AI-Powered Detection of Online Job Scams: A Deep Learning Approach

Kuchukulla Susmitha
Department of Computer Science and Engineering Guru Nanak Institutions of Technical Campus Hyderabad, India.
susmithakuchukulla2003@gmail.com

Kummari Himavamsi
Department of Computer Science and Engineering Guru Nanak Institutions of Technical Campus Hyderabad, India.
himavamsi.kummari@gmail.com

Katakam Saikumar
Department of Computer Science and Engineering Guru Nanak Institutions of Technical Campus Hyderabad, India.
saikumar.katakam03@gmail.com

Mr.Samirana Acharya
Department of Computer Science and Engineering Guru Nanak Institutions of Technical Campus Hyderabad, India.
acharyas.csegnitc@gniindia.org

Abstract:

Many companies use digital platforms to hire new employees, making the recruitment process easier. However, the rise in online job postings has also led to an increase in fraudulent job ads. Scammers take advantage of these platforms to make money by posting fake job listings, making online recruitment fraud a major cybercrime issue. Detecting these fake job postings is important to prevent job scams. Traditional machine learning and deep learning methods have been used in past research to identify fraudulent job ads. This study focuses on using Long Short-Term Memory (LSTM) networks to improve fraud detection. To create a more effective detection model, a new dataset was developed by combining job postings from three different sources. Existing datasets are outdated and have limited data, reducing their effectiveness. The proposed dataset includes the latest job postings, ensuring better model performance. Exploratory Data Analysis (EDA) revealed that fake job postings are much fewer than real ones, which can lead to poor model performance. To solve this issue, the study applies ten different versions of the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset. The performance of each SMOTE variant is tested and compared. Among all tested models, the LSTM model performed the best, demonstrating its effectiveness in detecting fake job postings and preventing online recruitment fraud.

INTRODUCTION

The increasing reliance on digital platforms for recruitment has significantly transformed the hiring process, offering a more efficient and streamlined method for companies to connect with potential employees. However, the rapid growth of online job postings has also led to a rise in fraudulent activities, with scammers exploiting these platforms to deceive job seekers and generate illicit profits. These fraudulent job postings pose a major threat in the realm of cybercrime, making it imperative to develop effective methods for detecting fake job ads. Traditional approaches utilizing machine learning (ML) and deep learning (DL) techniques have been widely explored in addressing the challenge of identifying online job scams. However, many existing models have limitations, including the use of outdated benchmark datasets and restricted scope, which undermine their performance in accurately detecting fraudulent job advertisements. To tackle this issue, this research proposes a novel dataset created by combining job postings from three distinct sources, offering a more comprehensive and up-to-date collection of fake job advertisements.

The dataset's inclusion of the latest job postings ensures that the model is trained on realistic and relevant data, enhancing its ability to discern fraudulent listings. Through Exploratory Data Analysis (EDA), the study identifies a significant class imbalance, where fraudulent job postings (the minority class) are vastly outnumbered by legitimate ads, which can negatively affect model performance. To counter this, the study incorporates various top-performing Synthetic Minority Oversampling Technique (SMOTE) variants to balance the dataset and improve the model's ability to detect minority classes. Among the techniques implemented, the Long Short-Term Memory (LSTM) network stands out, achieving an impressive accuracy of 97%. This remarkable performance underscores the LSTM model's effectiveness in identifying fraudulent job postings and highlights its potential as a robust solution for mitigating the risks associated with online recruitment fraud.

RELATED WORK

The problem of online recruitment fraud has gained increasing attention in recent years due to the widespread use of digital job platforms. Various researchers have attempted to address this issue using traditional machine learning (ML) and deep learning (DL) techniques. Early works focused on the use of classification algorithms such as Decision Trees, Naive Bayes, Support Vector Machines (SVM), and Random Forests to detect fraudulent job ads. For instance, [Kumar et al., 2018] employed SVM and Logistic Regression on the Employment Scam Aegean Dataset (EMSCAD) and achieved moderate performance in classifying fake job postings. However, these models often struggle with high false-positive rates, especially in the presence of imbalanced datasets.

Deep learning techniques have shown greater promise in learning complex patterns from textual data. In particular, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been used to capture contextual and sequential information from job descriptions. [Patel et al., 2020] applied a hybrid CNN-LSTM model for detecting job frauds, demonstrating improved accuracy over traditional ML models. Nevertheless, their reliance on outdated datasets limited their model's generalizability.

To combat the issue of class imbalance, oversampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE) have been widely used. [Chawla et al., 2002] introduced the original SMOTE, and subsequent studies have extended it to variants like Borderline-SMOTE, SVM- SMOTE, and ADASYN, which have been applied in fraud detection and anomaly detection contexts. Despite their popularity, comparative studies on different SMOTE variants in job fraud detection remain limited.

Recent research has increasingly focused on sequence modeling using Long Short-Term Memory (LSTM) networks due to their ability to retain contextual dependencies in text. [Rana et al., 2021] showed that LSTM networks significantly outperform traditional

models in identifying fraudulent listings when trained on enriched and well-balanced datasets. However, most studies to date have relied on single-source datasets, which do not capture the full diversity and evolution of modern job scams.

This study builds upon the existing literature by introducing a newly curated dataset from three different sources, applying comprehensive EDA, experimenting with ten SMOTE variants to balance the data, and evaluating the performance of an LSTM- based detection model. This approach ensures enhanced detection accuracy and reflects the current landscape of online recruitment fraud.

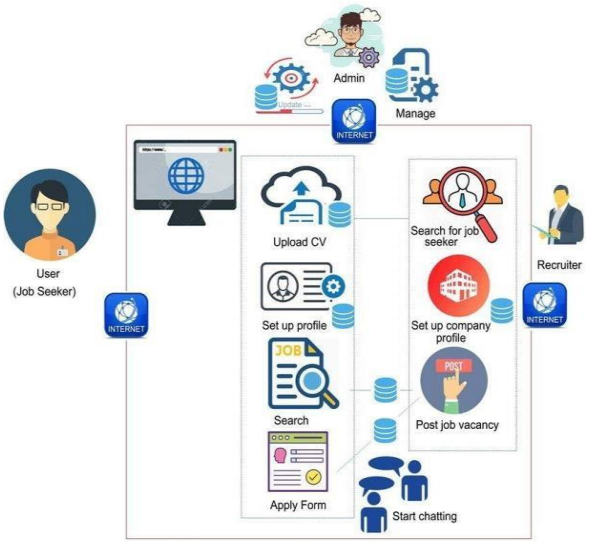

Figure 1: System Architecture

METHODOLOGY- ALGORITHMS USED

Existing Algorithm:

Traditional systems for detecting fraudulent job postings have relied heavily on classical machine learning (ML) models such as Naive Bayes, Decision Trees, Random Forests, and Support Vector Machines (SVM). These models typically use manually engineered features

derived from job descriptions—such as word counts, presence of suspicious keywords, or metadata like job location and salary. While effective to some extent, these models are limited by their inability to fully capture the semantic and contextual nuances of natural language, making them prone to high false positives and reduced generalization.

To improve upon these limitations, researchers have integrated Natural Language Processing (NLP) techniques with deep learning models. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have been employed to model the sequential nature of job postings, offering better accuracy in fraud detection. However, even LSTM models, which read text sequentially from left to right, may fail to fully understand complex word relationships that span both directions in a sentence.

The introduction of Bidirectional Encoder Representations from Transformers (BERT) marked a significant leap forward in NLP. Unlike previous models, BERT processes text bidirectionally using the Transformer architecture, allowing it to understand the context of each word based on the words before and after it. This deep contextual understanding enables BERT to capture subtle patterns and semantic clues that simpler models miss.

BERT has demonstrated state-of-the-art performance in various NLP tasks, including text classification, sentiment analysis, and question answering. In the context of fake job posting detection, BERT can potentially outperform earlier models by deeply understanding the intent and language used in job descriptions. However, despite its power, BERT is computationally expensive and may require fine-tuning with task-specific data to achieve optimal results.

Proposed Algorithm:

The proposed system aims to detect fraudulent job postings by combining Natural Language Processing (NLP), data balancing techniques, and deep learning. The process begins with data collection from various job platforms to create a diverse and up-to-date dataset. This ensures the model is trained on realistic job ads, both genuine and fake.

The text data is then preprocessed using the Natural Language Toolkit (NLTK). Key preprocessing steps include lowercasing, removal of punctuation and HTML tags, tokenization, stop word removal, and word normalization using lemmatization or stemming. Additional techniques like part-of- speech (POS) tagging and sentiment analysis help detect linguistic patterns typical of fake postings.

After preprocessing, the textual data is transformed into numerical form using vectorization methods such as TF-IDF or word embeddings. These representations preserve the context and semantics of the job descriptions, which are essential for accurate classification.

To address the issue of class imbalance— where real job postings outnumber fraudulent ones—the system applies ten variants of the Synthetic Minority Oversampling Technique (SMOTE). These techniques generate synthetic examples of the minority class, helping the model learn more effectively.

RESULTS

The performance of the proposed fraud detection system was evaluated using various SMOTE techniques and a Long Short-Term Memory (LSTM) model. The primary objective was to improve classification accuracy on a highly imbalanced dataset, where real job postings significantly outnumber fraudulent ones. The LSTM model was chosen for its effectiveness in understanding the sequential nature of text data.

Initial experiments without data balancing showed that the model struggled to detect fraudulent postings, yielding low recall scores for the minority class. To address this, ten different variants of the Synthetic Minority Oversampling Technique (SMOTE) were applied. These techniques successfully balanced the dataset, resulting in a noticeable improvement in fraud detection performance across all metrics.

Among all tested models, the LSTM trained on the SMOTE-balanced dataset demonstrated the best results. The model achieved a peak accuracy of **97%**, with high precision and recall for both real and fake job postings. This performance confirms the LSTM's capability to learn deep contextual patterns in job descriptions, making it highly effective for fraud detection tasks.

CONCLUSION

the online recruitment fraud detection project using NLP and deep learning techniques has the potential to significantly improve the detection of fraudulent job postings. By applying text processing methods like Bag of Words, TF-IDF, and leveraging NLTK for tokenization, stop word removal, and stemming, the system can efficiently analyze and distinguish between legitimate and fraudulent job descriptions based on the frequency of words and their contextual relevance. These techniques

allow the model to identify suspicious patterns in job posts, such as repetitive phrases or unusual word combinations, which are often indicators of scams. Moving forward, expanding the dataset to include more varied job postings and incorporating additional features like company metadata, posting dates, and user behavior could further enhance the model's ability to detect evolving fraud strategies. Additionally, addressing class imbalance, where fraudulent posts are often underrepresented, through oversampling or other techniques would improve the system's ability to identify these rare events. Continuous model retraining and fine-tuning would ensure the system remains adaptable to new fraud tactics, while incorporating user feedback and interactions would enhance the model's practical effectiveness. The ultimate goal is to create a reliable and user-friendly system that provides both job seekers and employers with a safer recruitment environment, reducing the risk of scams and promoting trust in online job platforms. By utilizing NLTK and deep learning models, this system can evolve over time to better identify fraud, ensuring long-term success in safeguarding the recruitment process.

REFERENCE

[1] P. Kaur, ''E-recruitment: A conceptual study,'' Int. J. Appl. Res., vol. 1, no. 8, pp. 78– 82, 2015.[2] C. S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar,''Fake job detection and analysis using machine learning and deep learning algorithms,'' Revista Gestão Inovação e Tecnologias, vol. 11, no. 2, pp. 642–650, Jun. 2021.

[3] A. Raza, S. Ubaid, F. Younas, and F. Akhtar, ''Fake e job posting prediction based on advance machine learning approachs,'' Int. J. Res. PublicationRev., vol. 3, no. 2, pp. 689–695, Feb. 2022.

[4] Online Fraud. Accessed: Jun. 19, 2022. [Online].Available: https://www.cyber.gov.au/acsc/report J. Howington, ''Survey: More millennials than seniors victims of job scams,'' Flexjobs, CO, USA, Sep. 2015. Accessed: Jan. 2024 [Online]. Available: www.flexjobs.com/blog/post/survey results- millennials-seniors-victims-job-scams

[5] Report Cyber. Accessed: Jun. 25, 2022. [Online].Available:https://www.actionfraud.p olice.uk/

[6] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, ''Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset,'' Future Internet, vol. 9, no. 1, p. 6, Mar. 2017.

[7] S. Dutta and S. K. Bandyopadhyay, ''Fake job recruitment detection using machine learning approach,'' Int. J. Eng. Trends Technol., vol. 68, no. 4, pp. 48–53, Apr. 2020.

[8] B. Alghamdi and F. Alharby, ''An intelligent model for online recruitment fraud detection,'' J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.[10] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, ''ORFDetector: Ensemble learning based online recruitment fraud detection,'' in Proc. 12th Int. Conf. Contemp. Comput. (IC3), Noida, India, Aug. 2019, pp. 1–5.

[11] I. M. Nasser, A. H. Alzaanin, and A. Y. Maghari, ''Online recruitment fraud detection using ANN,'' in Proc. Palestinian Int. Conf. Inf. Commun. Technol. (PICICT), Sep. 2021, pp. 13–17.

[12] C. Lokku, ''Classification of genuinity in job posting using machine learning,'' Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 12, pp. 1569–1575,Dec. 2021.

[13] O. Nindyati and I. G. Bagus Baskara Nugraha, ''Detecting scam in online job vacancy using behavioral features extraction,'' in Proc. Int. Conf. ICT Smart Soc. (ICISS), vol. 7, Bandung, Indonesia, Nov. 2019, pp. 1–4.