Fake job recruitment detection Using Machine learning

Sumesh KJ
Assistant Professor
Project Guide
Amrita Viswa Vidyapeetham
Amritapuri, India

Vijay Ram Challagundla
Computer Science and Engineering
Amrita Viswa Vidyapeetham
Amritapuri, India

Achyuta Vindhya
Computer Science and Engineering
Amrita Viswa Vidyapeetham
Amritapuri, India

Charmila T D
Computer Science and Engineering
Amrita Viswa Vidyapeetham
Amritapuri, India

I. Abstract

In order to avoid fraudulent job postings on the internet, the paper proposes an automated tool based on machine learning-based classification techniques. Various classifiers are used to detect fraudulent web posts, and the results of those classifiers are compared to determine the best employment scam detection model. It aids in the detection of fake job postings among a large number of postings. For the detection of fraudulent job postings, two major types of classifiers are considered: single classifiers and ensemble classifiers. However, experimental results show that ensemble classifiers outperform single classifiers in detecting scams.

II. Introduction

Employment scam

Scams involving employment are on the rise. According to CNBC, the number of job frauds more than doubled in 2018 over 2017. Unemployment is at an all-time high due to the current market condition. Economic stress and the impact of the coronavirus have resulted in a considerable reduction in work availability and job loss for many people. Scammers would love to take advantage of a situation like this. Many individuals are falling prey to these con artists who are preying on people's desperation as a result of an extraordinary event. The majority of fraudsters do this to obtain personal information from the person they are attempting to defraud. Addresses, bank account numbers, and social security numbers are examples of personal information. I am a university student who has received multiple scam emails of this nature.. Or they require investment from the job seeker with the promise of a job.

III. PROBLEM DEFINITION

There are a lot of job advertisements on the internet even on reputed job advertising sites which never seen fake but after selection the recruiters start asking money and bank details of candidates fall into their trap and lose lot of money.

The World Wide Web contains data in diverse formats such as documents, videos, audio, etc...The response that an article gets can be differentiated at a theoretical level to classify the article as real or fake.So, it is better to identify whether a job advertisement posted on the site is real or fake. Identifying it manually is very difficult and almost impossible. We can apply machine learning to train a model for fake job classification. It can be trained on the previous real and fake job advertisements and it can identify a fake job accurately

IV. PROBLEM STATEMENT

This project aims to create a classifier that will have the capability to identify fake and reals jobs. The final result will be evaluated based on two different models. Since the data provided has both numeric and text features on one model will be used on the text data and the other numeric data. The final output will be combination of two. The final model will take in any relevant job posting data and produce a final result determining whether the job is real or not.

Problem solution

The goal of this project is to offer a possible answer to this issue. To get the best results, the textual data is pre-processed, and relevant numerical fields are chosen as well. Multiple models' outputs are blended to give the best possible results.

V. DOMAIN BACKGROUND

Employment scams are on the rise. According to CNBC ,the number of employment scams doubled in 2018 as compared to 2017. Economic stress and the impact of the corona virus have significantly reduced job availability and the loss of jobs form any individuals.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 address, bank account details ,social security number etc.

VI. RELATED WORK

Spam Detection

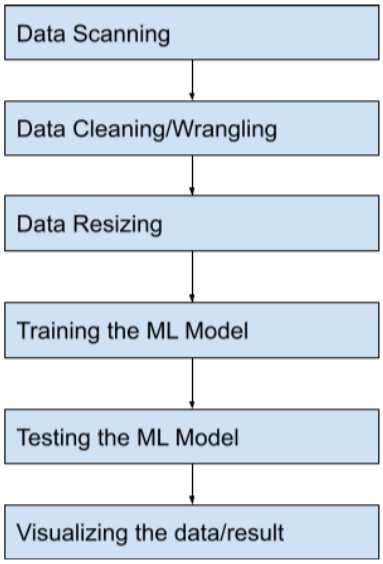
At least in the sphere of spam detection, the problem of recognising non-genuine information sources by content-based analysis is regarded to be solvable. To evaluate whether text is spam or real, statistical machine learning approaches are used. • These techniques include text pre-processing, feature extraction (i.e. a bag of words), and feature selection based on which features lead to the greatest performance on a test data-set. These features can be classified using Naive Bayes, Support Vector Machines, TF-IDF, or K-nearest Neighbor classifiers after they've been obtained. All of these classifiers are supervised machine learning models, which means that they need some labeled data to train the function.

Benchmark Data-set

It shows past work on fake news detection that is more directly connected to our goal of making a classification using simply text. The authors not only generate a new benchmark data-set of remarks, but also show that using meta-data (i.e. Speaker, party, etc.) to supplement the information provided by the text can significantly enhance fine-grained false news detection

VII. PROPOSED METHODOLOGY

As our data set was imbalanced so it became a biased data set, so for overcoming that problem we used under sampling for majority outcomes.Since the data provided has both numeric and text features models will be used on the text data and numeric data. The final output will be a combination of the two. The final model will take in any relevant job posting data and produce a final result determining whether the job is real or not.



System Modules:

Collecting Data-set: In the step data set which are considered as input has many fields in that except the last one all are called as features and the last cell is called as labels which are into 0,1 format.

Data set processing: Given the data-set is not in required format so we need to remove unwanted fields from the data-set and use them as features and process data for scalar format. And the data processing followed by checking null values for required for location, merging all relevant cols to one ‘text‘, and forming types of words are frequent in fraudulent and not fraudulent, Create our list of punctuation marks and list of stop-words, Load English tokenizer, tagger, parser, NER and word vectors, Apply to the DF series, remove weird numbers, All lowercase, Adding State last to keep it Caps,creating our bag of words

Pre-processing: We convert data into scalar format and then create new features which are passed to the algorithm and features are saved in x and labels in y.

Algorithm fit: In this step train features and labels are fit to the algorithm and the model is saved to a system which is used for prediction.

Prediction: In this step details are fed as input in the form of CSV of various profiles and prediction is performed.

Steps to Training the ML Model

Step 1 - Fetch the Data We fetch the data from CSV file and we store this in a data frame ’df’.

Step 2 - Split the Dataset We split the dataset into training dataset and test dataset. We use 75percent of our data to train and the rest 25percent to test. To do this, we will create a split parameter which will divide the data frame in a 75-25 ratio

Step 3 - Import the Libraries We start by importing the necessary libraries required t import the algorithm in Python. We import the numpy libraries for scientific calculation

VIII. TESTING THE ML MODELS

A) Naive Bayes: It is primarily used in text classification with a large training dataset. The Naive Bayes Classifier is a simple and effective Classification algorithm that aids in the development of fast machine learning models capable of making quick predictions.

B) Stochastic Gradient Descent: Stochastic gradient descent is a machine learning optimization algorithm that is commonly used to find the model parameters that correspond to the best fit between predicted and actual outputs. It’s an imprecise but effective technique. In machine learning applications, stochastic gradient descent is widely used.

C) Decision Tree Classifier: The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem, where the leaf node corresponds to a class label and attributes are represented on the tree’s internal node.

D) Random Forest: It is based on the concept of ensemble learning, which is the process of combining multiple classifiers to solve a complex problem and improve the model’s performance. Random forest(Oversampling and Under-sampling):One approach to addressing the problem of class imbalance is to randomly re-sample the training dataset. The two main approaches to randomly re-sampling an imbalanced dataset are to delete examples from the majority class, called under-sampling, and to duplicate examples from the minority class, called oversampling.

E) XGBoosting Classifier: The algorithm’s optimization techniques improve performance and thereby provide speed using the least amount of resources.

F) KNNeighbors Classifier: K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

G) GradientBoostingClassifier: It returns a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is used as the weak learner, the resulting algorithm is known as gradient-boosted trees; it typically outperforms random forest. A gradient-boosted trees model is constructed in the same stage-wise manner as other boosting methods, but it generalizes the other methods by allowing optimization of any differentiable loss function.

H) Support Vector Machine: The SVM algorithm’s goal is to find the best line or decision boundary for categoriz-

ing n-dimensional space so that we can easily place new data points in the correct category in the future. A hyperplane is the best decision boundary.SVM selects the extreme points/vectors that aid in the creation of the hyperplane. These extreme cases are referred to as support vectors, and the algorithm is known as the Support Vector Machine.

I) Logistic Regression: A categorical dependent variable's output is predicted using logistic regression. As a result, the result must be a categorical or discrete value. It can be Yes or No, 0 or 1, True or False, and so on, but instead of giving the exact values as 0 and 1, it gives the probabilistic values that fall between 0 and 1.

IX. EXPERIMENTAL RESULTS AND ANALYSIS

Module Description The Models will be evaluated based on two metrics:

1) Accuracy: This metric is defined by this formula

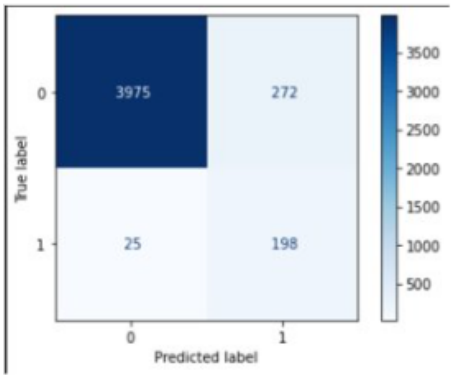
$$Accuracy = \frac{True\ Positve + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}$$

2) Recall Score: The Recall is the ratio tp/(tp + fn) where tp is number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

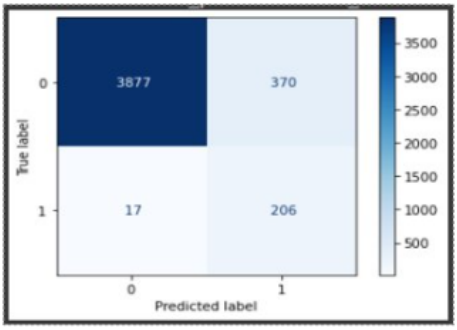
RESULTS

Tested every ML Algorithm with the processed dataset and find the Accuracy and Recall Score. And as for visualisation purpose confusion matrix is also drawn for every algorithm.

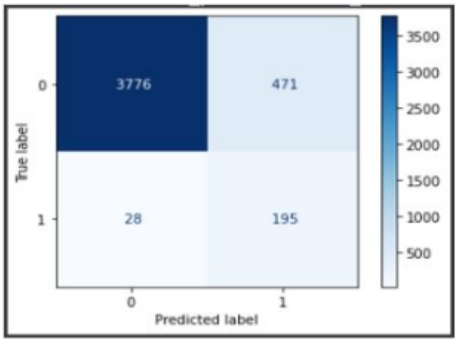
1. Naive Bayes



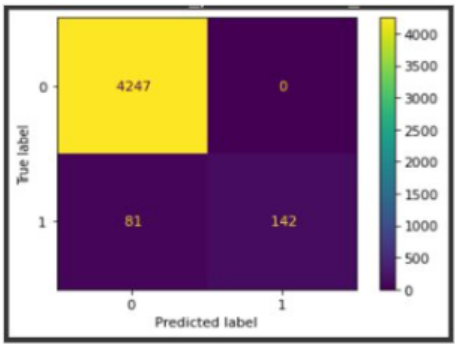
2. Stochastic Gradient Descent



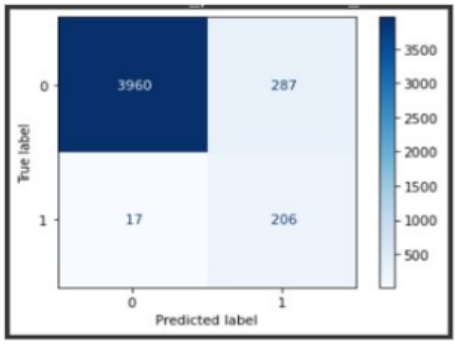
3. Decision Tree Classifier



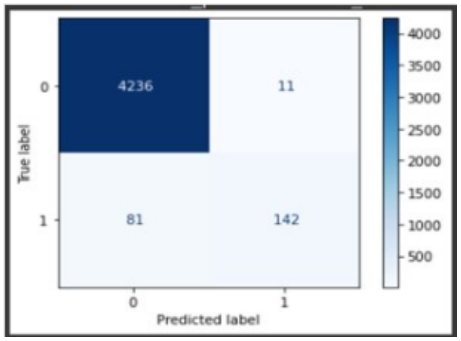
4. Random Forest Classifier



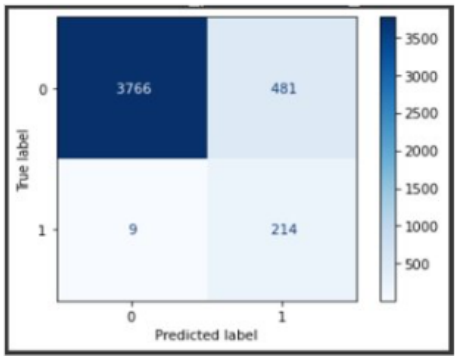
5. Random Forest Classifier Under Sampling



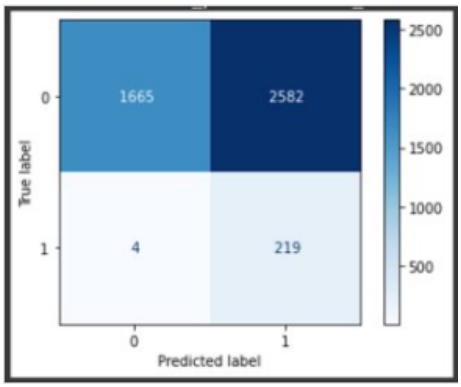
6. Random Forest Classifier Over sampling



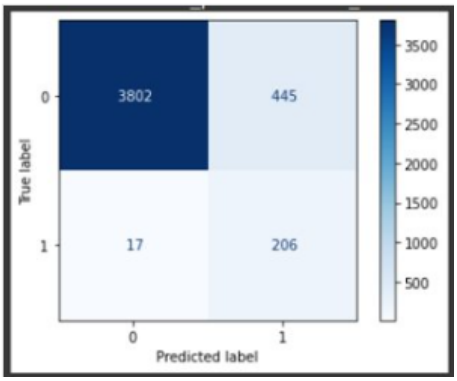
7. XGBoosting Classifier



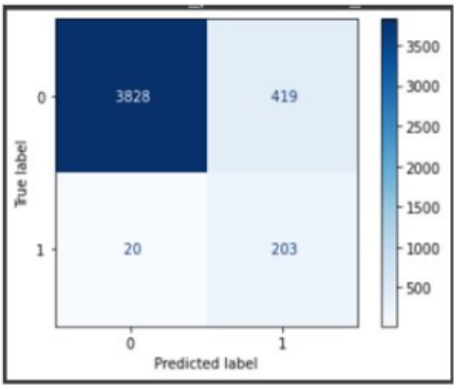
8. KNNeighbors Classifier



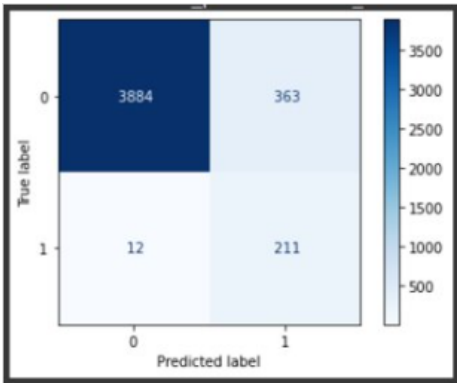
9. Gradient Boosting Classifier



10. Support Vector Machine



11. Logistic Regression

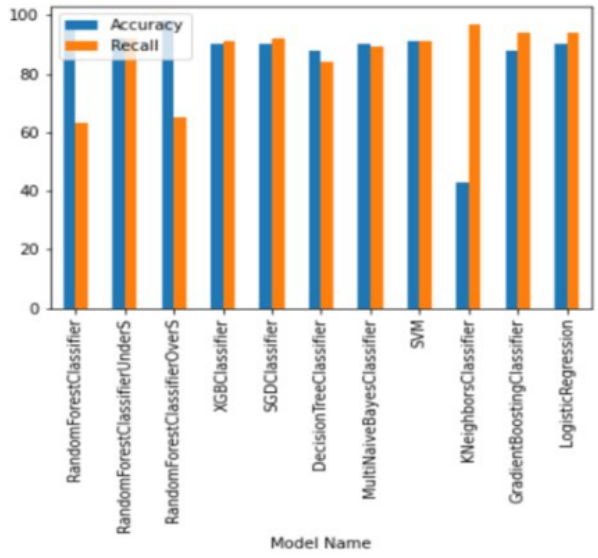


Comparative Visualisation of all algorithms:

As Table:

S.No	Algorithm	Accuracy	Recall Score
1	Multi Naive Bayes Classifier	0.90648	0.89686
2	Decision Tree Classifier	0.88389	0.84304
3	SGD Classifier	0.90425	0.92376
4	Random Forest Classifier	0.98187	0.63677
5	Random Forest Classifier Under-Sampling	0.92930	0.92376
6	Random Forest Classifier Over-Sampling	0.98031	0.65022
7	XGB Classifier	0.90357	0.91479
8	K Nearest Neighbors Classifier	0.43467	0.97309
9	Gradient Boosting Classifier	0.88926	0.94170
10	Support Vector Machine Classifier	0.91073	0.91928
11	Logistic Regression Classifier	0.90156	0.94170

As Bar Graph:



Evaluation Metrics

Since this data-set is meant to identify fraudulent classes, Recall is the appropriate metric. This is because the data-set will be heavily imbalanced. The formula for Recall or Sensitivity or True This analysis requires the minimization of false negatives. Recall is the model's ability to identify all points of interest in the data-set. This project can be later extended to find an appropriate balance between recall and precision by using F1-score. The F1-score will help identify the best model to minimize both false positives and false negatives. However, that is beyond the scope of the current analysis.

Project Design

Step1: Exploring the dataset and identifying the relevant columns. An appropriate method to deal with missing data also needs to be identified. Also, the relationship between the target variable and the other variables in the dataset will be explored.

Step2: The second step is to balance the two classes by using methods such as over-sampling and under-sampling. This is imperative because Machine learning algorithms tend to favor the class with the largest proportion of observations (known as majority class), which may lead to misleading accuracies. This is particularly problematic when we are interested in the correct classification of a "rare" category (also known as minority class). However, we find high accuracies, which are the product of the correct classification of the majority class (i.e., are the reflection of the underlying class distribution).

Step3: Comparing various models and selecting the one that performs the best based on the selected Metric.

Step4: Creating a web app for this analysis. The web app will be able to produce results on the authenticity of a job posting.

X. CONCLUSION

The dataset that is used in this project is very unbalanced. Most jobs are real, and few are fraudulent. Due to this, real jobs are being identified quite well. Certain techniques like under sampling, over sampling, SMOTE used to generate synthetic minority class samples. So the balanced dataset has performed to generate better results.

We perform all algorithms for our data set. We find accuracy and recall for them

We got similar values for all algorithms ,and we compared all the algorithms and choose the best one.

XI. FURTHER WORK

In Further we would like to improve our work to make this this project look some attractive by connecting this data to one WEB APPLICATION.

We would like to make this as user friendly, and we got to conclude that making a chrome extension would be better.

XII. ACKNOWLEDGMENT

We are grateful to our University's valued Chancellor, Sri Mata Amritanandamayi Devi and academic members for their assistance in preparing the research work. We would like to express gratitude to our project guide for her insights and inspired leadership. We would also like to thank our project coordinator for her guidance and consistent support during the project.

XIII. REFERENCES

REFERENCES

[1] Fake News Detection Using Machine Learning Vijaya Balpande Kasturi Baswe Kajol Somaiya Achal Dhande Prajwal Mire https://doi.org/10.32628/CSEIT12173115

[0] [2] Fake News Detection Using Machine Learning Ensemble Methods. Iftikhar Ahmad ,Muhammad Yousaf, Suhail Yousaf , and Muhammad Ovais Ahmad , https://doi.org/10.1155/2020/8885861

[3] Lftikhar Ahmad,1 Muhammad Yousaf,1 Suhail Yousaf,1 and Muhammad Ovais Ahmad Volume 2020 —Article ID 8885861 — https://doi.org/10.1155/2020/8885861

[4] N. Smitha and R. Bharath, "Performance Comparison of Machine Learning Classifiers for Fake News Detection," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 696-700, doi: 10.1109/ICIRCA48905.2020.9183072.

[5] An Intelligent Model for Online Recruitment Fraud Detection Bandar Alghamdi, Fahad AlharbyNaif Arab University (NAUSS), Riyadh, KSA.DOI: 10.4236/jis.2019.103009

[6] S. I. Manzoor, J. Singla and Nikita, "Fake News Detection Using Machine Learning approaches: A systematic Review," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 230-234, doi: 10.1109/ICOEI.2019.8862770.

[7] Fake News Detection Using Machine Learning Approaches To cite this article: Z Khanam et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1099 012040

[8] Fake News Detection using Machine Learning Ensemble Methods Volume 2020 —Article ID 8885861 — https://doi.org/10.1155/2020/8885861

[9] Samir Bandyopadhya April 2020DOI:10.14445/22315381/IJETT-V68I4P209S

[10] Fake News Detection DOI: 10.1109/SCEECS.2018.8546944

[11] Fake News Detection Using Machine Learning Approaches To cite this article: Z Khanam etal 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1099 012040

[12] Automating the Detection of Cyberstalking

[13] Fake News Detection DOI: 10.1109/SCEECS.2018.8546944

[14] SPAM Attacks Detection in the Remote Triggered WSN Experiments

[15] K. R. Vidya Kumari C. R. Kavitha (2018). Spam Detection Using Machine Learning in R (International Conference on Computer Networks and Communication Technologies pp 55–64)

[16] N. Kumar, S. Sonowal and Nishant, "Email Spam Detection Using Machine Learning Algorithms," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 108-113, doi: 10.1109/ICIRCA48905.2020.9183098.