FAKE JOB POST DETECTION USING MACHINE LEARNING

Mr. VAZIUDDIN MOHAMMED
Department of CSE (CS, DS)
MLR Institute of Technology
Hyderabad, India
vaziuddin.jits@gmail.com

GAJJALA AMULYA REDDY
Department of CSE (CS, DS)
MLR Institute of Technology
Hyderabad, India
gajjalaamulyareddy@gmail.com

JALAPATI UDAY KIRAN
Department of CSE (CS, DS)
MLR Institute of Technology
Hyderabad,India
udaykiranjalapati@gmail.com

A BHUVAN REDDY
Department of CSE (CS, DS)
MLR Institute of Technology
Hyderabad, India
bhuvanreddy19@gmail.com

SHARATH KUMAR CHAWLA
Department of CSE (CS, DS)
MLR Institute of Technology
Hyderabad, India
sharathchawla@gmail.com

Abstract- In recent years, due to advancements in modern technology and social communication, advertising new job posts has become a very common issue in the present world. So, fake job posting prediction tasks will be a great concern for all. Like many other classification tasks, fake job posing prediction leaves a lot of challenges to face. Here we use different data mining techniques and classification algorithms like KNN, decision tree, support vector machine, naive Bayes classifier, random forest classifier, multilayer perceptron and deep neural network to predict whether a job post is real or fraudulent. We have experimented on the Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. Deep neural network as a classifier, performs great for this classification task. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) in predicting a fraudulent job post.

Keywords :
Logistic Regression Algorithm,
Support Vector Machine, Naïve Bayes Algorithm.

I. INTRODUCTION

Employment swindling is one of the serious issues recently forwarded in the rule of Online Recruitment Frauds (ORF). In recent days, many parties have posted their vacancies connected to the internet so that these may be sure and timely for one task applicant. However, this intention grants permission be individual types of swindle by the deception community cause they offer employment to task-applicants in conditions of taking services from the ruling class. Fraudulent task advertisements can be situated against a believed association for violating their believe-ableness. These false task post detections draw a good consideration for acquiring an electrical tool for recognizing fake tasks. Reporting ruling class to people for preventing requests for aforementioned jobs. For this purpose, a machine intelligence approach is used that employs various categorization algorithms for perceiving fake posts. In this case, a classification form isolates fake task posts from the best set of task advertisements and alerts the user. To address the question of labelling scams on task posting, directed education invention as classification methods were deliberate originally. A classifier maps input changeable to aim classes by seeing the training dossier. Classifiers forwarded in the Project for labelling fake job posts from the remainder of something are specified concisely.

These classifiers can be widely categorized into –Single Classifier located Prediction and Ensemble Classifiers located Prediction. And authentic Job Recommendation system.

II. LITERATURE SURVEY

Vidro has labelled task scammers as fake online task messengers. They raise enumerations about many real and famous associations and adventures that caused fake task advertisements opening posts accompanying ill-motive. They investigated on EMSCAD dataset utilizing various classification algorithms like childlike Bayes classifier, haphazard thicket classifier, Zero R, One-R etc. Random Forest Classifier demonstrated the best conduct on the dataset accompanying 89.5% categorization accuracy. They erect logistic reversion operating very weak on the dataset. One-R classifier performed well when they equalized the dataset and experimented with that. They are reliable in their work to discover the problems in the ORF model (Online-Recruitment Fraud) and to resolve those questions utilizing miscellaneous dominant classifiers.

Alghamdi projected a model to discover deception exposure and connect to the internet conscription whole. They tested on the EMSCAD dataset using machine intelligence invention. They worked this dataset in three steps- dossier pre-processing, feature option and trickery discovery using the classifier. In the preprocessing step, they detached cacophony and HTML tags from the data so that the inexact document pattern remained maintained. They used feature selection techniques to weaken the number of attributes effectively and capably. Support Vector Machine was second hand for feature pick and an ensemble classifier utilizing random woodland was used to discover fake task posts from the test data. Random thicket classifier appeared as a forest-organized classifier which was processed as an ensemble classifier by way of most voting techniques. This classifier showed 97.4% categorization veracity to discover fake task posts.

Huynh proposed to use various deep interconnected system models like Text CNN, Bi-GRU- LSTM CNN and BiGRU CNN that are pre-trained accompanying handbook datasets. They processed classifying IT task datasets. They trained the IT task dataset on the Text CNN model containing a loop layer, combining tier and completely affiliated layers. This model prepared a dossier through spiral and pooling tiers. Then the prepared weights were levelled and given to the fully related coating. This model second-hand soft-max function for classification method. They likewise second-hand ensemble classifiers (Bi-GRU CNN, Bi-GRULSTM CNN) utilizing the majority vote method to increase categorization accuracy. They establish 66% categorization veracity utilizing Text CNN.

Kumar used machine intelligence methods to think of fake job posts, concentrating on look options to increase accuracy. Their approach was established to have a veracity of around 85%.

Zhang projected an automatic fake indicator model to equate valid and fake information (including items, inventors, and cases) using passage deal with. They had second-hand a custom dataset of revelation or articles situated by the PolitiFact site giggle account. This dataset was used to train the projected GDU long whole model. Receiving recommendations from multiple beginnings together, this prepared model performed well as a mechanical fake indicator model.

Malhotra grew a machine learning-located approach for detecting fake task posts utilizing diversified features established textbook, figure, and source believe-ableness. They are second-hand miscellaneous utilizing multiple facial characteristics established content,

image, and beginning believe-ableness. They second-hand miscellaneous or genuine established a preparation dataset of over 25,000 task posts. Their approach achieved an accuracy of 93% for detecting fake task posts on the dataset.

Tran projected a mixture of feature draft and deep learning approaches for recognizing fake job posts' established passage and metadata features. They secondhand Relief and Chi-Square feature draft orders to defeat the dimensionality of the feature room and work a Convolutional Neural Network (CNN) model for categorization. Their approach achieved an F score of 0.97 in detecting fake Job posts on the dataset.

Lee compared various machine intelligence and deep education models for recognizing fake job posts establishing diversified types of lineaments including passage, figure, and consumer nature. They used algorithms in the way that SVM, Random Forest, XGBoost, and CNN for classification and therefore judged the performance established differing verification to a degree of accuracy, accuracy, recall, and F-Score. They further distinguished the effect of different features inciting the model's conduct and erect that joining all types of features.

Chen projected a foundation that second-hand text excavating and machine intelligence methods to discover fraudulent task postings. The foundation handled cosine similarity to equate the task writing and the party profile to recognize some disagreements. The authors reported that their approach had an accuracy of 85% in labelling fake task posts.

Ravi projected a mixture approach that uses two together rule-located and machine intelligence techniques to identify fake task posts. The approach was proved to have an extreme veracity of over 90%.

Saini investigated the use of deep knowledge methods for anticipating fake job posts. They second-hand a Convolutional Neural Network (CNN) to categorize task postings as genuine or fake, realizing a veracity of 92%. The authors too distinguished their approach from other traditional machine knowledge algorithms and raised that the CNN outperformed bureaucracy.

III. EXISTING SYSTEM

K-Nearest Neighbours (KNN) Classifier:
The K-Nearest Neighbours algorithm is a plain and instinctive categorization order that belongs to the type of instance-located or inactive education algorithms. In KNN categorization, the class of a hidden dossier point is driven to establish the class labels of its k most forthcoming neighbours in the feature scope.
Key Characteristics:
Lazy Learner: KNN is frequently referred to as an inactive beginner cause it does not include specific preparation all along the model-construction point. Instead, it stores the entire preparation dataset and creates forecasts to establish the likeness between new instances and existent dossier points.
Distance Metric: The choice of distance rhythmical (for instance, Euclidean distance, Manhattan distance, etc.) plays an important duty in KNN categorization. It measures the similarity or closeness between dossier points in the feature room.
Hyperparameter k: One of the main challenges in KNN categorization is selecting the appropriate worth of k, which shows the number of most familiar neighbours deliberate for categorization. A tinier value of k grant permission brings about a more bendable resolution confine but power increases the influence of crash, while the best advantage of k grant permission influences a more flowing resolution boundary but takes care of disregarding local patterns.

ADVANTAGES OF THE EXISTING SYSTEM

- Simplicity
- Non-parametric
- Adaptability

DISADVANTAGES OF THE EXISTING SYSTEM

- Accuracy depends on the quality of the data.
- With large data, the prediction stage might be slow.
- Sensitive to the scale of the data and irrelevant features.
- Require high memory – need to store all of the training data.

IV. OBJECTIVES

To reinforce the accomplishment and applicability of the existent order promoting the K-Nearest Neighbours (KNN) classifier, various key objectives have been recognized. Firstly, skilled is a need to amend the hyperparameters of the KNN algorithm, specifically the worth of k, to help categorization accuracy and inference act. Additionally, surveying feature engineering methods will be important to embellish the bias power of the classifier and address the challenges formally by the extreme-spatial dossier. Given the prevalence of unstable datasets in honest-planet synopsises, developing designs to handle class shortcoming issues, in the way that adjusting class weights or engaging inspecting methods, will be essential for reconstructing the robustness of the classifier. Furthermore, exertions to reinforce the computational adeptness of the KNN treasure will be pursued, containing fact-finding approximate most familiar neighbour search algorithms and range reduction methods to allow scalability to the best datasets. Additionally, exploring arrangements to correct the classifier's strength to strident data and outliers, in addition to fact-finding ensemble approaches and transfer knowledge methods to further enhance allure acting, will have influence aims. Moreover, efforts should to reinforce the interpretability of the KNN classifier's determinations and explore time for merging it accompanying additional machine learning algorithms to forge mixture models that influence completing strengths. Finally, optimizing the classifier real-occasion forecast synopsises will be a key objective, involving the growth of methods to weaken query latency and correct openness in active atmospheres. Through addressing these aims, the aim search to considerably improve the effectiveness and flexibility of the existent order employing the KNN classifier across various machine intelligence uses.

V. PROPOSED SYSTEM

The target of this study is to detect whether a job post is fraudulent or not. Identifying and eliminating these fake job advertisements will help the seekers concentrate on only legitimate job posts. In this project, we are comparing three machine learning algorithms. Such as a Random forest Classifier and Decision Tree, and Add a boost to select the best predictive model for detecting fake Recruitments.

- < UNK> Using a machine learning algorithm, the whole data interpretation and analysis process is done by computer.
- No male intervention is required for the prediction or interpretation of data. The whole process of machine learning is machine starts learning and predicting the algorithm or program to give the best result.
- It can handle a variety of data, Even in an uncertain and dynamic environment,

it can handle a variety of data. It is multidimensional as well as a multitasker.

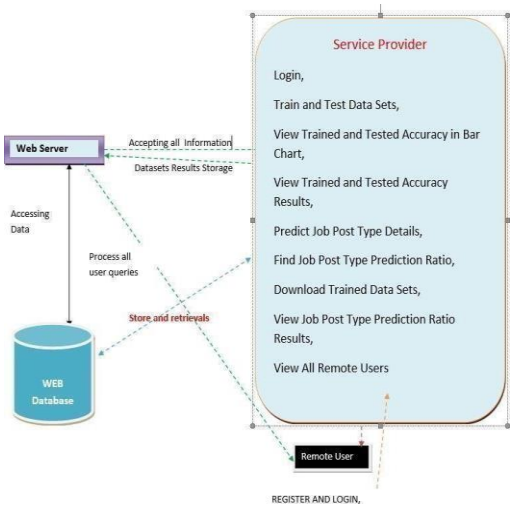

Fig 1: System Architecture of Proposed System

ADVANTAGES OF THE PROPOSED SYSTEM

- By using a machine learning algorithm, the whole process of data interpretation and analysis is done by computer.
- No male intervention is required for the prediction or interpretation of data. The whole process of machine learning is machine starts learning and predicting the algorithm or program to give the best result
- It can handle a variety of data, Even in an uncertain and dynamic environment, it can handle a variety of data. It is multidimensional as well as a multitasker.

VI. RESULTS AND DISCUSSIONS

Here are the screens and results of the proposed system.

Fig 2: Interface of Fake Job Post Prediction

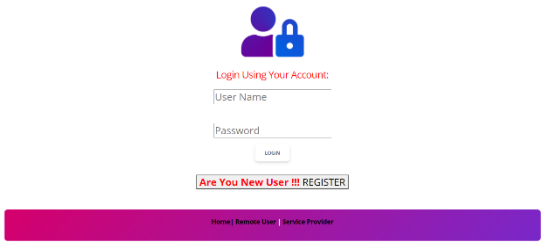

Fig 3: User login screen

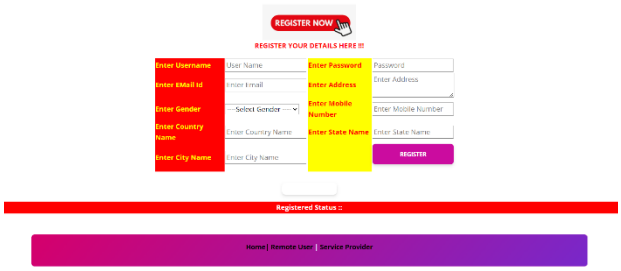

Fig 4: User registration screen

Fig 5: Service provider screen

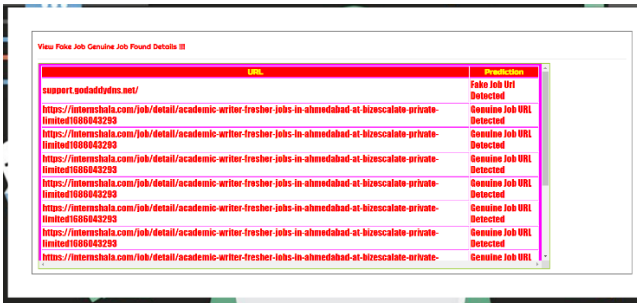

Fig 6: Job Prediction screen

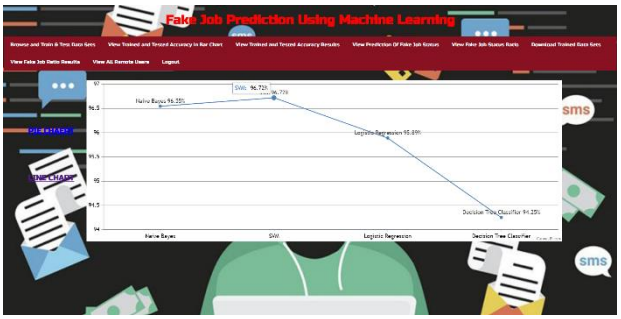

Fig 7: Test accuracy screen

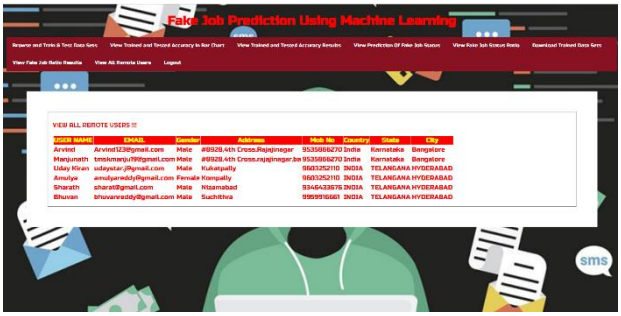

Fig 7: All remote users screen

VII. CONCLUSION AND FUTURE SCOPE

Job scam detection has become a great concern all over the world at present. In this paper, we have analyzed the impacts of job scams which can be a very prosperous area in the research field creating a lot of challenges to detect fraudulent job posts. We have experimented with the EMSCAD dataset which contains real-life fake job posts. In this paper, we have experimented with both machine learning algorithms (SVM, Naive Bayes, Random Forest and logistic classifier) and deep learning models (Deep Neural Network). This work shows a comparative study on the evaluation of traditional machine learning and deep learning-based classifiers. We have found the highest classification accuracy for logistic regression among traditional machine learning algorithms and 98 % accuracy.

There are several potential future enhancements for the project "A Comparative study on fake job post prediction using different data mining techniques." Here are a few ideas:

1. Expand the dataset: The study could be enhanced by including more data sources to increase the size and diversity of the dataset. This could involve scraping additional job postings from different job boards, and also including data from social media and other online platforms where fake job postings may be shared.

2. Incorporate more advanced machine learning techniques: While the study already compares different data mining techniques, incorporating more advanced machine learning algorithms, such as deep learning, could yield even better results.

3. Consider additional features: The current study uses a variety of features, such as job titles and company names, to predict fake job postings. However, there may be other

features that could be useful, such as the language used in the posting, the salary offered, or the location of the job.

4. Evaluate the impact of different algorithms on different types of job postings: It could be useful to evaluate how different algorithms perform on different types of job postings, such as those in different industries or with different levels of seniority.

5. Apply the model to real-world scenarios: Once the model is developed, it could be applied to real-world scenarios to test its efficacy. This could involve partnering with companies or job boards to help identify and remove fake job posts.

VIII.REFERENCES

[1]. S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset", FutureInternet 2017.

2. Alghamdi, F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", Journal of Information Security, 2019.

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 NewsDetection 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 byViolent Extremists", Security Informatics, 3, 5, 2014.

6. Y. Kim, "Convolutional neural networks for sentence classification," arXivPrepr. arXiv1408.5882, 2014.

7. T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.- T. Nguyen, "Hate Speech Detection on Vietnamese Social Media Text 2019.

8. P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," Neurocomputing.

9. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 2018

9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893.

10. K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification in biomedical texts," International Conference

on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1205-1209.

11. Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security& Its Applications, 8, 55- 72. https://doi.org/10.5121/imsa.2016.8405

12. Li, H.; Chen, Z.; Liu, B.; Wei, X.; Shao, J. Spotting fake reviews via collective positive-

13. unlabeled learning. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China.

14. Ott, M.; Cardie, C.; Hancock, J. Estimating the prevalence of deception in online review communities. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France.

15. Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of Fraudulent Emails by Employing Advanced Feature Abundance.