A RELATIVE STUDY ON DECEPTIVE JOB POST DETECTING USING NLP

Mrs. Mary Anitha T
Assistant Professor
The Oxford College of Engineering
Mary.anitha.charlton@gmail.com

Tejaswini K
Student of MCA
The Oxford college of Engineering
tejukrisnaa@gmail.com

ABSTRACT

The study "A relative study on deceptive job post using NLP" suggests employing machine learning-based categorization algorithms as part of an automated method to stop fraudulent job advertisements on the internet. The study's goal is to find misleading job postings among a lot of ads by using different classifiers. The study evaluates how well ensemble and single classifiers identify these phony postings. The findings of the experiment show that when it comes to spotting fraudulent job listings, ensemble classifiers outperform single classifiers by a large margin.

Keywords: Machine learning, collaborative approach, online recruitment, fake job.

INTRODUCTION

A laser security framework is a sort of security and alert framework that employments laser light and a light sensor to distinguish the nearness of an intruder or an object in a confined zone. The system consists of a laser diode that emits a continuous beam of laser light, a light-dependent resistor (LDR) that senses theintensity of the light, and a buzzer that produces a sound when the light is interrupted. The system can be castoff to protect homes, businesses, or other valuable assets from unauthorized access or theft. The topic "A Relative Study on
Deceptive Job Post Detection using NLP" is especially pertinent since it discusses how to spot and stop fraudulent job practices by utilizing natural language processing (NLP) techniques.

A. prediction based on single classifier: Classifiers are prepared to foresee obscure test occurrences. The taking after classifiers are utilized to distinguish fake work postings:

Naive bayes classifier:

The theorem [3] of conditional bayes Probability is a principle that is utilized by Trusting [2], a supervised class Bayes classifier fiction technique. Although its likelihood estimates are imprecise, the classifier's decision-making is extremely effective in practice. In the following scenarios—where the characteristics are independent or entirely functionally dependent—this classifier achieves an extremely promising result. Instead of feature dependencies, the correctness of this decoder is determined by the quantity of information lost in the class as an outcome of the independence assumption.

Multi-layer perceptron classifier:

Multi-layer perceptrons's [4] can be used as supervised classification tools by

adding the ideal training settings. The quantity of nodes in each layer and the number of hidden layers in a multilayer perceptron might vary for a particular problem. We consider the network design and the training data when selecting the parameters [4].

Decision tree classifier:

Known as lazy learners or K-neighbor classifiers [5], the K-nearest neighbor classifiers recognize objects by using the closest working-out examples in the feature planetary. The classifier reflects k objects as the nearest item while classifying an object. The secret to employing this categorization approach effectively is indicating the proper value for k [5].

Ensemble approach-based classifiers: Several machine learning algorithms can work together to progress the accuracy of the system as a whole thanks to the collective technique. Random forest (RF) [8] uses regression techniques and the concepts of ensemble learning for classification issues. By applying its weight to various dataset subsamples, each tree-like classifier in this classifier casts a vote for the class that best fits the input. By integrating several unstable learners into a single learner, boosting is a helpful strategy that improves classification accuracy [9]. The classifier sequence with a weighted majority vote is chosen by the boosting approach by applying the classification algorithm to the reweighted training data. An outstanding example of a boosting approach that produces better outcomes even when the mediocre learners' performance is below average is AdaBoost [9]**.** Boosting algorithms work wonders for fixing issues with spam filtering. Boosting gradients [10] is an additional boosting technique-based classifier that utilizes the decision tree concept. It also lessens the loss in predictions.

CONNECTED WORK

Fake news, email spam, and review spam have been the topic of numerous research in the pitch of online fraud detection. A study on "A relative study on deceptive job post detection using NLP" has also attracted attention, emphasizing the use of natural language processing methods to detect fraudulent job listings.

A. review spam detection:

Reviews of products are regularly posted on online forums by people. It may facilitate the decision-making process for other customers. There is a need to create strategies for identifying reviews that are manipulated by spammers in order to boost their income. To do this, features from the reviews can be extracted using natural language processing (NLP). Machine- learning algorithms are then used for these features. Lexicon-based techniques could be a suitable replacement for machine learning techniques that eliminate spam reviews using a corpus or vocabulary [1]**.**

B. Email spam detection:

Spam consists of unwanted mass emails that frequently find their way into users' inboxes, causing increased bandwidth use and possible storage problems. In order to combat this, email service providers such as Gmail, Yahoo Mail, and Outlook employ neural network-based spam filters. These filters take into account several adaptive spam filtering strategies, such as case-based, instance-based, heuristic- based, memory-based, and

content-based filtering approaches.

C. fake news detection:

Fake news on social media is often associated with echo chamber effects and malevolent user profiles. It's critical to take into account three aspects of false news detection research: the production of fake

Reviews of products are regularly posted on online forums by people. It may facilitate the decision-making process for other customers. There is a need to create strategies for identifying reviews that are manipulated by spammers in order to boost their income. To do this, features from the reviews can be extracted using natural language processing (NLP). Machine- learning algorithms are then used for these features. Lexicon-based techniques could be a suitable replacement for machine learning techniques that eliminate spam reviews using a corpus or vocabulary [1].

B. Email spam detection:

Spam consists of unwanted mass emails that frequently find their way into users' inboxes, causing increased bandwidth use and possible storage problems. In order to combat this, email service providers such as Gmail, Yahoo Mail, and Outlook employ neural network-based spam filters. These filters take into account several adaptive spam filtering strategies, such as case-based, instance-based, heuristic- based, memory-based, and content-based filtering approaches.

C. fake news detection:

Fake news on social media is often associated with echo chamber effects and malevolent user profiles. It's critical to take into account three aspects of false news detection research: the production of fake

news, its dissemination, and user interactions with fake news. The procedure of detecting false news entails examining social context in addition to content- related variables, which are subsequently utilized to train machine learning models for detection [12].

EXISTING METHODOLOGY

Much effort has been put into rectifying this issue by major job-providing sites such as Linked, Indeed Jobs, Glassdoor, etc. In addition to thorough fact-checks and inspections, they use carefully examined verification from the job posting businesses. However, a surprising amount of these fraud postings may be found on small, local platforms like monster.com, naukri.com, and shine.com. It typically takes some time for these postings to be removed. Numerous instances are noted where the advertisement and pamphlet are disseminated among individuals via social media platforms.

This challenge can be resolved by utilizing machine learning to forecast the likelihood that a job is fake, allowing the candidate to remain vigilant and make informed judgments when necessary. This is made possible by the wealth of data available on the internet and the growing capabilities of machine intelligence. The model can be trained and then quickly deployed to a local system using an app or web extension. It will be constructed using pre- existing data that has been made public by reliable Organization.

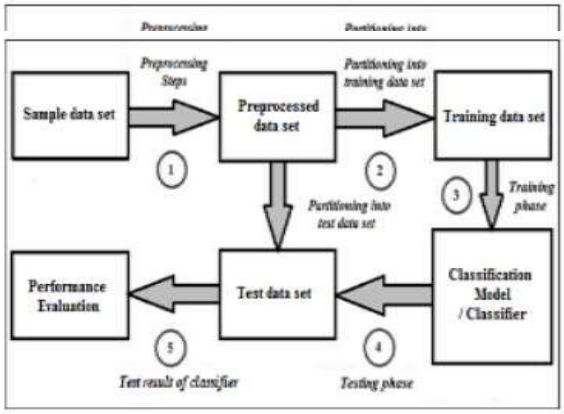

Fig.1 A detailed explanation of how classifiers operate.

PROPOSED METHODOLOGY

The word count, sentiment, and description of job posts were found to be strongly correlated with the probability of the job being fraudulent in the exploratory data analyses (D) and (E). Our use of natural language processing (NLP) to sentiment analysis, semantic pattern identification, and model building resulted from this realization.

Natural language processing (NLP) is concerned with extracting meaning from large volumes of natural language input, as well as with semantics and context. In order to build a machine learning model and find trends, this analysis is helpful. Knowing each word's logical and contextual meaning is essential to correctly identifying different job categories. Glove, Word2Vec, and Fast Text are well-liked word embedding methods that work well for tackling this problem.

We have chosen to employ the Glove (global vectors for word representation) model, a state-of-the-art method for natural language processing created by Stanford University academics. The foundation of Glove is the impression of word embeddings, which capture the context and meaning of words. Because it preserves word data for extended periods of time, this approach is preferred above others because it improves contextual knowledge across a worldwide corpus. The advantages of global matrix factorization and the local context window approach are integrated by combining a weighted least squares objective with a global log-bilinear regression model.

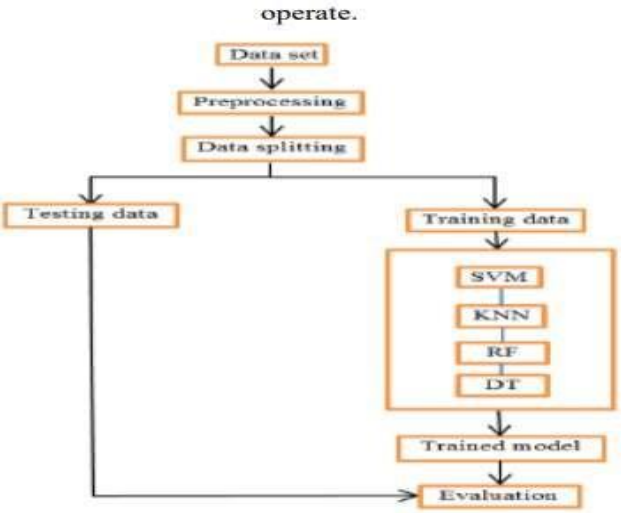

Fig.2 This frameworks classification models

EXPERIMENTAL RESULT

Every classifier that is described is trained and assessed using a dataset that includes both phony and real job postings. While Table 2 shows the outcomes for classifiers that employ ensemble approaches, Table 1 compares these classifiers using a range of assessment metrics. The total act of each classifier is shown in Figures 4 through 7, with regard to mean squared error (MSE), accuracy, Cohen's kappa score, and F1- score.

TABLE I

Act Judgement Chart For Single Classifier Based Calculation

Performance Measure Metric	Naive Bayes Classifiers	Multi-Layer Perception Classifier	K-Nearest Neighbor Classifier	Decision Tree Classifier
Accuracy	72.06%	96.14%	95.95 %	97.2%
F1-score	0.72	0.96	0.96	0.97
Cohen-Kappa score	0.12	0.3	0.38	0.67
mse	0.52	0.05	0 .04	0.03

EXPLORATORY DATA ANALYSIS

In demand to accomplish a number of goals, including shorter calculation times, noise reduction in the data, improved visual clarity for easier navigation, and increased model correctness, exploratory data analysis, or EDA, will be used.

A. Correlation Matrix

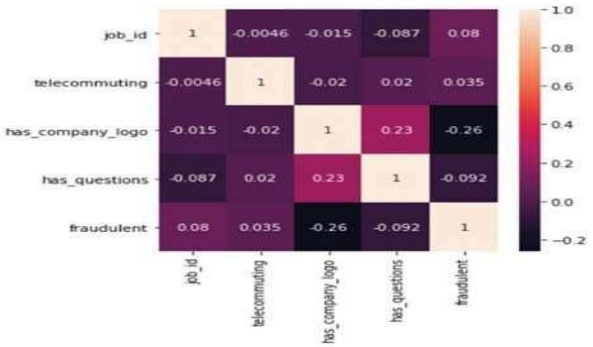

Fig. 1 correlation matrix for few data fields

B. Visualizing the amount of fraud job posts

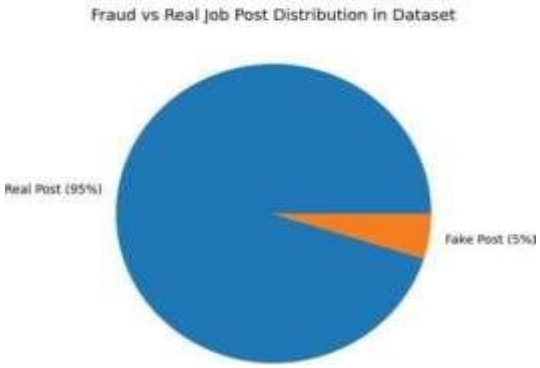

Fig. 2 Amount of fake vs real job posts

C. Total number of words in a real job post

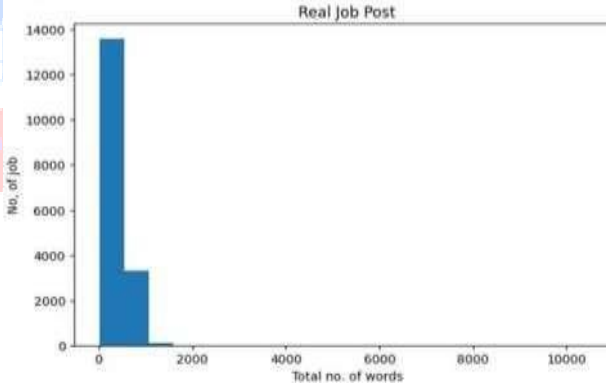

Fig.5 Histogram to depict range of words in all real job post

D. Total number of words in a fake job post

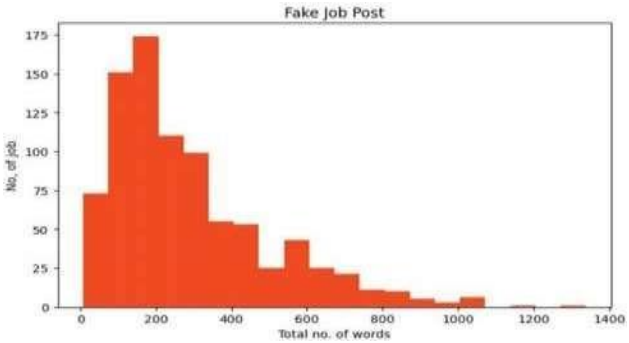

Fig. 6 Histogram to depict range of words in fraud job posts

CONCLUSION

We may infer from the data that the model performed exceptionally well, with an accuracy of over 97%. Furthermore, even with realistic LinkedIn data, the model's truth was greater than 98%, demonstrating its forcefulness and realism. The well-liked GloVe technique can be applied to practical NLP scenarios and is simple to implement. Therefore, users may easily use this realistic, easy-to-deploy model to gain extremely reliable predictions, alerts, and assistance.

FUTURE WORK

The outcomes of the Glove algorithm show great promise. Later, its effectiveness can be contrasted with that of Word2vec, an NLP algorithm that is also widely used for sentiment analysis and has a similar version. Improve the bogus job post detection project by using feedback loops for ongoing improvement and adaption to new scam strategies, incorporating sophisticated machine learning models such as BERT for contextual analysis, and putting real-time monitoring with automatic alerts in place. We intend to combine the two models, assigning a specific weight to each, and determine the ideal weight for both models to produce the greatest outcomes.

Reference

1) P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, “Semantic expansion using word embedding clustering and convolutional neural networks for improving short text classification,” Neurocomputing, vol. 174, pp. 806–814,2016.

2) Li, G. Zhan, and Z. Li, “News Text Classification Based on Improved BiLSTM-CNN,” in 2018 9th International Conference on Information Technology in Treatment and Schooling (ITME), pp. 2018, 890–893.

P. Wang, Xu, B. G. Xu, Xu. Tian, L.-C. Liu, and H. Hao, “Semantic expansion using word embedding clustering and convolutional neural network for improving short text. J. R. Scanlon and M. S. Gerber, “Automatic detection of cyber-recruitment by violent extremists,” Security Informatics, vol. 3, no. 1, p. 5, Dec. 2014, doi: 10.1186/s13388-014-0005-5.

3) S. Dutta and S. K. Bandyopadhyay, “Fake Job Recruitment Using Machine Learning Approach,” International Journal of Engineering Trends and Technology, vol. 68, no. 4, pp. 48–53, April 2020, doi: 10.14445/22315381/IJETT-V68I4P209S.