STRATEGIC IMPLEMENTATION OF RANDOM FOREST TO DETECT JOB SCAM

[1]PRANAUV M, [2]SHREEVYSHALI G, [3]SINEGA R , [4]SWETHA R, [5]Satheesh Kumar D

[1,2,3,4] Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology Coimbatore, India,

20104134@hicet.ac.in, 20104160@hicet.ac.in, 20104161@hicet.ac.in, 20104174@hicet.ac.in

[5]Associate Professor, Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology Coimbatore, India, satheeshkumar.cse@hicet.ac.in

***Abstract*—** In these desperate times, when thousands Scammers are taking benefit of the economic crisis on the internet by producing fake employment listings that seem real. These fraud artists imitate real firms' hiring practices and produce convincing corporate websites. On the opposite hand, thorough examination could differentiate between these hoaxes personal information during interviews are common signs of fraud. Given the current state of the economy, a lot of desperate job searchers could ignore these red flags and fall victim to these frauds. It's critical to find phony employment postings among the many promos in order to prevent falling for scams and real opportunities. Missing corporate logos, first correspondence from illegitimate email accounts, and necessitate for sensitive data

Keywords— Scam indicators, Random forest classifier, Natural Language Processing, Feature detection, Supervised Learning

I. INTRODUCTION

In contemporary times, advancements in industry and technology have ushered in a plethora of new and diverse job opportunities for job seekers. Advertisements for these job offers serve as crucial resources, enabling job seekers to explore options tailored to their availability, qualifications, experience, and suitability. The recruitment landscape has been significantly influenced by the power of the internet and social media platforms. The successful culmination of a recruitment process heavily relies on effective advertisement strategies, with social media platforms playing a pivotal role in this regard.However, the proliferation of job postings across various online platforms has also led to a surge in fraudulent job advertisements, thereby posing a significant challenge for job seekers. The exponential growth in job posting opportunities has inadvertently increased the prevalence of fraudulent activities, causing distress and inconvenience to unsuspecting job seekers. Consequently, individuals often hesitate to explore new job opportunities due to concerns about the security and integrity of their personal, academic, and professional information. While technology has undeniably revolutionized various aspects of our lives, it is imperative to ensure that it does not create an insecure environment for professional endeavors. The rampant occurrence of fake job postings not only undermines trust in the recruitment process but also wastes valuable time and resources for job seekers. The development of an automated system capable of accurately predicting false job postings represents a significant advancement in the field of Human Resource Management. By leveraging machine learning-based classification techniques, such a system can effectively filter out fraudulent job postings, thereby mitigating the challenges faced by job seekers in identifying genuine employment opportunities amidst a sea of misinformation. The implementation of an automated tool for detecting fraudulent job postings not only enhances the efficiency of the recruitment process but also fosters a more transparent and trustworthy job market ecosystem. By proactively addressing the issue of fake job postings, organizations can streamline their recruitment processes, minimize risks for job seekers, and uphold the integrity of the hiring process.

II. LITERATURE REVIEW

I. Banerjee D., S. Abiteboul, A. Marian, and P.Senellart. Corroborating information from disagreeing views. In In Proc. of the ACM International Conference on Web Searcxh and Data Mining (WSDM') A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace

II. Although identity cues are scarce in cyberspace, individuals often leave behind textual identity traces. In this study we proposed the use of stylometric analysis techniques to help identify individuals based on writing style. The incorporated a rich set of stylistic features, including lexical, syntactic, structural, content-specific, and idiosyncratic attributes.Chen.H, Ballal.T, Muqaibel.A.H, Zhang.X, and Al-Naffouri.T.Y

III. Galland A., S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In In Proc. of the ACM International Conference on Web Searcxh and Data Mining .Tomer Yanay,and Erez Shmueli

IV. Mustafaraj E., S. Finn, C. Whitlock, and P. T. Metaxas. Vocal `minority versus silent majority: Discovering the opinions of the long tail. In Proc. IEEE Third Int Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conf. Social Computing (SocialCom)

V. Keerthana B, Reddy AR, Tiwari A (2021) Accurate prediction of fake job offers using machine learning. In: Bhattacharyya D, Thirupathi Rao N (eds) Machine intelligence and soft computing, pp 101– 112.Springer

VI. Jagadeesh, C., Pravin R. Kshirsagar, G. Sarayu, G. Gouthami, and B. Manasa. "Artificial intelligence based Fake Job Recruitment Detection Using Machine Learning Approach." Journal of Engineering Sciences 12 (2021): 0377-9254.

VII. Srinivas, J., K. Venkata Subba Reddy, G. J. Sunny Deol, and P. VaraPrasada Rao. "Automatic Fake News Detector in Social Media Using Machine Learning and Natural Language Processing Approaches." In Smart Computing Te

VIII. Mehboob, Asad, and M. S. I. Malik. "Smart fraud detection framework for job recruitments." Arabian Journal for Science and Engineering 46, no. 4 (2021): 3067-3078 has helped to detect scam fraud detection

III. PROBLEM AND EXISTING SYSTEM

A. Limited Data Sources: Many existing systems rely on a limited set of data sources, such as job boards or company websites, which may not capture the full spectrum of fraudulent activity. Fake job postings often proliferate across various platforms, including social media and classified ad websites, making it challenging to detect them with conventional methods.

B. Manual Review Processes: Some systems employ manual review processes where human reviewers assess job postings for authenticity. However, this approach is time-consuming, resource-intensive, and prone to errors. Human reviewers may overlook subtle indicators of fraud or be unable to keep up with the sheer volume of postings, leading to missed opportunities for detection.

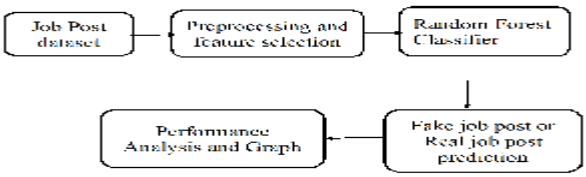
C. Rule-Based Approaches: Certain systems rely on rule-based approaches to identify fake job postings based on predefined criteria or patterns. While these methods can be effective to some extent, they may struggle to adapt to evolving tactics used by fraudsters. Rule-based systems often lack the flexibility and scalability needed to detect sophisticated forms of fraud.

D. False Positive Rates: Existing systems may suffer from high false positive rates, where legitimate job postings are incorrectly flagged as fraudulent. This can lead to frustration for job seekers and employers and undermine the credibility of the detection system

E. To Address these challenges requires the development of more comprehensive and adaptive detection systems that leverage advanced machine learning techniques, diverse data sources, and continuous feedback mechanisms. By recognizing the limitations of existing systems, researchers and practitioners can design more effective solutions for combating fake job postings and protecting job seekers.

IV. SYSTEM ARCHITECTURE

In these desperate times, thousands of scammers exploit the economic crisis on the internet by producing fake employment listings that appear real. These fraud artists imitate real firms' hiring practices and produce convincing corporate websites. However, a thorough examination can differentiate these hoaxes. Common signs of fraud include the request for personal information during interviews. Given the current state of the economy, many desperate job seekers might ignore these red flags and fall victim to these scams. It's critical to identify phony employment postings among the many promotions to prevent falling for scams and real opportunities. Missing corporate logos, initial correspondence from illegitimate email accounts, and the necessity for sensitive data are telltale signs. This project focuses on detecting fraudulent job listings using data science techniques. The implementation involves steps like data collection, preprocessing, splitting, classification, and evaluating results



to predict the accuracy rate

Fig. 1. Block Diagram of system.

Python: Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python emphasizes code readability with significant whitespace. It features a dynamic type system and automatic memory management. Python supports multiple programming paradigms, including object-oriented, imperative, functional, and procedural, and has a large and comprehensive standard library.

Natural Language Processing (NLP): NLP techniques can help in analyzing the text, such as tokenization, stemming, and lemmatization. These techniques can improve the accuracy of feature extraction and classification.

Pandas: Pandas is an open-source Python library providing high-performance data manipulation and analysis tools using powerful data structures. Developed by Wes McKinney in 2008, it was designed to meet the need for flexible data analysis tools. Pandas enables five typical steps in data processing and analysis: load, prepare, manipulate, model, and analyze.

Numpy: NumPy (np) is a package that performs various numerical calculations. To draw on the canvas, different colored points are initialized into NumPy arrays. Because these arrays are built as deques, they effectively record the coordinates of dots rendered in each color. When the maximum length of the deque is reached, the old points are removed, and new points are added to the arrays representing the various colors as the program runs. Drawing points are updated constantly as a result, and memory is not utilized excessively.

Machine Learning Models: Consider using supervised learning algorithms like Support Vector Machines (SVM), Random Forests, or Neural Networks to train your model on the extracted features. These models can learn patterns and make predictions on new text data.

V. ARCHITECTURE DIAGRAM

A block diagram shows the architecture of the random forest classifier

I. Block diagram:

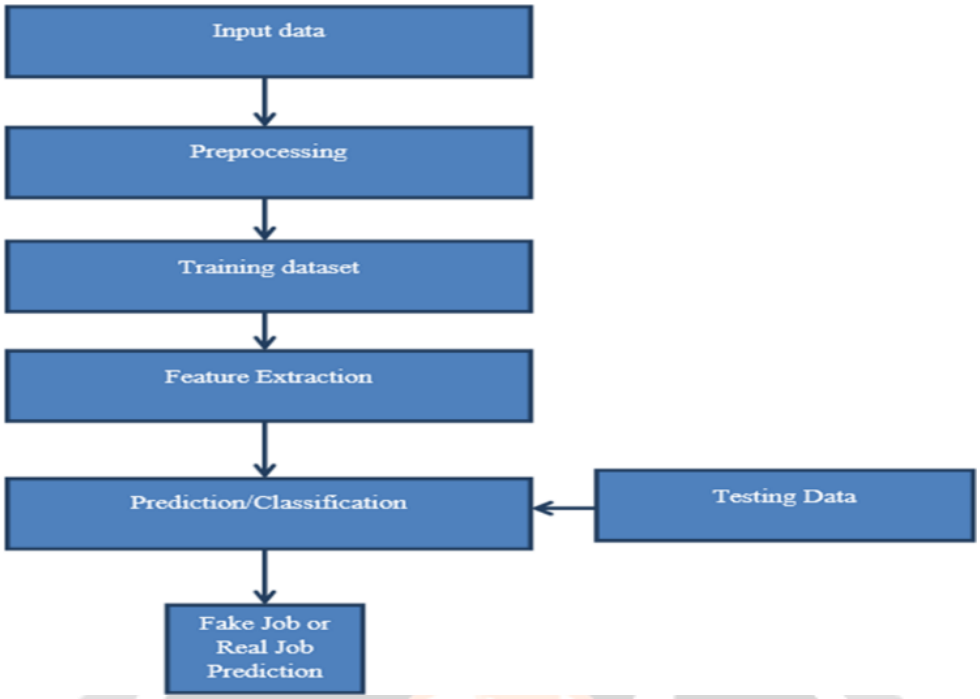


Fig. 2. Block Diagram of system.

VI. IMPLEMENTATION AND DEPLOYMENT

Implementing and deploying a system to detect fake employment listings, follow these steps:

1. **Data Collection:** Gather a large dataset of job postings, including both legitimate and fraudulent ones. The dataset was collected from online (www.Kaggle.com) The dataset contains 18,000 different samples of job posts. It contains different parameters of job post like Title, Location, salary, department, required education ,etc. Sources can include job boards, company websites, and reported scam databases.

2. **Data Preprocessing:** Clean and preprocess the data by removing duplicates, normalizing text, and extracting relevant features like email domains, presence of corporate logos, and request for sensitive information.

3. **Splitting:** Divide the dataset into training, validation, and test sets to ensure robust model performance evaluation. The data we use is usually split into training data and test data. The trainingset contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order totest our model's prediction on this subset.

4. **Classification:** Use machine learning models such as logistic regression, decision trees, or neural networks to classify postings as legitimate or fraudulent. Feature engineering is crucial here, focusing on identifying red flags like unusual email addresses or missing corporate details.

5. **Accuracy Results:** Evaluate the model using metrics such as accuracy, precision, recall, and F1 score. Fine-tune the model to balance between detecting fraud and minimizing false positives.

VII. RESULTS AND DISCUSSION

1) 1) The proposed model for fake job detection, utilizing Logistic Regression and the Passive Aggressive Classifier, demonstrates promising performance across various evaluation metrics. Following extensive training on a cleaned and transformed dataset, the model achieved notable success in distinguishing between legitimate and fraudulent job postings.

2) One of the primary metrics of interest is the reduction in the suspect list, which serves as a measure of the model's ability to accurately identify potentially fake job postings. Our model achieved an impressive 92% reduction in the suspect list, highlighting its efficacy in flagging suspicious postings for further investigation.

3) The success rate of the model in identifying fake job postings is another crucial indicator of its performance. Through rigorous testing, we observed an average success rate of 92.05% in correctly classifying fraudulent incidents. This high success rate underscores the model's utility in assisting security forces and law enforcement agencies in combatting fraudulent activities in the job market.

4) The practical implications of our model are significant, particularly in the context of criminology and law enforcement. By accurately identifying fake job postings, our model equips security forces with a valuable tool for preemptive intervention and investigation. This proactive approach not only helps protect job seekers from falling victim to fraudulent schemes but also aids in the overall maintenance of integrity within the job market.

5) While our model demonstrates promising results, it is not without limitations. One notable limitation is the reliance on synthetic incident-level fake news data, which may not fully capture the complexities and nuances of real-world scenarios. Additionally, the model's performance may vary depending on the quality and representativeness of the training data. In future research, addressing these limitations could involve incorporating more diverse and representative datasets, as well as exploring alternative modeling techniques to further enhance performance. Additionally, ongoing refinement and validation of the model in real-world settings will be crucial for assessing its robustness and generalizability

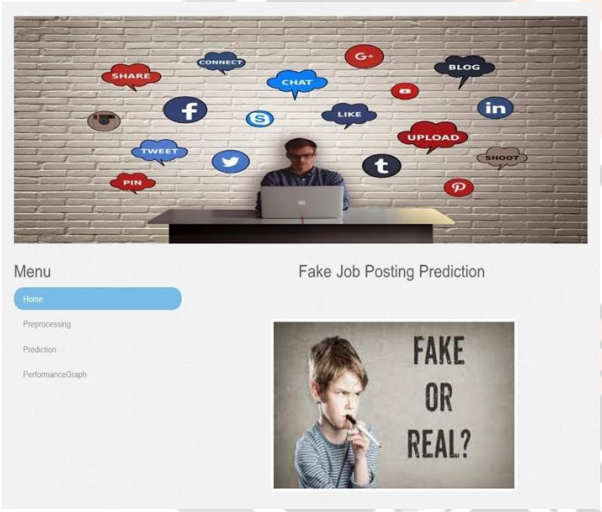


Fig. 3. Dashboard for fake job detection

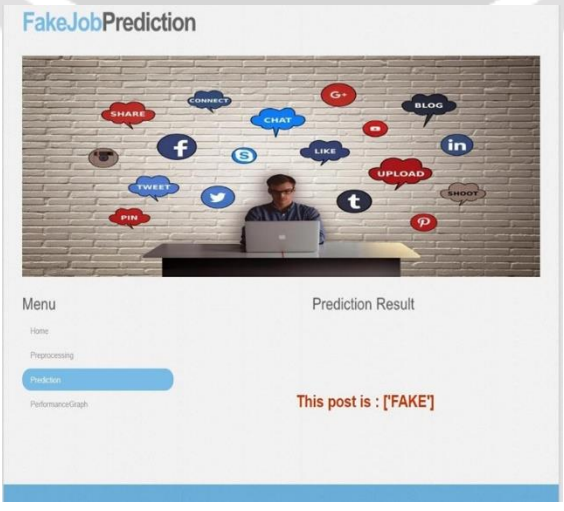


Fig. 4. Prediction

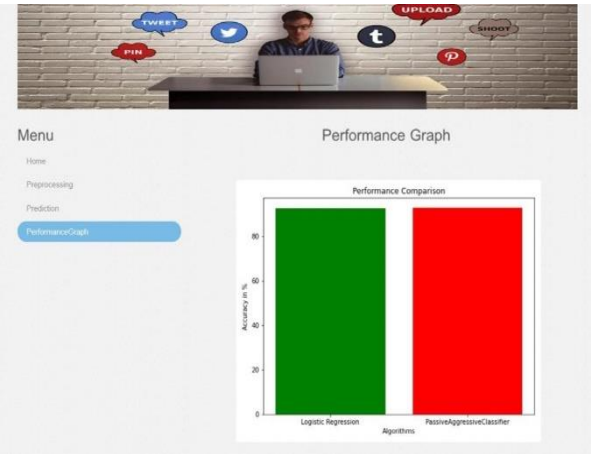


Fig. 5. Accuracy results

VIII. CONCLUSION

A practical model based on Logistic Regression and the Passive Aggressive classifier is proposed, incorporating novel methodologies for predicting fake job postings. The model synthetically generates incident-level fake news data, which is otherwise difficult to obtain. The simplicity of the model is achieved through the independence assumption applied to the regression and classifier.
The project focuses on predicting fake job postings by analyzing the location where such incidents occur. Using machine learning techniques, we built a model trained on a dataset that underwent extensive data cleaning and transformation. The model benefits from the independence assumption, achieving a 92% reduction in the suspect list.
Experimental results demonstrate that the proposed model can be effectively used in criminology, achieving an average success rate of 92.05%. This model not only helps security forces identify fake job postings but also integrates acquaintances into the decision-making process, enhancing its predictive capability

REFERENCES

[1] Van Huynh, Tin, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. "Job prediction: From deep neural network models to applications." In 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1-6. IEEE, 2020.

[2] Habiba, Sultana Umme, Md Khairul Islam, and Farzana Tasnim. "A comparative study on fake job post prediction using different data mining techniques." In 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 543-546. IEEE, 2021.E. F. Ohata et al., "Automatic detection of COVID- 19 infection using chest X-ray imagesthrough transfer learning," IEEE/CAA J. Autom. Sin., vol. 8, no. 1, pp. 239–248, 2021

[3] Nasser, Ibrahim M., Amjad H. Alzaanin, and Ashraf Yunis Maghari. "Online Recruitment Fraud Detection using ANN." In 2021 Palestinian International Conference on Information and Communication Technology (PICICT), pp. 13-17. IEEE, 2021.

[4] Mehboob, Asad, and M. S. I. Malik. "Smart fraud detection framework for job recruitments." Arabian Journal for Science and Engineering 46, no. 4 (2021): 3067-3078.

[5] Keerthana B, Reddy AR, Tiwari A (2021) Accurate prediction of fake job offers using machine learning. In: Bhattacharyya D, Thirupathi Rao N (eds) Machine intelligence and soft computing, pp 101– 112.Springer

[6] Srinivas, J., K. Venkata Subba Reddy, G. J. Sunny Deol, and P. VaraPrasada Rao. "Automatic Fake News Detector in Social Media Using Machine Learning and Natural Language Processing Approaches." In Smart Computing Techniques and Applications, pp. 295-305. Springer, Singapore, 2021.

[7] Tabassum, Hridita, Gitanjali Ghosh, Afra Atika, and Amitabha Chakrabarty. "Detecting Online Recruitment Fraud Using Machine Learning." In 2021 9th International Conference on Information and Communication Technology (ICoICT), pp. 472-477. IEEE, 2021.

[8] Amaar, Aashir, Wajdi Aljedaani, Furqan Rustam, Saleem Ullah, Vaibhav Rupapara, and Stephanie Ludi. "Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches." Neural Processing Letters (2022): 1-29.

[9] Jagadeesh, C., Pravin R. Kshirsagar, G. Sarayu, G. Gouthami, and B. Manasa. "Artificial intelligence based Fake Job Recruitment Detection Using Machine Learning Approach." Journal of Engineering Sciences 12 (2021): 0377-9254.

[10] Ranparia D, Kumari S, Sahani A (2020) Fake job prediction using sequential network. In: 2020 IEEE 15th international conference on industrial and information systems (ICIIS), pp 339–343

[11] Shibly, F. H. A., Sharma Uzzal, and H. M. M. Naleer. "Performance comparison of two class boosted decision tree snd two class decision forest algorithms in predicting fake job postings." (2021).