FAKE JOB POSTING PREDICTION USING MACHINE LEARNING APPROACH

Mrinal Kumari [1] , NSK Satya kala [2] , Nandini R [3] , Dilip HK [4] , Prof. Rashmi KT [5]

[1,2,3,4] Students, CSE Department, Sri Krishna Institute of Technology, B'lore-560090, India

[5] Faculty, CSE Department, Sri Krishna Institute of Technology, B'lore-560090, India

ABSTRACT — There is an increase in employment fraud. In 2018, there were twice as many job scams than there were in 2017, according to CNBC. High unemployment is a result of the current state of the market. Numerous people have lost their jobs as a result of economic stress and the coronavirus's effects, which have significantly decreased the number of jobs available. Such an instance offers fraudsters the ideal chance. Due to the desperation brought on by this exceptional tragedy, many individuals are becoming victims of these fraudsters. The purpose of most scammers doing this is to obtain personal information from their victims. Personal data may include an individual's address, financial account information, social security number, etc. As university students, we have encountered numerous instances of these scam emails. Users are offered a highly lucrative work opportunity by fraudsters, who then demand payment. Or they demand money from the job seeker in exchange for the promise of employment. Natural language processing (NLP) and machine learning approaches can be used to solve this hazardous issue.

Keywords: Fake Job Prediction, Real and Fake, NLP, Naïve Bayes, SGD Classifier.

I. INTRODUCTION

One of the most significant issues in the area of false recruiting that has lately come to light is employment fraud. Nowadays, a lot of businesses choose to publish job openings online so that candidates may apply quickly and easily. However, since scammers hire job seekers by defrauding them of their money, this intention could be fraudulent. False classified advertising can be posted against legitimate businesses and are untrustworthy. The goal of this fraudulent job detection is to develop automated technologies that can identify false job postings and alert applicants not to submit applications for them. In order to do this, a machine learning strategy is used, which employs a number of classification algorithms to identify fake posts. In this instance, the classification tool distinguishes fake job advertisements from a bigger collection of positions and notifies the user. First, supervised learning algorithms are taken into consideration as a classification technique to handle the issue of identifying false job listings. The classifier assigns the input variables to the desired class after taking into account the training data. Let's quickly review the classifier that was utilized in this study to separate the fake classified advertising from the others. A single classifier-based prediction and an ensemble classifier-based forecast can be used to broadly categorize this classifier-based forecast.

A. Single Classifier based Prediction-

In order to forecast the unknowable test instances, classifiers are trained. When identifying fake job postings, the following classifiers are applied:

a) Naive Bayes Classifier-

A statistical classification method based on the Bayes Theorem is called naive Bayes. One of the easiest supervised learning methods is this one. The quick, accurate, and dependable approach is the naive Bayes classifier. On large datasets, naive Bayes classifiers perform quickly and accurately. The Naive Bayes classifier makes the assumption that the impact of a specific feature on a class is unrelated to the impact of other characteristics. For instance, a loan applicant's suitability depends on factors including their income, history of loans and transactions, age, and geography. These traits are nevertheless taken into account separately even though they are interconnected. This assumption is regarded as naïve since it makes calculation easier. The term "class conditional independence" refers to this assumption.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.

- P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.

- P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.

- P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

b) SGD Classfier-

A straightforward yet very effective method for fitting linear classifiers and regressors under convex loss functions, such as (linear) Support Vector Machines and Logistic Regression, is stochastic gradient descent (SGD). SGD has been present in the machine learning field for a while, but in the context of large-scale learning, it has just lately attracted a lot of interest.

Large-scale and sparse machine learning issues that arise often in text categorization and natural language processing have been effectively tackled with SGD. The classifiers in this module are easily scalable to situations with more than 105 training instances and more than 10^5 features since the data is sparse.

The class SGD Classifier offers a straightforward stochastic gradient descent learning procedure that supports various classification loss functions and penalties. The decision boundary of an SGD Classifier that was trained using the hinge loss and is comparable to a linear SVM is shown below.

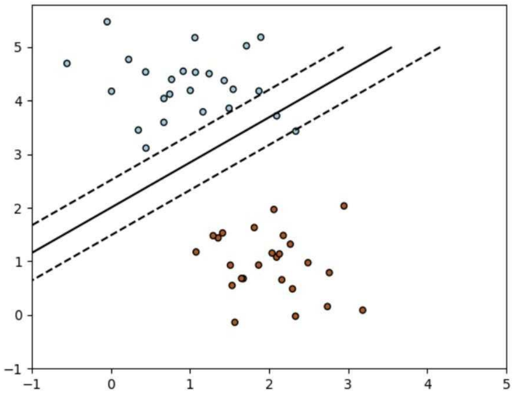

Figure 1: SGD Classifier Working

c) Natural Language Processing-

The field of computer science known as "natural language processing" (NLP) is more particularly the field of "artificial intelligence" (AI) that is concerned with providing computers the capacity to understand written and spoken words in a manner similar to that of humans. NLP integrates statistical, machine learning, and deep learning models with computational linguistics—rule-based modelling of human language. With the use of these technologies, computers are now able to interpret human language in the form of text or audio data and fully "understand" what is being said or written, including the speaker's or writer's intentions and state of mind.

II. METHODS AND METHODOLOGY

The project's data includes characteristics that characterize job postings. These job listings are either labelled as real or fraudulent. Only a very small portion of this dataset consists of fake job postings. That is expected. We are not expecting seeing many fake job advertisements. This project is divided into five stages. The project's five stages are as follows:

1. Problem Definition (Project Overview, Project statement, and metrics)
2. Data Collection
3. Data cleaning, exploring and pre-processing
4. Modeling
5. Evaluating

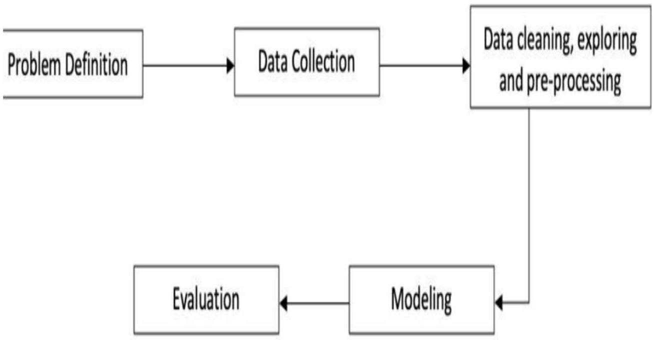

Figure 2: Methodology/ Workflow

A. Algorithms And Techniques

It is apparent from the preliminary study that the final modelling will make use of both text and numerical data. A final dataset is chosen before data modelling. For the final analysis of this research, a dataset with the following features will be used:

1. telecommuting
2. fraudulent
3. ratio: Based on region, the ratio of fraudulent to real employment opportunities
4. text: combination of title, location, company profile, description, benefits, necessary experience, education, industry, and function
5. character_count: Words in the textual data, in thousands a histogram of word counts

Textual data must first undergo further pre-processing before being used in any data modelling.

The algorithms and techniques used in project are:

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

A final model is selected after comparing the accuracy and F1-scores of Naive Bayes and SGD Classifier. The baseline model is naive bayes, and it is employed because encoding such probabilities is very helpful in computing the conditional probability of occurrence of two events depending on the probabilities of occurrence of each individual event.

Since SGD Classifier uses a straightforward stochastic gradient descent learning procedure and supports a variety of classification loss functions and penalties, it is used as a comparative model. When classified wrongly, this classifier will require severe penalties. The outcomes from these models are combined after being applied individually to the text and numerical data.

B. Methodology

The following steps are taken for text processing:

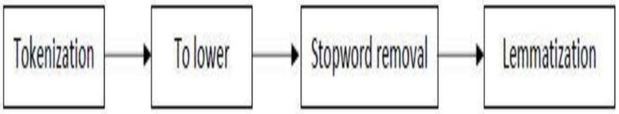

Figure 3: Pre-Processing of data set

- Tokenization: Smaller units of the textual information are divided. Words are used to divide the data in this instance.
- To Lower: A lowercase version of the split words is created.
- Stop word removal: Stop words are words that don't really add anything to sentences. For instance: the, a, an, he, have, etc. These words have been deleted.
- Lemmatization: The grouping of words with similar inflected forms is known as lemmatization.

C. Implementation

Following is a diagrammatic representation of how this project will be implemented. The dataset is broken up 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 towards classes with a majority, this is done. The test set is used to assess the trained model's performance. The final model for our analysis is chosen after comparing the accuracy and F1-score of the two models, naive bayes and SGD.

D. Refinement

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, various penalties are applied to the final model's evaluation. The differences in the results, nevertheless, were incredibly small.

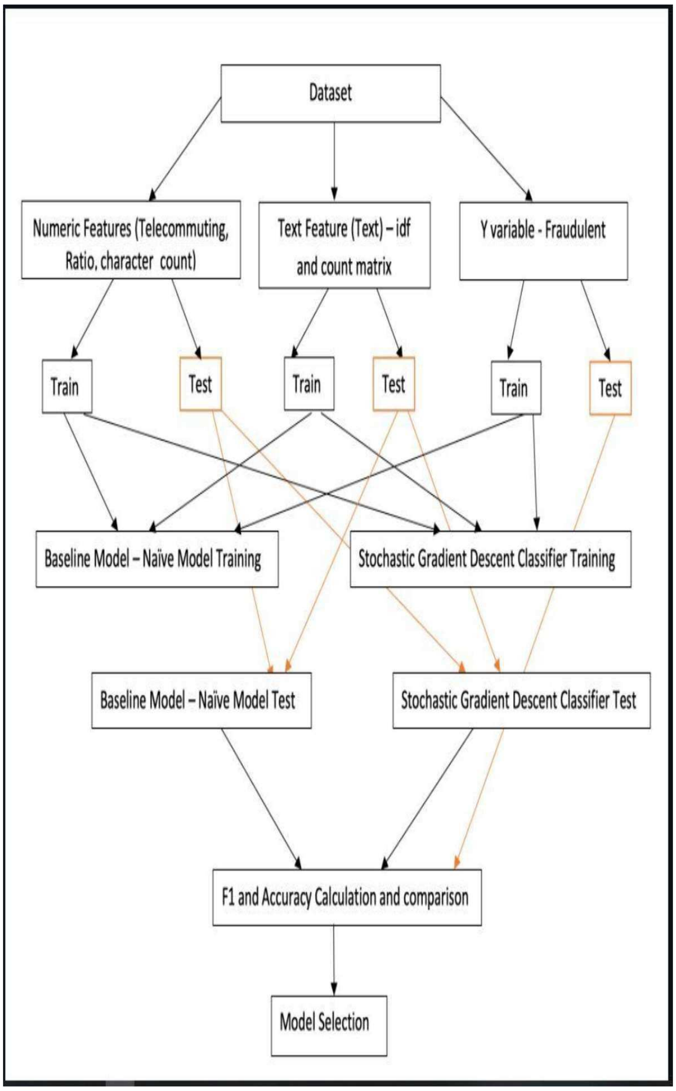
Dataset
Numeric Features (Telecommuting, Ratio, character count)
Text Feature (Text) – idf and count matrix
Y variable - Fraudulent
Train
Test
Train
Test
Train
Test
Baseline Model – Naive Model Training
Stochastic Gradient Descent Classifier Training
Baseline Model – Naive Model Test
Stochastic Gradient Descent Classifier Test
F1 and Accuracy Calculation and comparison
Model Selection

Figure4: Implementation of Algorithm

III. RESULT AND DISCUSSIONS

- This paper examined various approaches for predicting genuine and fraudulent employment using machine learning models.
- This survey research concludes that every method and algorithm presented has proven to be successful

in foretelling online recruitment fraud. The Fake Job Recruitment has proven to be extremely accurate.

- In the process of making predictions, a number of variables have been taken into account. These variables are thought to be the best ones for making precise predictions.

IV. CONCLUSION

Over a given dataset that includes both fake and real job posts, all of the aforementioned classifiers are trained and tested for spotting fake job postings. This study compares numerous algorithms employed by various methodologies for the accurate prediction of ORF, including machine learning approaches and algorithms like Naive Bayes and SGD Classifier. The methods and algorithms employed are seen to be effective in identifying the presence of fraudulent employment, which is demonstrated to be a crucial element in this process.

V. ACKNOWLEDGEMENT

We would like to thank **Prof. Rashmi K T, Asst. Professor, Dept. of CSE, SKIT and Dr. Shantharam Nayak, Dept. of CSE, SKIT** for their valuable suggestion, expert advice and moral support in the process of preparing this paper.

REFERENCES

[1] Mr. Gulshan P, Mr. Mukund T, Mr. Ajay A, Mr. Pankaj Kumar, Mrs. Aruna M G and Dr. Malatesh S H, "Fake Job Post Prediction Using Machine Learning Algorithms", International Journal of Innovative Research in Technology (IJIRT), IEEE, Volume 9 Issue 3, August 2022, ISSN: 2349-6002.

[2] C.S. Anita, P. Nagarajan, G. Aditya Sairam, P. Ganesh and G. Deepakkumar, "Fake Job Detection and Analysis Using Machine Learning and Deep LearningAlgorithms",IEEE,DOI:10.47059/revistageintec.v11i2.1701, Vol. 11 No. 2, April 2021, ISSN: 2237-0722.

[3] Karri Sai Suresh Reddy and Karri Lakshmana Reddy, "Fake Job Recruitment Detection", Journal of Emerging Technologies and Innovative Research (JETIR), IEEE, Volume 8, Issue 8, August 2021, ISSN: 2349-5162.

[4] Shawni Dutta and Prof. Samir Kumar Bandyopadhyay, "Fake Job Recruitment Detection Using Machine Learning Approach", International Journal of Engineering Trends and Technology (IJETT), IEEE, Volume 68 Issue 4, April 2020, ISSN: 2231-5381.

[5] Rakshitha M J, Impana B L, Pruthvini C K and Dr. Honnaraju B, "Fake Job Recruitment Detection Using Machine Learning", International Research Journal of Modernization in Engineering Technology and Science (IRJMETS), IEEE, Volume:04 Issue:07, July-2022, e-ISSN: 2582-5208.

[6] Priya Khandagale, Akshata Utekar, Anushka Dhonde and Prof. S. S. Karve, "Fake Job Detection Using Machine Learning", International Journal for Research in Applied Science & Engineering Technology (IJRASET), IEEE, Volume 10 Issue 04, April 2022, ISSN: 2321-9653.

[7] Ali Razaa, Saqib Ubaidb, Faizan Younasc and Farhan Akhtard, "Fake E Job Posting Prediction Based on Advance Machine Learning Approaches", International Journal of Research Publication and Reviews (IJRPR), IEEE, Vol 3, no 2, pp 689-695, February 2022, ISSN: 2582-7421.

[8] Devsmit Ranparia, Shaily Kumari and Dr. Aashish Sahani, "Fake Job Posting Prediction Using Sequential Network", International Conference on Industrial and Information Systems (ICIIS), IEEE, DOI: 10.1109/ICIIS51140.2020.9342738, May 2021, ISBN No: 978-1-7281-8524-8/20.

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

[10] S. Vidros, C. Kolias , G. Kambourakis and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset", IEEE, Future Internet, 2017, Volume 9 Issue 6, DOI: 10.3390/fi9010006.

[11] B. Alghamdi and F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", J. Inf. Secur., IEEE, vol. 10, no. 03, pp. 155–176, 2019, DOI: 10.4236/jis.2019.103009, ISSN: 2153-1242.

[12] I. Rish, "An empirical study of the Naive Bayes classifier", IEEE, January 2001, pp. 41–46, 2014.

[13] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, "Spam review detection techniques: A systematic literature review", Appl. Sci., IEEE, vol. 9,no. 5, pp. 1–26, 2019, DOI: 10.3390/app9050987.

[14] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media", ACM SIGKDD Explor. Newsl., IEEE, vol. 19, no. 1, pp. 22–36, 2017, DOI: 10.1145/3137597.3137600.

[15] E. G. Dada, J. S. "Machine learning for email spam filtering: review, approaches and open research problems",IEEE,May2019,DOI:10.1016/j.heliyon. 2019.e01802, ISSN: 2405-8440.

[16] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf and Muhammad Ovais Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods",IEEE,October2020,DOI:10.1155/2020/8885861.

[17] Arvin Hansrajh, Timothy T. Adeliyi and Jeanette Wing, "Detection of Online Fake News Using Blending Ensemble Learning", IEEE, July 2021, DOI: 10.1155/2021/3434458.

[18] T.Nandini, S.Gnana Chandrika, P.Mounika and V.Sandeeep Kumar, "Developing A Model to Detect Fraudulent Job Postings: Fake vs. Real", International Journal for Research Developments in Science and Technology (IJRDST), IEEE, March 2023, Volume 07, Issue 02, ISSN: 2581 – 4575.

[19] C. Jagadeesh, Dr. Pravin R Kshirsagar, G. Sarayu, G.Gouthami and B.Manasa, "Artificial Intelligence based Fake Job Recruitment Detection Using Machine Learning Approach", Journal of Energy Sciences (JES), IEEE, June 2021, Vol 12, Issue 06, ISSN: 0377-9254.

[20] Prof. Ajit Patil, Harshita Kaushik, Rajeshri Kalwale and Pranita Thorawase, "Fake News Detection Using Machine Learning Algorithms", GIS Science Journal, IEEE, Volume 8, Issue 12, ISSN: 1869-9391.