International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

FAKE JOB PREDICTION USING MACHINE LEARNING ALGORITHMS

Dr.B.V.S.Pavan Kumar[1],Gangarapu.Tejaswini[2], Kakumanu.Tharuni[3], Lakkam.Sruthi[4]

[1]Assistant Professor, School of CSE ,Malla Reddy Engineering College For Women(Autonomous Institution), Maisammaguda, Dhulapally,Secunderabad,Telangana-500100

[234]UG Student, Department of CSE,Malla Reddy Engineering College for Women, (Autonomous Institution), Maisammaguda,Dhulapally,Secunderabad,Telangana-500100
bvsppkumar@gmail.com

ABSTRACT

To avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed in the paper. Different classifiers are used for checking fraudulent post in the web and the results of those classifiers are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts. Two major types of classifiers, such as single classifier and ensemble classifiers are considered for fraudulent job posts detection. However, experimental results indicate that ensemble classifiers are the best classification to detect scams over the single classifiers.

Keywords: Fraudulent Job Posts, Employment Scam Detection, Machine Learning, Classification Techniques, Single Classifiers, Ensemble Classifiers, Fake Job Detection

I.INTRODUCTION

In the digital age, online job portals have become a primary medium for recruitment. They provide convenience and accessibility for employers and job seekers alike. However, this accessibility has also opened doors to exploitation by fraudsters, leading to a rise in Employment Scams. These scams, categorized under Online Recruitment Frauds (ORF), involve posting fake job advertisements, often impersonating reputable companies, with the intention of extorting money or sensitive information from job seekers. Such activities not only deceive individuals but also harm the credibility of legitimate organizations.

Addressing this issue is crucial to ensure a safe and trustworthy recruitment process. Fraudulent job advertisements typically exploit human vulnerabilities by promising lucrative opportunities in exchange for upfront fees or personal information. The scale and volume of job postings online make manual detection of these scams impractical. Consequently, automated detection systems have become essential for identifying and flagging fake job posts.Machine learning (ML) techniques offer a promising solution to this problem. ML models can analyze patterns and detect anomalies within job advertisements, distinguishing between legitimate and fraudulent posts. This study focuses on applying supervised learning algorithms to tackle employment scams. A classification tool is developed to segregate fraudulent posts from legitimate ones, thereby alerting users and preventing them from falling victim to such scams.

The classification approaches in this research are broadly divided into two categories: Single Classifier-based Prediction and Ensemble Classifier-based Prediction. Single

classifiers, such as Naive Bayes, Multi-Layer Perceptron (MLP), K-Nearest Neighbor (KNN), and Decision Tree, are trained individually to predict fraudulent job posts. Ensemble classifiers, on the other hand, combine multiple models to improve prediction accuracy and robustness.

By evaluating these models, the study aims to identify the most effective approach for detecting fraudulent job posts. The results highlight that ensemble classifiers generally outperform single classifiers, emphasizing the importance of combining models to achieve reliable and accurate detection. This research contributes significantly to creating safer online recruitment environments and protecting job seekers from scams.

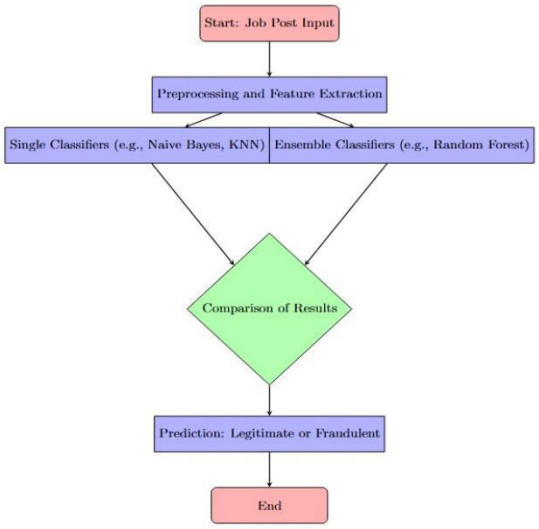

Fig 1: System Architecture

This architecture ensures efficient and accurate detection of fraudulent job posts by utilizing machine learning classification techniques, providing users with real-time monitoring and reporting capabilities.

II RELATED WORK

"An Intelligent Model for Online Recruitment Fraud Detection,"

This study research attempts to prohibit privacy and loss of money for individuals and organization by creating a reliable model which can detect the fraud exposure in the online recruitment environments. This research presents a major contribution represented in a reliable detection model using ensemble approach based on Random forest classifier to detect Online Recruitment Fraud (ORF). The detection of Online Recruitment Fraud is characterized by other types of electronic fraud detection by its modern and the scarcity of studies on this concept. The researcher proposed the detection model to achieve the objectives of this study. For feature selection, support vector machine method is used and for classification and detection, ensemble classifier using Random Forest is employed. A freely available dataset called Employment Scam Aegean Dataset (EMSCAD) is used to apply the model. Pre-processing step had been applied before the selection and classification adoptions. The results showed an obtained accuracy of 97.41%. Further, the findings presented the main features and important factors in detection purpose include having a company profile feature, having a company logo feature and an industry feature.

An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,

The naive Bayes classifier greatly simplify learn-ing by assuming that features are independent given class. Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classifiers. Our broad goal is to understand the data character-istics

which affect the performance of naive Bayes. Our approach uses Monte Carlo simulations that al-low a systematic study of classification accuracy for several classes of randomly generated prob-lems. We analyze the impact of the distribution entropy on the classification error, showing that low-entropy feature distributions yield good per-formance of naive Bayes. We also demonstrate that naive Bayes works well for certain nearly-functional feature dependencies, thus reaching its best performance in two opposite cases: completely independent features (as expected) and function-ally dependent features (which is surprising). An-other surprising result is that the accuracy of naive Bayes is not directly correlated with the degree of feature dependencies measured as the class-conditional mutual information between the fea-tures. Instead, a better predictor of naive Bayes ac-curacy is the amount of information about the class that is lost because of the independence assump-tion.

Bayes's Theorem and the Analysis of Binomial Random Variables

A very practical application of Bayes's theorem, for the analysis of binomial random variables, is presented. Previous papers (Walters, 1985; Walters, 1986a) have already demonstrated the reliability of the technique for one, or two random variables, and the extension of the approach to several random variables is described. Two biometrical examples are used to illustrate the method.

Multilayer perceptrons for classification and regression,

We review the theory and practice of the multilayer perceptron. We aim at addressing a range of issues which are important from the point of view of applying this approach to practical problems. A number of examples are given, illustrating how the multilayer perceptron compares to alternative, conventional approaches. The application fields of classification and regression are especially considered. Questions of implementation, i.e. of multilayer perceptron architecture, dynamics, and related aspects, are discussed. Recent studies, which are particularly relevant to the areas of discriminant analysis, and function mapping, are cited.

K -Nearest Neighbour Classifiers,

We analyze a Relational Neighbor (RN) classifier, a simple relational predictive model that predicts only based on class labels of related neighbors, using no learning and no inherent attributes. We show that it performs surprisingly well by comparing it to more complex models such as Probabilistic Relational Models and Relational Probability Trees on three data sets from published work.

III.ALGORITHM

Data Collection

Involves gathering job posts from various sources.This is typically done using web scraping techniques. Tools like BeautifulSoup and requests are commonly employed to extract relevant job postings from websites.

```
from bs4 import BeautifulSoup
import requests

def scrape_job_posts():
    response = requests.get('https://example.com/jobs')
    soup = BeautifulSoup(response.text, 'html.parser')
    # Code to parse and extract job post data goes here
    job_posts = []  # Placeholder for scraped data
    return job_posts
```

Data Preprocessing

Raw data from job posts often contains noise, such as special characters, redundant words, or irrelevant information.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

```
from sklearn.feature_extraction.text import TfidfVect

def preprocess_data(job_posts):
    vectorizer = TfidfVectorizer()
    features = vectorizer.fit_transform(job_posts)
    return features
```

data is tokenized, cleaned (stopwords and special characters removed), and transformed into a matrix of TF-IDF features. These numerical features allow machine learning algorithms to identify patterns within the data.

The RandomForestClassifier is a popular choice due to its robustness and ability to handle large datasets efficiently. The data is split into training and testing sets using train_test_split:

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier

def train_model(features, labels):
    X_train, X_test, y_train, y_test = train_test_split(featu
    model = RandomForestClassifier()
    model.fit(X_train, y_train)
    return model
```

```
from sklearn.metrics import accuracy_score

def evaluate_model(model, X_test, y_test):
    predictions = model.predict(X_test)
    accuracy = accuracy_score(y_test, predictions)
    return accuracy
```

```
def predict_fraudulent_posts(model, new_posts):
    features = preprocess_data(new_posts)
    predictions = model.predict(features)
    return predictions
```

```
def flag_posts(predictions):
    flagged_posts = [i for i, p in enumerate(predictions) if
    return flagged_posts
```

job posts predicted as fraudulent (p == 1) are flagged by recording their indices. These flagged posts can be presented to administrators or users for review.

RESULT

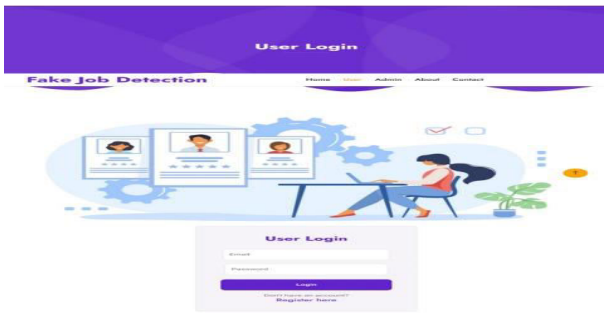

Fig 1: User Login

Fig 2: Admin Login

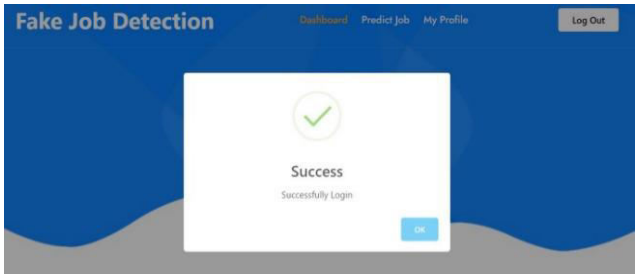

Fig 3:Login Success

Fig 4:Form Filling

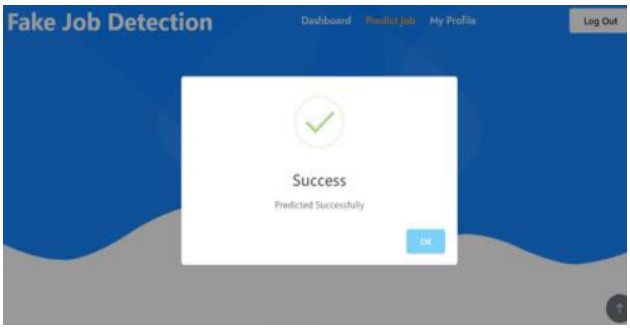

Fig 5:Predicted Successfull

CONCLUSION

Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, several machine learning algorithms are proposed as countermeasures in this paper. Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that Random Forest classifier outperforms over its peer classification tool. The proposed approach achieved accuracy 98.27% which is much higher than the existing methods

REFERENCES

[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,‖ Neurocomputing, 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. Classif. 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. 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.

[10] A. Natekin and A. Knoll, ―Gradient boosting machines, a tutorial,‖ Front. Neurorobot., 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 detection techniques: A systematic literature review,‖ 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-fake-fake jobposting-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