IJTRET "Voice of Innovative Minds"
doi.org/10.54473/IJTRET.2022.6103
International Journal of Trendy Research in Engineering and Technology
Volume 6 Issue 1 February 2022
ISSN NO 2582-0958

MACHINE LEARNING-BASED FAKE JOB RECRUITMENT DETECTION SYSTEM

[1]Arryan Sinha, [2]Dr. G. Suseela
Under Graduate student, Assistant Professor, School of Computing,
SRM Institute Of Science And Technology, Kattankulathur, India
Received 04 February 2022 Received in revised form 12 February 2022 Accepted 13 February 2022

ABSTRACT

In order to avoid fraudulent online job postings, we use an automated tool that uses natural language processing (NLP), and classification techniques based on machine learning are suggested on paper. Using the NLP library SpaCy in python we have performed various analyzes such as semantic, syntactic, tokenization of the task profile extracting features, and using a machine learning algorithm called Random Forest we have predicted its accuracy to classify a job profile as Real or Fake.

Keywords: Machine Learning, Random Forest Classification, Natural Language Processing, Python, etc

1. INTRODUCTION

Employment scam is one of the most serious issues in the recent history of cybercrime. Many organizations have recently decided to publicize their job openings so that job seekers may find them conveniently and quickly. However, this goal can be a scam for fraudsters because they hire job seekers by taking money from them. Fake job advertisements may be sent to a reputable company for breach of trust. This discovery of fake jobs highlights the importance of developing an automated method for detecting false jobs and disclosing them to the public while preventing job solicitations.

In order to detect fake jobs, a machine learning approach is used, which employs numerous categorization algorithms. The Classification Algorithm isolates the fake job profile from the larger dataset of job advertisements.

We introduced the Random Forest Classification approach of supervised learning with the NLP package SpaCy to handle the problem of detecting fraudulent employment.

2. RELATED WORK

Email spam identification and fake news detection, according to various research, have gotten a lot of interest in the field of online fraud detection[1-5].

In the field of online fraud detection, email spam detection and fake news identification have gotten a lot of attention.

2.1 Email SpamDetection-

Lots of unwanted emails, which are part of the Spam emails section, usually arrive in the user's inbox. This can lead to inevitable storage problems and bandwidth usage. To address this issue, Gmail, Yahoo Mail, and Outlook have included spam filters based on Neural Networks. When dealing with email spam detection problems, content-based filtering, status-based filtering, heuristic-based filtering, memory or sample-based filtering, and flexible spam filtering methods are considered.

2.2 Fake News Detection-

False news on social media exposes malicious user accounts, and echo chamber results. The basic research for the detection of false stories is based on three perspectives - how false stories are written, how false stories spread, and how the user is associated with false stories. Features associated with news and social content have been eliminated, and a machine learning model has been built up to detect misleading stories.

3. PROPOSEDMETHODOLOGY

The aim of this research is to discover whether the work is fraudulent or not. Identifying and ending these fake job advertisements will help job seekers focus only on official positions. In this context, a set of data from Kaggle is used that provides information about a job that you may or may not suspect. The data set has a schema as shown in Fig. 1.

job_id	int64
title	object
location	object
department	object
salary_range	object
company_profile	object
description	object
requirements	object
benefits	object
telecommuting	int64
has_company_logo	int64
has_questions	int64
employment_type	object
required_experience	object
required_education	object
industry	object
function	object
fraudulent	int64

Fig 1 Schema Structure of the dataset

This dataset contains 17,880 job posts. In order to better understand the intended purpose as a basis, a multi-step process is followed to obtain an equal database. Before entering this data into any category, some of the previous processing methods are applied to this database. Missing values, stop-words, irrelevant attribute deletion, and additional space are some of the pre-processing approaches. Fig 2 explains the architecture diagram.

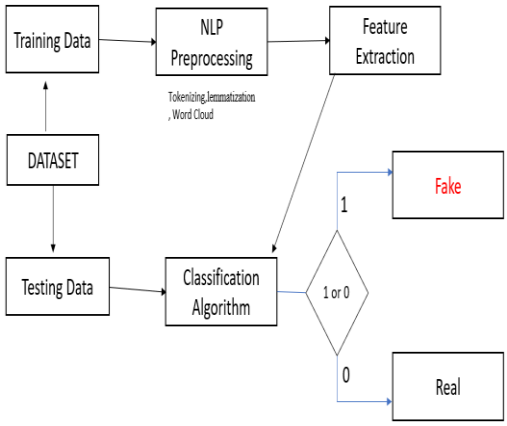

Fig 2 Architecture Diagram

3.1 Data Preprocessing

Pre-data processing involves converting raw data into well-structured data sets for data mining statistics. Raw data is frequently incomplete and formatted inconsistently. The adequacy or inadequacy of data correction has a direct impact on the effectiveness of any data analysis activity.

3.1.1 Data Visualization

Data visualization is a valuable skill because it allows us to obtain a qualitative knowledge of data. This is useful for studying and learning about the data set, as well as spotting patterns, corrupt data, and outliers.

In terms of bar graphs, pie charts, and other forms of data visualization, essential relationships can be expressed and demonstrated. Here we show the bar graph of country-wise job available Fig 3 and degree wise job available Fig 4

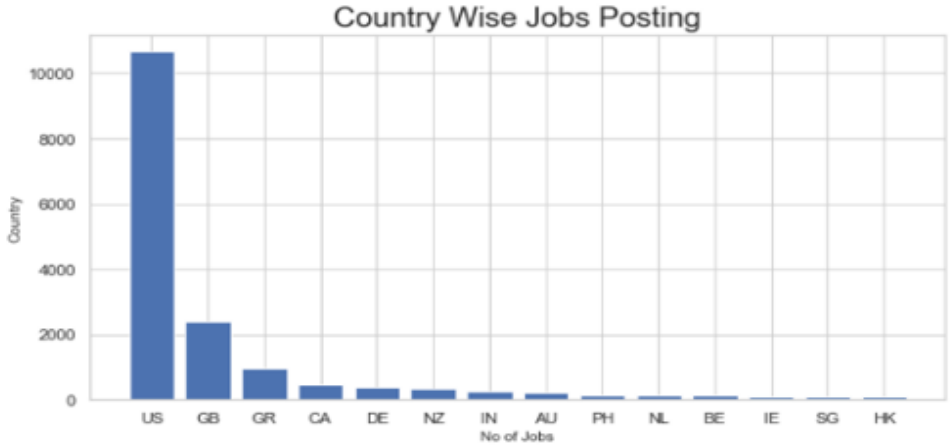

Fig 3 Country wise job posting

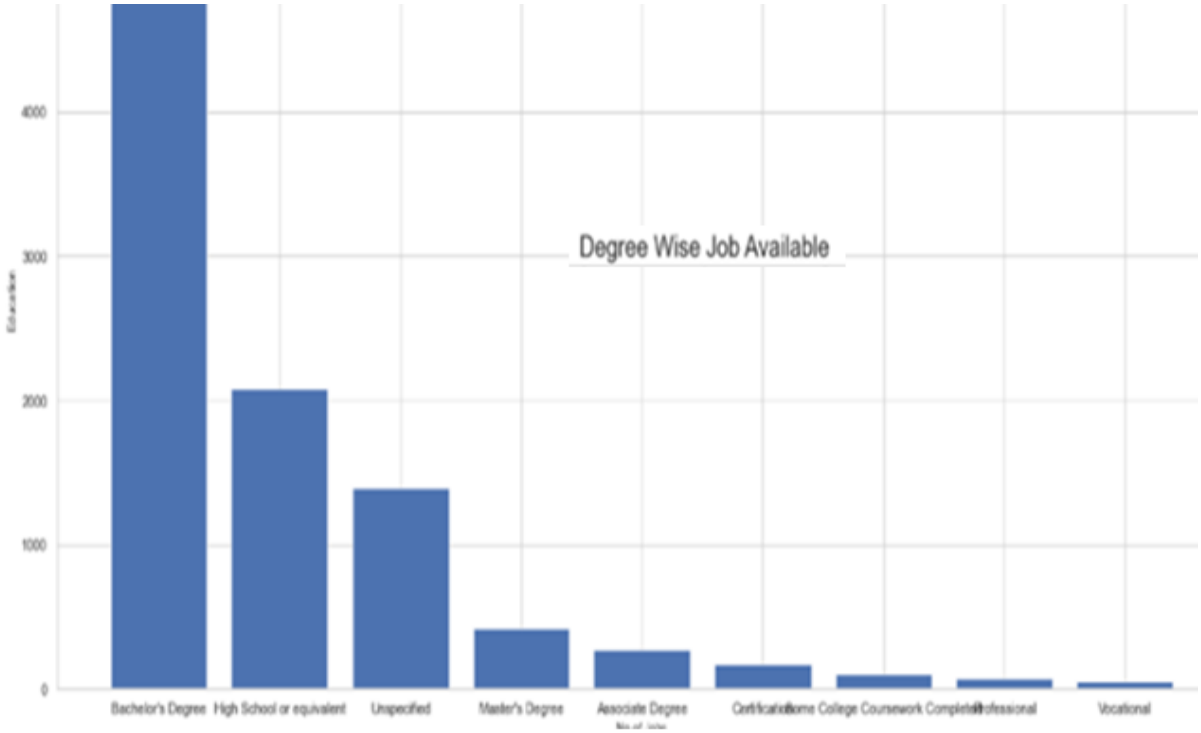

Fig 4 Degree wise job availability

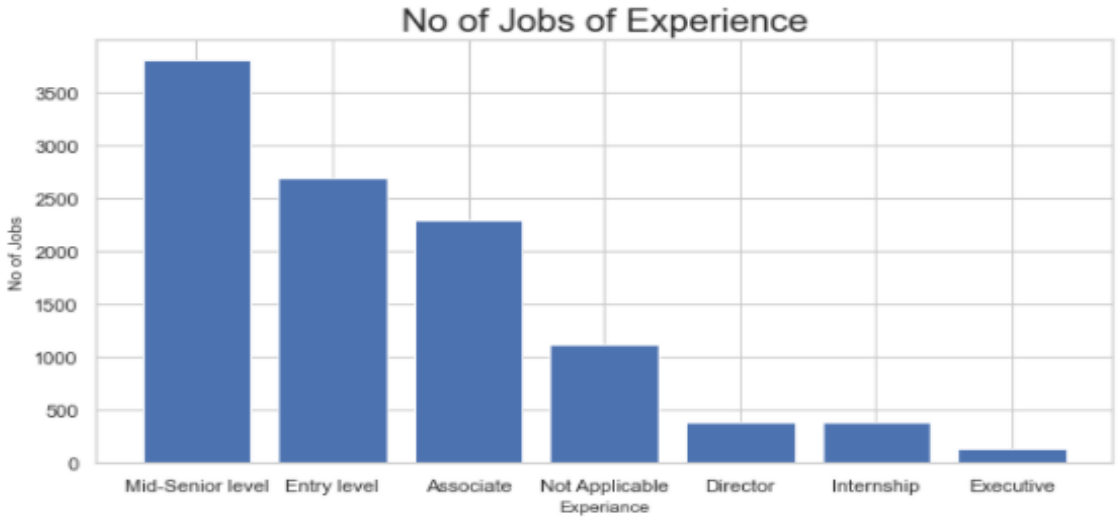

Fig 5 No of Jobs as per experience

3.1.2 Feature Selection

When building a predictive model, feature extraction is the process of minimizing the number of input variables. It is essential to limit the number of input variables in order to lower the model's calculation costs and, in some situations, to increase the model's performance.

3.2 NLP Preprocessing

Natural language processing (NLP)[6,7] is a collective term referring to the automatic processing of human languages. This includes both algorithms that take man-made text as input and algorithms that produce text that looks natural as output.

3.2.1 Word Cloud

It is a visualization technique for text data to identify the stopping words and extract the important word based on their frequency.

3.2.2 SpaCy

SpaCy is a free, open-source library for NLP in Python. It's written in Cython and is designed to build information extraction or natural language understanding systems. In this project, we did feature extraction through the Lemmatization process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word lemma, or dictionary form. And tokenization is used to break the sentence into separate words or tokens.

3.3 Implementation of Classifier

We use Random Forest Classifier which is a meta estimator that fits a number of decision tree classifiers [8,9]on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

The purpose of randomness is to decrease the variance of the forest estimator. Indeed, individual decision trees typically exhibit high variance and tend to overfit.

The injected randomness in forests yields decision trees with somewhat decoupled prediction errors. By taking an average of those predictions, some errors can cancel out.

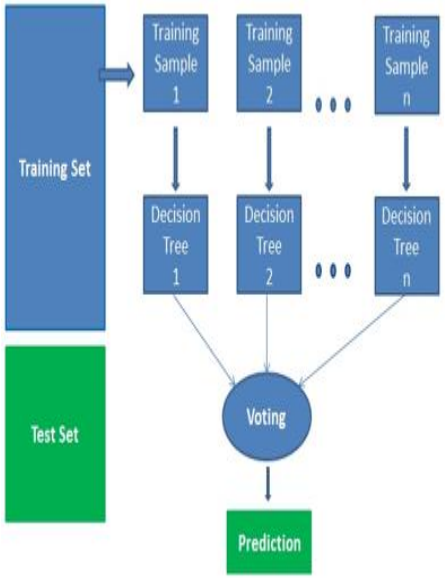

Fig 6 Random Forest classifier structure

Random forests achieve a reduced variance by combining diverse trees, sometimes at the cost of a slight increase in bias. In practice, the variance reduction is often significant hence yielding an overall better model

4. Experimental Analysis

Above mentioned dataset is trained and tested to find fake job vacancies in a given database that contains both false and official posts. The following table shows the Classification report of the prediction done through a random forest algorithm. Precision(total positive) shows of 97 %. F1 score for real jobs is 0.99 and for fake jobs 0.58

Classification Report

	precision	recall	f1-score	support
0	0.97	1.00	0.99	5104
1	1.00	0.40	0.58	260
accuracy			0.97	5364
macro avg	0.99	0.70	0.78	5364
weighted avg	0.97	0.97	0.97	5364

Fig 7 Classification report of Prediction

Although this random Forest Classifier has obtained a F1 score that closely resembles other competitors, this filter has shown significant performance in relation to other metrics.

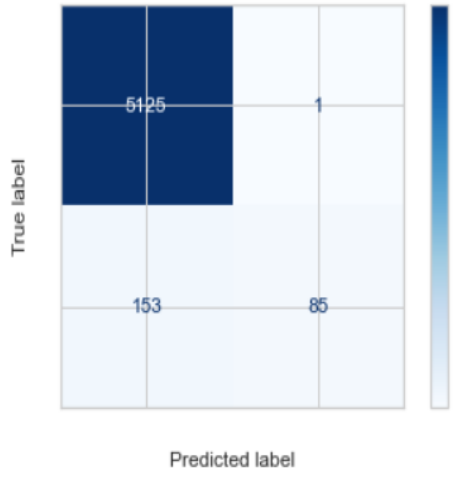

Fig 8 Confusion Matrix

5. 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. A supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that the Random Forest classifier outperforms its peer classification tool giving an accuracy of 97.1%.

References

[1] Bandyopadhyay Samir & Dutta, Shawni. 'Fake Job Recruitment Detection Using Machine Learning Approach'. International Journal of Engineering Trends and Technology. 68.(2020) 48-53,10.14445/22315381/IJETT- V68I4P209S.

[2] B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection," J. Inf. Secur., 10(3)(2019) 155–176, DOI: 10.4236/jis.2019.103009

[3] Shivam Bansal 'Real/Fake] Fake Job Posting Prediction, 2020 https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction

[4] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, ‐Spam review detection techniques: A systematic literature review,‖ Appl. Sci., 9(5)(2019)1–26, DOI:10.3390/app9050987

[5] K.Shu, A. Sliva, S.Wang, J.Tang,andH.Liu,‐Fake News Detection on Social Media,‖ACMSIGKDD Explor.News lett., 19(1)(2017)22–36, DOI 10.1145/3137597.3137600.

[6] Redd, Mallamma V., and M. Hanumanthappa. "Semantical and Syntactical Analysis of NLP." International Journal of Computer Science and Information Technologies 5(3) (2014) 3236 - 3238

[7] Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics.

[8] H. Sharma and S. Kumar, ―A Survey on Decision Tree Algorithms of Classification in Data Mining,‖ Int. J. Sci. Res., 5(4)(2016)2094–2097, DOI: 10.21275/v5i4.nov162954

[9] H. M and S. M.N, ‐A Review on Evaluation Metrics for Data Classification Evaluations,‖ Int. J. Data Min. Knowl. Manag. Process,5(2)(2015)01–11, 2015, DOI: 10.5121/ijdkp.2015.5201.