2022 International Conference on Futuristic Technologies (INCOFT)
Karnataka, India. Nov 25-27, 2022

An Evaluation of Intelligent Network Data Analytics Based on Machine Learning In 5G Data Networks

Parth Batra
Student, Department of CSE(Data Science and AI), Nanyang Technological University, Singapore and Noida Institute of Engineering and Technology, Greater Noida, India and Kendriya Vidyalaya,
Embassy of India, Kathmandu, Nepal Email id- parthbatra2510@gmail.com

Dr. Vikas Sagar
Assistant Professor, Department of CSE(AI),Nanyang Technological University, Singapore and Noida Institute of Engineering and Technology, Greater Noida, India and Kendriya Vidyalaya, Embassy of India,
Kathmandu, Nepal Email id- drvikas.sagar@niet.co.in

Kanishk Kandoi
Student, Department of Science, Nanyang Technological University, Singapore and Noida Institute of Engineering and Technology, Greater Noida, India and Kendriya Vidyalaya, Embassy of India, Kathmandu, Nepal Email id- kanishkkandoi52@gmail.com

2022 International Conference on Futuristic Technologies (INCOFT) | 978-1-6654-5046-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INCOFT55651.2022.10094324

\

***Abstract*—Big Data Analytics has developed as a judgment technique for unopened organizations to uncover hidden patterns, relationships, industry trends, and consumption patterns. Newline Of the most common sources of big data is Viral marketing data sets, as Web 2.0 null byte technology generate massive social corpora from our everyday routines. Newline in basic language Web 2.0 technology applications including Internet newline data analytics, relationship management, Text Analytics, and opinion line break mining depend heavily on processing.**

Even before compared to prestige cellular networks, 5G cellular networks have many major updates, such as network data analytics-based network data analytics, which will allow network administrators to either implement their own machine learning (ML)-based data analytics methodologies or incorporating third-party solutions into their networks. This study originally presents the structure and protocols of network data analytics based on the 3rd Generation Partnership Project (3GPP) standard standards. Then, based on the fields specified by the 3GPP specification, a cell-based artificial data set for 5G networks is built.

Network slice, a major 5G technology, divides a physical network into many virtual end-to-end networks, each of which may receive logically separate network resources to offer richer services. 5G mobile data and sensor data are combining to produce an increasing network traffic. Traffic explosion has grown into a mixed network type, involving network viruses, worms, network theft, and hostile assaults. How to differentiate traffic kinds, restrict fraudulent traffic, and make optimal use of sensor data in the context of a 5G network slice, as well as the importance of this research

Additionally, some abnormalities are added to this collected data (for examples, unexpected increased volume in a particular tissue), and so these irregularities are categorized within each compartment, subscriber's category, and remote controller.

Keywords*—*5G cellular networks, machine learning, technology of 5G, network viruses, Traffic explosion.

I. INTRODUCTION

Wireless cellular networking is a rapidly evolving field that is predicted to have a dramatic rise in the number of users and endpoints as a result of recent technological advances. Fourth-generation (4G) cellular networks, as standardized by the 3rd Generation Partnership Project (3GPP), have allowed users to achieve speeds of several hundred megabits per second (Mbps), opening the door to high-bandwidth services like high-definition television.

The expectations of users are expected to grow rapidly, and 4G cannot keep up. Cisco's Visual Networking Index, in instance, predicts that by 2020 there will be 12.3 billion mobile-connected devices and that the typical smartphone would produce 132 GB of data each year. Furthermore, the fast rise of applications needing M2M-type communication (e.g., Internet of things (IoT)) has brought additional requirements that are not handled by the earlier cellular technologies created for H2H communications. Consequently, 5G cellular networks have arisen in the recent decade. Using the current 4G infrastructure, 3GPP initially issued the specifications for the non-standalone (NSA) mode of 5G access, allowing for compatibility between the various cellular technologies now in use.

This report creates an associated experimental research to help build and execute the system. The experiment begins with the collection of the small training dataset, followed by statistical analysis of the data's features, filtering, cleaning, and the application of machine learning algorithms such as decision tree, random forest, and regression to the trained small test dataset. To evaluate a model's efficacy, we take samples. This study utilizes the package program to collect real-world traffic data and feed it into the model for evaluation and judgment, gauging the model's efficacy by its ability to produce accurate results. This work discusses the feature selection processes of two distinct machine learning algorithms and examines the performance differences between them via experimental analysis. Chi-square filtering has been shown to improve accuracy and overall performance in experiments.

A. Related concepts and work

Network slices every user's needs may be met by customizing their own "slice" of the network. Through the use of network "slicing," network operators may differentiate between the needs of various tenant categories and provide resources accordingly. One of 5G's most important innovations is network slicing, which relies heavily on network functions virtualization (NFV). NFV separates itself from the rest of the network's hardware and software through centralized server deployment; each network function (NF) is responsible for its own software. To better satisfy the requirements of 5G's everything-connected nature, network slices let several virtual independent sub-networks to share a single physical network communication link and complete data exchange.

B. Machine learning

Machine learning, an implementation method, is crucial to the study of artificial intelligence. What follows is a breakdown of the many points in its development. The 1950s may have been a watershed year for the birth of machine learning. However, the suggested Perceptron can only handle linear classification problems and not XOR logic. With an emphasis on inductive learning methods, technology based on logical representations of symbols flourished in the 1960s and 1970s. In the 1980s, decision trees began to emerge as a popular tool for expressing complicated data linkages in terms of knowledge acquisition, and by the mid-1990s, their usage had become widespread in classification algorithms. This simplicity, however, comes at the price of the learning process having to cope with too broad and complicated assumptions. Statistical learning and deep learning developed as feasible answers to the challenge of coping with huge volumes of data starting in the mid-1990s and continuing to the current day.

C. Big Data and Big Data Analytics

Huge Information is the information accumulated from the large numbers of associations among individuals and the innumerable communications between the machines. It is gathered from different sources, for example, Web 2.0, weblogs, gigantic machines (air plans), RFIDs and GPS. The Huge Information has three properties to be specific, Volume, Speed, and Assortment.

Fig. 1. Big Data Analytics Applications

Different components like reality based, serious areas of strength for navigation framework, business, and IT procedures, business needs, logical devices, and representatives' abilities assume a vital part in the progress of Enormous Information Examination. Fig.1. delineates Large 4 information investigation assists associations with lessening costs, to pursue better and quicker choices, and to make new items or administrations for measuring up to clients' assumptions. With the fast improvements in correspondence innovations, the connections among people, gatherings, and networks Social Enormous Information Examination has arisen as a fruitful exploration region.

II. DATA PREPROCESSING

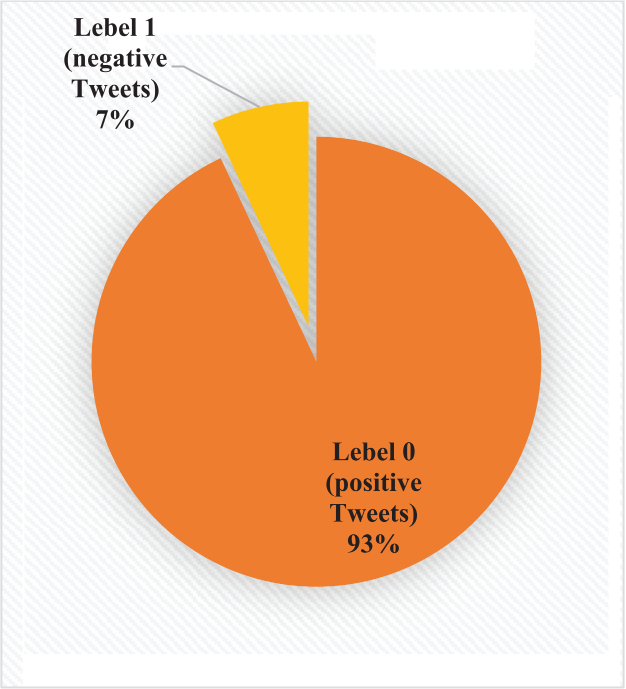


Fig. 2. Data Pre-processing

TABLE I.

Lebel 0 (positive Tweets)	Lebel 1 (negative Tweets)
93.0%	7.0%

III. NETWORK DATA ANALYTICS FUNCTION NWDAF

A newly defined data analytics function in 5G cell networks that gives network examination upon demand from other NFs. As its information source, NWDAF can utilize some other NF. Consequently, there is a two-way connection among NWDAF and NFs as portrayed in Fig. 3. Note that Nnwdaf addresses the assistance based point of interaction of NWDAF, and Nnf addresses the help based connection point of any NF (e.g., Npcf addresses the assistance based point of interaction of PCF). As found in the figure, NWDAF can either give network examination information to other NFs (i.e., investigation data) or NFs can demand membership from NWDAF for information conveyance (i.e., occasion's membership) by utilizing Nnwdaf interface.

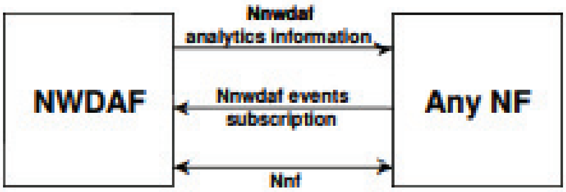


Fig. 3. Data collection and network data analytics exposure architecture

IV. RELATED WORK

(Kuraeva Anna and Kazantsev Nikolay, 2015) Both public and private affiliations working and orchestrating tremendous data projects need to take on a little by little technique for spreading out the right targets and sensible presumptions. Accomplishment depends upon their ability to consolidate and research, and backing dynamic through assessment. The improvement of gigantic data has begun to emerge and should augment recognizable quality soon in the public region of the Russian Association.

(Mohamad Hardyman Barawi and Yet Yong Seng, 2013) The objective information is real factors, and dynamic information changes depending on the situation. Isolating profound information is a perplexing communication, where Feeling Assessment systems will be applied. Feeling Examination is a usage of Ordinary Language Dealing with, which is used to perceive the close to home information from electronic diversion data.

(Shoukry and Rafea, 2012) To restrict the effect of this commonality on our portrayal we will pre-process presents earlier on using them. For example, blunders might be words that offer viewpoint anyway are erroneously spelled, and along these lines ought to be found and overhauled to evaluate feeling all the more exactly. Following the work done on Egyptian Arabic Samir Tartir and Ibrahim AbdulNabi performed pre-taking care of endeavors on Arabic posts, which integrate Ejection of URLs, Fix botches, wipe out stop words.

(Shaojian Zhuo et al., 2014) Ongoing investigation into feeling examination has over and over shown that adding an extremity dictionary to the cycle may extraordinarily improve the exactness of the subsequent characterizations. Following popular assessment and sentiments online has developed progressively subject to client produced data from web-based entertainment and informal communities. The fluffy semantics numerical models.

(Stefano Ferilli et al., 2015) The pre-taking care of tasks consolidate ejection of planning orders, tokenization, clearing of everything aside from words, expressive highlight engravings and emoticons, replacement of shortenings and conversational enunciations with their lengthy and standard transformation, extraction, and decision of n-grams, stopword departure, normalization, and assurance of most gigantic terms.

V. METHOD AND MATERIAL

First, we present the overall process flow of our system. We next outline the topology and traffic circumstances of the network we use for our analysis. In this research, we focus on a system's workflow, which includes a user equipment (UE), user data, 5G system base antennas (SBA), machine learning models, a network workflow data analysis framework (NWDA.

A. Workflow

In the 5G SBA, which houses the NWDAF and other NFs, information gathered from UE is sent, as displayed in Fig. 4. Through the help based interface (SBI), NWDAF is connected to other NFs, and the two might trade data with each other. To gauge the presentation of the organization load and to recognize network load irregularities, NWDAF assembles information from a few NFs and fits numerous ML models.

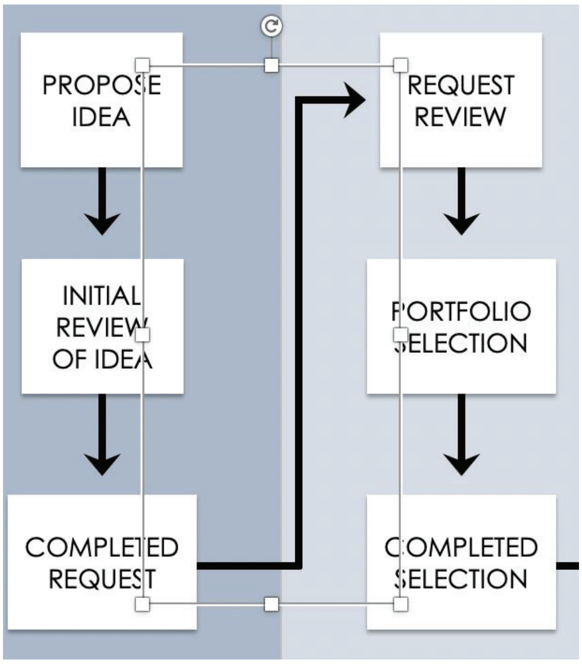


Fig. 4. The high-level workflow of the proposed system

B. Topology

We use a good cell geology, which includes a legitimate plan of RRU cells, a nice course of action of ally classes, and a respectable course of action of individual stuff (i.e., contraption) types. For straightforwardness, regardless of the way that our structure model can maintain topographies that contain a colossal number of RRU cells, endorser classes, and individual stuff types, we contemplate an association topography that involves five RRU cells in our propagations. In all of these cells, there are three ally classes, where these endorser characterizations address platinum, gold, and silver participations. A model depiction of our association geology ought to be noticeable in Fig. 5.

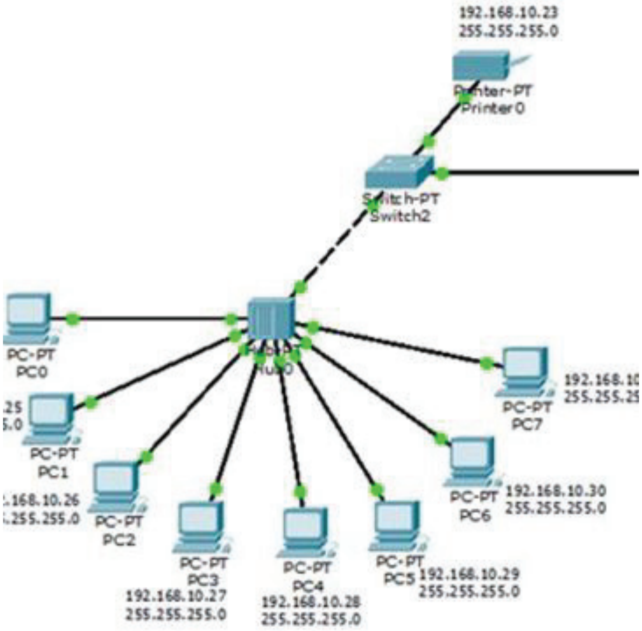


Fig. 5. Sample network topology representation

TABLE II.

Time of day	IOT	Vehicle	Cell phone	Smart Watch	Table computer
00.01	1%	12%	2.6%	2.8%	1%
01.02	5%	1%	2%	3%	2.6%
11.00	1.2%	2.9%	5.%	2.9%	1.9%

C. *Traffic*

As demonstrated by the proposed model, each individual stuff inside each endorser order and RRU cell integrates a fated proportion of traffic load close to the beginning of the propagation. Hence, we can say that association traffic is splashed from the very beginning to the farthest furthest reaches of the re-authorization. Then, at that point, for every re-order time step (Δt), some piece of the pile handovers from one cell (i.e., source) to another phone (i.e., target), which is neigh exhausting the source cell. The handover cycle furthermore occurs considering destined extents.

VI. DATA GENERATION

There are various sorts of man-made intelligence/ML models all through software engineering history. Among these numerous computer based intelligence/ML models, the calculations behind these models work in an unexpected way. As a general rule, one can say that ML calculations can be sorted under three unique parts, in particular, directed, unaided, and support learning. Since we consider administered ML calculations in this paper, marked information becomes essential in this specific situation.

TABLE III.

Category	IOT device	Vehicle	Cell phone	Smart Watch	Table computer
Platinum (4)	2Gbps	10 Gbps	11 Gbps	16 Gbps	10 Gbps
Gold (5)	5 Gbps	2 Gbps	23 Gbps	6 Gbps	11 Gbps
Silver (8)	6 Gbps	2 Gbps	25 Gbps	9 Gbps	2 Gbps

VII. MACHINE LEARNING MODELS

With the ongoing improvements in innovation and the going with huge amount of information open and important to be handled, ML calculations have become progressively well known and are considered as a practical answer for the overwhelming majority different kinds of issues. From authentic information, ML models extricate the particular necessary subtleties. A wide assortment of ML calculations exist, each customized to a particular class of issues. Segment V makes sense of that there are three kinds of ML models; the ones we use in this study are administered learning calculations. Considering that the 5G informational index made is accurately named, we might take utilization of regulated learning strategies. We tackle two issues in this review.

Late mechanical improvements have brought about a gigantic amount of information that must be handled, making ML calculations more appealing as a potential answer for a large number of issues. Out of earlier information, ML models are used to give explicit fundamental data. You might find an extensive variety of ML calculations, every one custom-made to a specific issue class. All of the ML models utilized in this study are instances of administered learning calculations, one of the three sorts of ML models examined in Segment V. Considering that the 5G informational collection made is accurately named, we might take utilization of regulated learning techniques. In the review, we focus in on two essential issues.

A. *ML Modelling*

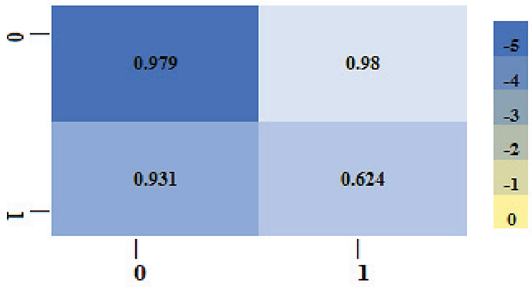
The evaluation metric used in the competition is F1-Score; therefore, we will try to maximize the same.

Confusion matrix

Logistic Regression

Training Scores:	**Accuracy=0.979**	**F1-Score=0.98**
Validation Scores:	**Accuracy=0.931**	**F1-Score=0.624**

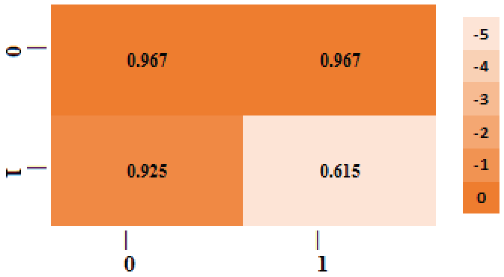
This is confusion matrix



Naive Bayes Classifier

Training Scores:	**Accuracy=0.967**	**F1-Score=0.967**
Validation Scores:	**Accuracy=0.925**	**F1-Score=0.615**

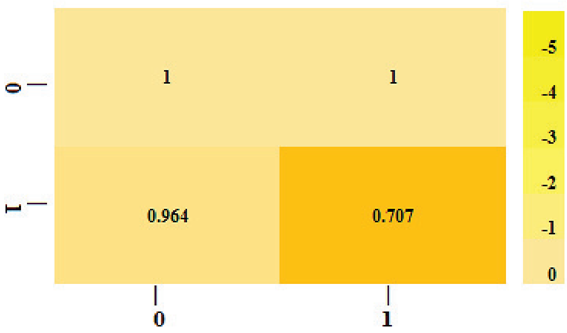
This is confusion matrix



Random Forest Classifier

Training Scores:	**Accuracy=1.0**	**F1-Score=1.0**
Validation Scores:	**Accuracy=0.964**	**F1-Score=0.707**

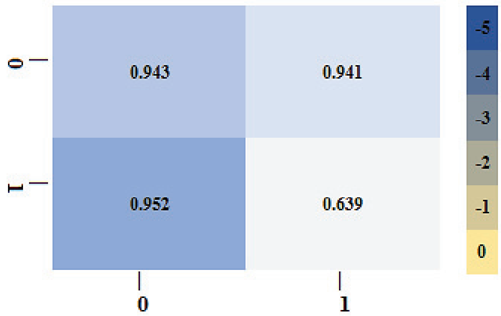
This is confusion matrix



Extreme Gradient Boosting Classifier

Training Scores:	Accuracy=0.943	F1-Score=0.941
Validation Scores:	Accuracy=0.952	F1-Score=0.639

This is confusion matrix



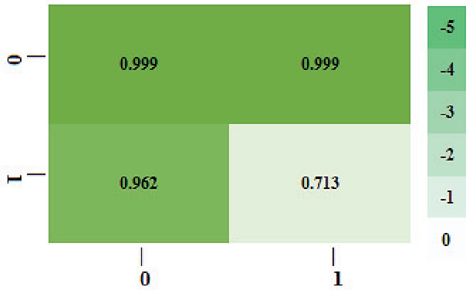
Hyper parameter Tuning

The Irregular Timberland Classifier played out the best out of the four referenced previously. The ongoing second-best model is XGBoost. While looking at the two classifiers, we can see that arbitrary woods is over fitting while XGBoost is under fitting. Along these lines, I evaluated various settings for the hyper boundaries of the arbitrary timberland and XGBoost classifiers.

Random Forest Classifier

Training Scores:	Accuracy=0.999	F1-Score=0.999
Validation Scores:	Accuracy=0.962	F1-Score=0.713

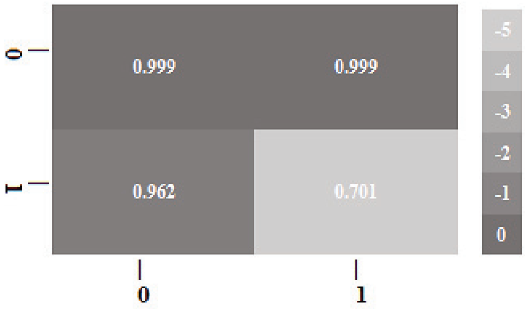
This is confusion matrix



Extreme Gradient Boosting Classifier

Training Scores:	Accuracy=0.999	F1-Score=0.999
Validation Scores:	Accuracy=0.962	F1-Score=0.701

This is confusion matrix



VIII. CONCLUSION

In this research, we provide a unique approach to achieving smart network analytics in 5G cellular networks. In order to achieve this goal, we first detail NWDAF inside the service-based architecture of 5G cellular networks, and then utilize a number of ML approaches to solve two critical issues. Time series analysis, and more especially linear regression, LSTM, and RNN models, are used to make predictions about network load in the first challenge. In the second issue, XGBoost, a state-of-the-art tree-based gradient boosting approach, is combined with logistic regression models to categorize network abnormalities. In addition, we provide a method for the systematic development of a cell-based data set to test the efficacy of network data analytics in 5G cellular networks by using the parameters specified in the 5G standard specification.

Through our tests, we have shown that neural network models are superior to linear regression models at estimating network congestion. Comparatively, when it comes to network anomaly classification, tree-based XGBoost performs better than logistic regression. Finally, we use several widely-used ML models to demonstrate how NWDAF may be put to good use.

IX. FUTURE WORK

Our next effort will also include detecting communication patterns using exploratory data analysis based on the provided data set. Lastly, our work may be expanded upon by using other AI/ML models and putting more attention on additional NWDAF capabilities.

REFERENCES

[1] 3GPP, “5G System; Network data analytics services; Stage 3,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.520), September 2019, version 16.1.0.

[2] 3GPP, “5G System; Unified data management services; Stage 3,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.503), Sept. 2019, version 16.1.0.

[3] Hernandez-Chulde and C. Cervello-Pastor, “Intelligent optimization and machine learning for 5G network control and management,” in Proc. PAAMS, June 2019, pp. 339–342.

[4] Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, “Five disruptive technology directions for 5G,” IEEE Commun. Mag., vol. 52, no. 2, pp. 74–80, Feb. 2014.

[5] 3GPP, “Study on latency reduction techniques for LTE,” 3rd Generation Partnership Project (3GPP), Technical Report (TR 36.881), July 2016, version 14.0.0.

[6] Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. NIPS, 2014, pp. 2672–2680.

[7] S. Sevgican, M. Turan, K. Gökarslan, H. B. Yilmaz, and T. Tugcu, “Synthetic 5G cellular network data for NWDAF,” 2019.

[8] 3GPP, “5G System; Usage of the unified data repository service for policy data, application data and structured data for exposure; Stage 3,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.519), Oct. 2019, version 16.2.0

[9] Meera, R., Nair, Ramya, G., R., and Bagavathi Sivakumar, P.(2017). Usage and analysis of Twitter during 2015 Chennai flood towards disaster management. Procedia Computer Science., 115: 350–358.

[10] Marianela García Lozano, Jonah Schreiber and Joel Brynielsson(2017). Tracking geographical locations using a geo-aware topic model for analyzing social media data. Decision Support Systems., 99: 18–29.

[11] Caseiro and Arnaldo Coelho (2018). The influence of Business Intelligence capacity, network learningand innovativeness on startups performance. Journal of Innovation & Knowledge., 4(3): 139-145.

[12] Mario Marchand and Marina Sokolova(2005). Learning with Decision Lists of Data-Dependent Features. Journal of Machine Learning Research., 6: 427–451.

[13] Rada Mihalcea, ”Unsupervised Large-VocabularyWord Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling,” Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, PP. 411–418, 2005.

[14] P. Edmonds, E., Agirre(2006). Word Sense Disambiguation: Algorithms and Applications. Springer Verlag. Text, Speech and Language Technology Series., HAL Id: artxibo-00080512.

[15] Daniel Jurafsky and James, H., Martin(2006). Computational Lexical semantics. Speech and Language Processing., 354-368.

[16] Ping Chen, Wei Ding, Chris Bowes, David Brown, “A Fully unsupervised word sense disambiguation method using dependency knowledge,” Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, PP. 28–36, 2009. ISBN 978- 1-932432-41-1.

[17] Yathiraju, D. . (2022). Blockchain Based 5g Heterogeneous Networks Using Privacy Federated Learning with Internet of Things. Research Journal of Computer Systems and Engineering, 3(1), 21–28. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/37

[18] Degambur, L. -., Mungur, A., Armoogum, S., & Pudaruth, S. (2021). Resource allocation in 4G and 5G networks: A review. International Journal of Communication Networks and Information Security, 13(3), 401-408. doi:10.54039/IJCNIS.V13I3.5116

[19] Goswami, H., & Choudhury, H. (2021). Security of IoT in 5G cellular networks: A review of current status, challenges and future directions. International Journal of Communication Networks and Information Security, 13(2), 278-289. doi:10.54039/ijcnis.v13i2.4955

[20] Venu, S., Kotti, J., Pankajam, A., Dhabliya, D., Rao, G. N., Bansal, R., . . . Sammy, F. (2022). Secure big data processing in multihoming networks with AI-enabled IoT. Wireless Communications and Mobile Computing, 2022 doi:10.1155/2022/3893875

[21] Akıncı, R., Akdoğan, E., & Aktan, M. E. (2022). Comparison of machine learning algorithms for recognizing drowsiness in drivers using electroencephalogram (EEG) signals. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 44-51. doi:10.18201/ijisae.2022.266

[22] Veeraiah, D., Mohanty, R., Kundu, S., Dhabliya, D., Tiwari, M., Jamal, S. S., & Halifa, A. (2022). Detection of malicious cloud bandwidth consumption in cloud computing using machine learning techniques. Computational Intelligence and Neuroscience, 2022 doi:10.1155/2022/4003403