www.revistageintec.net
ISSN: 2237-0722

revista GEINTEC
Gestão, Inovação e Tecnologias

Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms

C.S. Anita[1]; P. Nagarajan[2]; G. Aditya Sairam[3]; P. Ganesh[4]; G. Deepakkumar[5]
[1]Professor, Department of CSE, R.M.D. Engineering College, Kavaraipettai, Tamil Nadu, India.
[1]csa.cse@rmd.ac.in
[2]Professor, Department of ECE, Rajalakshmi Institute of Technology, Chennai, Tamil Nadu, India.
[2]nagarajan.p@ritchennai.edu.in
[3]Associate Software Engineer, Flex, India.
[4]Program Analyst, Cognizant Technology Solutions.
[5]Associate Engineer, Sutherland Global Services Pvt. Ltd.

Abstract

With the pandemic situation, there is a strong rise in the number of online jobs posted on the internet in various job portals. But some of the jobs being posted online are actually fake jobs which lead to a theft of personal information and vital information. 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. In this paper, machine learning and deep learning algorithms are used so as to detect fake jobs and to differentiate them from real jobs. The data analysis part and data cleaning part are also proposed in this paper, so that the classification algorithm applied is highly precise and accurate. It has to be noted that the data cleaning step is a very important step in machine learning project because it actually determines the accuracy of the machine learning as well as deep learning algorithms. Hence a great importance is emphasized on data cleaning and pre-processing step in this paper. The classification and detection of fake jobs can be done with high accuracy and high precision. Hence the machine learning and deep learning algorithms have to be applied on cleaned and pre-processed data in order to achieve a better accuracy. Further, deep learning neural networks are used so as to achieve higher accuracy. Finally all these classification models are compared with each other to find the classification algorithm with highest accuracy and precision.

Key-words: Convolutional Neural Network (CNN), Bi-directional Long Short Term Memory (LSTM), Machine Learning, Data Cleaning, Logistic Regression, Random Forest, Ensemble Modeling.

Received: 16.03.2021 – Accepted: 18.04.2021

1. Introduction

The growth of the internet has made the process of recruitment a far easier process. In addition to this the current pandemic has also played a major role towards the shift in trend of job recruitment these days. Online recruitment has paved the way to more candidates along with streamlined processes and has served a great purpose in bridging the gap between the recruiters and the potential candidates. Candidates can now apply to a large number of jobs according to their specialization on the internet with just the click of a button. Businesses use various internet-based solutions with the help of E-recruitment [10]. Online recruitment helps users to expand their job search and to recruit the most qualified candidates. This helps them to reach out to qualified candidates from all over the world. When online recruitment is relied on the client ends up in hiring the best candidate. Social media like Facebook and LinkedIn is used to learn more about candidates. Various tools such as pre-employment screening, personality assessment and testing for screening the candidates allows companies to select the qualified candidates, and thereby improve the efficiency. This process of course has minimal human intervention. The online recruitment is cost effective advantage in terms of communication.

But some of these jobs being posted are not actual jobs but just fake jobs set as traps to steal potential data. When candidates apply for these jobs the potential information is stolen or in some cases, their laptops are hacked to steal vital information. Cybercriminals pool the victim's information and sell their data on the dark web for the use of others or use it even after years. With proper dataset, and good amount of analysis and cleaning, and with the proper application of machine learning and deep learning algorithms, these fake jobs can be identified and data theft can be prevented.

2. Related Work

There is a number of works on the detection of fake jobs as well as in related topics such as detection of fake news [5][4] and spam mails. Most of these papers use machine learning algorithms for detection of fake jobs from a pool of real jobs. Some of the common methodology used in various research papers for the detection of fake jobs are:

- Using publicly available dataset for the detection of fake jobs[3].
- Using machine learning algorithms in order to classify the fake jobs and real jobs from a pool of job posts[1][2].

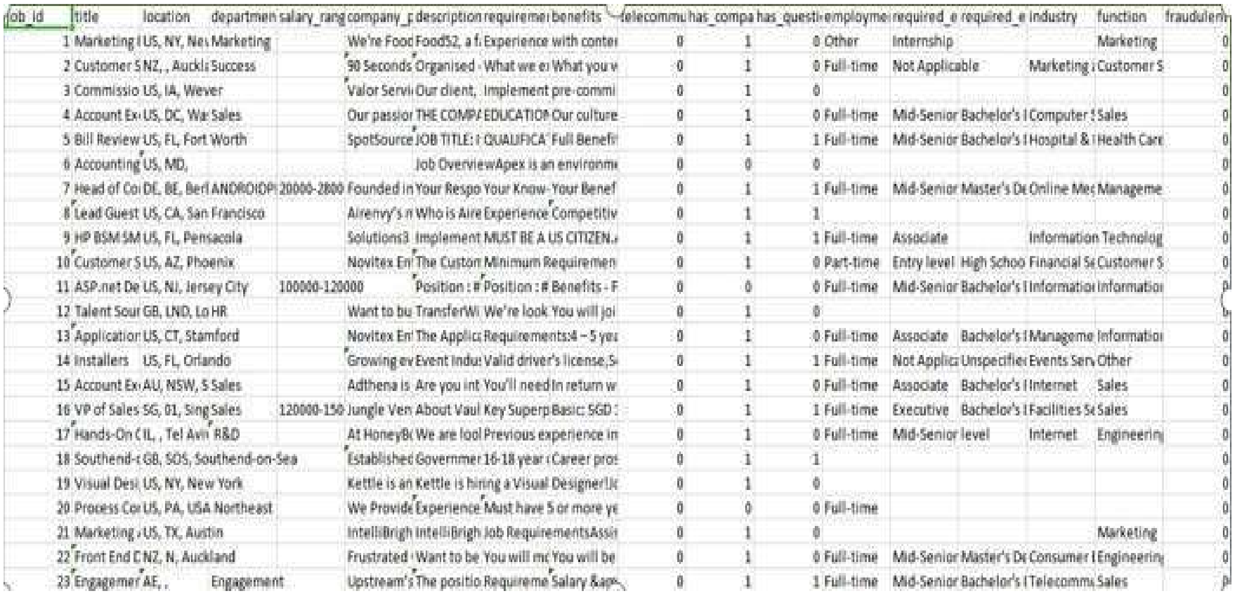
- Using ensemble classification modeling to increase the accuracy of machine learning algorithms[2].
- Semantic analysis of textual data is done for relation bridging and to analyze data for fake jobs[8] as well as fake news[7].
- Using N-gram analysis for the detection of and classification of catagorical data.[6]
- Semantic analysis and natural language processing is used for feature extraction of textual data and to increase the efficiency [8].
- Using complex classification algorithms such as the XGBoost algorithm for getting a higher accuracy of classification[9]

3. Technical Work

3.1 Data Collection

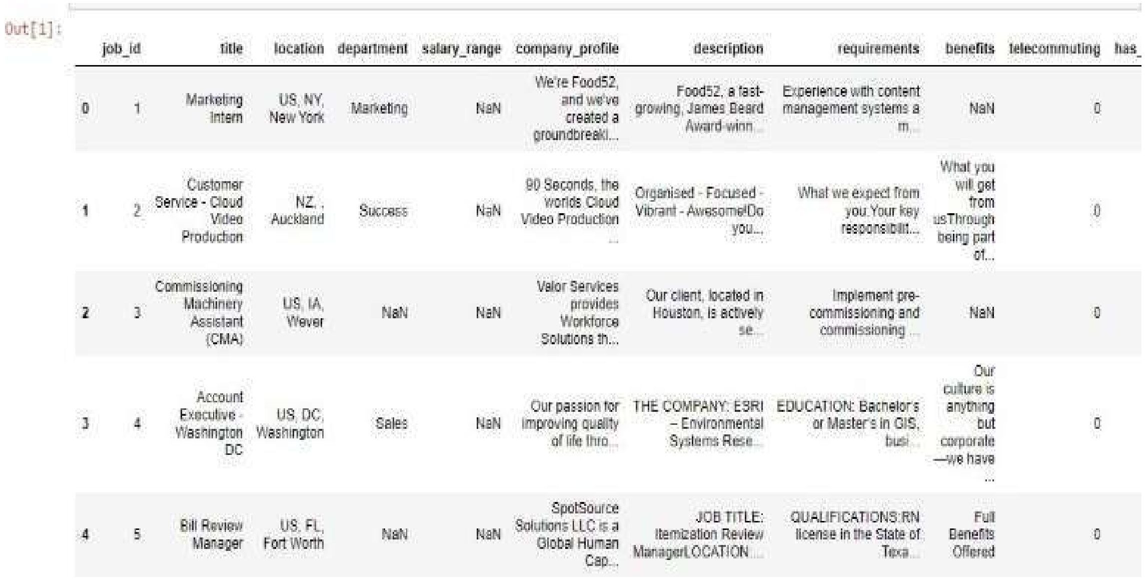
We have trained and tested the dataset obtained from Kaggle as well as from the University of Aegon in order to detect the fake jobs. The data we have collected consists of 17 columns and nearly 18000 rows of textual data as well as numerical data for training and testing the machine learning and deep learning algorithms. These columns are related to the headings and the data present in various job posts being posted in online job hiring sites such as internshala, naukri, etc. These data give a complete image of how the job is being posted online. Figure 1 shows the sample data being collected.

Fig. 1 - Different Columns Present in the Dataset



job_id	title	location	departmen	salary_rang	company_p	description	requireme	benefits	telecommu	has_compa	has_questi	employme	required_e	required_e	industry	function	fraudulen
1	Marketing I	US, NY, Ne	Marketing		We're Food	Food52, a f	Experience with conte		0	1	0	Other	Internship			Marketing	0
2	Customer S	NZ, , Auckl	Success		90 Seconds	Organised -	What we e	What you w	0	1	0	Full-time	Not Applicable		Marketing a	Customer S	0
3	Commissio	US, IA, Wever			Valor Servi	Our client,	Implement pre-commi		0	1	0						0
4	Account Ex	US, DC, Wa	Sales		Our passior	THE COMPA	EDUCATION	Our culture	0	1	0	Full-time	Mid-Senior	Bachelor's I	Computer S	Sales	0
5	Bill Review	US, FL, Fort Worth			SpotSource	JOB TITLE: I	QUALIFICA	Full Benefi	0	1	1	Full-time	Mid-Senior	Bachelor's I	Hospital & I	Health Care	0
6	Accounting	US, MD,				Job Overview	Apex is an environm		0	0	0						0
7	Head of Co	DE, BE, Berl	ANDROIDPI	20000-2800	Founded in	Your Respo	Your Know-	Your Benef	0	1	1	Full-time	Mid-Senior	Master's De	Online Med	Manageme	0
8	Lead Guest	US, CA, San Francisco			Airenvy's n	Who is Aire	Experience	Competitiv	0	1	1						0
9	HP BSM SM	US, FL, Pensacola			Solutions3	Implement	MUST BE A US CITIZEN.		0	1	1	Full-time	Associate		Information Technolog		0
10	Customer S	US, AZ, Phoenix			Novitex En	The Custon	Minimum Requiremen		0	1	0	Part-time	Entry level	High Schoo	Financial Se	Customer S	0
11	ASP.net De	US, NJ, Jersey City		100000-120000		Position : #	Position : #	Benefits - F	0	0	0	Full-time	Mid-Senior	Bachelor's I	Informatio	Informatio	0
12	Talent Sour	GB, LND, Lo	HR		Want to bu	TransferWi	We're look	You will joi	0	1	0						0
13	Applicatior	US, CT, Stamford			Novitex En	The Applica	Requirements:4 – 5 yea		0	1	0	Full-time	Associate	Bachelor's I	Manageme	Informatio	0
14	Installers	US, FL, Orlando			Growing ev	Event Indu	Valid driver's license,S		0	1	1	Full-time	Not Applica	Unspecifie	Events Serv	Other	0
15	Account Ex	AU, NSW, S	Sales		Adthena is	Are you int	You'll need	In return w	0	1	0	Full-time	Associate	Bachelor's I	Internet	Sales	0
16	VP of Sales	SG, 01, Sing	Sales	120000-150	Jungle Ven	About Vaul	Key Superp	Basic: SGD	0	1	1	Full-time	Executive	Bachelor's I	Facilities Se	Sales	0
17	Hands-On C	IL, , Tel Avi	R&D		At HoneyBo	We are lool	Previous experience in		0	1	0	Full-time	Mid-Senior level		Internet	Engineering	0
18	Southend-c	GB, SOS, Southend-on-Sea			Established	Governmer	16-18 year	Career pros	0	1	1						0
19	Visual Desi	US, NY, New York			Kettle is an	Kettle is hiring a Visual Designer!Jo			0	1	0						0
20	Process Cor	US, PA, USA Northeast			We Provide	Experience	Must have 5 or more ye		0	0	0	Full-time					0
21	Marketing A	US, TX, Austin			IntelliBrigh	IntelliBrigh	Job Requirements	Assi	0	1	0					Marketing	0
22	Front End C	NZ, N, Auckland			Frustrated	Want to be	You will mc	You will be	0	1	0	Full-time	Mid-Senior	Master's De	Consumer E	Engineering	0
23	Engagemer	AE, ,	Engagement		Upstream's	The positio	Requireme	Salary &ap	0	1	1	Full-time	Mid-Senior	Bachelor's I	Telecommu	Sales	0

Received: 16.03.2021 – Accepted: 18.04.2021



Out[1]:

	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting	has_
0	1	Marketing Intern	US, NY, New York	Marketing	NaN	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...	NaN	0	
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	NaN	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - Awesome!Do you...	What we expect from you:Your key responsibilit...	What you will get from usThrough being part of...	0	
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	NaN	Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement pre-commissioning and commissioning ...	NaN	0	
3	4	Account Executive - Washington DC	US, DC, Washington	Sales	NaN	Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION: Bachelor's or Master's in GIS, busi...	Our culture is anything but corporate—we have ...	0	
4	5	Bill Review Manager	US, FL, Fort Worth	NaN	NaN	SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICATIONS:RN license in the State of Texa...	Full Benefits Offered	0	

3.2 Data Analysis

Once the data is collected, data analysis has to be done in order to have an insight about the data we are dealing with. The python modules and libraries such as pandas, numpy, matplotlib and seaborn helps us to get a visual insight of the distribution of the data and provides a basic insight about the real jobs and fake jobs. From the analysis phase, we get an image of how unclean our data is and hence it requires data cleaning. Figure 2 shows the data analysis phase. Figure 2(a) shows the number of fake jobs and real jobs in the dataset. Figure 2(b) shows the number of characters used in fake and real jobs. Figure 2(c) shows the distribution of columns such as employment type, required experience and required education asked in real and fake jobs.

Fig. 2(a) - Number of Fake Jobs and Real Jobs in the Dataset
X-axis: Classification of fake and real jobs; Y-axis: Number of fake and real jobs.

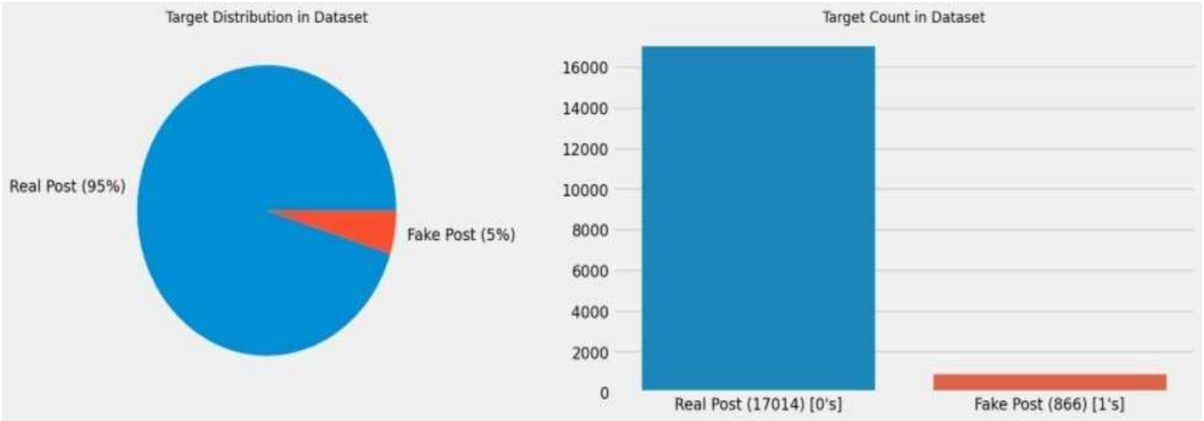


Fig. 2(b) - Number of Characters Being Used in Fake Jobs and Real Jobs
X- axis: Number of characters; Y axis: Number of jobs that uses the number of characters

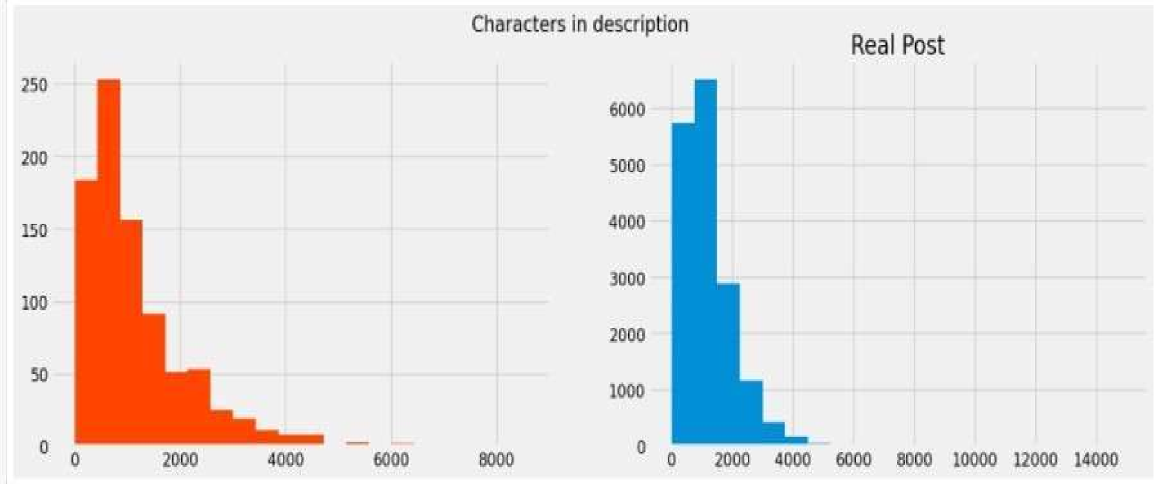
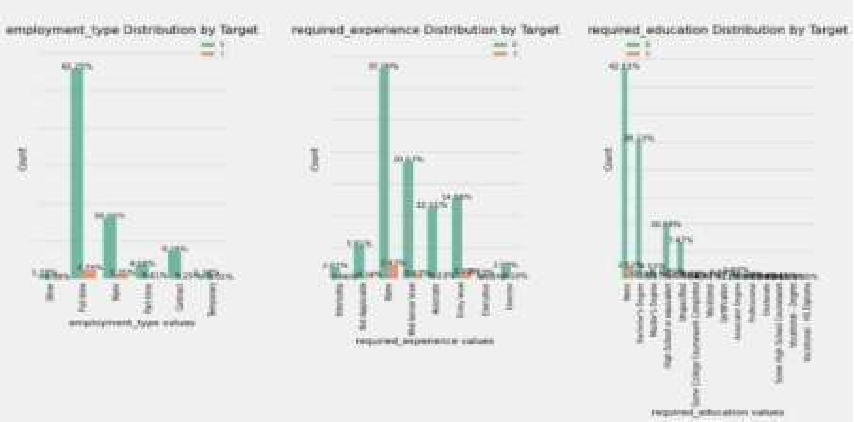


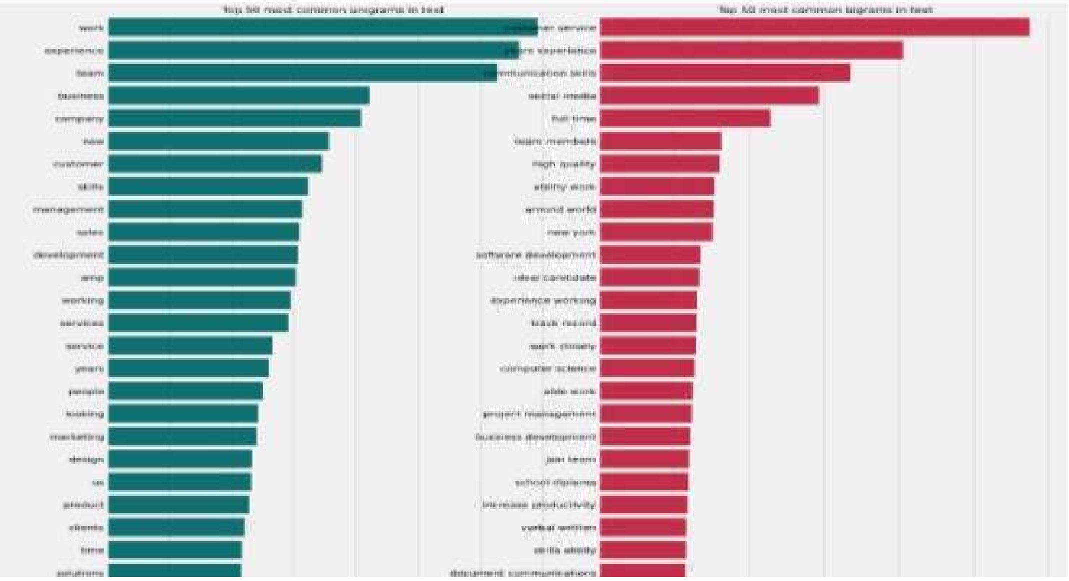
Fig. 2(c) - Distribution of Columns such as Employment Type, Required Experience and Required Education n Real and Fake Jobs
X-axis: fake and real jobs as green and red respectively; Y-Axis: Number of posts with that particular distribution



3.3 Data Cleaning and Preprocessing

After the data analysis is performed on the obtained data, we get that there are a lot of null values and textual data which needs to be cleaned. Hence we first have a look at all the null values present in each column and remove the columns which have a large number of null values. After this we check for stopwords. Stopwords are all the unnecessary words which do not contribute for the detection of fake jobs. Figure 3 shows the cleaned data. The graph shows the distribution of unigrams and bigrams in the cleaned dataset. The graph to the left show unigrams and the graph to the right show the bigrams.

Fig. 3 - Cleaned Data-. X-axis: Number of Unigrams and Bigrams; Y-axis: The Unigrams and Bigrams



Once the stop words are removed from the data, all the textual data are combined together into a single column so that it is in a form suitable for the application of machine learning as well as deep learning algorithms. All the data cleaning process are done using the python libraries such as pandas and NLP packages such as textblob.

3.4 Application of Classification Algorithms

Once the data is clean and processed, the machine learning and deep learning algorithms can be applied for the detection of fake jobs.

3.5 Application of Logistic Regression

Logistic regression is the simplest, easy to understand machine learning classification algorithm. The logistic regression algorithm is applied on the cleaned data and the accuracy and the efficacy of the algorithm is recorded.

3.6 Application of K-Nearest Neighbor (KNN) Algorithm

The KNN algorithm is a simple supervised machine learning algorithm used for both classification and regression problems. This algorithm analyzes the data and prediction is based on the distance between the query example and current example of the data.

3.7 Application of Random Forest Algorithm

Random forest is a supervised learning algorithm which builds an ensemble of decision trees. The concept of bagging is that combination of learning modules increases the overall result. Thus random forest algorithm is applied and the results are recorded.

3.8 Application of Deep Learning Algorithm (Bi-LSTM)

The Bi-direction LSTM is a highly accurate sequence processing model that consists of two LSTMs. Input is taken in the forward direction by one LSTM and the other LSTM takes the input in the backward direction. The amount of information available to the network is increased, and hence the accuracy is also increased.

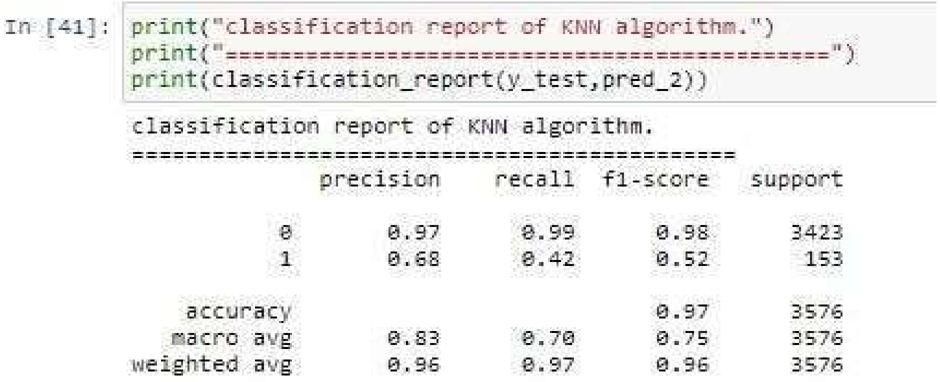
LSTM networks are a type of recurrent neural network that is used in various domains such as machine transition, speech recognition and more.

4. Results and Discussion

4.1 Evaluation and Comparison of Machine Learning and Deep Learning Models

Once the machine learning and deep learning algorithms are applied, we have to evaluate and compare so as to determine the best model for the classification of fake jobs from a pool of job posts. The classification reports of all the used models are printed in order to get a basic insight of the accuracy and efficiency of various models used. Figure 4(a),4(b),4(c) and 4(d) shows the classification reports of the various algorithms.

Fig. 4(a) - Classification Report of KNN Algorithm



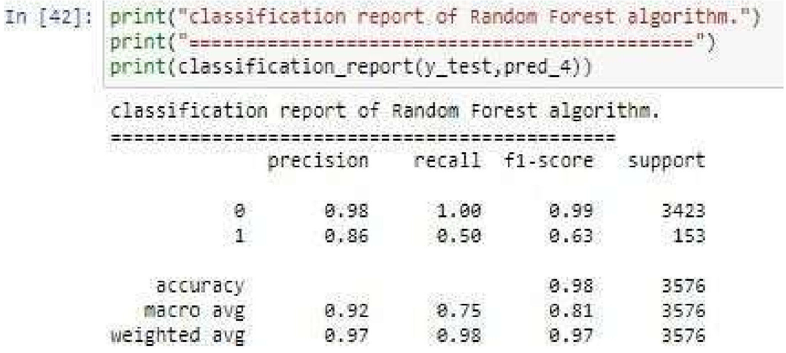
```
In [41]: print("classification report of KNN algorithm.")
         print("=============================================")
         print(classification_report(y_test,pred_2))

         classification report of KNN algorithm.
         =============================================
                       precision    recall  f1-score   support

                    0       0.97      0.99      0.98      3423
                    1       0.68      0.42      0.52       153

             accuracy                           0.97      3576
            macro avg       0.83      0.70      0.75      3576
         weighted avg       0.96      0.97      0.96      3576
```

Fig. 4(b) - Classification Report of Random Forest Algorithm



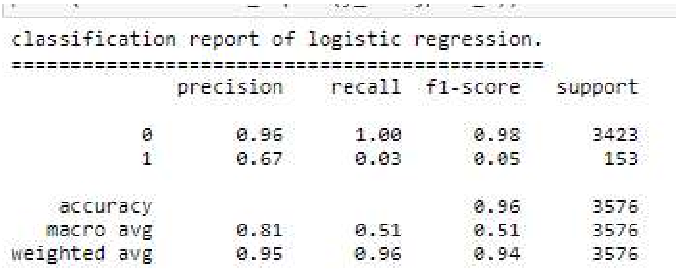
```
In [42]: print("classification report of Random Forest algorithm.")
         print("=============================================")
         print(classification_report(y_test,pred_4))

         classification report of Random Forest algorithm.
         =============================================
                       precision    recall  f1-score   support

                    0       0.98      1.00      0.99      3423
                    1       0.86      0.50      0.63       153

             accuracy                           0.98      3576
            macro avg       0.92      0.75      0.81      3576
         weighted avg       0.97      0.98      0.97      3576
```

Fig. 4(c) - Classification Report of Logistic Regression

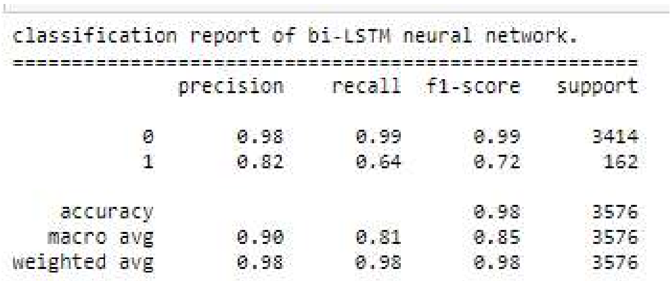


```
classification report of logistic regression.
=============================================
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      3423
           1       0.67      0.03      0.05       153

    accuracy                           0.96      3576
   macro avg       0.81      0.51      0.51      3576
weighted avg       0.95      0.96      0.94      3576
```

Fig. 4(d) - Classification Report of Bi-LSTM Neural Network



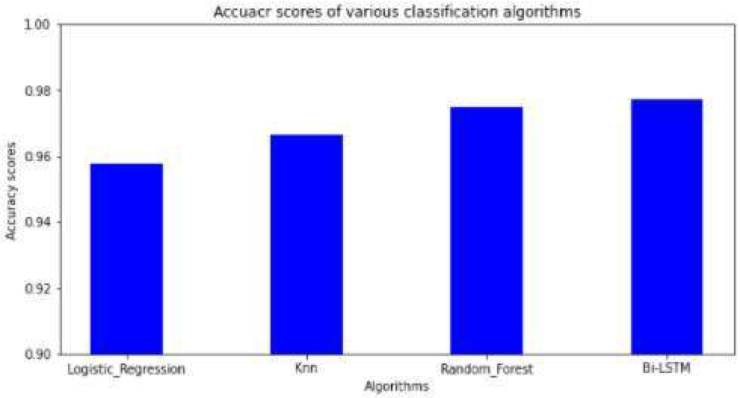
```
classification report of bi-LSTM neural network.
====================================================
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      3414
           1       0.82      0.64      0.72       162

    accuracy                           0.98      3576
   macro avg       0.90      0.81      0.85      3576
weighted avg       0.98      0.98      0.98      3576
```

Fig. 5 - Accuracy Comparison of Various Algorithms

X-axis: Various algorithms used; Y-axis: Accuracy values of all the algorithms and neural network used.



5. Conclusion

In this paper, we have applied machine learning and deep learning algorithms to classify and detect fake jobs from real jobs in a large dataset of job posts. Machine learning algorithms such as logistic regression, KNN classifier and random forest algorithm are used for classification purpose. Deep learning algorithm, Bi-Directional LSTM is used to train the neurons for classification. Additionally, it can be inferred that, data cleaning is a major step in any machine learning as well as deep learning algorithm. The accuracy of the algorithms can be fine-tuned by cleaning and preprocessing the data in a proper way. By applying the classifiers we come to know that the Bi-Directional LSTM gives the most accurate result in detecting the fake jobs compared to other classification algorithms.

References

Shwani, D., & Samir, K.B. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. *International Journal of Engineering Trends and Technology (IJETT)*, *68*(4), 48-53.

Alghamdi, B., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. *Journal of Information Security*, *10*(03), 155.

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet*, *9*(1), 6. https://doi.org/10.3390/fi9010006

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. *In International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, 127-138.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-36.

Arthur, D., & Vassilvitskii, S. (2006). k-means++: The Advantages of Careful Seeding. *In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). Society for Industrial and Applied Mathematics.*

Bhattacharjee, S.D., Talukder, A., & Balantrapu, B.V. (2017). Active learning based news veracity detection with feature weighting and deep-shallow fusion. *In IEEE International Conference on Big Data (Big Data)*, 556-565.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, 675-684.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.

Giovanni, L.C., Prashant, S., Luis, M.R., Johan, B., Filippo, M., & Alessandro, F. (2015). Computational Fact Checking from Knowledge Networks. *Plos One*, *10*(10), e0141938. https://doi.org/10.1371/journal.pone.0128193

Received: 16.03.2021 – Accepted: 18.04.2021