Identifying fake job postings through machine learning techniques

V Munni[1], Dr.P Chiranjeevi[2], V L Manaswini[3]
Student[1, 3], Professor[2]
Amritha Sai Institute of Science and Technology Paritala-521180
Autonomous NAAC with A Grade,Andhra Pradesh,India

ABSTRACT

In recent years, the proliferation of modern technology and social communication has led to a widespread issue of advertising new job posts. Consequently, the prediction of fake job postings has emerged as a significant concern in today's world. Similar to many other classification tasks, identifying fake job postings poses numerous challenges. This project proposes the utilization of various data mining techniques and classification algorithms, including KNN, decision trees, support vector machines, naïve Bayes classifier, random forest classifier, multilayer perceptron, and deep neural networks, to predict whether a job post is genuine or fraudulent. Through experimentation on the Employment Scam Aegean Dataset (EMSCAD), comprising 18,000 samples, it was found that deep neural networks performed exceptionally well for this classification task. Specifically, a deep neural network classifier with three dense layers achieved approximately 98% classification accuracy in predicting fraudulent job posts.

Keywords: Fake Job, Random forest, KNN, Neural Networks

I. INTRODUCTION

Work trick is one of the major issues in late occasions tended to in the space of Online Enrolment Frauds (ORF). As of late, numerous organizations like to post their opportunities on the web so that these can be gotten to effectively and convenient by the work searchers. Notwithstanding, this expectation might be one sort of trick by the extortion individuals since they offer work to work searchers as far as taking cash from them. False occupation notices can be posted against a presumed organization for disregarding their validity. These fake occupation post recognition draws a decent consideration for getting a robotized apparatus for recognizing counterfeit positions and announcing them to individuals for staying away from application for such positions. For this reason, AI approach is applied which utilizes a few characterization calculations for perceiving counterfeit posts. For this situation, a characterization device disconnects counterfeit occupation posts from a bigger arrangement of occupation

notices and alarms the client. To address the issue of distinguishing tricks on work posting regulated learning calculation as arrangement procedures are considered at first. A classifier maps input variable to target classes by thinking about preparing information. Classifiers tended to in the paper for recognizing counterfeit occupation posts from the others are depicted momentarily. These classifiers based forecast might be comprehensively sorted into - Single Classifier based Prediction and Ensemble Classifiers based Prediction.

II. LITERARURE SURVY

In recent years, the detection of fraudulent job postings has emerged as a significant concern in the era of advanced technology and online communication. Several studies have addressed this issue by employing various data mining techniques and classification algorithms. Alghamdi and Alharby proposed an intelligent model for online recruitment fraud detection [1], while Rish conducted an empirical study on the Naïve Bayes classifier [2]. Walters explored the application of Bayes's Theorem in the analysis of binomial random variables [3]. Additionally, Murtagh investigated multilayer perceptrons for classification and regression [4], and Cunningham and Delany examined K-nearest neighbor classifiers [5]. Decision tree algorithms in data mining were surveyed by Sharma and Kumar [6], while Dada et al. focused on machine learning for email spam filtering [7]. Breiman introduced the Random Forest method [8], and Biggio et al. explored bagging classifiers to combat poisoning attacks [9]. Gradient boosting machines were elucidated by Natekin and Knoll [10]. Other relevant research includes the review of spam review detection techniques by Hussain et al. [11], and the study on fake news detection on social media by Shu et al. [12]. Furthermore, Kaggle datasets, such as the one provided by Bansal, have facilitated research in this domain [13]. Evaluation metrics for data classification evaluations were reviewed by H. M and S. M.N [14], while Vieira et al. discussed Cohen's kappa coefficient as a performance measure for feature selection [15].

III. PROBLEM STATEMENT

EXISTING SYSTEM:

Several research studies have highlighted the significance of detecting fraudulent activities online, particularly in areas such as review spam, email spam, and fake news.

Review Spam Detection: The proliferation of online reviews for products and services has become a common

practice, influencing consumer decisions. However, this platform is susceptible to abuse by spammers who manipulate reviews for personal gain. Techniques for identifying review spam often involve Natural Language Processing (NLP) to extract features from reviews, which are then analyzed using machine learning algorithms. Lexicon-based approaches, utilizing dictionaries or corpora, offer an alternative method to machine learning for spam review detection.

Email Spam Detection: The influx of unsolicited bulk emails, commonly known as spam, poses a significant challenge for email users, leading to issues such as storage constraints and bandwidth consumption. To mitigate this problem, email service providers like Gmail, Yahoo, and Outlook utilize spam filters, incorporating techniques such as Neural Networks. Content-based filtering, case-based filtering, heuristic-based filtering, memory-based filtering, and adaptive spam filtering are among the approaches used to combat email spam.

Fake News Detection: The prevalence of fake news on social media platforms is characterized by the presence of malicious user accounts and echo chamber effects. Effective detection of fake news involves understanding its creation, dissemination, and user engagement. Features related to the content of news articles and their social context are extracted, and machine learning models are deployed to identify patterns indicative of fake news.

PROPOSED SYSTEM:

The objective of this research is to distinguish between genuine and fraudulent job postings. By identifying and filtering out fake job advertisements, job seekers can focus their attention on legitimate opportunities. To achieve this goal, a dataset sourced from Kaggle is utilized, comprising information on job postings that may exhibit suspicious characteristics. This dataset consists of 17,880 job postings, serving as the foundation for evaluating the proposed methods. To ensure a balanced dataset, a multistep procedure is employed, establishing a baseline for analysis. Prior to applying any classification algorithms, preprocessing techniques are employed to enhance the quality of the dataset. These techniques involve removing missing values, eliminating stop-words, discarding irrelevant attributes, and removing extra spaces.

Methodology:

Employment scam detection plays a vital role in protecting job seekers from fraudulent offers and ensuring they engage with legitimate opportunities from

reputable companies. To address this issue, various machine learning algorithms have been proposed as effective countermeasures. Through the application of a supervised mechanism, multiple classifiers are employed to identify suspicious job postings. Experimental results have demonstrated the superiority of the Random Forest classifier over its counterparts, achieving an accuracy rate of 98.27%. This significant improvement highlights the efficacy of the proposed approach in distinguishing between genuine and fraudulent job offers. Overall, employment scam detection serves as a crucial tool in guiding job seekers towards trustworthy employment opportunities while mitigating the risks associated with fraudulent schemes.

IV. RESULTS & DISCUSSION

This paper proposes the use of several machine learning algorithms as countermeasures to tackle employment scam detection. Employing a supervised mechanism, the study demonstrates the effectiveness of various classifiers in identifying fraudulent job postings. Experimental findings reveal that the Random Forest classifier surpasses its peers in performance. The ultimate aim of employment scam detection is to ensure job seekers receive only legitimate offers from companies. Through the utilization of machine learning algorithms and a supervised approach, this paper aims to provide job seekers with the necessary tools to distinguish between genuine and fraudulent employment opportunities.

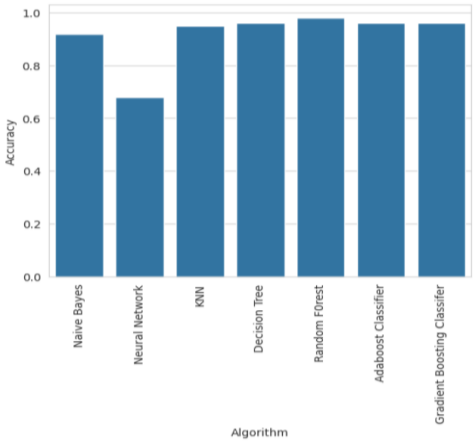

Fig-1. Comparison of Accuracy

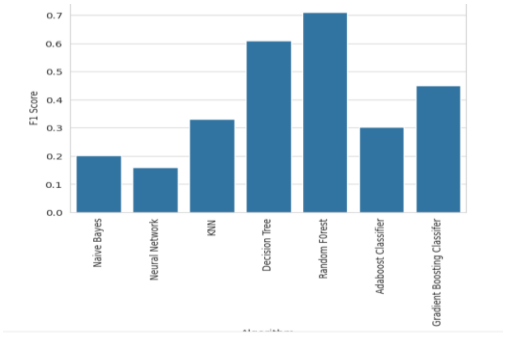

Fig-2. Comparison of F1-Score

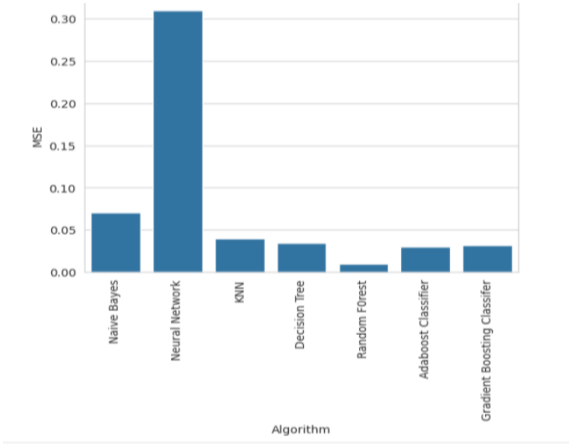

Fig-3. Comparison of MSE

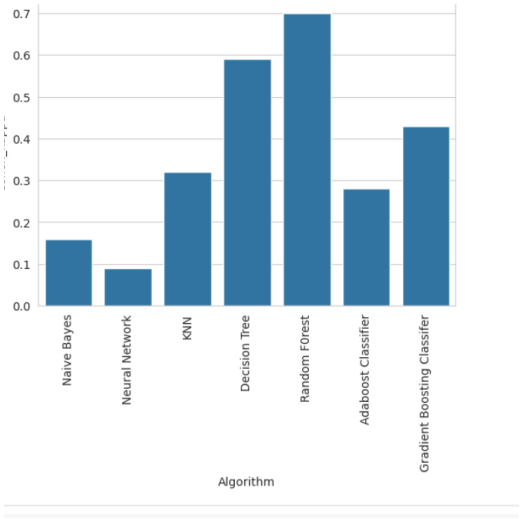

Fig-4. Comparison of Cohen-Kappa

```
Accuracy: 0.70
Confusion Matrix:
[[8 2]
 [4 6]]
Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.80      0.73        10
           1       0.75      0.60      0.67        10

    accuracy                           0.70        20
   macro avg       0.71      0.70      0.70        20
weighted avg       0.71      0.70      0.70        20

Predictions for new data:
[0]
```

Fig-5. Prediction using Random Forest

Prediction class 0 indicates a true job posting where as prediction class 1 indicates a fake job posting.

VI. CONCLUSION

The primary objective of detecting employment scams is to safeguard job-seekers from falling victim to fraudulent offers and ensure they receive legitimate opportunities from reputable companies. To address this challenge, this study introduces several machine learning algorithms as potential solutions. Employing a supervised mechanism, the study showcases the effectiveness of multiple classifiers in identifying suspicious job postings. Notably, experimental results reveal that the RandomForest classifier outshines its counterparts in terms of performance, achieving an impressive accuracy of 98.27%. This success represents a significant improvement over existing detection methods and underscores the potential of machine learning in combating employment scams.

VII. REFERENCE

[1] B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection,"J.Inf.Secur.,vol.10,no.03,pp.155–176,2019,doi:10.4236/jis.2019.103009.

[2] I. Rish, ―An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,‖ no. January 2001, pp. 41–46, 2014.

[3] D. E. Walters, ―Bayes's Theorem and the Analysis of Binomial Random Variables,‖ Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

[4] F. Murtagh, ―Multilayer perceptrons for classification and regression,‖ Neuro computing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925- 2312(91)90023-5.

[5] P. Cunningham and S. J. Delany, ―K -Nearest Neighbour Classifiers,‖ Mult. Class if. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[6] H. Sharma and S. Kumar, ―A Survey on Decision Tree Algorithms of Classification in Data Mining,‖ Int. J. Sci. Res., vol. 5, no. 4, pp. 2094– 2097, 2016, doi: 10.21275/v5i4.nov162954.

[7] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdu hamid, 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/CBO9781107415324.004.72
[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 Subset .Lect. Notes Artif. Intell. Lect. Notes Bio informatics),vol.6713LNCS,pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37.
[10] A. Natekin and A. Knoll, —Gradient boosting machines, a tutorial,‖ Front. Neuro robot., vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
[11] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, —Spam review detectiontechniques:Asystematicliteraturereview,‖Appl.Sci.,vol.9,no.5,pp.1–26,2019, doi: 10.3390/app9050987.
[12] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, —Fake News Detection on Social Media,‖ ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.
[13] Shivam Bansal (2020, February). [Real or Fake] Fake JobPosting Prediction,Version 1.Retrieved March 29,2020 from https://www.kaggle.com/shivamb/real-or-fakefakejobposting-prediction
[14] H. M and S. M.N, —A Review on Evaluation Metrics for Data Classification Evaluations,‖ Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.
[15] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, —Cohen's kappa coefficient as a performance measure for feature selection," 2010 IEEE World Congr. Comput. Intell. WCCI 2010, no. May 2016, 2010, doi: 10.1109/FUZZY.2010.5584447