Fake Job Detection Using

Machine Learning

Shawni Dutta#1 and Prof.Samir Kumar Bandyopadhyay\*2

*#1Lecturer, Department of Computer Science Department University*

*2Academic Advisor, the Bhowanipur Society College*

*The Bhowanipur Education Society College, 5, Elgin Rd, Sreepally, Bhowanipore, Kolkata, West Bengal*

*700020, India*

***Abstract*** — *To avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed in the paper. In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by the job-seekers. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them. Many people fall for this scam and lose a lot of money. By performing an exploratory data analysis on the data, based on the insights we would gain, we can find out which job posting are fake and which are not. For this purpose, machine learning approach is applied which employs several classification algorithms for recognizing fake posts. By using previous data of fake and legit job posting the system would train the model to classify jobs as real or fake. To address the problem of identifying scams on job posting, supervised learning algorithm as classification techniques can be considered initially. It is going to use two or more machine learning algorithms and will choose the one which would yield higher accuracy score in the prediction which can predict whether a job posting headline is fake or legit.*

**Keywords *—*** *Fake Job, Online Recruitment, Machine Learning, Ensemble Approach.*

# I. INTRODUCTION

Economic stress and the coronavirus’s impact have significantly reduced job availability and job loss for many individuals. A case like this presents an appropriate opportunity for scammers. Many people are falling prey to these scammers using the desperation that is caused by an unprecedented incident. Most scammers do this to get personal information from the person they are scamming. Personal information can contain addresses, bank account details, social security numbers, etc. The scammers provide users with a very lucrative job opportunity and later ask for money in return. Or they require investment from the job seeker with the promise of a job. Now-a-days, there are a lot of job scams because of unemployment.

There are a lot of websites which connect a recruiter to a suitable candidate. Sometimes, fake recruiters post a job posting on the job portal with a motive to get money. This problem occurs with many job portals. Later, people shift to a new job portal in search of real job but the fake recruiters join this portal as well. So, it is important to detect real and fake jobs. Employment scam is one of the serious issues in recent times addressed in the domain of Online Recruitment Frauds (ORF). In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by the job-seekers. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them.

This is a dangerous problem that can be addressed through Machine Learning techniques and Natural Language Processing (NLP). For this purpose, machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user. To address the problem of identifying scams on job posting, supervised learning algorithm as classification techniques are considered initially. A classifier maps input variable to target classes by considering training data. Classifiers addressed in the paper for identifying fake job posts from the others are described briefly. These classifiers based prediction may be broadly categorized into – Single Classifier based Prediction and Ensemble Classifiers based Prediction.

# II. LITERATURE SURVEY

Online recruitment fraud detection is a relatively new field and there is not much work done in this area. There are some indirect methods to solve Online recruitment fraud to little extent such as Email Spam filtering that restricts to send email to user that has some advertising content, anti-phishing techniques to identify fake websites, countermeasures against opinion fraud to help identify the posting of deceptive and misleading fake reviews. According to several studies, Review spam detection, Email Spam detection, Fake news detection have drawn special attention in the domain of Online Fraud Detection.

## A. Review Spam Detection:

People often post their reviews online forum regarding the products they purchase. It may guide other purchaser while choosing their products. In this context, spammers can manipulate reviews for gaining profit and hence it is required to develop techniques that detects these spam reviews. This can be implemented by extracting features from the reviews by extracting features using Natural Language Processing (NLP). Machine learning techniques are applied on these features.

## B. Email Spam Detection:

Unwanted bulk mails, belong to the category of spam emails, often arrive to user mailbox. This may lead to unavoidable storage crisis as well as bandwidth consumption. To eradicate this problem, Gmail, Yahoo mail and Outlook service providers incorporate spam filters using Neural Networks. While addressing the problem of email spam detection, content based filtering, case-based filtering, heuristic based filtering, memory or instance-based filtering, adaptive spam filtering approaches are taken into consideration.

## C. Fake News Detection:

Fake news in social media characterizes malicious user accounts, echo chamber effects. The fundamental study of fake news detection relies on three perspectives- how fake news is written, how fake news spreads, how a user is related to fake news. Features related to news content and social context are extracted and a machine learning models are imposed to recognize fake news.

To the best of knowledge only Vidros. propose a method to detect the fraud jobs. However, they worked only with balanced dataset and the performance of prediction algorithms on an imbalanced dataset is not known.  Hence it is important to evaluate the prediction models on imbalanced dataset.The ORF Detector for online fraud detection proposed is an ensemble-based model. They have taken three baseline classifiers J48, Logistic Regression and Random Forest and applied average vote, Majority vote and Maximum vote on the classifers. But the main drawback of this approach is it only worked on balanced dataset and also yields less accuracy.

**Single Classifier based Prediction (Models implemented) –**

Classifiers are trained for predicting the unknown test cases. The following classifiers are used while detecting fake job posts-

## A] NAIVE BAYES –

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. The accuracy of this classifier is not related to feature dependencies rather than it is the amount of information loss of the class due to the independence assumption is needed to predict the accuracy.

## B] SUPPORT VECTOR MACHINE –

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. SVMs have their unique way of implementation as compared to other machine learning algorithms. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

## C] LOGISTIC REGRESSION -

Logistic Regression is a ‘Statistical Learning’ technique categorized in ‘Supervised’ Machine Learning methods dedicated to ‘Classification’ tasks. Logistic regression model takes a linear equation as input and use logistic function and log odds to perform a binary classification task. Logistic regression is a classification algorithm used to assign observations to a discrete set of classes It transforms its output using the logistic sigmoid function to return a probability value.

**Ensemble Approach based Classifiers (Random Forest)-**

Ensemble approach facilitates several machine learning algorithms to perform together to obtain higher accuracy of the entire system. Random forest exploits the concept of ensemble learning approach and regression technique applicable for classification based problems.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if its true otherwise the whole dataset is used to build each tree. This classifier assimilates several tree-like classifiers which are applied on various sub-samples of the dataset and each tree casts its vote to the most appropriate class for the input.

# III. PROPOSED METHODOLOGY

The main goal of this system is to determine whether a job posting news is fake or legit. Identifying and deleting fraudulent job postings will allow job seekers to focus solely on actual employment openings. In this system we intend to take a dataset from Kaggle that provides information regarding the job which includes several fields like job\_id, title, location, department etc. Then comes data preprocessing removing trivial spaces, null entries, stopwords and many more. After preprocessing and cleaning the data, making it prediction ready the data is fed to the classifier for predictions.

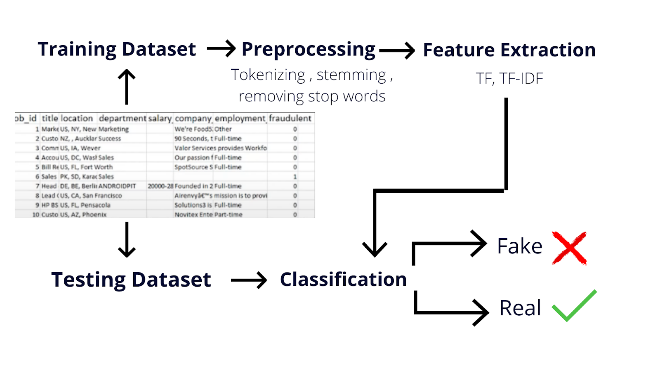


Fig. 1. Fake Job Detection Framework

## A. Dataset Details

This dataset from kaggle includes 17,880 data entries of job postings. Before fitting this data into any of the machine learning models or classifiers we firstly need to preprocess this data to get it ready for prediction purposes. Before fitting this data to any classifier, some pre-processing techniques are applied to this dataset. Pre-processing techniques include missing values removal, stop-words elimination, irrelevant attribute elimination and extra space removal. This prepares the dataset to be transformed into categorical encoding in order to obtain a feature vector.

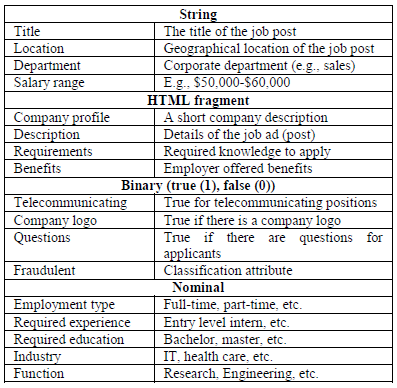


Fig. 2. Detailed description of Data

## B. Data Preprocessing

Preprocessing the data is a concept that includes changing the raw data into a clean data set. The dataset is preprocessed in order to check missing values, noisy data, and other inconsistencies before executing it to the algorithm. It also includes extracting noise and uninformative characters and words in the text, stop-words elimination, irrelevant attribute elimination and extra space. A pre-processing step was required with the data set due to its nature, before being applied to the classifier. Now here the data is textual we need to convert it into some numerical format to actually get some predictions.

Here NLP is used. NLP (Natural Language Processing) **Natural language processing** (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. A pipeline is created to make the textual data machine understandable.

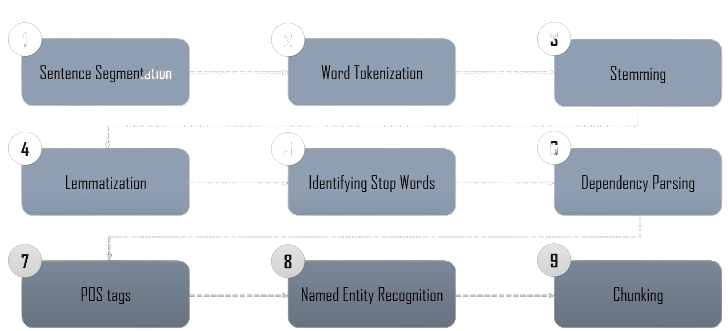
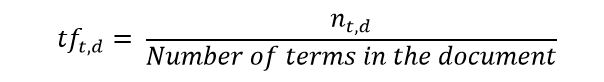


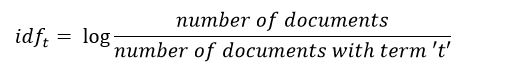
Fig. 3. NLP Pipeline

## C. Feature Extraction

Feature extraction is a part of the dimensionality reduction process, in which an initial set of the raw data is divided and reduced to more manageable groups. The most important characteristic of these large data sets is that they have a large number of variables. These variables require a lot of computing resources to process. So Feature extraction helps to get the best feature from those big data sets by selecting and combining variables into features, thus, effectively reducing the amount of data. These features are easy to process, but still able to describe the actual data set with accuracy and originality. In this study TFI-DF is used for feature extraction. TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.



IDF (Inverse Document Frequency)  is a measure of how important a term is. We need the IDF value because computing just the TF alone is not sufficient to understand the importance of words:



## D. Implementation of Classifiers

In this section classifiers are trained using appropriate parameters. This framework utilized Logistic Regressor, SVM, Naive Bayes models for predictions. SVM has several important features because of this it acquires publicity and has hopeful experimental performance. SVM creates a hyper level in authentic input space to divide the data points.

Whereas Naive Bayes predict the probabilityof different classes based on data and logistic regressor understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. The Classification algorithm used was the Random forest ensemble classifier, was constructed based on a combination of tree-structured Classifiers.

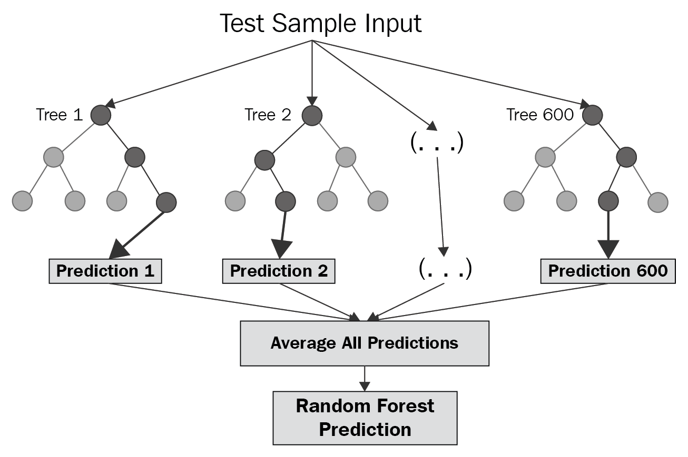


Fig.4. Random Forest classifier concept

This random forest model was built on 100 numbers of estimators on which the boosting is terminated. After constructing these classification models, training dataset is used for prediction purpose, after prediction is done performance evaluation is done.

## E. Performance Evaluation Metrics

When performing classification predictions, there are four types of outcomes that could occur: TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative). We have used four metrics for evaluating the performance of Fake Job detection system which are:

**Accuracy**: Accuracy is one metric which gives the fraction of predictions our model got right.

Accuracy = TP+TN/TP+FP+FN+TN

**Recall**: Recall gives the fraction you correctly identified as positive out of all positives.

Recall = TP/TP+FN

**Precision**:  Precision gives the fraction of correctly identified as positive out of all predicted as positives.

Precision = TP/TP+FP

**F1** **Score**: It is defined as the harmonic mean of the model’s precision and recall.

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

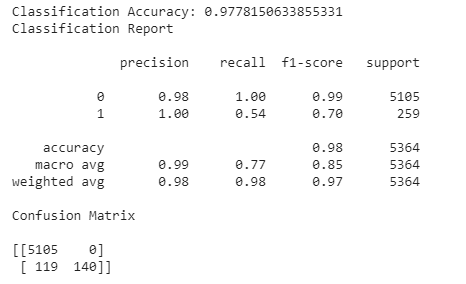


Fig.5. Classification Report

# IV. EXPERIMENTAL RESULTS

Considering the fraud detecting problem, the situation of not detecting the job as fraud (low sensitivity) could be threatening for job-seekers. Whilst the low specificity (predicting legitimate job as fraud) may only cause a further inspection by a human given the fact that real jobs would be obvious to realize. However, the problem lies in tricking people with fraud jobs that may look like real ones. Table 6 shows a comparison between the proposed model.

# 

Fig.6. Accuracy of different Classifiers

# V. MODEL DEPLOYMENT

To make our model available for end users we are going to deploy our model using Python Flask on Heroku.

Flask:

Flask is a web application framework written in Python. It has multiple modules that make it easier for a web developer to write applications without having to worry about the details like protocol management, thread management, etc.

Heroku:

Heroku is a cloud platform that supports several programming languages in which we can deploy our applications.

## VI. CONCLUSIONS

Only reputable business offers will be sent to you. Several machine learning methods are proposed for detecting employment scams. In this work, we discuss counter measures. Supervised mechanism is used to demonstrate the utilization of many mechanisms. Classifiers for detecting job scams. The results of the experiments show that Random Forest is effective. The classifier exceeds its peers in classification. The proposed method had a 97 percent accuracy rate. Which is significantly greater than current approaches?

## REFERENCES

[1] 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,” 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 (accessed Jun. 19, 2021).

[4] A. Pagotto, “Text Classification with Noisy Class 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,” in 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1–5.

[8] J. Kim, H.-J. Kim, and H. Kim, “Fraud detection for job placement using hierarchical clusters-based deep neural networks,” Appl. Intell., vol. 49, no. 8, pp. 2842–2861, 2019.

 [9] T.-J. Yang, “Neural network simplification using a progressive barrier based approach.” Massachusetts Institute of Technology, 2018. [10] C. X. Ling and C. Li, “Data mining for direct marketing: Problems and solutions.,” in Kdd, 1998, vol. 98, pp. 73–79. [11] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: one-sided selection,” in Icml, 1997, vol. 97, pp. 179–186.