A Survey on Fake Job Recruitment Detection Using Machine Learning Algorithms

Jotham N. Wanniang [#1], **Varun Arora** [#2], **Ankur Dey** [#3],
[#1,2,3] B.Tech Students,

Department of Computer Science & Engineering,
Sikkim Manipal Institute of Technology, Majitar, East Sikkim-737136, India.

jothamwanniang@gmail.com [1], ervarunarora@yahoo.com [2], deyankur2000@gmail.com [3]

Abstract

The project proposes an application that uses machine learning-based categorization approaches to avoid bogus job postings on the internet. Different classifiers are employed to check for fraudulent posts on the web, and their findings are compared to determine the best employment scam detection model. It aids in the detection of bogus job postings among a large number of postings. The article examines the various strategies used to tackle the bogus job posting on the internet. A survey for each and every approach selected for locating this fake job positing communications from the internet, and finally, we attempt to identify the problem gap between each and every works that has already been published on this topic. We collected several well-known papers related to our topic from 2004 to 2022, taking into account the similarities between each and every approach for effective detection, and finally attempting to determine which mechanism is best in providing fake job detection. For the identification of fake job postings, two basic types of classifiers are considered: single classifiers and ensemble classifiers. However, experimental results show that ensemble classifiers outperform single classifiers in detecting scams.

Keywords: Fake Job Posting, Scam Detection Model, Machine Learning Categorization, Ensemble Classifier.

1. INTRODUCTION

Nowadays, most recruitment is conducted online through websites such as naukri.com and monster.com. Organizations post job advertisements with the requisite competencies on these portals. These websites allow job searchers and applicants to post their resumes and skill details. Companies can now scan the profiles of potential applicants and contact them, and candidates can apply to job profiles in which they are interested. Companies contact the shortlisted candidates for additional processing and recruit suitable candidates after the initial screening. Online recruitment is advantageous to both candidates and employers. In December 2016, Naukri.com had a database of approximately 49.5 million registered users, with 11000 resumes being updated every day [1]. This demonstrates the influence that these online job platforms have on users. Both recruiters and candidates benefit from online recruitment. However, in recent years, scammers have entered the internet

recruitment market, giving rise to a new sort of fraud known as internet Recruitment Fraud (ORF). ORF spammers make enticing employment offers to prospects while stealing their money and personal information. ORF is not only harmful to users, but it is also problematic for businesses [2]. As a result, it harms firms' reputations and creates an unfavourable impression in the minds of job searchers about the particular company. It shows some of the news excerpts that are highlighting the damage caused by the ORF situation [3].

This is a clip from the news media indicating that job seekers had lost approximately 2 crore rupees as a result of ORF. It shows an ORF warning sign provided by one of the MNCs (ABB business). News clips illustrating the ORF issue. The amount of loss occurred as a result of ORF, with one of the MNCs issuing a warning against ORF. ORF detection is a critical subject that has received little attention from the scientific community and is now a largely untapped field. Detecting fraudulent job [4] offers among real job offers is a technically difficult task. The biggest obstacle is the problem of class inequality, as the number of fraudulent jobs is relatively low in comparison to real jobs. This facilitates learning the features of fraud jobs for automated prediction a challenge.

2. LITERATURE REVIEW

In this section, we attempt to define a list of many models or strategies that are discussed in relation to proposed work and how to prevent fake job recruitments on the internet. We gathered more than ten research papers that will describe numerous models or strategies used for fake job recruitment.

Mr. P. Gulshan et al., (2022) [5] proposed an article on "Fake Job Post Prediction Using Machine Learning Algorithms".In this article, the authors proposed There is a significant increase in the number of online jobs posted on various employment portals during the pandemic. As a result, the task of predicting bogus job postings will be significant. Everyone has problems. Thus, utilizing modern deep learning and machine learning classification algorithms, these fraudulent jobs may be precisely recognized and classified from a pool of job ads containing both false and actual postings. To predict if a job posting is genuine or fake, this article proposes using several data mining approaches and classification algorithms such as KNN, decision tree, support vector machine, Naive Bayes classifier, random forest classifier, multilayer perceptron, and deep neural network. We conducted experiments using EMSCAD, which contained 18000 employee samples. For this, we employed three dense layers.

Sultana Umme Habiba et al., (2021) [6] proposed an article "A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques". In this article the authors discussed on advances in modern technology and social communication in recent years, advertising new job openings has become a very typical issue in today's society. As a result, the problem of predicting bogus job postings will be of major importance to all. Fake job-posting prediction, like many other classification problems, presents a number of obstacles. This research suggests using several data mining techniques and classification algorithms such as KNN, decision tree, support vector machine, nave bayes classifier, random forest classifier, multilayer perceptron, and deep neural network to predict if a job

posting is legitimate or fake. We tested with the Employment Scam Aegean Dataset (EMSCAD), which has 18000 samples. Deep neural networks perform admirably as classifiers in this categorization assignment.

Hridita Tabassum et al., (2021) [7] proposed an article on "Detecting Online Recruitment Fraud Using Machine Learning". In this article, the authors proposed a method for detecting ORF. We presented our results based on the previous model and methodology utilized to develop the ORF detection model using our own dataset. We based our dataset on the Bangladesh job field and a publicly available dataset as a reference. Furthermore, the techniques of Logistic Regression, AdaBoost, Decision Tree Classifier, Random Forest Classifier, Voting Classifier, Light GBM, and Gradient Boosting have been applied. We determined the accuracy of various prediction models, with Light GBM (95.17%) and Gradient Boosting (95.17%) providing the highest accuracy. We attempted to develop an accurate method for detecting fraudulent job postings in this research.

Ibrahim M. Nasser et al., (2021)[8] proposed an article on "Online Recruitment Fraud Detection using ANN ".In this article the authors concentrated on In this paper, an Artificial Neural Network-based approach for detecting fraudulent job postings is developed. For training and testing the suggested model, the public Employment Scam Aegean Dataset (EMSCAD) is used with appropriate text preparation techniques. The precision, recall, and f-measure of our model are 91.84%, 96.02%, and 93.88%, respectively. The results reveal that the proposed ANN-based model outperforms comparable current algorithms in detecting employment fraud.

Syed Mahbub et al., (2020) [9] proposed an article on "Online Recruitment Fraud Detection: A Study on Contextual Features in Australian Job Industries". In this article, the authors proposed the extraction of such contextual features was also automated. The study concludes that including contextual variables improves the automated online recruiting fraud detection model's performance metrics. The study's practical implications are twofold. To begin, the contextual feature space-generating engine may be applied to any dataset with minimal localization work. Second, such learning models can be utilized to detect and prevent online recruiting fraud at the back end of online job recruitment platforms. The study not only shows the benefits of employing contextual elements in fraud detection using a real-world dataset, but it also shows how these contextual features can be collected automatically from the web using localized corporate registries.

Tejasva Bhatia et al., (2021) [10] proposed an article on "Detection of Fake Online Recruitment Using Machine Learning Techniques". In this research, we employed machine learning to detect fraudulent job listings. In our suggested ML strategy, we used two data balancing strategies, "Adaptive Sympathetic" and "Synthetic Minority Oversampling Technique," in conjunction with a feature extraction method called "Term Frequency-Inverse Document Frequency." However, some research has employed "Bag of Words" for feature extraction, which could just count the number of times the word appeared, however, the technique used in this research work (i.e. TF-IDF) also offers the importance of words. The public "Employment Scam Aegean Dataset" (EMSCAD) was used, which contained

approximately 18000 job postings, 800 of which were fraudulent. We employed two machine-learning algorithms, Random Forest and k-nearest neighbor.

Sangeeta Lal et al., (2019) [11] proposed an article on "ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection". In this article, the authors proposed that online recruiting fraud (ORF) is an emerging cyber security challenge. Scammers in ORF make attractive employment offers to job seekers and then steal their money and personal information. Scammers in India have stolen millions of dollars from unsuspecting job seekers. As a result, it is critical to discover a solution to this challenge. OR Detector, an ensemble learning-based model for ORF detection, is proposed in this study. We put the suggested model to the test using a publicly available dataset of 17,860 annotated tasks. The proposed model is shown to be effective, with an average f1-score of 94% and an accuracy of 95.4. Furthermore, it improves specificity by 8% when compared to baseline classifiers.

Yuan Zhang et al., (2015) [12] proposed an article on "Incentive Mechanism Design for Smartphone Crowdsensing".In this article the authors discussed smartphones and their importance due to their pervasiveness and strong sensing capability, smartphones are increasingly recognized as outstanding carriers for mobile sensing crowd-sourcing. However, it is unclear how to appropriately motivate phone users to participate. In this research, we investigate incentive schemes in which each user bids her cost for participation and job owners pay the users she has chosen for sensing. We explore a bidding technique that is relatively widespread in real mobile phone sensing applications but has received no attention in previous research on mobile phone sensing mechanism design. We demonstrate how the job owner might accurately pick users while also lowering her overall payment, even if users submitted false expenses.

Cheng Chen et al., (2013) [13] proposed an article on "Battling the Internet water army: Detection of hidden paid posters". While sponsored posters are an attractive method in corporate marketing, they can have a substantial detrimental impact on online communities because the information provided by paid posters is usually untrustworthy. When two competing companies pay paid posters to spread fake news or critical remarks about each other, ordinary netizens may become overwhelmed and find it difficult to trust the information they obtain from the Internet. Based on real-world trace data, we comprehensively explore the behavioral pattern of online paid posters in this research. To identify potential online paid posters, we construct and verify a new detection mechanism that employs both non-semantic and semantic analysis. Our studies with real-world datasets reveal very promising results.

J. Rusiru Boteju et al., (2011) [14] proposed an article on "AskME: A database abstraction for ad-hoc networks". In this article, I'll go deeper into the impact of crowd-sourcing on web-service security. I will concentrate on the future good and negative implications of crowd-sourcing platforms. First, I will explore how crowd sourcing might assist us in addressing complex issues such as dealing with fraudulent online identities in online social networks. I'll talk about some recent work on crowd-sourced Sybil detection that I've done with a big user survey and various ground-truth datasets of fraudulent and real people. Human workers can be highly accurate in detecting authentic and fraudulent identities under the correct conditions, according to the findings. We can, in fact, create

scalable systems for crowd-sourced Sybil detection, and user testing demonstrate that it can produce extremely accurate results with very low error rates.

3. EXISTING SYSTEM & ITS LIMITATIONS

One of the recent major challenges addressed in the field of Online Recruitment Fraud (ORF) is employment fraud. Many organizations now choose to list their job openings online so that job seekers can access them readily and quickly. However, this could be a form of scam perpetrated by fraudsters who offer jobs to job seekers in exchange for money. To undermine a reputable company's credibility, fraudulent job adverts can be issued. These fraudulent job post detections generate a lot of interest in developing an automated solution for recognizing bogus jobs and alerting people so that they don't apply for them.

LIMITATIONS OF THE EXISTING SYSTEM

1. In recent years, many organizations have preferred to list their job openings online so that job seekers may access them readily and quickly.

2. However, this may be a form of scam by the fraud people because they offer jobs to job searchers in exchange for money. Notifies the user.

3. To solve the challenge of spotting job advertising scams, supervised learning algorithms as classification techniques are originally examined.

4. Using training data, a classifier maps input variables to target classes. Classifiers mentioned.

4. PROPOSED SYSTEM & ITS ADVANTAGES

For detecting bogus posts, a machine learning approach is used, which employs numerous classification algorithms. In this scenario, a classification tool detects and warns the user when it detects bogus job postings among a bigger set of job adverts. To begin addressing the challenge of spotting job posting scams, supervised learning algorithms as classification techniques are being examined. A classifier uses training data to map input variables to target classes. The classifiers covered in the research for distinguishing phony job postings from others are briefly outlined. These classifier-based predictions can be roughly classified into two types: single classifier-based predictions and ensemble classifier-based predictions. In this study, the Nave Bayes Algorithm produced the best results.

ADVANTAGES OF THE PROPOSED SYSTEM

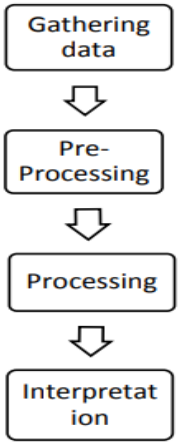
1. For detecting bogus posts, a machine learning approach is used, which employs numerous categorization algorithms. In this scenario, a classification tool detects and warns the user when it detects bogus job postings among a bigger set of job adverts.
2. In this scenario, a classification tool detects and warns the user when it detects bogus job postings among a bigger set of job adverts. To begin addressing the challenge of spotting job posting scams, supervised learning algorithms as classification techniques are being examined.
3. A classifier uses training data to link input variables to target classes. Classifiers were addressed.

5. PROPOSED MACHINE LEARNING ALGORITHMS

The proposed system contains ML Algorithms and we try to compare several ML classification algorithms in order to identify the fake job recruitments and then try to predict the future occurrence based on the input data. There are totally 4 modules present in this current application:

1) Gathering data,
2) Data Pre-Processing
3) Apply ML Models
4) Data Visualization

The whole approach is depicted by the following flowchart.



1) DATA GATHERING

Here we try to load the data set from kaggle repository and once dataset is downloaded we try to load the dataset to the system for performing the operations.

```
[ ] from google.colab import files
    files.upload()

    Choose Files  No file chosen    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
    Saving kaggle.json to kaggle.json
    {'kaggle.json': b'{"username":"b131258","key":"03e26e822f85937c720993918d2d78d7"}'}

[ ] !pip install -q kaggle

[ ] !mkdir ~/.kaggle
    !cp kaggle.json ~/.kaggle
    !chmod 600 ~/.kaggle/kaggle.json

    ! kaggle datasets download -d shivamb/real-or-fake-fake-jobposting-prediction

    Downloading real-or-fake-fake-jobposting-prediction.zip to /content
     87% 14.0M/16.1M [00:00<00:00, 67.7MB/s]
    100% 16.1M/16.1M [00:00<00:00, 66.5MB/s]
```

2) DATA PRE-PROCESSING

Data pre-processing is a technique that is used to convert raw data into a clean dataset. The data is gathered from different sources is in raw format which is not feasible for the analysis. Once the dataset is pre-processed now we can see the list of fields available in the dataset as follows:

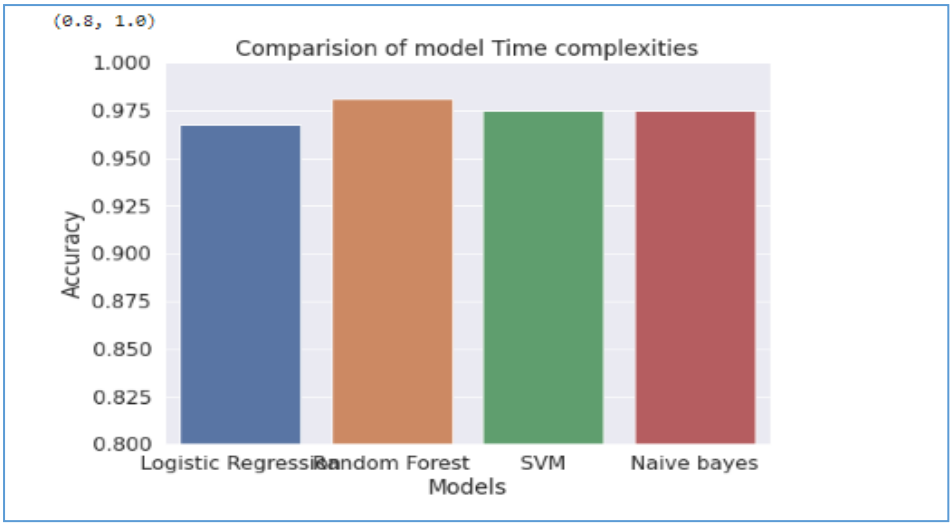
```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   job_id               17880 non-null  int64
 1   title                17880 non-null  object
 2   location             17534 non-null  object
 3   department           6333 non-null   object
 4   salary_range         2868 non-null   object
 5   company_profile      14572 non-null  object
 6   description          17879 non-null  object
 7   requirements         15185 non-null  object
 8   benefits             10670 non-null  object
 9   telecommuting        17880 non-null  int64
 10  has_company_logo     17880 non-null  int64
 11  has_questions        17880 non-null  int64
 12  employment_type      14409 non-null  object
 13  required_experience  10830 non-null  object
 14  required_education   9775 non-null   object
 15  industry             12977 non-null  object
 16  function             11425 non-null  object
 17  fraudulent           17880 non-null  int64
dtypes: int64(5), object(13)
memory usage: 2.5+ MB
```

3) APPLY ML ALGORITHMS

Once data is divided into test and train folders now we can apply well known ML Algorithms on the training data and then check the performance of each and every ML algorithm in order to predict the fake job recruitment and then check which algorithm gives accurate and efficient result.

4) DATA VISUALIZATION

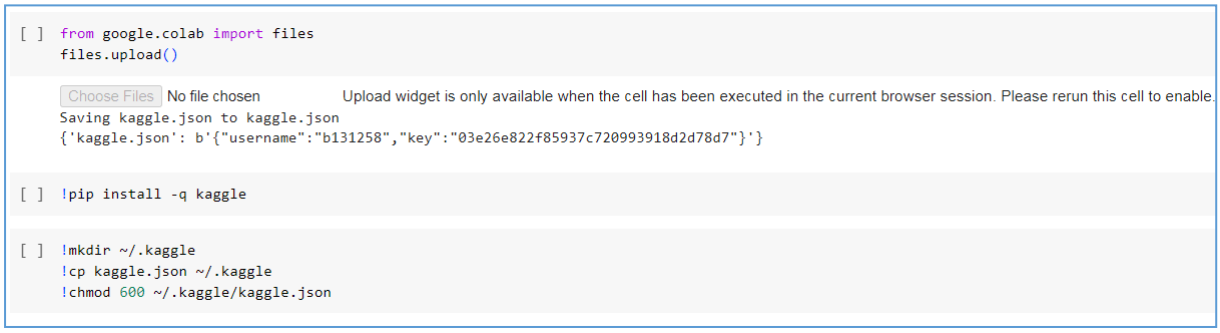
The data set used for is further spitted into two sets consisting of two third as training set and one third as testing set. Here we apply several ML algorithms such as Naïve Bayes,SVM,Logistic Regression, Random Forest to predict the fake job posts and finally came to an conclusion that Random Forest is best among all algorithms to predict the fake job postings messages.



From the above window we apply several ML algorithms such as Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest to predict the fake job posts and finally came to an conclusion that Random Forest is best among all algorithms to predict the fake job postings messages.

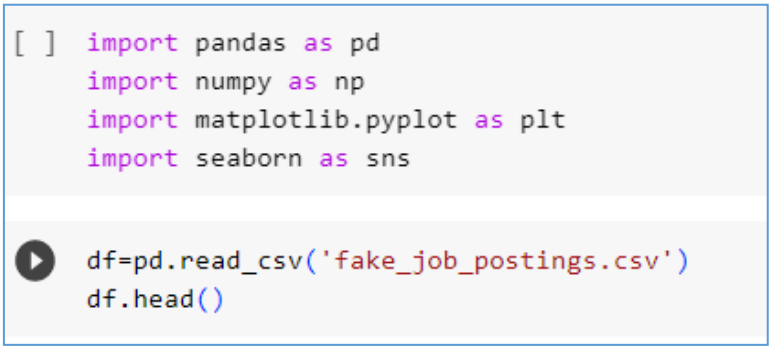
6. RESULT ANALYSIS

LOAD DATASET



From the above window we can clearly see the dataset is loaded, and we try to load the dataset from kaggle website using json file.

IMPORT NECCESSARY LIBRARIES



From the above window we can clearly see the necessary libraries are loaded and hence the .csv file is loaded into the application.

DESCRIBE THE INFORMATION

```
[ ] df.describe()
```

	job_id	telecommuting	has_company_logo	has_questions	fraudulent
count	17880.000000	17880.000000	17880.000000	17880.000000	17880.000000
mean	8940.500000	0.042897	0.795302	0.491723	0.048434
std	5161.655742	0.202631	0.403492	0.499945	0.214688
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	4470.750000	0.000000	1.000000	0.000000	0.000000
50%	8940.500000	0.000000	1.000000	0.000000	0.000000
75%	13410.250000	0.000000	1.000000	1.000000	0.000000
max	17880.000000	1.000000	1.000000	1.000000	1.000000

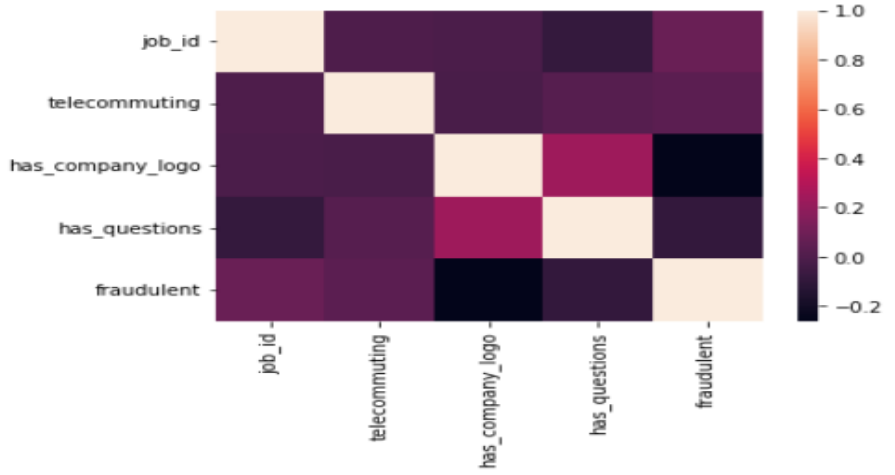
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   job_id               17880 non-null  int64
 1   title                17880 non-null  object
 2   location             17534 non-null  object
 3   department           6333 non-null   object
 4   salary_range         2868 non-null   object
 5   company_profile      14572 non-null  object
 6   description          17879 non-null  object
 7   requirements         15185 non-null  object
 8   benefits             10670 non-null  object
 9   telecommuting        17880 non-null  int64
 10  has_company_logo     17880 non-null  int64
 11  has_questions        17880 non-null  int64
 12  employment_type      14409 non-null  object
 13  required_experience  10830 non-null  object
 14  required_education   9775 non-null   object
 15  industry             12977 non-null  object
 16  function             11425 non-null  object
 17  fraudulent           17880 non-null  int64
```

From the above window we can clearly see the information is retrieved and also numbers of columns are also displayed.

HEAT MAP

```
[ ]  sns.heatmap(df.corr())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f91c1bd2a90>
```



From the above window we can clearly see the heat map is generated for the input dataset.

GENERATE WORDCLOUD

```
wc = WordCloud(width = 800, height = 800, background_color ='white',collocations=True,min_font_size = 10).generate(sentence_maker(df['clean_text']))
plt.figure(figsize=(15,15))
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

From the above window we can clearly see the word cloud is generated based on the main keywords which are present in that job dataset.

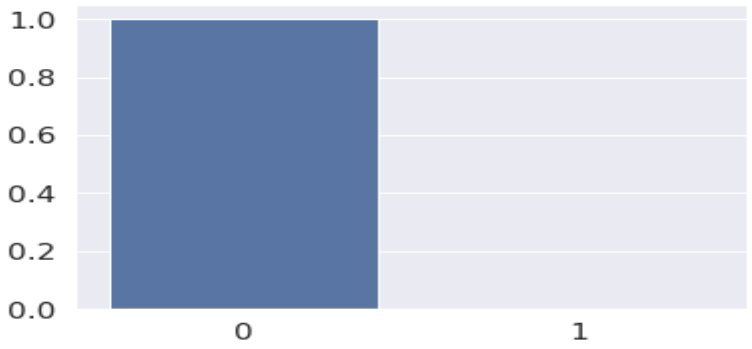
TEST THE INPUT ON SAMPLE JOB PROFILE

```
[ ] test='My name is Michael. I've made it my job to help people succeed online. I'm constantly on the lookout for the best ways and means to make your job simple

[ ]
    text_clean=clean_text(test)
    text=' '.join(text_clean[0])
    # lstm_tokenizer.fit_on_texts(list(text_clean))
    test_token =cv.transform([text])
    test_token.shape

    (1, 102581)

    # test_text
    prob= naive_bayes.predict_proba(test_token)
    pred=naive_bayes.predict(test_token)
    results=pd.DataFrame(data=prob)
    import seaborn as sns
    sns.barplot(data=results)
    plt.show()
```

From the above window we can clearly see the sample input is tested with Naive bayes algorithm and now we can see the resultant result as whether the post is genuine or fraud.

PLOT THE RESULT

```
# test_text
prob= naive_bayes.predict_proba(test_token)
pred=naive_bayes.predict(test_token)
results=pd.DataFrame(data=prob)
import seaborn as sns
sns.barplot(data=results)
plt.show()
```



From the above window we can clearly see the '0' indicates fake and '1' indicates normal post and based on the above test string the system predicts the result as fake job recruitment.

7. CONCLUSION

This article offered an overview of all prior study efforts done in order to detect fake job profiles which are almost available on the internet. In this article, we attempt to integrate machine learning with certain advanced learning approaches in order to identify various forms of fake job recruitment over the internet. Several publications describe the process of fake job recruitment identification and how they suffer as a result of it. In certain papers, we used deep learning models to perform a particular operation on a big dataset in an accurate manner to achieve an accurate output. Our experimental results clearly state that our proposed approach is very accurate in identifying fake job recruitment using random forests.In the future, we hope to expand the research to include some additional issues and draw conclusions based on all of this research.

8. REFERENCES

1] B. Alghamdi and F. Alharby, *—An Intelligent Model for Online Recruitment Fraud Detection,"* J. Inf. Secur., vol. 10,no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009.

[2] I. Rish, *—An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,‖* no.January 2001, pp. 41–46, 2014.

[3] D. E. Walters, *—Bayes's Theorem and the Analysis of Binomial Random Variables,‖* Biometrical J., vol. 30, no. 7,pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

[4] F. Murtagh, *—Multilayer perceptrons for classification and regression,‖* Neurocomputing, vol. 2, no. 5–6, pp. 183–197,1991, doi: 10.1016/0925-2312(91)90023-5.

[5] P. Cunningham and S. J. Delany, ―*K -Nearest Neighbour Classifiers*,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007,doi: 10.1016/S0031-3203(00)00099-6.

[6] S. U. Habiba, M. K. Islam and F. Tasnim, "A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques," *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, DHAKA, Bangladesh, 2021, pp. 543-546, doi: 10.1109/ICREST51555.2021.9331230.

[7] H. Tabassum, G. Ghosh, A. Atika and A. Chakrabarty, "Detecting Online Recruitment Fraud Using Machine Learning," *2021 9th International Conference on Information and Communication Technology (ICoICT)*, Yogyakarta, Indonesia, 2021, pp. 472-477, doi: 10.1109/ICoICT52021.2021.9527477.

[8] I. M. Nasser, A. H. Alzaanin and A. Y. Maghari, "Online Recruitment Fraud Detection using ANN," *2021 Palestinian International Conference on Information and Communication Technology (PICICT)*, Gaza, Palestine, State of, 2021, pp. 13-17, doi: 10.1109/PICICT53635.2021.00015.

[9]S. Mahbub, E. Pardede and A. S. M. Kayes, "Online Recruitment Fraud Detection: A Study on Contextual Features in Australian Job Industries," in *IEEE Access*, vol. 10, pp. 82776-82787, 2022, doi: 10.1109/ACCESS.2022.3197225.

[10] T. Bhatia and J. Meena, "Detection of Fake Online Recruitment Using Machine Learning Techniques," *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India, 2022, pp. 300-304, doi: 10.1109/ICAC3N56670.2022.10074276.

[11] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur and R. Mourya, "ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection," *2019 Twelfth International Conference on Contemporary Computing (IC3)*, Noida, India, 2019, pp. 1-5, doi: 10.1109/IC3.2019.8844879.

[12] Y. Zhang, Y. Fang and S. Zhong, "Incentive Mechanism Design for Smartphone Crowdsensing," *2015 IEEE Fifth International Conference on Big Data and Cloud Computing*, Dalian, China, 2015, pp. 287-292, doi: 10.1109/BDCloud.2015.55.

[13] C. Chen, K. Wu, V. Srinivasan and X. Zhang, "Battling the Internet water army: Detection of hidden paid posters," *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, Niagara Falls, ON, Canada, 2013, pp. 116-120, doi: 10.1145/2492517.2492637.

[14] J. RusiruBoteju and C. I. Keppitiyagama, "AskME: A database abstraction for ad-hoc networks," *2011 International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, 2011, pp. 121-121, doi: 10.1109/ICTer.2011.6075038.