www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 16-21

e-ISSN : 2583-1062

Impact Factor : 2.625

REAL AND FAKE JOB POSTING USING MACHINE LEARNING

A. Mukil Chockalingam[1], Dr. P. Dinesh kumar[2], Dr. S. Geetha[3]

[1]Final Year M. Tech-CFIS, Department of Computer Science and Engineering, Dr. M.G.R Educational and Research Institute, Chennai 600 095, Tamilnadu, India.

[2]Professor, Department of Computer Science and Engineering, Dr. M.G.R Educational and Research Institute, Chennai 600 095, Tamilnadu, India.

[3]Head of the Department, Department of Computer Science and Engineering, Dr. M.G.R Educational and Research Institute, Chennai 600 095, Tamilnadu, India.

ABSTRACT

Everything is being online now it allows people to reduce their manual efforts all of the job postings are posted online now so it gives the company a wide range of area to gather the talented candidates and for the people who are searching they can also know the information most companies can directly post the job. All job postings which are posted are not true there are fraudulent job postings also. So, we try to classify the fraudulent posting from real. The aim is to predict machine learning based techniques for real or fake job prediction results in best accuracy. The analysis of dataset is done by supervised machine learning technique (SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given data set.

1. INTRODUCTION

The existing data-driven approaches typically capture credibility-indicative representations from relevant articles for fake news detection, such as skeptical and conflicting opinions. However, these methods still have several drawbacks: Due to the difficulty of collecting fake news, the capacity of the existing datasets is relatively small; and there is considerable unverified news that lacks conflicting voices in relevant articles, which makes it difficult for the existing methods to identify their credibility. Especially, the differences between true and fake news are not limited to whether there are conflict features in their relevant articles, but also include more extensive hidden differences at the linguistic level, such as the perspectives of emotional expression (like extreme emotion in fake news), writing style (like the shocking title in clickbait), etc., It difficult to fully capture these differences. This method is built a machine learning model to classify the real or fake job posting to overcome this method to implement machine learning approach. The dataset is first preprocessed and the columns are analyzed to see the dependent and independent variable and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

2. LITERATURE SURVEY

This process has to avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed in the paper. Different classifiers are used for checking fraudulent post in the web and the results of those classifiers are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts.

Two major types of classifiers, such as single classifier and ensemble classifiers are considered for fraudulent job posts detection. However, experimental results indicate that ensemble classifiers are the best classification to detect scams over the single classifiers. In recent years, due to advancement in modern technology and social communication, advertising new job posts has become very common issue in the present world. So, fake job posting prediction task is going to be a great concern for all. Like many other classification tasks, fake job posing prediction leaves a lot of challenges to face. This paper proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier, multilayer perception and deep neural network to predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples.

Deep neural network as a classifier, performs great for this classification task. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post. The process of searching jobs is one of the most problematic issue freshers faces, this process is used by various scamsters to lure freshers into scams and profit from the students. In order to avoid this, this paper proposes a system with deep learning and flask for front-end, that can identify fake jobs. The deep learning

algorithm extracts specific features from the website's article and based on those features predicts if the job is genuine or not.

The proposed system makes use of a deep learning-based system and a web page to help non-technical users to analyze these fake scams and secure their jobs. While browsing for jobs online we saw that many scamsters demanded money for booking slots to interviews that did not exist and also extort money from students with promise of giving them jobs in return, this served as motivation for this proposal. The objectives that are to be considered are: Prediction of real or fake job. And a front-end page to allow non-technical user to use the model The proposed system is basically an ANN classification model based on Multinomial Naive Bayes algorithm to determine fake job posting or real one. The model is trained to be as efficient as possible by making the dataset to be a part of double-blind study and also considering the various formats of posting jobs in professional websites and other sites too. This therefore makes searching of jobs much more efficient and also allows the users to be worry free when they search for jobs online.

During the pandemic, there is strong rise in the number of online jobs posted on various job portals. So, fake job posting prediction task is going to be big problems for all. Thus, these fake jobs can be precisely detected and classified from a pool of job posts of both fake and real jobs by using advanced deep learning as well as machine learning classification algorithms. This paper proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier, multilayer perception and deep neural network to predict a job post if it is real or fraudulent. We have experimented on EMSCAD which containing 18000 employee samples. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post. Index Terms - Random Forest, KNN, Naive Bayes, Real and Fake, support vector machine, deep learning, and classification.

Some of the biggest sources of spreading fake jobs or rumours are social media websites such as Google Plus, Facebook, Twitters, and other social media outlet [1]. Even though the problem of fake jobs is not a new issue, detecting fake jobs is believed to be a complex task given that humans tend to believe misleading information and the lack of control of the spread of fake content [2]. Fake jobs have been getting more attention in the last couple of years, especially since the US election in 2016. It is tough for humans to detect fake jobs. It can be argued that the only way for a person to manually identify fake jobs is to have a vast knowledge of the covered topic. There are many jobs' adverts on the internet, even on reputable job posting sites, that never appear to be false. However, following the selection, the so-called recruiters begin to seek money and bank information. Many candidates fall into their traps and lose a lot of money as well as their existing job. As a result, it is preferable to determine whether a job posting submitted on the site is genuine or fraudulent. Manually identifying it is extremely difficult, if not impossible! An automated online tool (website) based on machine learning-based categorization and algorithms are presented to eliminate fraudulent job postings on the internet. It aids in the detection of bogus job postings among the vast number of postings on the internet.

3. EXISTING SYSTEM

In today's world, fake jobs on social media are a universal trend and has severe consequences. There has been a wide variety of countermeasures developed to offset the effect and propagation of Fake jobs. The most common are linguistic-based techniques, which mostly use deep learning (DL) and natural language processing (NLP). Even government-sponsored organizations spread fake jobs as a cyberwar strategy. The proposed methods for fake jobs detection to learn various representations and discover explainable comments and sentences. Experimental results on real-world datasets show that the proposed method is effective and explainable.

3.1 Drawbacks

- Accuracy is low.
- Machine Learning is not implemented.
- Deployment is not implemented.

4. PROPOSED SYSTEM

The proposed model is to build a machine learning model that is capable of classifying whether the job is fake or not. The fake jobs are considered to be widespread and controlling them is very difficult as the world is developing toward digital everyone now has access to internet and they can post whatever they want. So, there is a greater chance for the people to get misguided. The machine learning is generally built to tackle this type of complicated task like it takes more amount of time to analyze these types of data manually. The machine learning can be used to classify whether

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 16-21

e-ISSN : 2583-1062

Impact Factor : 2.625

the job is fake or not by using the previous data and make them to understand the pattern and improve the accuracy of the model by adjusting parameters and use that model as the classification model. Different algorithms can be compared and the best model can be used for classification purpose.

4.1 Advantages

- Accuracy will be improved.
- Machine Learning algorithms are used.
- Project will be deployed.

4.2 Design Architecture

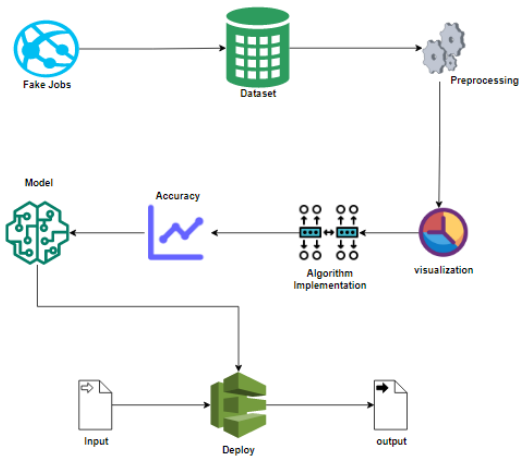


Figure 1. Design Architecture

It has a set of 2400 datasets are there that are all a job description. These data used to preprocess function. This process used to find the any error or missing values to find it separately while using this model then next process is used to data visualization. This process from preprocessing to separate the fake and real description from the first phase model it separate mistakes from visualized image. Important thing is using algorithm implementation then, am using knn, logistic regression, support vector machine. These algorithms used by finding the accuracy it gives us an accurate model to find the version of deploy it is used to implement the web pages used by flask it gives us a description using giving a solid output.

4.2 List of Modules

- Data Pre-processing
- Data Analysis of Visualization
- KNN (K-Nearest Neighbor)
- Logistic regression
- Support vector machine
- Deployment

Data Pre-Processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters.

Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. A number of different data cleaning tasks using Python's Pandas library and specifically, it focusses on probably the biggest data cleaning task, missing values

and it able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modeling. Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.

```
data = pd.read_csv('review_dataset.csv')
data.drop(columns='Unnamed: 0', axis=1, inplace=True)
```

```
data
```

```
data.head()
```

Data Visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end. Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.

KNN (K-Nearest neighbour)

- K-Nearest neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Logistic Regression

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.
- In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).
- Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 16-21

e-ISSN : 2583-1062

Impact Factor : 2.625

Before diving into the implementation of logistic regression, we must be aware of the following assumptions about the same −

- In case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1.
- There should not be any multi-collinearity in the model, which means the independent variables must be independent of each other.
- We must include meaningful variables in our model.
- We should choose a large sample size for logistic regression.

K-Nearest Neighbour

```
: #import library packages
  import pandas as pd
  import matplotlib.pyplot as plt
  import seaborn as sns
  import numpy as np

: import warnings
  warnings.filterwarnings("ignore")

: df = pd.read_csv('rf.csv', usecols = ['description','fraudulent'])

: df
```

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.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper plane.
- SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Logistic regression

```
#import library packages
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

import warnings
warnings.filterwarnings("ignore")

df = pd.read_csv('rf.csv', usecols = ['description','fraudulent'])

df = df.dropna()
df
```

DEPLOYMENT Flask (Web Frame Work)

Flask was designed to be **easy to use and extend**. The idea behind Flask is to build a solid foundation for web applications of different complexity. From then on you are free to **plug in any extensions** you think you need. Also, you are free to build your own modules. Flask is great for all kinds of projects. It's especially good for prototyping. Flask depends on two external libraries: the Jinja2 template engine and the Werkzeug WSGI toolkit. Still the question remains why use Flask as your web application framework if we have immensely powerful Django, Pyramid, and don't forget web mega-framework Turbo-gears? Those are supreme Python web frameworks BUT out-of-the-box Flask is pretty impressive too with its:

- Built-In Development server and Fast debugger
- integrated support for unit testing
- RESTful request dispatching
- Uses Jinja2 Templating
- support for secure cookies
- Unicode based
- Extensive Documentation
- Google App Engine Compatibility

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

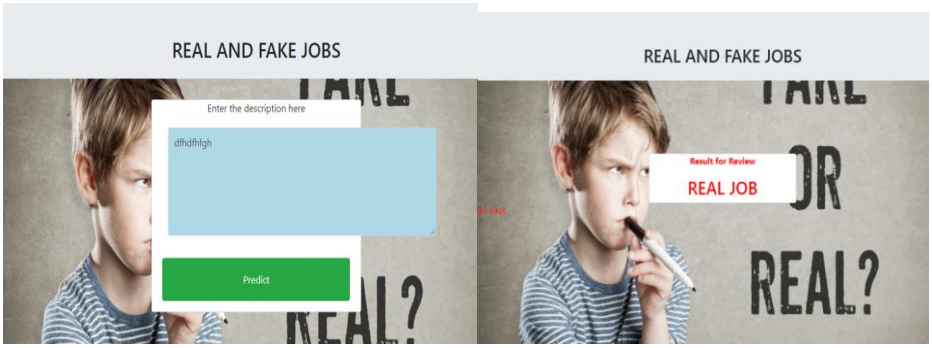
Vol. 03, Issue 04, April 2023, pp : 16-21

e-ISSN : 2583-1062

Impact Factor : 2.625

- Extensions available to enhance features desired

Output Screenshot



5. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set of higher accuracy score algorithm will be find out. The founded one is used in the application which can help to find the Real and fake jobs.

6. REFERENCES

[1] Priya Khandagale, Akshata Utekar, Anushka Dhonde, Prof. S. S. Karve. DOI Link: https://doi.org/10.22214/ijraset.2022.41641.

[2] Enhanced RSA Algorithm using Fake Modulus and Fake Public Key Exponent Raghunandhan K R, Ganesh Aithal, Surendra Shetty, Rakshith N.2018.

[3] Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset. Sokratis Vidros, Constantinos Kolias, Georgios Kambourakis and Leman Akoglu, 2017.

[4] Machine Learning and Job Posting Classification: A Comparative Study Ibrahim M. Nasser and Amjad H. Alzaanin, 2020.

[5] Fake Job Recruitment Detection Using Machine Learning Approach, Samir Bandyopadhyay, Shawni Dutta,2020.

[6] Predicting of Job Failure in Compute Cloud Based on Online Extreme Learning Machine: A Comparative Study Chunhong Liu, Jingjing Han, Yanlei Shang, Chuanchang Liu, Bo Cheng, and Junliang Chen, 2017.

[7] Enhanced RSA Algorithm using Fake Modulus and Fake Public Key Exponent, Raghunandhan K R, Ganesh Aithal, Surendra Shetty, Rakshith N 2018.

[8] Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset.SokratisVidros, Constantinos Kolias, Georgios Kambouraki and Leman Akoglu 2017.