FAKE JOB DETECTION

A PROJECT REPORT

Submitted by

Vijayakumaran S (21ADR061)

Rohith S J (21ADR040)

Rishi Raghav G (21ADR038)

Vibeesh N (21ADR059)

Vignesh T (21ADR060)

for

FOUNTATION OF ARTIFICIAL INTELLIGENCE
AND
DATA SCIENCE

DEPARTMENT OF ARTIFICIAL INTELLIGENCE

KONGU ENGINEERING COLLEGE
(Autonomous)

PERUNDURAI ERODE – 638 060

Abstract

Employment fraud is becoming more prevalent. According to CNBC, there were twice as many employment frauds in 2018 than there were in 2017 and still increasing more. The state of the market today has resulted in substantial unemployment. The availability of work has been drastically reduced, and many people have lost their jobs as a result of economic hardship and the coronavirus's effects. Such a situation offers con artists the ideal opening. Due to a rare incidence, many people are becoming victims of scammers who feed on their despair. Most con artists use this technique to obtain personal information from their victims. Address, bank-account details, social privacy number, safety-pin , and other personal details are examples. As a student, I have encountered several of these fraudulent emails. Users are offered a highly lucrative job opportunity by the con artists, who then demand payment. Or In exchange for the promise of work, they demand money from the job applicant. NLP(natural language processing) and machine-learning approaches may be used for solving the severe dilemma (NLP). This project makes use of information from Kaggle. The characteristics of a job ad are contained in this data. These job listings are either labelled as genuine or bogus. Only a very minor portion of this collection consists of fake job listings. That is acceptable. We don't anticipate seeing many phoney job postings. We are gonna use different Machine Learning approach to this Fake Job detection problem like Naïve Bayes , Stochastic Gradient Descent (SGD), The Random forest ensemble classifier. This project is divided into five stages. The project's five stages are as follows: i. Problem definition ii. Data collection iii. Data Cleaning, Exploring, and Pre-processing iv. Modeling v.Evaluating

Keywords

Fake Job, Natural Language Processing, Online Recruitment, Machine Learning, SDG, Navie Bayas, Ensemble Approach.

I. INTRODUCTION

Economic difficulty and the effects of the coronavirus have significantly decreased the availability of jobs and led to employment loss for many people. Scammers would love to exploit a circumstance like this. These con artists are relying on people's desperation as a result of an unprecedented incident, and many people are falling victim to them. Many scammers will do this in order to get privacy details from the target of their fraud. Personal data includes things like address, bank account details, and social privacy number. Scammers entice victims with a great job offer before requesting money in return. As an alternative, they can require the job seeker to make a monetary investment in return for the assurance of employment. These days, there are a lot of job scams because of unemployment. Several websites can help recruiters identify qualified candidates. Sometimes, only to make money, fake recruiters may advertise a job on a job board. Many job boards experience this problem. Later, those looking for legitimate jobs visit a new job portal; however, phoney recruiters also move to this portal. Therefore, being able to tell the difference between real and fake employment possibilities is crucial. Employment fraud is one of the most important problems that has been addressed recently in the domain of internet recruiting scams. Nowadays, a lot of businesses like to post their job positions online so that candidates may easily and quickly find them. But the con artist can be using this as one of their fraudulent methods. However, this could be a hoax carried out by con artists who provide work to job seekers in exchange for cash. NLP(natural language processing) and machine-learning approaches may be used for overcome this hazardous issue (NLP). A machine learning strategy that makes use of several categorization algorithms is employed to detect fake job post. In this event, a classification method separates fake-job postings from a larger-pool of legitimate job postings and alerts the user. We are gonna use different Machine Learning approach to this Fake Job detection problem like Naïve Bayes , Stochastic Gradient Descent (SGD), The Random forest ensemble classifier.

II. LITERATURE SURVEY

In the manner of various studies, fake news, review spam, and email spam detection in the area of Online Fraud Detection have received particular attention.

A. Review Spam Detection :

In internet forums, people routinely post reviews of the products they purchase. It might serve as a guidance for other buyers when they select their products. In this situation, spammers might alter reviews to their financial-advantage, necessary as result the development of algorithms to identify these fake evaluations. This can be accomplished by utilising Natural Language Processing to extract attributes from the reviews (NLP). Following that, these features are subjected to machine learning techniques. Lexicon-based methods could be a substitute for machine learning methods that rely on dictionaries or corpora to weed out bogus reviews.

B. Email Spam Detection :

Spam emails, which are unwanted bulk emails, frequently reach users' mailboxes. Both bandwidth use and an inevitable storage issue could result from this. Gmail, Yahoo Mail, and Outlook service providers have implemented neural network-based spam filters to address this issue. When addressing the issue of email spam detection, a variety of methodologies are taken into account, including content-based filtering, case-based filtering, heuristic-based filtering, memory-based filtering, instance-based filtering, and adaptive spam filtering.

C. Fake News Detection :

In social media, Echo chamber effects and fraudulent user accounts are characteristics of fake news. Three perspectives are crucial to the basic study of false news detection: how fake news is Volume 8 in 2021, Issue 8 of JETIR (August 2021). Features related to news content and social environment are gathered and used to machine learning models in order to detect bogus news. You can access the data for this project at Kaggle.

17983 observations and 21 characteristics make up the dataset.

The data is a mix of binary, textual, and integer datatypes. Below is a quick definition of the variables.

LITERATURE REVIEW

PROBLEM STATEMENT- Fake Job Detection using various ML algorithms.

TITLE OF THE PAPER	PROBLEM DEFINITION	OBJECTVE	PROPOSED SYSTEM	PROS	CONS
1. Fake Job Detection Using Machine Learning(Random Forest ensemble classifier model)	These days, a lot of businesses want to post their vacant positions online so that job hunters may easily find them. However, this could just be a ruse used by con artists to get others to labour for them in exchange for money. This hoax deceives many individuals, who end up losing a lot of money.	In order to stop fake job advertising on the internet, this study suggests an automated solution based on machine learning-based categorization algorithms..	Random forest classifier model	The classification technique used was the Random forest ensemble classifier, which was constructed from a number of tree-structured classifiers..	Given the difficulty in identifying fraud, it may be dangerous for job searchers if the job is not flagged as fraudulent.
2.Identifying the fake job recruitment using knn.	Now a days job posting on the internet have grown popular and its is easy to find job for job seekers. But there also some fake job post are available in internet which are all posted by scammers. Some people have lost money due to scammers .so we are going to identify	To detect the fake job recruitment post and identifying the scammers. This increased awareness of bogus job postings encourages the development of an automated system for spotting hoaxes and alerting individuals to them so they won't apply for them..	Using text processing, an automatic fake detector model can tell the difference between real and false news (including articles, authors, and subjects). They have employed a unique dataset of news or items shared on Twitter via the PolitiFact website account.	It can able to detect fake job post and report it to the social media. It will aler the job seekers.	Sometimes it may detect the original job post also a fake job post and report it.
3.Detection of Fake job recruitment.	There are many job offers in the internet. The offered jobs may be fake or fraudulent. By this many people lost their money by	The main objective of this is identifying the fake job recruitment by detecting it. To avoid internet job postings that are fake.	supervised learning algorithms are originally taken into consideration as classification approaches to address the issue of recognising scammers on job	From a wider collection of job advertising, a classification tool separates out bogus job postings and notifies the	Sometimes by over usage it may not identify the fake one or alerts the user.

	investing in it.		postings.	user.	
4. Detection of Fake Job Recruitment Using Machine Learning.	For a variety of unethical motives, a false job posting is a (rarely) well-designed sort of fraud that targets job searchers. Nevertheless, a sceptical person scanning through the enormous pool of employment may believe that these frauds are legitimate.	The article suggests an automated application that uses machine learning-based categorization approaches to prevent fraudulent job postings online. To check for fraudulent posts on the web, many classifiers are applied, and the results of those classifiers are compared to find the optimum employment scam detection model..	The article suggests an automated application that uses machine learning-based categorization approaches to prevent fraudulent job postings online.	High Accuracy, Low Time, User Friendly.	Not Safe.

DATASET USED	**PREFERRED MATRICES USED**	**PERFORMANCE**
1. Job-related characteristics from the Kaggle dataset, including job id, title, location, and department	ACCURACY %	ACCURACY : 97.2%
2. Fake Job Description Prediction	ACCURACY%	ACCURACY : 97%
3. The dataset used are in 0 1 format and it's called as lables.	ACCURACY %	ACCURACY : 98.27%
4. Kaggle[13] dataset is used.	ACCURACY %, PRECISION, RECALL, MEAN SQUARED ERROR.	ACCURACY: 98%

III. METHODOLOGY

- DATASET :

You can access the data for this project at Kaggle - https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction.

This dataset contains 17,983 observations and 21 characteristics. The data is a mix of binary, textual, and integer datatypes. Below is a quick definition of the variables.

The data in the dataset is combination of integer, binary and textual datatypes. An insight about the dataset is given below.

#	Variable	Datatype	Description
1	job_id	int	Identification number given to each job posting
2	title	text	A name that describes the position or job
3	location	text	Information about where the job is located
4	department	text	Information about the department this job is offered by
5	salary_range	text	Expected salary range
6	company_profile	text	Information about the company
7	description	text	A brief description about the position offered
8	requirements	text	Pre-requisites to qualify for the job
9	benefits	text	Benefits provided by the job
10	telecommuting	boolean	Is work from home or remote work allowed
11	has_company_logo	boolean	Does the job posting have a company logo
12	has_questions	boolean	Does the job posting have any questions
13	employment_type	text	5 categories – Full-time, part-time, contract, temporary and other
14	required_experience	text	Can be – Internship, Entry Level, Associate, Mid-senior level, Director, Executive or Not Applicable
15	required_education	text	Can be – Bachelor's degree, high school degree, unspecified, associate degree, master's degree, certification, some college coursework, professional, some high school coursework, vocational
16	Industry	text	The industry the job posting is relevant to
17	Function	text	The umbrella term to determining a job's functionality
18	Fraudulent	boolean	The target variable → 0: Real, 1: Fake

A summary statistic is not required in this case because the majority of the datatypes are either Booleans or text. Job id is the lone integer, which is irrelevant for our investigation. To find null values, the dataset is further examined.

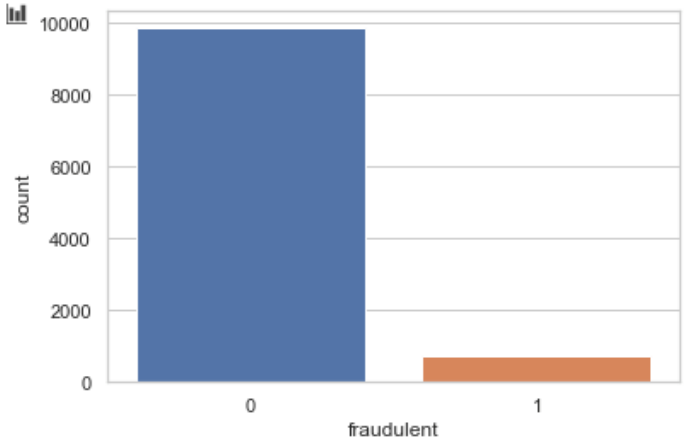
job_id	0
title	0
location	346
department	11547
salary_range	15012
company_profile	3308
description	1
requirements	2695
benefits	7210
telecommuting	0
has_company_logo	0
has_questions	0
employment_type	3471
required_experience	7050
required_education	8105
industry	4903
function	6455
fraudulent	0

The values for many variables, including department and salary range, are missing. The remaining analysis skips these columns.

After a preliminary analysis of the dataset, it became clear that the job listings were in various languages because they were taken from various nations. This initiative uses data from US-based sites, which make up almost 60% of the information, to streamline the procedure. To make sure that all of the data is in English for simple interpretation, this was done.

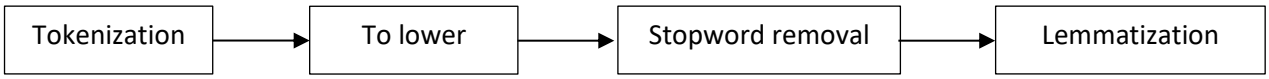
For additional study, the location is divided into two, the state and the city. The final dataset has 21 features and 10583 observations.

With 9891 (91% of the jobs) being actual and just 776 (or 9% of the jobs) being fraudulent, the dataset is significantly imbalanced. The gap may be very clearly seen in a countplot of the same data.



DATA PREPROCESSING :

TEXT PROCESSING:



Tokenization : The textual data in the dataset is divided into smaller pieces. The data's are divided into words in this instance.

To Lower : Then Lowercase letters are used to the separated words.

Stopword removal : Stopwords are phrases that lack significant further meaning. For instance, "the," "a," "an," "he," "had," etc. This phrase has been deleted.

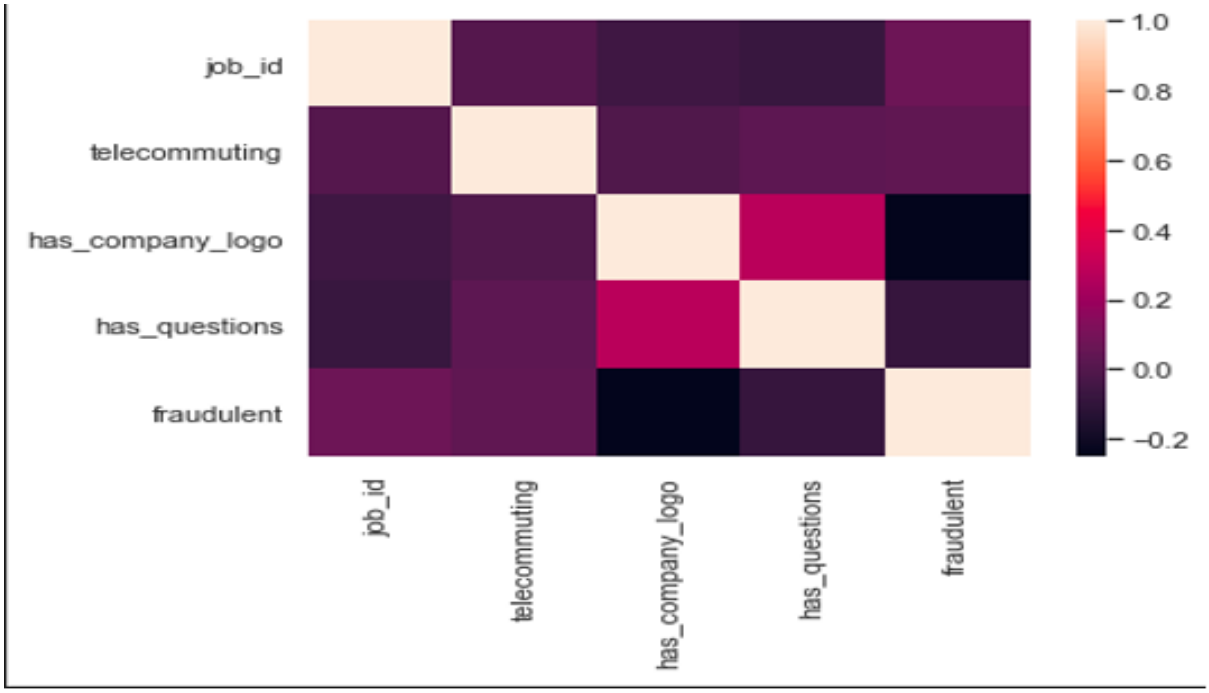
Lemmatization : The grouping of words with similar inflected forms is known as lemmatization.

IMPLEMENTATION :

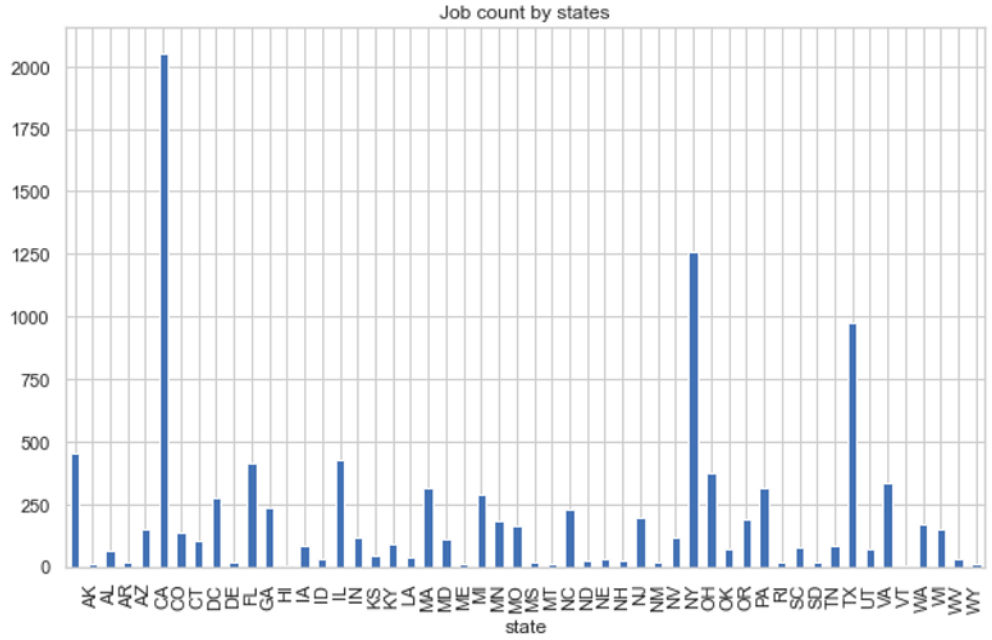
The dataset is broken down into text, numeric, and y-variable categories. For subsequent analysis, the text dataset is transformed into a term-frequency matrix. The datasets are then divided into test and train datasets using sci-kit learn. The train set, which comprises 70% of the dataset, is used to train the baseline model Naive Bayes and another model SGD. The models' combined results from the two test sets, text and numeric, are used to determine if a job posting is fake if both models indicate that a certain piece of data is not fraudulent. To lessen the bias of machine learning algorithms toward classes with a majority, this is done. The test set is used to assess the trained model's performance.

To enhance the model's outcomes, the independent variables have undergone a variety of modifications. Features have been added and removed to do this. Additionally, several penalties are applied to the final model's evaluation. The differences in the results, nevertheless, were incredibly small.

To investigate the link between the numerical data, the first stage in this project's visualisation of the dataset is the creation of a correlation matrix.There are no significantpositive or negative correlations between the numerical data in the correlation matrix.

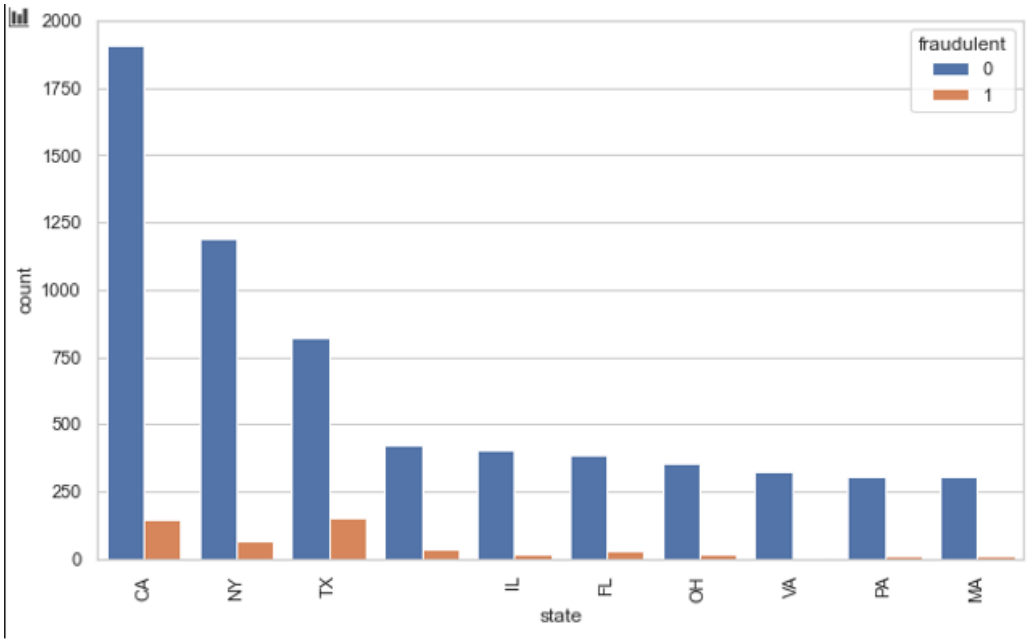


Regarding the Boolean variable telecommuting, a fascinating pattern was discovered. A 92% possibility that the job is fake exists when both of these variables have values of zero. The linguistic aspects of this dataset are examined after the numeric features. We begin this investigation by going to the site.

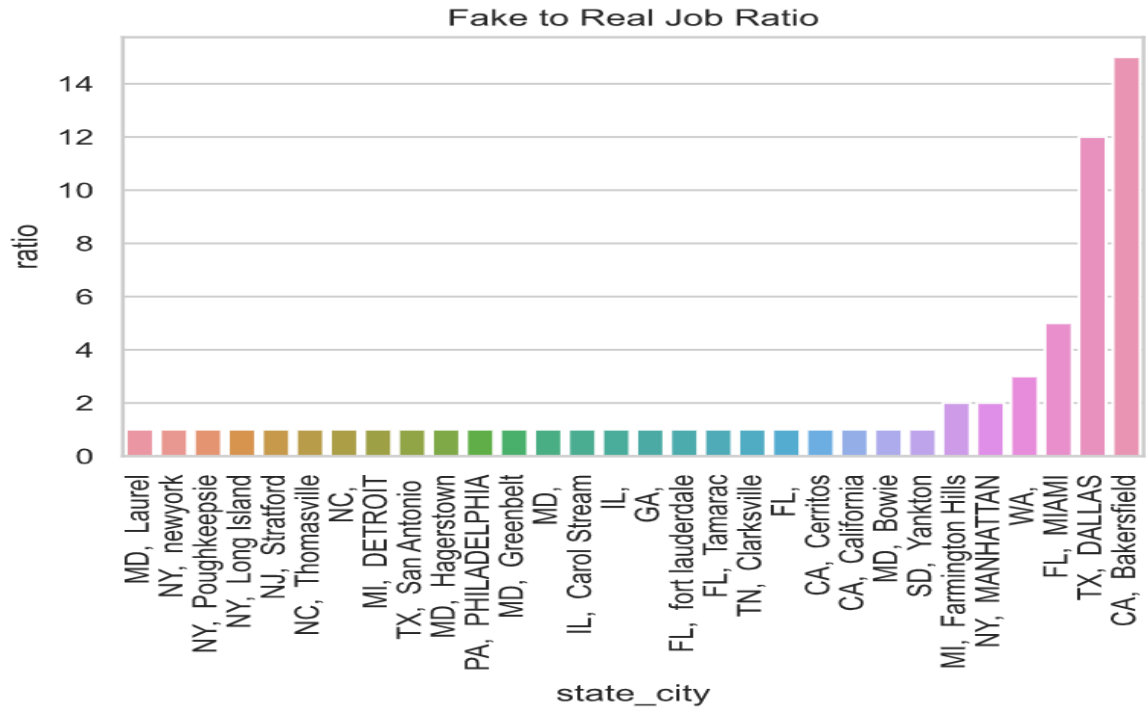


Texas, New York, and California have the most job listings.

The distribution of fake and actual employment in the top 10 states is depicted in this barchart.



According to the graph above, Texas and California are more likely than other states to have bogus jobs. Now we are going to plot ratio of fake,



The California has highest fake job ratio 15:1 and Texas has 13:2.There is a strong likelihood that any job listings from these websites are fake.

ALGORITHMS:

The Methods used in project are:

1. Random Forest
2. SGD Classifier
3. Naïve Bayes Algorithm
4. Natural Language Processing

The accuracy and F1 values of Naive Bayes and SGD classifiers are compared to select the final model. Naïve Bayes is the base model and is used because it can calculate the conditional probability that two events will occur based on the probability that each individual event will occur.

The SGD classifier is used to implement a simple stochastic gradient descent training routine that supports different classification loss functions and penalties.

IV. PERFORMANCE EVALUATION

SGD is the final model applied in this investigation. The table below displays the SGD outcomes as well as the baseline model.

Model	Accuracy	F1-score
Naive Bayes (base model)	0.971	0.743
SGD	0.975	0.79
Random Forest	0.974	0.76

According to metrics the SGD model has greater performance than baseline model.

V. CONCLUSION

Job fraud detection ensures job seekers only to receive legitimate offers from companies. Several machine learning techniques have been proposed for detecting employment fraud. Our model SGD is best suited for classifying fake job post and the scammers.

It has an accuracy 97%, so it can classify and detect whether the posted job is real or fake.

References

[1] S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar, “Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms,” Rev. GEINTECGESTAO Inov. E Tecnol., vol. 11, no. 2, pp. 642–650, 2021.

[2] B. Alghamdi and F. Alharby, “An intelligent model for online recruitment fraud detection,” J.Inf. Secur., vol. 10, no. 03, p. 155, 2019.

[3] “Report | Cyber.gov.au.” https://www.cyber.gov.au/acsc/report (accessedJun. 19, 2021).

[4] A. Pagotto, “Text Classification with NoisyClass Labels.” Carleton University, 2020.

[5] “Employment Scam Aegean Dataset.” http://emscad.samos.aegean.gr/ (accessed Jun. 19, 2021).

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

[7] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur and R. Mourya, "ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection," 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1-5, doi: 10.1109/IC3.2019.8844879..

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