IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

DOI:10.48047/IJFANS/V11/I12/202

Predicting Fake Job Posts with a Voting Classifier of Multiple Classification Models

Ch.Vijayananda Ratnam[1], Assistant Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

B.Nithya Kranthi Sri[2], **D.Dhanwanth Sai**[3], **A.Preetham Paul**[4], **Ch.Leela Aditya**[5]
[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
[1,2,3,4,5] vijayanandaratnam@vvit.net, nithyakranthi6@gmail.com,
dasaridhanwanthsai@gmail.com,
preethampaul001@gmail.com,adityach054@gmail.com

Abstract

The detection of fake job posts is becoming increasingly important in the modern job market. With the rise of online job postings, scammers and fraudulent actors are taking advantage of unsuspecting job seekers by posting fake job listings that appear legitimate. This paper proposes a machine learning approach to detect fake job posts using a combination of textual and categorical data. We extract various features from the job post text, such as the presence of certain keywords, as well as features from the job post, such as the job title, employment type, required experience. Models like Logistic regression, SVM, Decision tree, Random forest, Gradient boosting, XGBoost, and MLP with Adam optimizer are compared using various metrics like accuracy, F1 score, ROC AUC score, and more after training. This research can be used to build automated systems to detect fake job posts, helping to protect job seekers from scams and fraudulent activities in the job market.

Keywords: Machine learning, Logistic regression, Decision tree, Random forest, Gradient boosting, XGBoost, MLP, Voting classifier.

Introduction

Currently, there are numerous online platforms that facilitate job searching and recruitment, and among them, LinkedIn stands out as one of the most widely used and popular platforms. LinkedIn has become a popular platform for online job searching and recruitment. Companies can post job openings on LinkedIn and search for potential candidates based on their profiles, skills, and experience. Job seekers can also create their profiles and apply for jobs posted by companies. With over 740 million users worldwide, LinkedIn provides a vast network of professionals and job opportunities.

However, the prevalence of fake job postings on LinkedIn has become a growing concern for job seekers. Scammers and fraudulent companies often use LinkedIn to post fake job openings to collect personal information or scam people out of money. The high number of

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

job postings and the ease of creating a profile on LinkedIn makes it challenging to distinguish between legitimate and fake job postings.

The impact of social media and the internet on the hiring process has greatly increased the number of opportunities for job details to be shared on LinkedIn. It has also led to the increase of fraudulent job postings, which can be frustrating and discouraging for job seekers. As such, LinkedIn has implemented various measures to protect its users from fraudulent job postings, such as verifying job postings and allowing users to report suspicious activity. Both the recruiter and the candidates benefit from online hiring. The influence of social media and the internet on the hiring process has increased. Social media and electronic media advertisements have greatly impacted this because a recruiting process's effectiveness depends on how well it is advertised. This has led to an increasing number of opportunities for job details to be shared.

A huge step forward in finding new employees will be made if job postings can be accurately filtered to detect fake job postings. A new door is opened to deal with challenges in the area of human resource management through an automated system that predicts fake job postings.

Critical thinking abilities, research, and common sense are all necessary for finding phoney job postings. Some telltale signs of phoney job advertisements include claims of big remuneration for little effort, demands for advance payments, or a lack of firm information. Moreover, spelling and language mistakes in the job description can suggest that the post was written by a non-professional recruiter. It's crucial to conduct extensive research on the business and the position to spot bogus job postings. Check the website of the business for information, see if there are any contact details, and make sure the job posting is real. Also, you can read reviews written by previous customers and look for news stories about the business online. Moreover, if an employer contacts you, use caution and hold off on giving any personal information until you have confirmed the validity of the employer.

Literature Survey

An important field of research that has attracted more interest recently is the detection of fake job postings. The following are some of the most important studies and articles in this field:

1. Sultana Umme Habiba et al., [1] trained SVM, KNN, Naive Bayes, Random Forest, and MLP models comparatively and achieved maximum accuracy using Random Forest and Maximum average accuracy using MlP classifier.

2. E.Baraneetharan et al., [2] has used KNN, SVM, XGBoost over all the attributes of the data set, out of three XGBoost model has achieved an accuracy of 98.53% which is greater than others with minimum training time.
3. Huynh et al., [3] proposed various deep neural network models, namely Text CNN, Bi-GRU-LSTM CNN, and BiGRU CNN, which are pre-trained using text datasets. To improve the accuracy, they also used an ensemble classifier that combined Bi-GRU CNN and Bi-GRULSTM CNN using a majority voting approach. Their findings revealed that TextCNN had a classification accuracy of 66%, while Bi-GRU-LSTM CNN had a classification accuracy of 70%. However, the ensemble classifier outperformed the individual models and delivered the best results for the classification challenge.
4. E. G. Dada et al., [4] has surveyed some of the publicly available datasets and performance metrics that used several machine learning approaches like KNN, Naïve Bayes, SVM, Decision Trees, Adaboost, Random forests and some deep learning algorithms like CNN for Email Classification for spam filtering.
5. F. Murtagh et al.,[5] has introduced the multi-layer perceptron and its applications as a supervised classification method with some examples and pointed out that any optimization technique requires careful consideration of implementation issues.
6. S. Vidros et al., [6] has analyzed the possible aspects of employment scam by introducing the EMSCAD, a publicly available dataset containing both real life legitimate and fraudulent job ads. By experimenting with the dataset they had shown the results of text mining in conjunction with metadata can provide a preliminary foundation for job scam detection algorithms.

In general, these papers and publications offer insightful information about the application of machine learning and NLP methods for identifying bogus job advertisements. They show the potential of these methods to enhance the precision and effectiveness of bogus job post identification, which is crucial for safeguarding job seekers from fraud and scams.

Problem Identification

The issue of fake job post detection is brought on by the fact that several people and businesses post false job vacancies online, frequently with the goal of misleading job seekers. These fictitious job postings may be used to sell goods, services, or training courses or to gather personal data. Fake job postings may occasionally be used to entice job seekers into dubious schemes like pyramid schemes or investment fraud.

In general, the issue of detecting false job postings is complicated and multifaceted, necessitating a combination of technical, legal, and educational solutions. Advanced fraud detection algorithms, more education and awareness campaigns for job seekers,

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

and tougher legislation and enforcement mechanisms for companies and job sites are some possible answers. [7-15]

Methodology

We have trained various supervised machine learning models using the EMSCAD dataset which consists of around 17880 job postings after cleaning and preprocessing the dataset.

A. Multi-Layer Perceptron

MLP is a type of neural network commonly used for classification tasks. It is made up of many layers of nodes (neurons), where each node gets input from the layer above, adds the inputs using a weighted formula, and then uses an activation function to create an output. It is customary for the activation function to be a non-linear function, such as ReLU, which brings non-linearity into the model and enables it to learn intricate correlations between inputs and outputs.

Backpropagation is used to train the network, which entails computing the gradients of the loss function concerning the network's weights and biases and updating them by gradient descent or a similar technique. It can be regularised to reduce overfitting and boost generalization performance using strategies like dropout, early halting, and weight decay. It is a strong model that can identify intricate patterns in the data and perform well on classification tasks, but it needs a lot of data and computer power to train and improve.

B. Other Supervised Machine Learning Algorithms

We used classifiers including Logistic Regression, Support Vector Machines, Decision Tree, Random Forest, Gradient Boosting, and XGBoost for comparative analysis of our dataset. Then, in order to increase accuracy, we deployed a soft voting classifier and selected three classifiers based on metrics like accuracy score, F1 score, and ROC AUC score.

C.DataSet

The Employment Scam Aegean Dataset (EMSCAD) dataset, which is made available to the public by the University of the Aegean Laboratory of Information & Communication Systems Security, was the dataset used in this work. There are 17,880 true job posts in this dataset, 17,014 of which are genuine and 866 of which are not. This dataset has been further processed, published to Kaggle, and is freely accessible.

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Implementation

A. Collect and preprocess data:

The employment Scam Aegean Dataset (EMSCAD) dataset which we have used has an initial size of 17,880 rows out of which 17,014 are real and 866 are fake. Most of the references have used either Text or Categorical whereas we have used both text and categorical values.

We have dropped features "job_id" which is unique for each job post, along with "location", "benefits" which has low feature score, "department", and "salary_range" which have many empty fields."fraudulent" is used as the class label, it takes on the value 1 if the job posting is fraudulent, and 0 otherwise. We have removed the rows which have more than 5 features as "nan" and filled the remaining "NaN" values with an empty string.The count of Null values is represented in Fig.1.Then we cleaned the text data from HTML tags, URLs, punctuations, digits, underscore, single characters, and multiple spaces using regular expressions and then applied SnowballStemmer to stem the text data.

```
job_id                   0
title                    0
location               346
department           11547
salary_range         15012
company_profile       3308
description              1
requirements          2695
benefits              7210
telecommuting            0
has_company_logo         0
has_questions            0
employment_type       3471
required_experience   7050
required_education    8105
industry              4903
function              6455
fraudulent               0
dtype: int64
```

Fig.1.Count of Null values in dataset

TfidfVectorizer is a tool that, when used with nltk stopwords on text features, can transform the text into a vector that can be used for analysis. It works by combining two key ideas: Term Frequency (TF) and Document Frequency (DF). Term frequency refers to

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

how often a particular word appears within a given document. One-hot encoding is used for categorical values which converts them into numeric values.

text_features=['title','company_profile', 'description',' requirements']

categorical_features=['telecommuting','has_company_logo','has_questions','employment_type','required_experience','required_education','industry']

B.Model training and evaluation:

In Fig.2. Models like Logistic regression, SVM, Decision tree, Random forest, Gradient boosting, XGBoost, and MLP with Adam optimizer are compared using various metrics like accuracy, F1 score, ROC AUC score, and more after training.

Model/Metrics	Accuracy	Precision	F1 Score	ROC-AUC
Logistic Regression	98.15	95	76	81
SVC	98.01	98	73	79
Decisionv Tree	98.35	80	80	89
Random Forest	**98.41**	100	**79**	83
Gradient Boosting	97.98	98	76	81
XGBoost	**98.69**	95	**84**	88
MLP	**98.71**	95	**85**	89

Fig.2.Performance of various Classifiers

Random Forest, XGBoost, and MLP with adam optimizer are found to be consistent across all the metrics, Although the Decision tree has more roc_auc score and recall score it has a very low precision score. So, we haven't considered a decision tree for voting. So we have applied a Soft Voting classifier on Random Forest, XGBoost, and MLP with adam optimizer and we observed improvement in accuracy and f1 scores.

C. Interface Creation:

Streamlit is an open-source framework for building web applications in Python that allows data scientists and machine learning engineers to quickly create interactive and intuitive user interfaces for their models. With Streamlit, one creates a simple web application that takes inputs from the user, runs them through your machine learning model, and displays the output in real-time. To use Streamlit, we first defined the layout of your application using Python code. This can include text inputs, drop-down menus, sliders, and other interactive widgets that allow the user to input data.Next, we wrote the code that loads your trained machine learning model and uses it to make predictions based on the user inputs.

To take input from the user to predict if the job posting is fraudulent or not we have chosen streamlit which is an open-source app framework in python language and is mainly used to build and share machine learning web apps.

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

For the text fields used, text_area takes input and for categorical fields, the select box is used then input data is converted into a pandas data frame and the same process which is used in datacleaning is used to clean.

Results and Conclusion

Fake job post prediction at an early stage can save job seekers and make them only apply for legitimate jobs among the postings available. In this work, we have developed a fake job post prediction software by applying various supervised machine learning algorithms to classify a given post taken as input from the user, as real or fake.

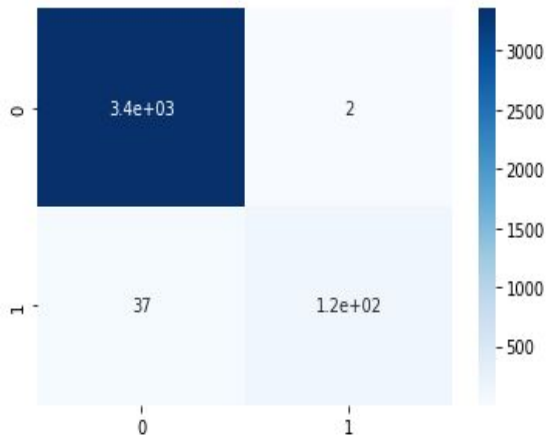

Fig.3. Confusion Matrix for Voting Classifier

This paper experimented with different algorithms such as Logistic Regression, Support Vector Machine(SVM), Decision Trees, Random Forest, Gradient boosting, XGBoost, and Multi-layer Perception(MLP).In Fig.3.Confusion matrix for Voting Classifier is given.

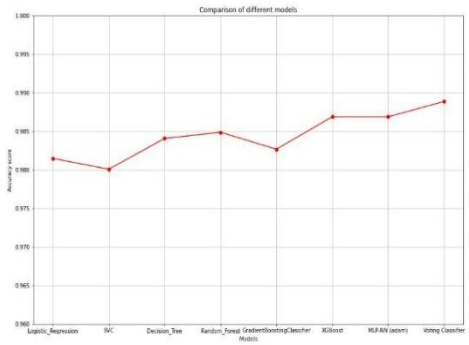

Fig.4.Accuracy Score comparison among various classifiers

In Fig.4. Accuracy Score is compared among various classifiers. So, we considered three of the mentioned algorithms with better-evaluating metrics for soft voting to further

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

improve the overall performance of the system. It is represented in Fig.5. This proposed approach achieved an accuracy of 98.89% which is much higher than the existing methods.

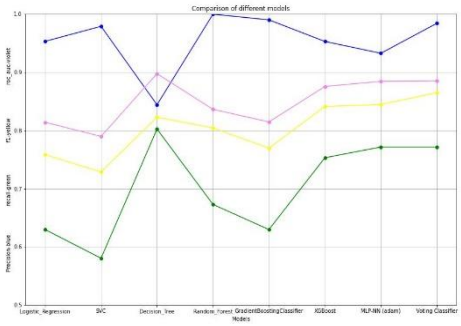

Fig.5.Comparison of various performance metrics among various classifiers

Future Scope

Several directions can be looked at for future work. Most existing fake job post detection methods rely on textual analysis. However, incorporating other modalities such as images and audio could improve the accuracy of the detection system. Different industries may have their jargon and language. As a result, developing domain-specific fake job post detection models can aid in more accurately detecting fake job postings.

References

[1] Sultana Umme Habiba,Md. Khairul Islam,Farzana Tasnim:"A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques"2021 2nd International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST) IEEE|DOI:10.1109/ICREST51555.2021.9331230

[2] E. Baraneetharan: "Detection of Fake Job Advertisements using Machine Learning algorithms",Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)

[3] 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 using the Bi-GRU-LSTM-CNN Model,” arXiv Prepr. arXiv1911.03644, 2019.

[4] 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.

[5] 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.

[6] 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

[7] Sri Hari Nallamala, et al., "A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 729 – 732, March 2018.

[8] Sri Hari Nallamala, et.al., "An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 542 – 545, March 2018.

[9] Sri Hari Nallamala, et.al, "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", International Journal of Advanced Trends in Computer Science and Engineering, (IJATCSE), ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2, Page No: 259 – 264, March / April 2019.

[10] Sri Hari Nallamala, et.al, "Breast Cancer Detection using Machine Learning Way", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3, Page No: 1402 – 1405, July 2019.

[11] Sri Hari Nallamala, et.al, "Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment", International Journal of Scientific and Technology Research, (IJSTR), ISSN: 2277-8616, Vol. 8, Issue 11, Page No: 2168 – 2173, November 2019.

[12] Kolla Bhanu Prakash, Sri Hari Nallamala, et al., "Accurate Hand Gesture Recognition using CNN and RNN Approaches" International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3216 – 3222.

[13] Sri Hari Nallamala, et al., "A Review on 'Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management'", Vol.83, May - June 2020 ISSN: 0193-4120 Page No. 11117 – 11121.

[14] Nallamala, S.H., et al., "A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems", IOP Conference Series: Materials Science and Engineering, 2020, 981(2), 022008.

[15] Nallamala, S.H., Mishra, P., Koneru, S.V., "Breast cancer detection using machine learning approaches", International Journal of Recent Technology and Engineering, 2019, 7(5), pp. 478–481.