for. *Metadata* features are related to the employer location, questions, logo etc. Table 1, lists all the 21 features used in this work.

C. *Predictive Model Building:* We build ORFDetector model using 3 base classifiers (J48, LR, and RF) and 3 ensemble techniques (AV, MV, and MXV). Using these, we create following 3 combinations:

ORFDetector$_{AV}$: It is created by applying AV ensemble technique on J48, LR, and RF.

ORFDetector$_{MV}$: It is created by applying MV ensemble technique on J48, LR, and RF.

ORFDetector$_{MXV}$: It is created by applying MXV ensemble technique on J48, LR, and RF.

D. *Prediction Phase:* ORFDetector is used to predict the label of new jobs advertisement. For this we extract, all the 21 features from the target job advertisement and create the feature vector. For example, we extract spam words present in the company_profile, description, requirements, benefits. The feature vector is then given the ORFDetector to predict the label. Equation 1 shows the criteria for fraud or legitimate job prediction.

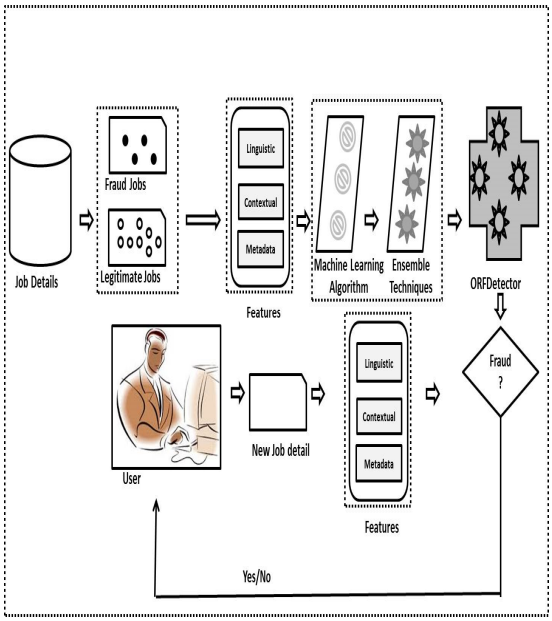


Fig. 2. ORFDetector Framework

Table I: Features used in this study for fraud job prediction as mentioned by Vidros et al. [5]

S. No	Feature	Type
1	contains-spamwords	Linguistic
2	has-consecutive-punctuation	
3	contains-money-in-title	
4	contains-money-in-description	
5	has-no-company-profile	Contextual
6	has-short-company-profile	
7	has-no-long-company-profile	
8	has-short-description	
9	has-short-requirements	
10	contains-email-link	
11	prompts-for-external-application	
12	addresses-lower-education	
13	located-in-us	Metadata
14	is-telecommuting	
15	has-no-company-logo	
16	has-no-questions	
17	has-emphasized-description	
18	has-emphasized-requirements	
19	has-emphasized-benefits	
20	has-no-html-lists-in-requirements	
21	has-no-html-lists-in-benefits	

V. EXPERIMENTAL DETAILS

A. *Dataset Details*

We conduct all our experiments on publicly available dataset share by Vidros et al. [5]. This dataset consists of 17,860 annotated jobs out of which 866 (4.84%) are fraudulent and 16994 (95.15%) are legitimate (refer to Table II for more details). This dataset is highly imbalance in nature as the ratio of fraudulent and legitimate jobs is 1:21(approx.).

Table II: Experimental dataset detail

Field	Value
Total Instances	17860
Fraud Jobs	866
Legitimate Jobs	16994
Time period	2011-2014

B. *Evaluation metrics*

We use six metrics for evaluating the performance of ORFDetector with the four different metrics:

1. **Accuracy**: The percentage of data points correctly predicted out of the total data points.
2. **Precision:** The percentage of data points correctly predicted as fraud out of the total instances that are predicted as fraud.

3. **Recall**: The percentage of data points predicted as fraud out of the total data points that are fraud.
4. **F1-measure:** It is the harmonic mean of precision and recall.
5. **Sensitivity**: It is the proportion of true positive cases that were classified as positive.
6. **Specificity**: It is the proportion of false positive cases that were classified as false.

VI. RESULTS

The ORF prediction is a binary class problem, i.e., there are two cases- either a job is fraud (positive class) or legitimate (negative class). For carrying out our experiments the positive class is labeled as 1 and the negative class is labeled as 0. The ORF dataset is imbalanced in nature, i.e., number of positive class instances is much lower as compared to negative class instances. The previous approach applies dataset balancing technique before ORF prediction. In contrast, in this works, we tested the performance of predictors on the imbalanced dataset. As in real worlds the dataset is often imbalanced in nature and hence, it is important to design predictors that can give good performance on imbalanced dataset. In this work, we use three base classifiers and 3 ensemble techniques. We use WEKA implementation of all the algorithms and use default WEKA parameters for all the algorithms. In the following subsection we present results of the experiments.

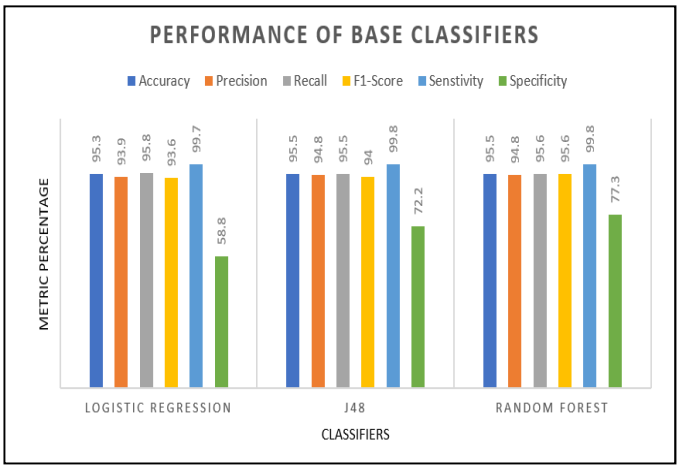


Fig. 3. Performance of base classifiers

Table III: Results of baseline Classifiers

Metrics	Classifiers			
	Logistic Regression	J48	Random Forest	Avg.
Accuracy	95.3%	95.5%	95.5%	95.4%
Precision	93.9%	94.8%	94.8%	94.5%
Recall	95.3%	95.5%	95.6%	95.6%
F1-Score	93.6%	94%	95.6%	94.4%
Sensitivity	99.7%	99.8%	99.8%	99.7%
Specificity	58.8%	72.2%	77.3%	69.4%

A. *RQ 1: What is the performance of base classifiers for the task of fraud job prediction?*

Motivation & Approach: In this RQ, we first test the performance of baseline classifier, i.e., LR, J48, and RF, to identify their performance for ORF prediction on the imbalanced dataset. It is important to compute the performance of baseline classifiers to find how accurately they can predict the fraud jobs. We tested the performance of base classifiers using six different parameters, i.e., accuracy, precision, recall, f1-measure, specificity and sensitivity.

Results: Table III and Fig.3 shows the results obtained by base classifiers. Table 3 shows that the average accuracy achieved by the base classifiers is 95.4%. The average values of precisions, recall, F1-measure, sensitivity and specificity are 94.5, 95.6, 94.4, 99.4, 69.4. This shows that in spite of giving good accuracy the value of specificity is low, i.e., only 69%. That means that a large number of legitimate companies are classified as fraud. This can have negative impact on the reputation the company. Additionally, this can harm several job seekers as they will miss a legitimate job opportunity. Hence, it is important to increase the value specificity.

B. *RQ 2: What is the performance of ORFDetector classifiers for fraud job prediction?*

Motivation & Approach: In this RQ, we test the performance of the proposed ORFDetector model to find its performance for the task ORF prediction. The results obtained by this RQ will be beneficial in identifying the effectiveness of ensemble techniques in ORF prediction. Ensemble based techniques have been found useful in increasing the performance of predictions in various applications. However, there performance for the task of ORF prediction is unexplored yet.

Results: Figure 4 and Table IV shows the results obtained by the proposed ORFDetector framework. The proposed ORFDetector give the average accuracy, precision, recall, F1-measure, sensitivity, and specificity of 95.5%, 94.9%, 95.6%, 94%, 95.7%, and 77.8%. These results indicate that ensemble based techniques are useful in improving

the performance of baseline classifier. ORFDetector framework give f1-score 94% which is nearly equal to the F1-score given by baseline classifiers. Additionally, it improved the value of specificity by 8% as compared to baseline classifiers.

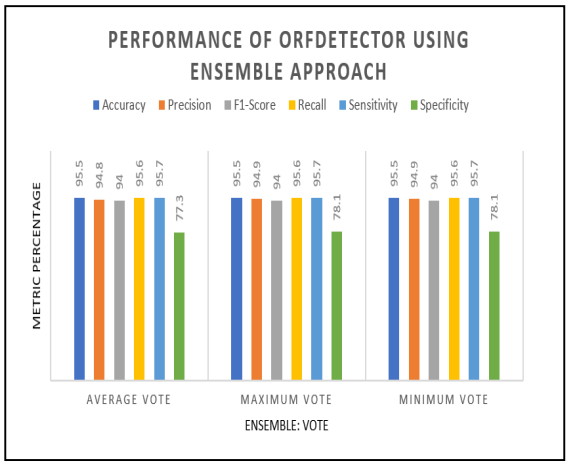


Fig. 4. Results of ORFDetector Framework

Table IV: Results of ORFDetector Framework

Metrics	**Ensemble Approach**			
	Average Vote	Maximum Vote	Minimum Vote	Average
Accuracy	95.5%	95.5%	95.5%	95.5%
Precision	94.8%	94.9%	94.9%	94.9%
Recall	95.6%	95.6%	95.6%	95.6%
F1-Score	94%	94%	94%	94%
Sensitivity	95.7%	95.7%	95.7%	95.7%
Specificity	**77.3%**	**78.1%**	**78.1%**	**77.8%**

VII. CONCLUSIONS AND FUTURE WORK

Online recruitment is an easy and efficient way to recruit good candidates for an organization. However, it has given rise to a new type of fraud in cyber world, i.e., ORF. The ORF is a big problem in the cyber world. In ORF cyber criminals steal personal information and money of innocent candidates by giving them lucrative job offers. ORF is an important but is has not received much attention from the research community. In this work, we proposed an ensemble-based model ORFDetector for ORF detection. The proposed model is found to be effective and give average f1-score and accuracy of 94% and 95.4, respectively. Additionally, it increases the specificity by 8% as compared to the baseline classifiers.

In future, we plan to extend the work presented in this paper to several novel and interesting future directions. First, we will work on identifying more feature to differentiate fraud and legitimate jobs to improve our current feature set. Second, we will evaluate the proposed model on more datasets to test the generalizability of the propose model.

References

[1] Naukri.com, https://www.naukri.com/ [accessed on 21/4/2018]

[2] monster, http://www.monsterindia.com/ [accessed on 21/4/2018]

[3] Bengalureans lose up to Rs 2 crore to online job scamsters, https://timesofindia.indiatimes.com/city/bengaluru/bengalureans-lose-up-to-rs-2-crore-to-online-job-scamsters/articleshow/59361782.cms [accessed on 22/4/2018]

[4] ABB, http://new.abb.com/ [accessed on 22/4/2018]

[5] Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. Future Internet, 9(1), 6.

[6] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18.

[7] Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.

[8] Naukri.com, https://en.wikipedia.org/wiki/Naukri.com [accessed on 04/06/2018]

[9] CareerBuilder. Think You Can Spot a Fake Resume? 2015. Available online: http://thehiringsite. careerbuilder.com/2012/05/04/think-you-can-spot-a-fake-resume [accessed on 7 May 2015]

[10] Vidros, S.; Kolias, C.; Kambourakis, G. Online recruitment services: Another playground for fraudsters. Comput. Fraud Secur. 2016, 2016, 8–13

[11] Indeed. Job Forum. 2015. Available online: http://www.indeed.com/forum (accessed on 8 May 2015).

[12] Mashable. 10 Signes a Job Is a Scam. 2015. Available online: http://mashable.com/2013/10/05/10-signs-ajob-is-a-scam (accessed on 8 May 2015).

[13] Agrawal, B.; Kumar, N.; Molle, M. Controlling spam emails at the routers. In Proceedings of the ICC 2005—2005 IEEE International Conference on Communications, Seoul, Korea, 15–19 May 2005; Volume 3, pp. 1588–1592.

[14] Zhang, Y.; Hong, J.I.; Cranor, L.F. Cantina: A content-based approach to detecting phishing web sites. In Proceedings of the 16th international conference on World Wide Web, Banff, AB, Canada, 8–12 May 2007; ACM: New York, NY, USA, 2007; pp. 639–648.

[15] Wenyin, L.; Huang, G.; Xiaoyue, L.; Min, Z.; Deng, X. Detection of phishing webpages based on visual similarity. In Proceedings of the Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, Chiba, Japan, 10–14 May 2005; ACM: New York, NY, USA, 2005; pp. 1060–1061.

[16] Ott, M.; Choi, Y.; Cardie, C.; Hancock, J.T. Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, Oregon, 19–24 June 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 309–319.