IJARSCT

ISSN (Online) 2581-9429

Impact Factor 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 1, July 2022

A Comparative Study on the Prediction of Fake Job Posts using Various Data Mining Techniques

Kesireddy Samara Simha Reddy[1], Jangiti Dwarakamai[2], Motha Rahul[3], CH Yogesh Chowdary[4], Mr. M. Srinivasa Reddy[5]

UG Students, Department of Information Technology[1,2,3,4]

Assistant Professor, Department of Information Technology[5]

Malla Reddy Engineering College, Hyderabad, India

Abstract: *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 datamining techniques and classification algorithm like KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier , multi-layer 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) containing18000 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. Index Terms--false job prediction, deep learning, data mining.*

Keywords: Multi-Layer Perceptron, Data Mining, KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier

I. INTRODUCTION

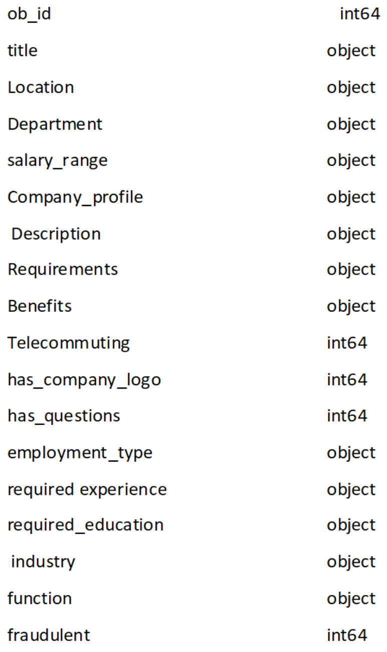
In modern time, the development in the field of industry and technology has opened a huge opportunity for new and diverse jobs for the job seekers. With the help of the advertisements of these job offers, job seekers find out their options depending on their time, qualification, experience, suitability etc. Recruitment process is now influenced by the power of internet and social media. Since the successful completion of a recruitment process is dependent on its advertisement, the impact of social media over this is tremendous, social media and advertisements in electronic media have created newer and newer opportunity to share job details. Instead of this, rapid growth of opportunity to share job posts has increased the percentage of fraud job postings which causes harassment to the job seekers. So, people lack in showing interest to new job postings due to preserve security and consistency of their personal, academic and professional information. Thus, the true motive of valid job postings through social and electronic media faces an extremely hard challenge to attain people's relict and reliability. Technologies are around us to make our life easy and developed but not to create unsecured environment for professional life. If jobs posts can be filtered properly predicting false job posts, this will be a great advancement for recruiting new employees. Fake job posts create inconsistency for the job seeker to find their preferable jobs causing a huge waste of their time. An automated system to predict false job post opens a new window to face difficulties in the field of Human Resource Management.

II. LITERATURE SURVEY

According to several studies, Review spam detection, Email Spam detection, Fake news detection have drawn special attention in the domain of Online Fraud Detection. A Review Spam Detection-People often post their reviews online forum regarding the products they purchase. It may guide other purchaser while choosing their products. In this context, spammers can manipulate reviews for gaining profit and hence it is required to develop techniques that detects these spam reviews. This can be implemented by extracting features from the reviews by extracting features using Natural Language Processing (NLP). Next, machine learning techniques are applied on these features. Lexicon based approaches may be one alternative to machine learning techniques that uses dictionary or corpus to eliminate spam reviews [11]. B. Email Spam Detection-Unwanted bulk mails, belong to the category of spam emails, often arrive to user mailbox. This may lead to unavoidable storage crisis as well as bandwidth consumption. To eradicate this problem, Gmail, Yahoo mail and Outlook service providers incorporate spam filters using Neural Networks. While addressing the problem of email spam detection, content-based filtering, case-based filtering, heuristic-based filtering, memory or instance-based filtering, adaptive spam filtering approaches are taken into consideration [7]. C. Fake News Detection-Fake [17] news in social media characterizes malicious user accounts, echo chamber effects. The fundamental study of fake news detection relies on three perspectives- how fake news is written, how fake news spreads, how a user is related to fake news. Features related to news content and social context are extracted and a machine learning model are imposed to recognize fake new.

III. PROPOSED METHODOLOGY

The target of this study is to detect whether a job post is fraudulent or not. Identifying and eliminating these fake job advertisements will help the jobseekers to concentrate on legitimate job posts only. In this context, a dataset from Kaggle [13] is employed that provides information regarding a job that may or may not be suspicious. The dataset has the schema as shown in Fig. 1.



```
ob_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
```

Fig. 1. Schema structure of the dataset

This dataset contains 17,880 number of job posts. This dataset is used in the proposed methods for testing the overall performance of the approach. For better understanding of the target as a baseline, a multistep procedure is followed for obtaining a balanced dataset. Before fitting this data to any classifier, some pre-processing techniques are applied to this dataset. Pre-processing techniques include missing values removal, stop-words elimination, irrelevant attribute elimination and extra space removal. This prepares the dataset to be transformed into categorical encoding in order to obtain a feature vector. This feature vectors are fitted to several classifiers. The following diagram Fig. 2 depicts a description of the working paradigm of a classifier for prediction.

DOI 10.48175/IJARSCT-5733

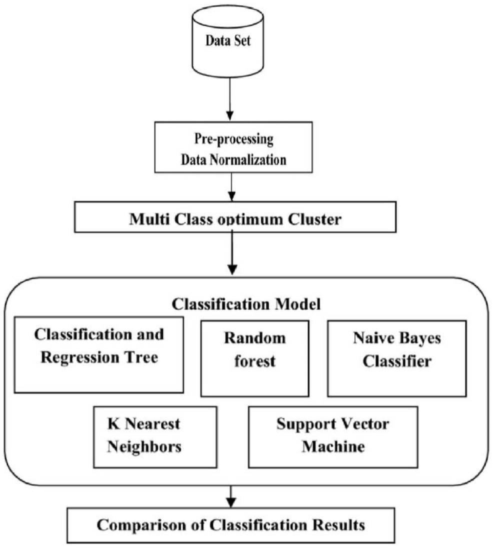


Fig. 2. Detailed description for working of Classifiers

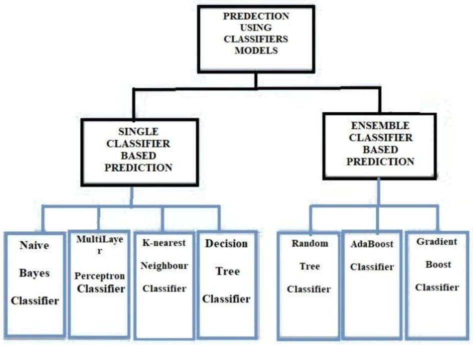

PREDECTION USING CLASSIFIERS MODELS
SINGLE CLASSIFIER BASED PREDICTION
ENSEMBLE CLASSIFIER BASED PREDICTION
Naive Bayes Classifier
MultiLaye r Perceptron Classifier
K-nearest Neighbour Classifier
Decision Tree Classifier
Random Tree Classifier
AdaBoost Classifier
Gradient Boost Classifier

Fig. 3. Classification models used in this framework

3.1 Implementation of Classifiers

In this framework classifiers are trained using appropriate parameters. For maximizing the performance of these models, default parameters[13] may not be sufficient enough. Adjustment of these parameters enhances the reliability of this model which may be regarded as the optimised one for identifying as well as isolating the fake job posts from the job seekers.

This framework utilised MLP classifier as a collection of 5 hidden layers of size 128, 64, 32, 16 and 8 respectively. The K-NN classifier gives a promising result for the value k=5 considering all the evaluating metric. On the other hand, ensemble classifiers, such as, Random Forest, AdaBoost and Gradient Boost classifiers are built based on 500 numbers of estimators on which the boosting is terminated. After constructing these classification models, training data are fitted into it. Later the testing dataset are used for prediction purpose. After the prediction is done, performance of the classifiers are evaluated based on the predicted value and the actual value[15].

3.2 Performance Evaluation Metrics

While evaluating performance skill of a model, it is necessary to employ some metrics to justify the evaluation. For this purpose, following metrics are taken into consideration in order to identify the best relevant problem-solving approach. Accuracy [14] is a metric that identifies the ratio of true predictions over the total number of instances considered. However, the accuracy may not be enough metric for evaluating model‘s performance since it does not consider wrong predicted cases. If a fake post is treated as a true one, it creates a significant problem. Hence, it is necessary to consider false positive and false negative cases that compensate to misclassification. For measuring this compensation, precision and recall is quite necessary to be considered [12].

IV. EXPERIMENTAL RESULTS

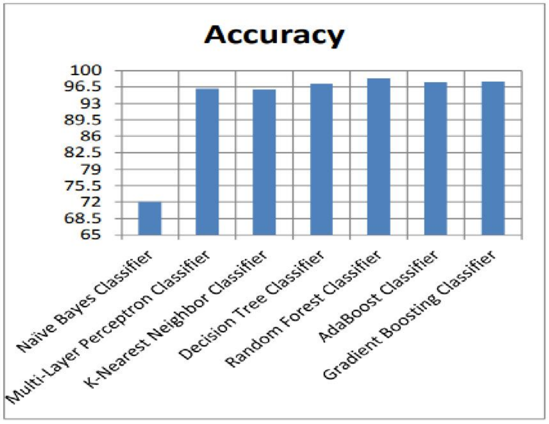
All the above-mentioned classifiers are trained and tested for detecting fake job posts over a given dataset that contains both fake and legitimate posts. The following Table 1 shows the comparative study of the classifiers with respect to evaluating metrics and Table 2 provides results for the classifiers that are based on ensemble techniques. Fig. 4 to Fig. 7 depict overall performance of all the classifiers in terms of accuracy, f1-score, Cohen-kappa score, MSE[16] respectively.

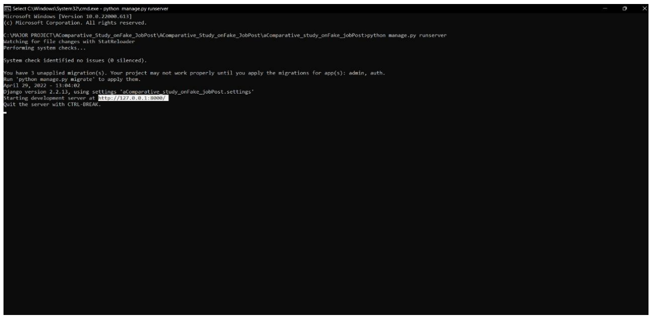
TABLE I
PERFORMANCE COMPARISON CHART FOR SINGLE CLASSIFIER BASED PREDICTION

Performance Measure Metric	Naïve Bayes Classifier	Multi-Layer Perceptron Classifier	K-Nearest Neighbor Classifier	Decision Tree Classifier
Accuracy	72.06%	96.14%	95.95%	97.2%
F1-Score	0.72	0.96	0.96	0.97
Cohen-Kappa Score	0.12	0.3	0.38	0.67
MSE	0.52	0.05	0.04	0.03

Table II: Performance Comparison Chart for Ensemble Classifier Based Prediction

Performance Measure Metric	Random Forest Classifier	AdaBoost Classifier	Gradient Boosting Classifier
Accuracy	98.27%	97.46%	97.65%
F1-Score	0.97	0.98	0.98
Cohen-Kappa Score	0.74	0.63	0.65
MSE	0.02	0.03	0.03





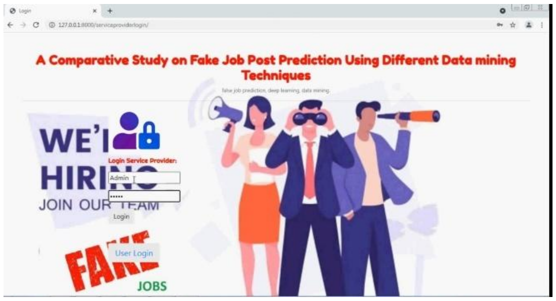
```
Select C:\Windows\System32\cmd.exe - python  manage.py runserver
Microsoft Windows [Version 10.0.22000.613]
(c) Microsoft Corporation. All rights reserved.

C:\MAJOR PROJECT\AComparative_Study_onFake_JobPost\AComparative_Study_onFake_JobPost\aComparative_study_onFake_jobPost>python manage.py runserver
Watching for file changes with StatReloader
Performing system checks...

System check identified no issues (0 silenced).

You have 3 unapplied migration(s). Your project may not work properly until you apply the migrations for app(s): admin, auth.
Run 'python manage.py migrate' to apply them.
April 29, 2022 - 13:04:02
Django version 2.2.13, using settings 'aComparative_study_onFake_jobPost.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```

Fig.4. Run command


127.0.0.1:8000/serviceproviderlogin/
A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques
WE'RE HIRING
JOIN OUR TEAM
Login Service Provider
Login
User Login
FAKE
JOBS

Admin

Fig.5. Admin login

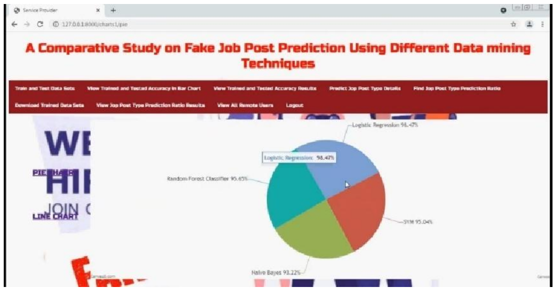
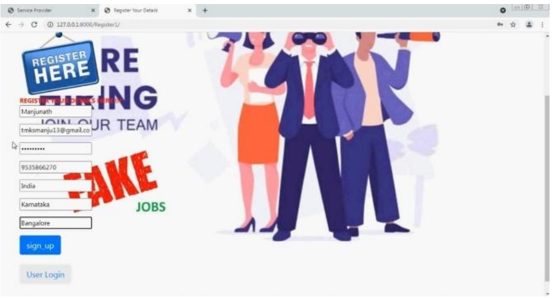
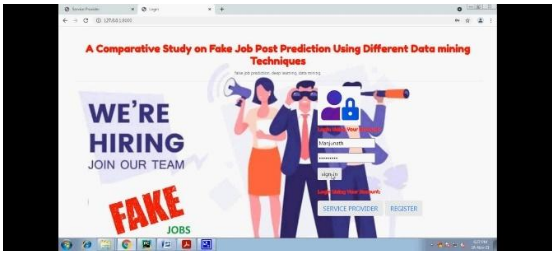


Fig.6. Pie chart of EMDSCAN data set



Manjunath
tmksmanju13@gmail.co
9535866270
India
Karnataka
Bangalore

Fig.7. User registration Page



A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

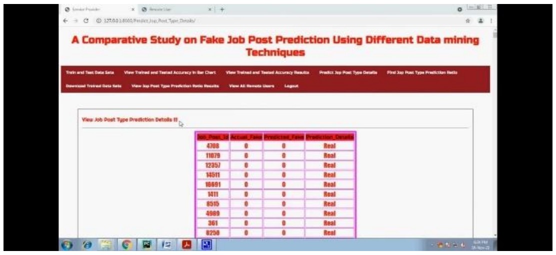
fake job prediction, deep learning, data mining

Manjunath

sign in

SERVICE PROVIDER REGISTER

Fig.8. User login



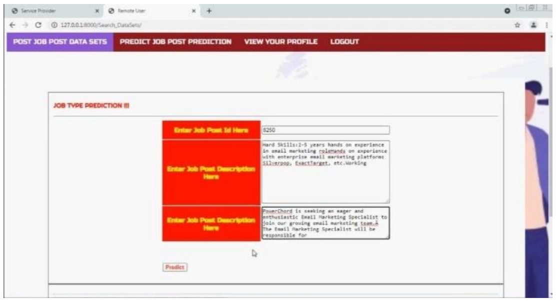
A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Jop Post Type Details
Find Jop Post Type Prediction Ratio
Download Trained Data Sets
View Jop Post Type Prediction Ratio Results
View All Remote Users
Logout

View Job Post Type Prediction Details !!!

Job_Post_Id	Actual_Fake	Predicted_Fake	Prediction_Details
4708	0	0	Real
11079	0	0	Real
12357	0	0	Real
14511	0	0	Real
16691	0	0	Real
1411	0	0	Real
8515	0	0	Real
4989	0	0	Real
361	0	0	Real
8250	0	0	Real

Fig.9. After data cleaning true and fake jobs posts

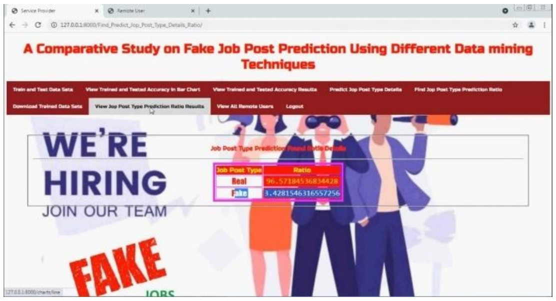


JOB TYPE PREDICTION !!!

Enter Job Post Id Here	8250
Enter Job Post Description Here	Hard Skills:2-5 years hands on experience in email marketing roleHands on experience with enterprise email marketing platforms Silverpop, ExactTarget, etc.Working
Enter Job Post Description Here	PowerChord is seeking an eager and enthusiastic Email Marketing Specialist to join our growing email marketing team.Â The Email Marketing Specialist will be responsible for

Predict

Fig.10. checking the job post



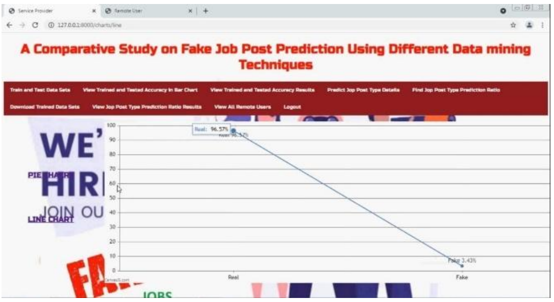
A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Jop Post Type Details
Find Jop Post Type Prediction Ratio
Download Trained Data Sets
View Jop Post Type Prediction Ratio Results
View All Remote Users
Logout
WE'RE
HIRING
JOIN OUR TEAM
FAKE
JOBS

Job Post Type Prediction Found Ratio Details

Job Post Type	Ratio
Real	96.57184536834428
Fake	3.4281546316557256

Fig.11.Distinguishing between true and fake job



A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Jop Post Type Details
Find Jop Post Type Prediction Ratio
Download Trained Data Sets
View Jop Post Type Prediction Ratio Results
View All Remote Users
Logout
PIE CHART
LINE CHART
Real: 96.57%
Real 96.57%
Fake 3.43%
100
90
80
70
60
50
40
30
20
10
0
Real
Fake

Fig.12. Graphical representation of true and fake jobs

DOI 10.48175/IJARSCT-5733

V. CONCLUSION

Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, several machine learning algorithms are proposed as countermeasures in this paper. Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that Random Forest classifier outperforms over its peer classification tool. The proposed approach achieved accuracy 98.27% which is much higher than the existing methods.

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]. H. Sharma and S. Kumar, ―A Survey on Decision Tree Algorithms of Classification in Data Mining,‖ Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.

[7]. E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems,‖ Heliyon, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[8]. L. Breiman―ST4_Method_Random_Forest,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[9]. B.Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, ―Bagging classifiers for fighting poisoning attacks in adversarial classification tasks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37.

[10]. A. Natekin and A. Knoll, ―Gradient boosting machines, a tutorial,‖ Front. Neurorobot., vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[11]. N.Hussain, H.T.Mirza, G.Rasool, I.Hussain, and M.Kaleem, Spam review detection techniques: A systematic literature review,‖ Appl. Sci., vol. 9, no. 5, pp. 1–26, 2019, doi: 10.3390/app9050987.

[12]. Mandru, D.B., Aruna Safali, M., Raghavendra Sai, N., Sai Chaitanya Kumar, G. (2022). Assessing Deep Neural Network and Shallow for Network Intrusion Detection Systems in Cyber Security. In: Smys, S., Bestak, R., Palanisamy, R., Kotuliak, I. (eds) Computer Networks and Inventive Communication Technologies . Lecture Notes on Data Engineering and Communications Technologies, vol 75. Springer, Singapore. https://doi.org/10.1007/978-981-16-3728-5_52

[13]. Mandru, Deena Babu. "An Improved K-Means Algorithm for Web Page Clustering."

[14]. Krishna, Paruchuri Jeevan, M. Deena Babu, and G. Manoj Someswar. "Design & Development of an improved PACK System using TRE Technique for Cloud Computing Users." *International Journal of Research* 3.2 (2016): 384-393.

[15]. Rakesh, Ganya, M. D. Babu, and G. Manoj Someswar. "A Novel Integrated Attestation Graph Analysis Scheme for Enhancing Result Quality and Higher Attacker Pinpointing Accuracy." *International Journal of Research* 3.2 (2016): 214-225.

[16]. Mandru, Deena Babu, and Y. K. Krishna. "Enhanced Cluster Ensemble Approach Using Multiple Attributes in Unreliable Categorical Data." *International Journal of Psychosocial Rehabilitation* 23.1 (2019).

[17]. Mandru, Deena Babu, and Y. S. Krishna. "Multi-view Cluster Approach to Explore Multi-Objective Attributes based on Similarity Measure for High Dimensional Data." *International Journal of Applied Engineering Research* 13.15 (2018): 12289-12295.