IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006171|

Text Classification with BERT

Jeevan Rohith A M[1], Jesuriya D[2], Manoj Michael Raj A[3], Mohamed Aarif S[4]

UG Students, Department of Computer Science and Engineering, Saranathan College ofEngineering, Trichy, Tamil Nadu, India[1-4]

ABSTRACT: Employment scams are one of the serious issues in this modern world. Many Multinational Companies started their hiring process online during the pandemic and also post-pandemic. Though the online hiring process has several advantages the major drawback is some groups can scam during the hiring process. Their intention may be to borrow many from the user or to reduce the credibility of the company. To overcome such a situation, we have proposed a project which classifies the given job to the predefined label fake/ real jobs. This classifier uses the BERT model which is trained by google. The classifier takes any job detail and differentiates the fake jobs from the real jobs using text classification.

KEYWORDS: Text Classification, BERT, Fake Job Postings, Machine Learning, Job Predictor

I.INTRODUCTION

Text classification which is also known as text tagging or text categorization is the process of classifying the given text into predefined labels. It is also one of the hot topics in natural language processing. BERT stands for Bidirectional Encoder Representation from transformers and is a pre-trained model which was developed by Google.
BERT is trained with English Wikipedia which is 1500M words and BookCorpus which is 80M words. Many developers prefer BERT because it has high accuracy when compared to other classifying models andfrequent updates. As the name suggests, BERT uses the bidirectional nature of the encoders with transformers to perform several Natural Language Processing tasks like sentiment analysis, Q&A, Ham, Spam detection, etc. BERT has two types BERT Base and BERT Large which differ by the number of encoding transformers. Employment Scam is one of a serious issue in recent times, particularly during the pandemic.
Both IT and Non-IT industries started their job hiring process online. Giving an advertisement poster to a job posting web page is common nowadays. But there is a serious issue where people are scammed by fake job advertisements. Their intention mainly focuses on borrowing money from job seekers and also reducing the credibility of the reputed companies. To overcome such dangerous issues this project gives a solution to classify the job posting into predefined labels (real or fake job).

II.RELATED WORKS

According to several reviews and articles Fake Job Posting has been one of the hot topics among Online Fraud detection.

A. LITERATURE SURVEY

[1] Text Classification Research Based on Bert Model and Bayesian Network by Haijun Tao, Shiling Feng, Songsong Liu, the Bert model is a pre-training model based on deep learning. It has refreshed the best performance of 11 NLP missions as soon as it appears, and it also has a wide range of applications. The text data of people's livelihood governance is huge, and there is a large amount of unstructured data, which makes the traditional text analysis and mining technology increase in the space-time complexity of computing; so, it is very important to choose appropriate text classification technology. Here we propose to use the Bert model in combination with the Bayesian network to achieve one new classification of text. That is, the Bayesian network is used to perform the classification of two categories first, and we can get the approximate category range of each text, and then the Bert model is used to classify the text into specific categories. The combination of these two methods can greatly reduce the errors caused by the classification defects of using only one of the methods. Thereby achieving an improvement in the accuracy of text classification.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006171|

This paper mainly realizes the method of the Bayesian network combined with the Bert model to deal with the classification of the social governance texts.

[2] Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge by Da Luo, JindianSu, Shanshan Yu, general language model BERT pre-trained on cross-domain text corpus, BookCorpus and Wikipedia, achieves excellent performance on a couple of natural language processing tasks through the way of fine-tuning in the downstream tasks. But it still lacks of task-specific knowledge and domain-related knowledge for further improving the performance of BERT model and more detailed fine-tuning strategy analyses are necessary. To address this problem, a BERT-based text classification model BERT4TC is proposed via constructing auxiliary sentence to turn the classification task into a binary sentence-pair one, aiming to address the limited training data problem and task-awareness problem. The architecture and implementation details of BERT4TC are also presented, as well as a post-training approach for addressing the domain challenge of BERT. Finally, extensive experiments are conducted on seven public widely-studied datasets for analyzing the fine-tuning strategies from the perspectives of learning rate, sequence length and hidden state vector selection. After that, BERT4TC models with different auxiliary sentences and post-training objectives are compared and analyzed in depth. The experiment results show that BERT4TC with suitable auxiliary sentence significantly outperforms both typical feature-based methods and fine-tuning methods, and achieves new state-of-the-art performance on multi-class classification datasets. For binary sentiment classification datasets, our BERT4TC post-trained with suitable domain-related corpus also achieves better results compared with original BERT model.

This paper proposes a BERT-based text classification model by constructing auxiliary sentence to turn the task into a sentence-pair one, aiming to incorporate more task-specific knowledge and address task-awareness challenge.

[3]In the paper Identifying Real and Fake Job Posting-Machine Learning Approach by Devi.A P, Gayathri.R, Sandhiya.S,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.

The main focus of this paper is to support for the idea that deep learning could be useful in a way for the task of classifying fake jobs. Patterns that our model has found to indicate fake jobs include generalizations, colloquialisms and exaggerations.

[4] Fake Job Recruitment Detection Using Machine Learning Approach by Samir Bandyopadhyay, Shawni Dutta, 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. For this purpose, machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user. To address the problem of identifying scams on job posting, supervised learning algorithm as classification techniques are considered initially. A classifier maps input variable to target classes by considering training data. Classifiers addressed in the paper for identifying fake job posts from the others are described briefly.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006171|

It mainly focuses on Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection and if the use of multiple classifiers enhances the detection of fake jobs.

[5] A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques by Habiba, Sultana & Islam, Md &Tasnim, Farzana, 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 perceptron 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.

This paper has experimented both machine learning algorithms (SVM, KNN, Naive Bayes, Random Forest and MLP) and deep learning model (Deep Neural Network). It shows a comparative study on the evaluation of traditional machine learning and deep learning-based classifiers.

[7] The agricultural industry is still one of the most important sectors on which the majority of Indians rely. Detection of illnesses in these crops is thus crucial to economic progress. Tomatoes and apples are two major crops that are grown in enormous quantities. As a result, the resolution of this research is to detect and identify various illnesses in tomato and apple crops. To categorise tomato and apple leaf diseases from the Plant Village dataset, the suggested methodology employs a CSSA with Bi-LSTM networkmodel. As part of the future work, other learning rates and optimizers could be employed to experiment with the suggested model. It could also contain experimenting with newer architectures to improve the model's performance on the train set. As a result, the aforementioned model can be used as a decision tool to assist and provision farmers in recognizing diseases that can be discovered in tomato plants. With an accuracy of 96 percent, the suggested technology can detect leaf diseases with low computational work.

[8] This paper identifies the covid-19 patient roaming in public places during quarantine time, which has been identified by using IOT and AI techniques with core concept of Human face detection. This is the new idea in this covid -19 conditions for Human face detection. In district wise, day by day regular covid -19 positive cases are stored in the cloud by using an IoT mechanism. The storable data's such as name, mobile number, address with photos (with different poses). These personal details are properly stored and retrieved from the cloud database. The store and retrieve process are handled by using IoT with Raspberry Pi. In CCTV with face detection system is used to observe the actual situation and detect any human presence on the video. We set up the cameras in important places that are connected to the cloud server to forward the covid 19 affected and non-covid person's faces. In this recognition, processes are handled by using AI techniques and the classification of covid positive and normal case by using convolution neural network (CNN). The roaming persons are captured in the camera continuously, AI technique will match and classify the face with already stored database (testing center data). In this classification process, If AI recognizes the covid positive patient, raspberry pi will follow the classified personal data that will directly send a message to the government health care unit, they will take legal action against the person. This experiment, we have conducted by using OpenCV with python platform. This proposed model will minimize the covid 19 spread in public, and also decrease the mortality rate due to covid disease.

[9] This paper provides a solution for monitoring high-risk MHF based on IoT sensors, data analysis-based feature extraction, and an intelligent system based on the Deep Convolutional Generative Adversarial Network (DCGAN) classifier. Various clinical indicators such as heart rate of MF, oxygen saturation, blood pressure, and uterine tonus of maternal are monitored continuously. Many data sources produce large amounts of data in different formats and ratios. The smart health analytics system proposes to extract several features and measure linear and non-linear dimensions. Finally, a DCGAN has been proposed as a predictive mechanism for the simultaneous classification of MFH status by considering more than four possible outcomes. The results showed that the proposed system for mobile monitoring between MFH is a practical solution based on the IoT.

[10] This paper aims at implementing computer vision which can take the sign from the users and convert them into text in real time. The proposed system contains four modules such as: image capturing, preprocessing classification and

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006171|

prediction. By using image processing the segmentation can be done. Sign gestures are captured and processed using OpenCV python library. The captured gesture is resized, converted to grey scale image and the noise is filtered to achieve prediction with high accuracy. The classification and predication are done using convolution neural network.

B. EMAIL SPAM DETECTION

Email spam, also known as junk mails are unwanted emails or emails from unknown sources that comes in bulk to the spams folder in the mailbox. It affects the storage and causes bandwidth problems. To overcome this problem spam filters have been introduced by google and other mail service providers using Neural Network services.[6]

C. FAKE JOB PREDICTOR

The Existing system of this text classification with is approached by using 3 techniques and algorithms, namely Natural Language Processing,Naive Bayes Algorithm, SGD Classifier. Here, Naive Bayes Algorithm and SGD Classifier's F1 scores and accuracies are compared and a final model is chosen. This existing system is done by using Naive Bayes as the baseline model. This system is implemented in a way that the dataset is split into two sets and one set is trained using both the Naive Bayes and the SGD models
which is 70% and the remaining 30% of the dataset is given to test sets. The results from the two models of the test sets are compared and the final model is selected for analysis. Naive Bayes model acquires an accuracyof 0.971 and an F1 score of 0.743 and SGD acquires an accuracy of 0.974 and an F1 score of 0.79. Although the accuracy is quite high in both the models, it has a major drawback. This model predicts most of the jobs as real since the dataset is unbalanced.

III. PROPOSED SYSTEM

In our system, we have used BERT model as the baseline model. It was developed and published by Google. This model predictsreal and fake jobs from the information provided by the user.

A. DATA OVERVIEW

We have used a dataset from Kaggle which consists of 17881 data and 18 fields provided by the university of Aegean.

Table 1 Feature Description

#	Variable	Datatype	Description
1	job_id	int	Identification number given to each job posting
2	title	text	A name that describes the position or job
3	location	text	Information about where the job is located
4	department	text	Information about the department this job is offered by
5	salary_range	text	Expected salary range
6	company_profile	text	Information about the company
7	description	text	A brief description about the position offered
8	requirements	text	Pre-requisites to qualify for the job
9	benefits	text	Benefits provided by the job
10	telecommuting	boolean	Is work from home or remote work allowed
11	has_company_logo	boolean	Does the job posting have a company logo
12	has_questions	boolean	Does the job posting have any questions
13	employment_type	text	5 categories – Full-time, part-time, contract, temporary and other
14	required_experience	text	Can be – Internship, Entry Level, Associate, Mid-senior level, Director, Executive or Not Applicable
15	required_education	text	Can be – Bachelor's degree, high school degree, unspecified, associate degree, master's degree, certification, some college coursework, professional, some high school coursework, vocational

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006171|

16	Industry	text	The industry the job posting is relevant to
17	Function	text	The umbrella term to determining a job's functionality
18	Fraudulent	boolean	The target variable → 0: Real, 1: Fake

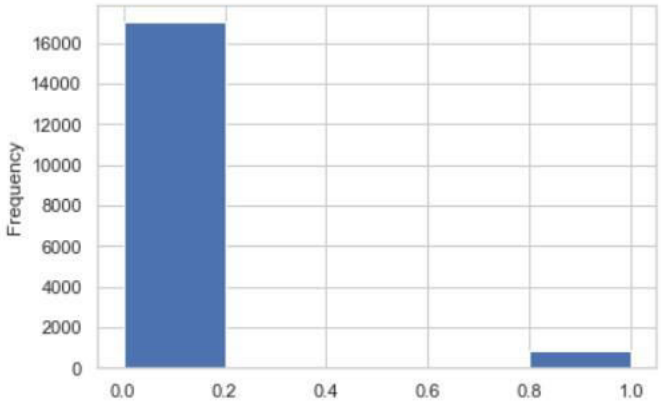

Fig. 1. Shape of the Dataset

```
job_id                  int64
title                  object
location               object
department             object
salary_range           object
company_profile        object
description            object
requirements           object
benefits               object
telecommuting           int64
has_company_logo        int64
has_questions           int64
employment_type        object
required_experience    object
required_education     object
industry               object
function               object
fraudulent              int64
dtype: object
```

Fig. 2. Datatypes of the Fields

Since the dataset is unbalanced the existing system (Fake Job Predictor) does a good predicting the real jobs. But in order to predict both the real and fake jobs the dataset should be balanced during sampling.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006171|

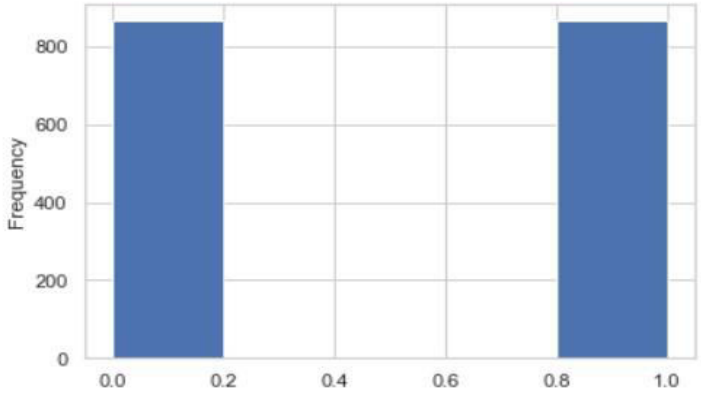

Fig. 3 Shape of the dataset after balancing

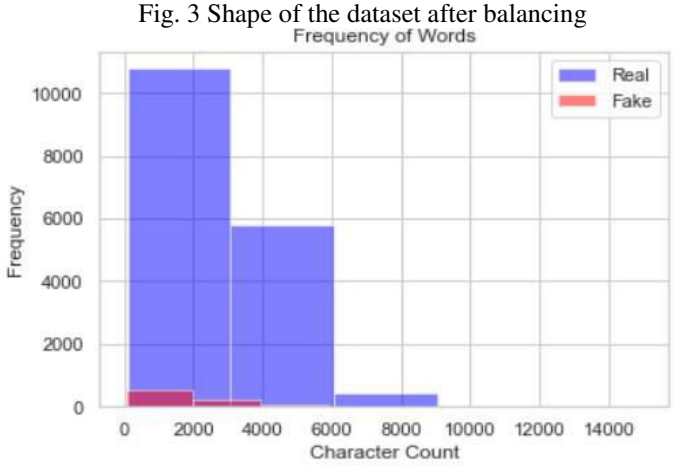

Fig. 3. Frequency of words before balancing the dataset

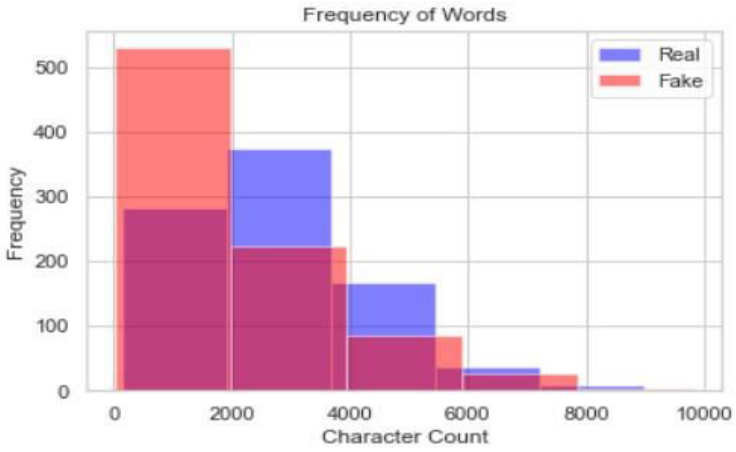

Fig. 4. Frequency of words after balancing the dataset

From this dataset we will take different sets of samplings to train the model. The data is analyzed and pre-processed in order to get the accuracy by classifying the text. Initially, data analysis phase has been done by using the bar graph diagram. Further data cleaning phase is performed in which removal of stop words, punctuations, numerical values and finally the text has been converted to lower case.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006171|

B. IMPLEMENTATION

The text which is used in the model is the combination of four fields namely company profile, description, requirements, benefits and other fields are dropped. Then BERT preprocessor is used to preprocess the text and the preprocessed text is given as the input to the BERT encoder. The output from the BERT encoder will be an encoded matrix. We will be using two neural network layers one is the dropout layer and another one is the dense layer. Dropout layer is used to prevent the model overfitting and dense layer which has only one neuron is used to get the output between 0 and 1. Now the trained and test dataset has been fit into the model with 50 epochs training.

Pre-Training

BERT is a pre-trained model on English language using a Masked Language Modeling (MLM) objective. The BERT model was pretrained on BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables and headers). This pre-trained model is trained with our dataset.

Pre-Processing and Encoding

The data is pre-processed and encoded into a special token. It is achieved through BERT preprocessor and BERT encoder. The BERT preprocessor uses TensorFlow operators from TF.text package to converts raw text into numeric input tensors expected by the encoder. The BERT encoder generates the tokens and additionally it has special tokens like [CLS], [SEP],[PAD] respectively at the beginning, in middle and at the end.

Classification

These special tokens are given as a input to the BERT model for classification. It forms word embeddings with respect to the text in the context and classifies into two labels fraudulent 0 and 1. If the output is closer to 0 it is classified as a real job, otherwise it is classified as a fake job.

IV. PERFORMANCEEVALUATION METRICS

	precision	recall	f1-score	support
0	0.80	0.92	0.85	615
1	0.90	0.76	0.82	597
accuracy			0.84	1212
macro avg	0.85	0.84	0.84	1212
weighted avg	0.85	0.84	0.84	1212

Fig. 5. Training Accuracy

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006171|

	precision	recall	f1-score	support
0	0.77	0.85	0.81	251
1	0.84	0.77	0.81	269
accuracy			0.81	520
macro avg	0.81	0.81	0.81	520
weighted avg	0.81	0.81	0.81	520

Fig. 6. Test Accuracy

V. ACKNOWLEDGEMENT

First and foremost, we want to convey our gratitude and deep appreciation to Dr. S.Venkatasubramanian, M.E., Associate Professor, Department of Computer Science and Engineering, Saranathan College of Engineering, Trichyfor his kind guidance, as well as our college, family and friends for their continuous support during out studies.

V.CONCLUSIONS

In this paper, we have discussed a Fake/Real Job Classifier using BERT model. Among the various tasks involved increasing the accuracy of the model was one of the difficult tasks due to the constraints in the dataset. We are Partially satisfied with the evaluation results. This trained model can be improved a lot with the help of better cost-effective algorithms. Improving and enhancing this idea will be a very useful application. This application will be useful for job searchers and also job posters to create a better standard for improving job postings and avoiding Online Fake Job Postings and Job Scams.

REFERENCES

[1] S. Liu, H. Tao and S. Feng, "Text Classification Research Based on Bert Model and Bayesian Network," 2019 Chinese Automation Congress (CAC), 2019, pp. 5842-5846, doi: 10.1109/CAC48633.2019.8996183.
[2] S. Yu, J. Su and D. Luo, "Improving BERT-Based Text Classification with Auxiliary Sentence and Domain Knowledge," in IEEE Access, vol. 7, pp. 176600-176612, 2019, doi: 10.1109/ACCESS.2019.2953990.
[3] Devi.A P, Sandhiya.S, Gayathri.R "Identifying Real and Fake Job Posting-Machine Learning Approach," Vol. 8, Issue 8, August 2021 DOI: 10.17148/IARJSET.2021.8857.
[4] Bandyopadhyay, Samir & Dutta, Shawni. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. International Journal of Engineering Trends and Technology. 68. 10.14445/22315381/IJETT-V68I4P209S.
[5] Habiba, Sultana & Islam, Md &Tasnim, Farzana. (2021). A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques. 10.1109/ICREST51555.2021.9331230.
[6] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A.O. Adetunmbi, and O. E. Ajibuwa, "Machine learning foremail spam filtering: review, approaches and open researchproblems,‖Heliyon, vol. 5, no. 6, 2019, doi:10.1016/j.heliyon.2019.e01802.
[7] Venkatasubramanian.S., "A Chaotic Salp Swarm Feature Selection Algorithm for Apple and Tomato Plant Leaf Disease Detection", International Journal of Advanced Trends in Computer Science and Engineering, 10(5), pp.3037–3045,2021. https://doi.org/10.30534/IJATCSE/2021/161052021

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006171|

[8]S.Venkatasubramanian,K., Senthil Kumar & J, Gnana & M, Ayeesha. "IoT and AI Based Recognition and Classification of Covid 19 Persons in Public Place",
Turkish Online Journal of Qualitative Inquiry. 12. pp.7098-7110, 2021
[9] S.Venkatasubramanian, "Ambulatory Monitoring of Maternal and Fetal using Deep Convolution Generative Adversarial Network for Smart Health Care IoT System"
International Journal of Advanced Computer Science and Applications(IJACSA), 13(1), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130126
[10] R. Harini, R. Janani, S. Keerthana, S. Madhubala and S. Venkatasubramanian, "Sign Language Translation," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 883-886, doi: 10.1109/ICACCS48705.2020.9074370

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details