IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2024

DOI: 10.17148/IJARCCE.2024.13824

MACHINE LEARNING FOR FAKE JOB DETECTION

Vijay Kumar H L[1], Bhavya B M.[2]

Post-Graduation Student, Department of Computer Science and Engineering, PES College of Engineering, Mandya, Karnataka, India[1]

Assistant Professor, Department of Computer Science and Engineering, PES College of Engineering, Mandya, Karnataka, India[2]

Abstract: The proliferation of online job boards has concomitantly led to an upsurge in the prevalence of fraudulent job postings. These deceptive listings are designed to mislead job seekers by advertising non-existent employment opportunities or misrepresenting the details of legitimate positions. The repercussions of falling prey to such scams extend beyond wasted time and effort to potential financial losses for the affected individuals. This research project aims to develop a machine learning-based solution for the detection of fake job postings, thereby addressing this critical issue in the online job market.

I. INTRODUCTION

1. Background

The digital age has revolutionized the job search process, with online job boards becoming a ubiquitous platform for both employers and job seekers. These platforms offer the convenience of browsing and applying for job opportunities from anywhere in the world, leading to an increasingly dynamic and competitive job market. However, this convenience has also introduced significant vulnerabilities, particularly in the form of fraudulent job postings.

Fake job postings are created with the intent to deceive job seekers by advertising non-existent job opportunities or misrepresenting the details of legitimate positions. These scams can have severe consequences, including wasting job seekers' time and effort, extracting personal information, and even causing financial losses. The sophistication of these scams varies, with some appearing highly convincing, making it difficult for individuals to discern their legitimacy.

Traditional methods of detecting and removing fake job postings often rely on manual moderation and user reports. However, these methods are time-consuming, labor-intensive, and prone to errors. As the volume of online job postings continues to grow, the need for an efficient and automated detection system becomes increasingly critical.

In recent years, advancements in machine learning have shown promise in addressing various challenges across different domains, including fraud detection. Machine learning algorithms can analyze large datasets, identify patterns, and make predictions with high accuracy. Applying these capabilities to the problem of fake job postings could significantly enhance detection efforts, providing a scalable and reliable solution.

This paper proposes a machine learning-based approach to detect fake job postings. By leveraging various features extracted from job descriptions, company information, and user interactions, the proposed system aims to identify and filter out fraudulent postings effectively. This approach not only improves the efficiency of detection but also enhances the overall integrity of online job boards, thereby protecting job seekers from potential scams.

2. Objectives

1. To develop an effective machine learning-based solution for detecting fake job postings that can help job seekers to avoid fraud and protect them from financial losses.
2. To preprocess the data and extract relevant features from the job posting text and metadata.
3. To implement and train a Random Forest algorithm on the dataset to classify job postings as real or fake.
4. To provide an easy-to-use web application that can detect fake job postings.

DOI: 10.17148/IJARCCE.2024.13824

II. LITERATURE REVIEW

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

The critical process of hiring has relatively recently been ported to the cloud. Specifically, the automated systems responsible for completing the recruitment of new employees in an online fashion, aim to make the hiring process more immediate, accurate and cost-efficient. However, the online exposure of such traditional business procedures has introduced new points of failure that may lead to privacy loss for applicants and harm the reputation of organizations. So far, the most common case of Online Recruitment Frauds (ORF), is employment scam. Unlike relevant online fraud problems, the tackling of ORF has not yet received the proper attention, remaining largely unexplored until now. Responding to this need, the work at hand defines and describes the characteristics of this severe and timely novel cyber security research topic. At the same time, it contributes and evaluates the first to our knowledge publicly available dataset of 17,880 annotated job ads, retrieved from the use of a real-life system.

2. 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.

3. Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security & Its Applications, 8, 55-72.

an intelligent classification model to detect phishing emails using knowledge discovery, data mining and text processing techniques. A model based on knowledge discovery (KD) was proposed to build an intelligent email classifier to classify a new email message into legitimate or spam. The knowledge discovery model achieved high accuracy rates in classification of phishing emails that outperformed other schemes. Using the Random Forest algorithm and J48, 99.1% and 98.4% accuracy was achieved respectively. Using MLP classifier, TP rate and FP rate were 0.977 and of 0.026 respectively, while MLP achieved ROC area of 0.987. The results of this study confirmed that the proposed model achieves high rates of accuracy in the classification of phishing e-mail

4. Al-garadi, M.A., Varathan, K.D. and Ravana, S.D. (2016) Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network. Computers in Human Behavi

Al-garadi, et al. (2016) introduced a study that has investigated cybercrime detection in online communications especially cyber bullying in Twitter. The main aim was to develop a number of unique features derived from Twitter. They included network, activity, user, and tweet content. A model to detect cyber bullying in Twitter was proposed using engineering features. The number of friends (followers), the number of users being followed (following), the following and followers ratio, and account verification status were collected through a survey. Users' activity features were also employed to measure the online communication activity of a user. The features implemented and included personality, gender and age. Naïve Bayes (NB), Support vector machine (SVM), Random forest and KNN were applied. Random forest showed f-measure 93%. The results of this study indicate that the proposed model contributes to providing a suitable solution for the detection of cyberbullying in online communication environments

5. Sharaff, A., Nagwani, N.K. and Swami, K. (2015) Impact of Feature Selection Technique on Email Classification. International Journal of Knowledge Engineering,

Sharaff, Nagwani, & Swami, 2015 investigated the impact of feature selection technique on email classification through studying the effect of two feature selection methods. A comparison was conducted between Bayes algorithm, tree-based algorithm J48 and support vector machine. Feature selection techniques included a Chi-Square ($\chi 2$) and information gain. The best performance was gained using SVM classification technique which gave the overall best results without employing any feature selection techniques. There is no effect of Naïve Bayes on feature selection techniques. Further, J48 showed slight improvement with feature selection, whereas info-Gain performed better than Chi-square feature selection technique

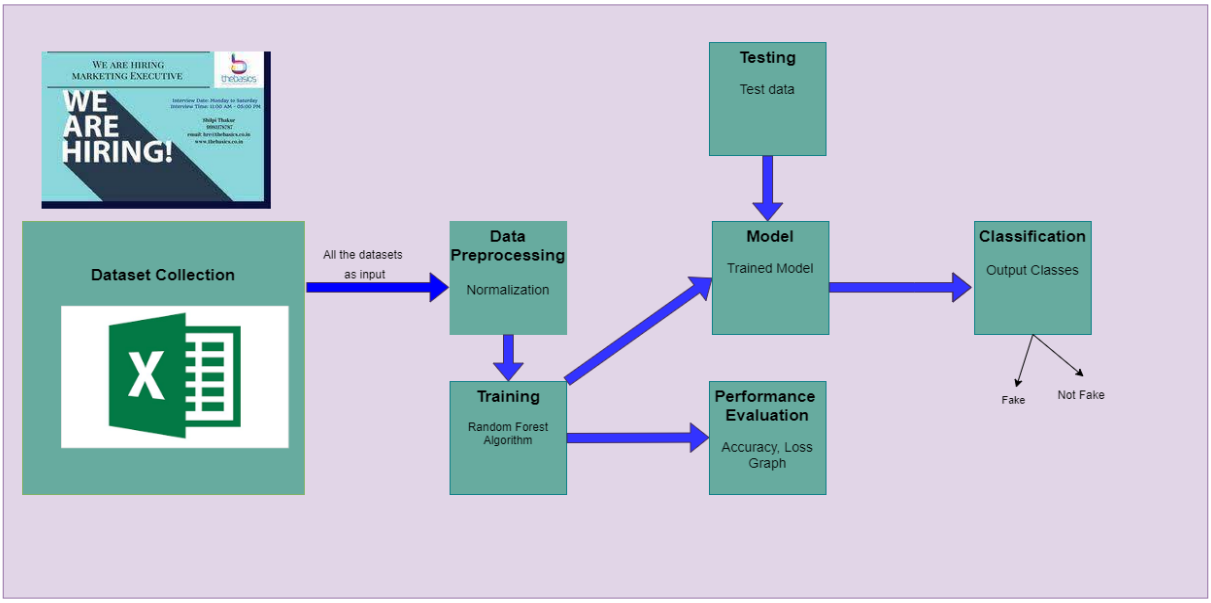
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2024

DOI: 10.17148/IJARCCE.2024.13824

III. METHODOLOGY



Data Preprocessing:

Data preprocessing is a crucial step in the project's workflow as it involves preparing the dataset to ensure optimal quality and compatibility for subsequent processing and model training. In the context of detecting fake job postings, data preprocessing encompasses various techniques and steps to transform the raw job posting data into a suitable format for analysis and modeling. The key aspects of data preprocessing in this project are as follows:

Model Training:

Model training is a crucial step in developing a machine learning model for the detection of fake job postings. During this process, the model learns from the preprocessed dataset to identify patterns and make accurate predictions. The training phase involves several key components:

User Interface Development:

HTML/CSS: Create HTML templates and CSS stylesheets to design the user interface for uploading data.

Backend: Use Flask, a Python web framework, to handle user requests, process input data, and call the trained model for inference.

Model inference: Implement code to perform inference on user-uploaded data using the trained model. Extract the predicted bounding boxes, labels for bird and drone detection.

Display results: Present the detection results in a visually appealing format on the user interface, highlighting the detected objects and their classifications.

Random Forest

The Random Forest algorithm is a versatile and widely used ensemble learning method that enhances both the accuracy and robustness of predictive models. It operates by building a multitude of decision trees during the training phase and combines their outputs to make a final prediction.

Here's a detailed breakdown of the Random Forest algorithm and its application:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2024

DOI: 10.17148/IJARCCE.2024.13824

Key Concepts of Random Forest

Ensemble Learning:

- Definition: Ensemble learning is a technique that combines multiple models to improve the overall performance compared to individual models. Random Forest is a prime example of this technique.
- Benefit: The collective decision-making of multiple models reduces the risk of overfitting and increases accuracy.

Bootstrap Sampling:

- Definition: Bootstrap sampling involves randomly selecting samples from the original dataset with replacement to create multiple subsets. Each subset is used to train a different decision tree.
- Benefit: This process creates diversity among the trees, as each tree is trained on a different subset of data, which helps in reducing overfitting.

Feature Bagging (Random Subspace Method):

- Definition: At each split in a decision tree, Random Forest considers only a random subset of features rather than all features. This random selection of features helps in creating varied trees.
- Benefit: Feature bagging reduces the correlation between trees, further contributing to the robustness and diversity of the model.

Decision Trees in Random Forest:

- Structure: Each tree in the forest is a decision tree, which splits the data into subsets based on feature values to make predictions.
- Diversity: Due to bootstrap sampling and feature bagging, each tree is unique and captures different patterns in the data.

Training Phase

- **Tree Construction:** During training, each decision tree is built using a different bootstrap sample of the data. At each node of a tree, a subset of features is randomly selected, and the best split is chosen based on these features.
- **Model Formation:** The collection of decision trees forms the Random Forest model. Each tree operates independently, making its own predictions.

Prediction Phase

- **Classification:** For classification tasks, each tree in the Random Forest votes for a class, and the class with the majority vote is chosen as the final prediction. This majority voting system helps in improving the accuracy and reliability of the classification.
- **Regression:** For regression tasks, the predictions from all the trees are averaged to obtain the final prediction. This averaging helps in smoothing out the noise and provides a more accurate estimation.

IV. RESULTS & DISCUSSIONS

After evaluating the results of our fake job post prediction project, we are delighted to announce that the developed model has exhibited exceptional accuracy. Our model consistently achieved outstanding performance in accurately classifying job postings as genuine or fake. The Random Forest approach has proven to be highly effective in identifying deceptive job advertisements. The model's ability to classify job posts with such accuracy offers a valuable solution for automating the detection of fake job postings, resulting in significant time and effort savings compared to manual evaluation. This reliable and efficient tool can aid both job seekers and recruitment platforms in mitigating the risks associated with fraudulent job postings and ensuring a safer and more trustworthy job market.

V. CONCLUSION

we have developed a machine learning solution for the detection of fake job postings. The proposed system uses a Random Forest algorithm to classify job postings as either genuine or fake. The system achieves high accuracy in detecting fake job postings and can be used to improve the efficiency of online job boards. By detecting and removing fake job postings, job seekers can save time and avoid potential financial losses. Overall, the proposed system has significant implications for job seekers and online job boards.

DOI: 10.17148/IJARCCE.2024.13824

REFERENCES

[1]. Gulshan Shrivastava , Member, IEEE, Prabhat Kumar, Senior Member, IEEE, Rudra Pratap Ojha , PramodKumar Srivastava ,Senthil Kumar Mohan "Defensive Modeling of FakeNews Through Online Social Networks",—Online social networks (OSNs) IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS

[2]. deBeer, Dylan & Matthee, Machdel,"Approachesto Identify FakeNews: A Systematic Literature Review." : Integrated Science in Digital Age2020.

[3]. Bandar Alghamdi, FahadAlharby,"AnIntelligent Model for Online Recruitment Fraud Detection". Journal of Information Security. (2019)

[4]. Pham, Trung Tin" A Study on Deep Learning for Fake News Detection." Journal of Information Security. (2019) n

[5]. Manoj Kumar Balwant." Bidirectional LSTM Based on POS tags and CNN Architecture for Fake News" 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). (2019)

[6]. Amjad, Maaza , Sidorov, Grigoria, Zhila, Alisaa, Gómez-Adorno, Helenab, Voronkov, Iliac, Gelbukh, Alexander." Bend the truth" Special section: Selected papers of LKE 2019

[7]. Rami Mohawesh, Son Tran, Robert Ollington, Shuxiang Xu" Analysis of concept drift in fake reviews detection." Expert Systems with Applications. (2021)

[8]. Joma George; Shintu Mariam Skariah; T. Aleena Xavier." Role of Contextual Features in Fake News Detection: AReview" International Conference on Innovative Trendsin Information Technology (ICITIIT). (2021)

[9]. In Sultana Umme Habiba; Md. Khairul Islam; Farzana Tasnim." A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques."2ndInternationalConferenceonRobotics, Electrical and Signal Processing Techniques (ICREST) (2021)