A COMPARATIVE STUDY ON FAKE JOB POST PREDICTION USING DIFFERENT DATA MINING TECHNIQUES

Mrs. N. Baby Rani
Assiatant Professor
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda, Hyd-500100, Telangana, India.

P. Yasaswini
Student
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda, Hyd-500100, Telangana, India.

P. Anjani
Student
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda, Hyd-500100, Telangana, India.

P.Kanaka Durga Bhavani
Student
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda, Hyd-500100, Telangana, India.

Abstract—The study suggests an automated way of preventing bogus job postings online that uses categorization techniques based on machine learning. To determine the most effective model for identifying job scams, the output of multiple classifiers was compared. In order to verify false internet postings, these classifiers are used. In the midst of several other ads, it aids in identifying fraudulent job listings. The two fundamental categories of classifiers considered for the purpose.

INTRODUCTION

In modern time, the development in the field of industry and technology has opened a huge opportunity for new and diverse jobs for the job seekers. With the help of the advertise-ments of these job offers, job seekers find out their options depending on their time, qualification, experience,suitability etc. Recruitment process is now influenced by the power of internet and social media. Since the successful completion of a recruitment process is dependent on its advertisement, the impact of social media over this is tremendous. Social media and advertisements in electronic mediahave created newer and newer opportunity to share job details. Instead of this, rapid growth of opportunity to share job posts has increased the percentage of fraud job postings which causes harassment to the job seekers. So, people lacks in showing interest to new job postings due to preserve security and consistency of their personal, academic and professional information. Thus, the true motive of valid job postings through social and electronic media faces an extremely hard challenge to attain people's belief and reliability. Technologies are around us to make our life easy and developed but not to create unsecured environment for professional life. If jobs posts can be filtered properly predicting false job posts, this will be a great advancement for recruiting new employees. Fake job posts create inconsistency for the job seeker to find their preferable jobs causinga huge waste of their time. An automated system to predict false job post opens a new window to face difficulties in the field of Human Resource Management.

Online job advertisements which are fake and mostly willing to steal personal and professional information of job seekers instead of giving right jobs to them is known as job scam. Sometimes fraudulent people try to gather money illegally from job seekers. A recent survey by Action Fraud from UK has shown that more than 67% people are at great risk who look for jobs through online advertisements but unaware of fake job posts or job scam. In UK, almost 700000 job seekers complained to lose over $500000 being a victim of job scam. The report showed almost300% increase over the last two years in UK. Students, fresh graduates are being mostly targetedby the frauds as they usually try to get a secured job for which they are willing to pay extra money. Cybercrime avoidance or protection techniques fail to decrease this offence since frauds change theirway of job scam very frequently. Fraudsters who want to gain other people's personal information like insurance details, bankdetails, income tax details, date of birth, national id create fake job advertisements. Advance fee scams occur when frauds ask for money showing reasons like admin charges, information security checking cost, management cost etc.

EXISITING SYSTEM

Many researches occurred to predict if a job post is real or fake. A good number of research works are to check online fraud job advertiser. They experimented on EMSCAD dataset using several classification algorithms like naive bayes classifier, random forest classifier, Random Forest Classifier showed the best performance on the dataset with 89.5% classification accuracy.

They found logistic regression performing very poor on the dataset. One R classifier performed well when the balanced the dataset and experimented on that. They tried in their work to find out the problems in ORF model (Online Recruitment Fraud) and to solve those problems using various dominant classifiers.

They applied feature selection technique to reduce the number of attributes effectively and efficiently.

Disadvantages-

The system is implemented by Conventional Machine Learning.

The system doesn't implement for analyzing large data sets.

PROPOSED SYSTEM

The system has used EMSCAD to detect fake job post. This dataset contains 18000 samples and each row of the data has 18 attributes including the class label. The attributes are job_id, title, location, department, salary_range, company_profile, description, requirements, benefits, telecommunication, has_company_logo, has_questions, employment_type, required_experience, required_education, industry, function, fraudulent (class label). Among these 18 attribute, we have used only 7 attributes which are converted into categorical attribute. Telecommuting, has_company_logo, has_questions, employment_type, required experience, required_education and fraudulent are changed into categorical value from text value. For example, "employment_type" values are replaced like this- 0 for "none", 1 for 'full-time", 2 for "part-time" and 3 for "others", 4 for "contract' and 5 for "temporary". The main goal to convert these attributes into categorical form is to classify fraudulent job advertisements without doing any text processing and natural language processing. In this work, we have used only those categorical attributes.

ADVANTAGES

The proposed has been implemented EMSCAD technique which is very accurate and
Sfast.

The system is very effective due to accurate detection of Fake job posts which creates inconsistency for the job seeker to find their preferable jobs causing a huge waste of their time.

SYSTEM REQUIREMENTS

H/W System Configuration:

Processor : Pentium –IV
RAM : 4 GB (min)

Hard Disk : 20 GB

Key Board : Standard Windows Keyboard

Mouse: Two or Three Button Mouse

Monitor : SVGA

SOFTWARE REQUIREMENTS

Operating system : Windows 7 Ultimate

Coding Language : Python.

Front-End : Python.

Back-End : Django-ORM

Designing : Html, css, javascript.

Data Base : MySQL (XAMP Server).

FEASIBILITY

An important outcome of preliminary investigation is the determination that the system requestis feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employeesconnected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

Algorithm:

Random Forest is a popular machine learning algorithm that belongs to the supervised learningtechnique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complexproblem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

IMPLEMENTATION

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designedto provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error is in the input are controlled by the input design. The application has been developed in user-friendly manner.The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. After completion of a project, a new project may be assigned to the client. User authentication procedures are maintained at the initial stages itself

CONCLUSION

Job scam detection has become a great concern all over the world at present. In this paper, we have analyzed the impacts of job scam which can be a very prosperous area in research filed creating a lot of challenges to detect fraudulent job posts. We have experimented with EMSCAD dataset which contains real life fake job posts. In this paper we have experimented both machine learning algorithms (SVM, KNN, Naïve Bayes Random Forest and MLP) and deep learning model (Deep Neural Network). This work shows a comparative study on the evaluation of traditional machine learning and deep learning based classifiers. We have found highest classification accuracy for Random Forest Classifier among traditional machine learning algorithms and 99 % accuracy for DNN (fold 9) and 97.7% classification accuracy on average for Deep Neural Network.

REFERENCES

[1] S.Vidros, C. Kolias , G. Kambourakis ,and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and aPublic Dataset", Future Internet 2017, 9, 6; doi:10.3390/fi9010006.

[2]B. Alghamdi, F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", Journal of Information Security, 2019, Vol 10, pp. 155176,https://doi.org/10.4236/iis.2019.103009 .

[3]Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia-Tuan Nguyen, "Job Prediction: From Deep Neural Network Models to Applications", RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.

[4]Jiawei Zhang, Bowen Dong, Philip S. Yu, "FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network", IEEE 36th International Conference on Data Engineering (ICDE), 2020.

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

[6]Y. Kim, "Convolutional neural networks for Sentence classification," arXiv Prepr. arXiv1408.5882, 2014.

[7] 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.

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

[9] 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.