Evaluation of the Explainable AI-NLP Framework for Text Categorization

Amar Preet Singh,
Assistant Professor, Department of Electrical & Electronics Engineering, Noida Institute Of Engineering and Technology,
Greater Noida, Uttar Pradesh, India,
Email Id- amarpreet.en@niet.co.in

Gaurav Gupta,
Assistant Professor, Department of Computer Science and Engineering, Tula's Institute,
Dehradun, India,
Email Id- gauravgupta@tulas.edu.in

Soumya K,
Assistant Professor, Department of Computer Science and Information Technology, Jain (Deemed to be University),
Bangalore, India,
Email Id- soumya.k@jainuniversity.ac.in

2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) | 979-8-3503-6684-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC3SE62002.2024.10593585

***Abstract*: Although there has been a decrease in interpretability, significant progress has lately been achieved in developing state-of-the-art models. This assessment provides a thorough analysis of Explainable AI (XAI) within the framework of Natural Language Processing (NLP). This study examines the fundamental classification of descriptions in addition to the many techniques for producing and displaying explanations. The purpose of the survey is to provide a comprehensive study of the many approaches and strategies used to clarify the models' predictions, making it a helpful tool for people who develop NLP models. Additionally, we draw attention to the gaps in the current literature and urge further investigation into potential avenues for this crucial field of study. An experiment is also conducted and the results are discussed in this article.**

Index Terms: Natural Language Processing (NLP), Machine learning, Explainability strategies, Model predictions, Explainability approaches, Interpretation techniques

I. INTRODUCTION

Over the past few years, there have been notable developments in Natural Language Processing (NLP) systems. Because they were easily comprehended by humans, "white box techniques" like rules, DT, HMMs, and logistic regressions were commonly used to design such systems. Explainability has decreased as model performance has increased due to the advent of "black box techniques," such as deep learning models and feature-based language integration. People's faith in different AI systems that are utilised in daily life, such chatbots, recommendation engines, and information retrieval systems, may be damaged by a lack of openness regarding the methodologies employed by models to create their results.[1]

The relevance of explainability has been recognised by the wider AI community, which has led to the creation of a new discipline called Explainable AI (XAI). Similar to how different academic subjects have different demands when it comes to the application of certain methodologies, these fields also have different environments in which to study explainability. In particular, the XAI presentations given at the most esteemed NLP conferences during the last seven years are examined in this study. To the best of our knowledge, no poll has looked specifically into XAI in relation to NLP before.

This poll places a strong focus on the concept of explainability, particularly when viewed through the eyes of an end user who want to comprehend the decision-making process. This research focuses on the challenge known as the "outcome explanation problem." Explanations are essential to fostering user confidence in NLP-based AI systems because they enable users to understand the predictions made by the models. Additionally, having a solid grasp of how the model works allows users to offer insightful feedback to engineers, which improves the model's overall quality.[2]

Model prediction explanations have traditionally been divided into two categories: those that deal with specific forecasts, those that deal with the model's general prediction process, and those that need post-processing. It's important to consider other factors when characterising explanations, particularly the methods employed to produce or display them.

The corpus of NLP literature that is currently available on these subjects is examined in this study. It looks for methods and procedures for explainability and visualisation that are often employed in the process of creating explanations.[3-6]

The research continues by examining the inadequacies and difficulties in the creation of efficient explainability approaches in the field of natural language processing. It also examines popular assessment methods used to score the quality of explanations.[8–12]

This assessment centres on the latest developments in natural language processing (NLP) [34] to pinpoint the procedures utilised to achieve XAI and assess the efficacy of these techniques. The "explanator" and the "black-box model" it explains are distinguished by the categorization method that is used.

His strategy works particularly well when a surrogate model is utilised and a black-box model requires an explanation. But the importance of this disparity is mitigated in most contemporary NLP research, where a single neural network is used for both prediction and explanation. The primary objectives of this survey are as follows: This study has three goals in mind: to (1) give informative details about the current state of XAI in NLP; (2) show off current methods to developers who are interested in building explainable NLP models; and (3) to draw attention to areas in which research is lacking, especially the formal definitions and criteria needed to assess how explainable NLP models are.

A web-based platform has been designed to provide readers an interactive experience and detailed information on every item featured in this poll. [13]

II. CATEGORIZATION OF EXPLANATIONS

Other research have already proven that explanations are often categorised based on two fundamental qualities. The first distinguishes between explanations that are tailored for individual model predictions (called local explanations)[35-36] and explanations that cover the model's whole prediction process. The second element makes a distinction between explanations that spontaneously arise during the prediction

979--83503-6684-6/24/$31.00 ©2024 IEEE

process (self-explanatory) and those that need further processing after the fact (post-hoc). We discuss these aspects in the following sections and provide a brief synopsis of the four resulting categories.[14, 37] A local explanation is a claim or justification offered in favour of a model's forecast for a specific input. 46 articles altogether, out of the 50 papers that were included in our comprehensive research, fall into this specific category.

Global Explanation: Independent of any particular input, a global explanation offers a thorough justification by elucidating the fundamental mechanics of the model's prediction process. The last four papers that have been reviewed for this survey fall into this category. Given that the primary goal of this survey is to clarify predictions rather than comprehend the general behaviour of the model—a subject outside the purview of this study—the limited prevalence of "global explanations" is to be expected.

Regardless of whether they are local or global, explanations can be further categorised based on whether they are integral to the "prediction process" or need additional post-processing steps after the model has produced its forecast [38].

The directly interpretable strategy, sometimes referred to as the self-explanatory technique, entails producing explanations alongside forecasts. This is accomplished by employing the model's knowledge all through the prediction process [4].

III. ASPECTS OF EXPLANATIONS

While the previously given taxonomy offers a useful and comprehensive categorization for explanations, it leaves out further important features. Two more explanation components are presented in this section: (1) the methods used to derive explanations and (2) the way in which the explanations are communicated to the end user.

In this talk, we will look at the common explainability methods, the basic operations that make explainability possible, and the visualisation techniques that are usually used to show the outcomes of various related explainability techniques.[15–20] For each of the four previously mentioned "high-level" explanation classes, the common combinations of explainability strategies, processes, and visualisation techniques are determined. These pairings as well as excellent essays.[21].

Explainability techniques and visualisations have a complicated connection since, although they are closely connected, they differ greatly from one another, that warrant examination on their own. The explanation derivation process is mostly concerned with the mathematical justifications for a model's output; this is an area that artificial intelligence scientists and engineers frequently investigate.

The method uses a range of explainability techniques to provide "raw explanations," which are similar to attention scores. On the other hand, the process of explaining these "raw explanations" to end users—ideally overseen by UX engineers—focuses on determining the most effective way to do so by utilising appropriate visualisation tools, such as saliency heat maps [22–31].

IV. EXPERIMENT

The PCMag Review Dataset and the Skytrax User Reviews Dataset are the two separate datasets used in this study's trials. The PCMag website, which provides reviews for a range of electronic items including computers, cellphones, and cameras, is the source of the PCMag Review Dataset.

Every item in this dataset includes a lengthy review text, three brief comments that summarise the review from a neutral, positive, and negative standpoint, as well as the product's overall rating score. The overall rating score has a range of 1.0 to 5.0. Items with review texts longer than 70 words or comments with more than 75 tokens are eliminated in an effort to streamline the focus on thorough information visualisation. Next, the dataset is randomly split into training, development, and test sets using the distribution of the overall rating..

The airline review part of the Skytrax website is where the Skytrax User Reviews Dataset can be found. Each item in this dataset consists of a review text, five sub-field scores that indicate how users rated the in-flight experience, cuisine, cabin amenities, seat comfort, and ticket value.

The total score goes from 1 to 10, while the sub-field values vary from 0 to 5. Like the PCMag dataset, the dataset is randomly divided into training, development, and test sets with the distribution of total rating scores. Items with review texts longer than 300 tokens are sorted out.

Several embeddings (GloVe for PCMag and random for Skytrax), hidden dimensions, and batch sizes are used in the experimental setups for both datasets. Furthermore, certain filter sizes and numbers are given for the CNN model. Consistency is ensured and meaningful comparisons between various models and datasets are made possible by these experimental conditions. [31]

V. RESULTS

The quality of these explanations is measured using BLEU scores, and using a model improved with the proposed framework (CVAE+GEF) compared to the basic model (CVAE) alone shows significant gains. Figure 1 displays The BLEU scores for the positive, negative, and neutral explanations are shown; these show improvements in a number of different areas when CVAE+GEF is used. Furthermore, the models' performance in classifying text data is assessed using classification accuracy metrics; CVAE+GEF outperforms CVAE in these metrics, indicating that fine-grained information from explanations improves overall classification performance (see figure 2).

Better results compared to CVAE, suggesting that fine-grained information from explanations enhances overall classification performance shown in figure 2.

Numerical explanations are subjected to further examination, which shows gains in accuracy when using the suggested framework. The findings of the human evaluation show that, when compared to the basic model (CVAE), explanations produced by CVAE+GEF are thought to be more closely linked with categorization results. In order to address these concerns, the section ends by noting possible flaws in text explanation generation, such as the propensity for generated explanations to be shorter and the inclusion of

unknown words (UNKs).Figure 3 shows Accuracy for different models on the Skytrax User Reviews Dataset.

VI. CONCLUSION AND FUTURE SCOPE

This research goal is to look into recent advancements in XAI, especially those that concern NLP. The study was based on papers that were presented at significant NLP workshops throughout the previous seven years. The fundamental categories of explanations have been discussed, as have the processes for producing and presenting explanations. Additionally, the common methods and techniques for creating explanations in the context of NLP models have been introduced.

This survey's main goal is to give developers with relevant data so they may build better interpretable NLP models. It also aims to motivate researchers to address the current issues with XAI within the framework of natural language processing.[13–17] This survey has led to numerous important lines of inquiry. First things first, it's critical to create clear and consistent language for the idea of explainability and to create a deeper understanding of it.

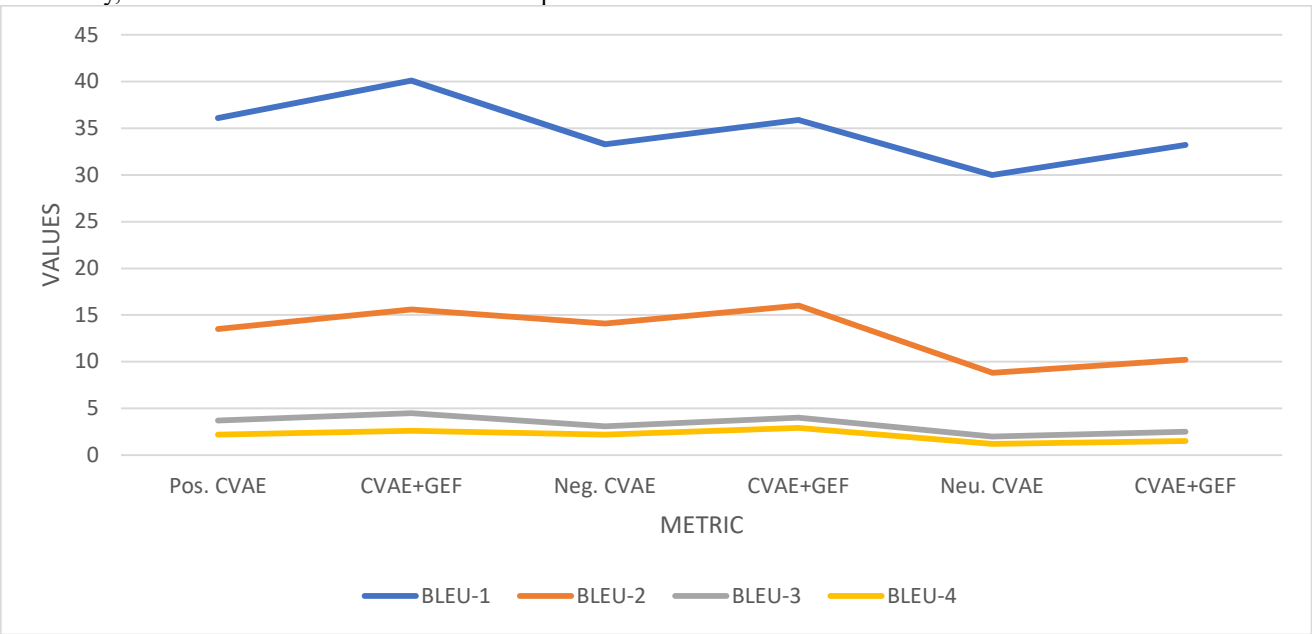

Fig. 1. BLEU ratings for explanations that are generated

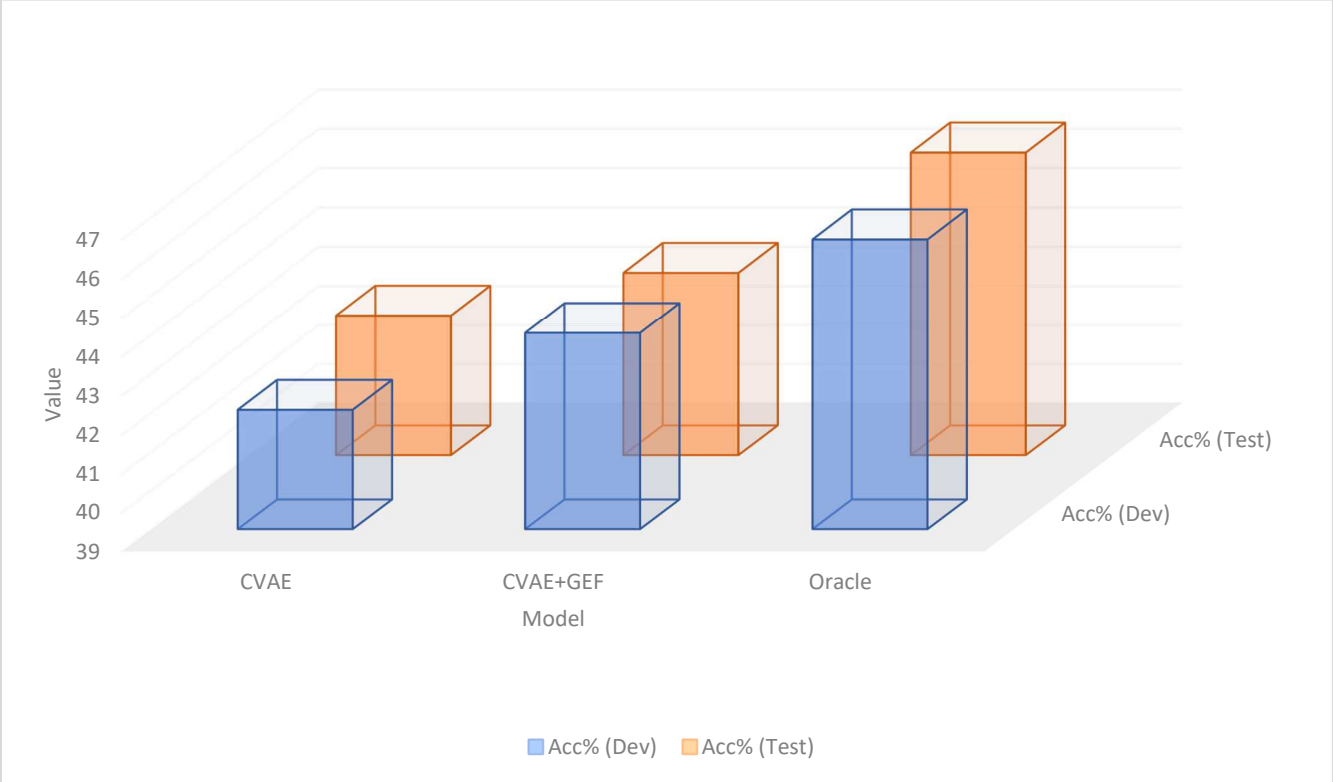

Fig. 2. Accuracy of categorization

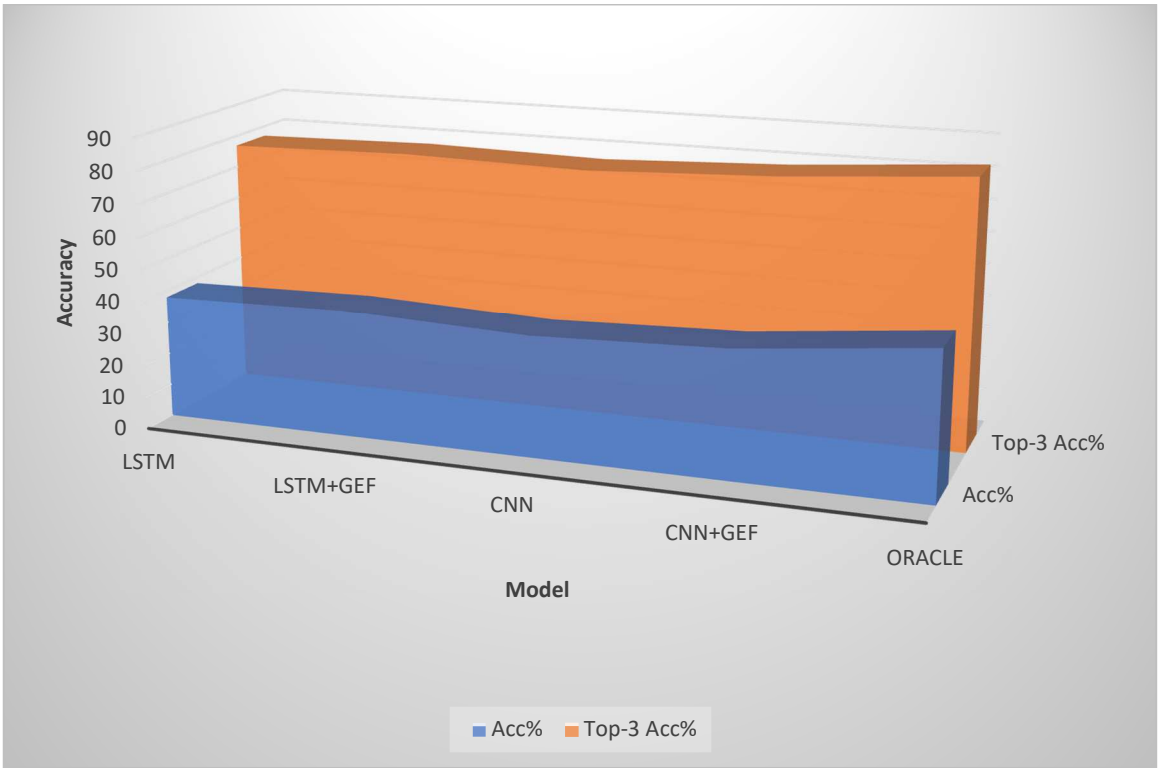

Fig. 3. Skytrax User Reviews' accuracy A collection of models' datasets

Moreover, the analysis of trade-offs between different model properties, such their interpretability and prediction accuracy, is an important area of research that requires the development of standardised assessment metrics [18–22]. In order to ensure that proposed explanations adequately elucidate the model's conclusions, it is necessary to conduct a comprehensive analysis of the issue of causality or integrity in explanations.[33]

One of the survey's most notable conclusions is how little emphasis was placed on global explanations—just four research were included in this category. White box models are widely used in NLP, which contributes to its explainability on a global scale.

White box models are widely used in NLP, which contributes to its explainability on a global scale. Explaining black box models is the primary objective of the XAI discipline, with a special emphasis on local explanations. Even though they are typically not presented as fundamentally interpretable or explicable, white box models are still in use. Still, it's important to consider their potential importance while assessing explanation techniques.

REFERENCES

[1] Nikita Bhutani, Kun Qian, Yunyao Li, H. V. Jagadish, Mauricio Hernandez, and Mitesh Vasa. 2018. Exploiting structure in representation of named entities using active learning. In Proceedings of the 27th International Conference on Computational Linguistics, pages 687–699, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[2] J. Kim, S. Hur, E. Lee, S. Lee and J. Kim, "NLP-Fast: A Fast, Scalable, and Flexible System to Accelerate Large-Scale Heterogeneous NLP Models," 2021 30th International Conference on Parallel Architectures and Compilation Techniques (PACT), Atlanta, GA, USA, 2021, pp. 75-89, doi: 10.1109/PACT52795.2021.00013.

[3] "[iSAI-NLP 2020 Keynote Speakers - 3 abstracts]," 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Bangkok, Thailand, 2020, pp. i-iii, doi: 10.1109/iSAI-NLP51646.2020.9376827.

[4] Thushan Ganegedara; Andrei Lopatenko, Natural Language Processing with TensorFlow: The definitive NLP book to implement the most sought-after machine learning models and tasks , Packt Publishing, 2022.

[5] L. Kryeziu, V. Shehu and A. Chaushi, "Evaluation and Verification of NLP Datasets for the Albanian Language," 2022 International Conference on Artificial Intelligence of Things (ICAIoT), Istanbul, Turkey, 2022, pp. 1-5, doi: 10.1109/ICAIoT57170.2022.10121823.

[6] G. Pradeepa and R. Devi, "Malicious Domain Detection using NLP Methods — A Review," 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 1584-1588, doi: 10.1109/SMART55829.2022.10046882.

[7] N. Patel and D. Patel, "Implementation Approach of Indian Language Gujarati Grammar's Concept “sandhi” using the Concepts of Rule-based NLP," 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2021, pp. 481-485.

[8] A. Dash, A. Mohanty and S. Ghosh, "Advanced NLP Based Entity Key Phrase Extraction and Text-Based Similarity Measures in Hadoop Environment," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-6, doi: 10.1109/ISCON57294.2023.10112121.

[9] A Ambikapathy, Jyotiraditya Sandilya, Ankit Tiwari, Gajendra Singh, Lochan Varshney”Analysis of Object Following Robot Module Using Android, Arduino and Open CV, Raspberry Pi with OpenCV and Color Based Vision Recognition”Advances in Power Systems and Energy Management: Select Proceedings of ETAEERE 2020, 2021, 365-377, Springer singapore.

[10] A.Ambikapathy,Gajendra Signh and PrabhakarTiwari ”Multi- level (13) inverter for Grid connected PV system with PID controller International journal of control theory and Applications, ISSN 0974-5572,vol 10, issue 6, 215-222, 2017.

[11] K Logavani, A Ambikapathy, GArun Prasad, Ahmad Faraz, Himanshu singh”Smart Grid, V2G and Renewable Integration” Electric Vehicles: Modern Technologies and Trends, 175-186, Springer singapor Ae, 2020

[12] Arunprasad Govindharaj, Anitha Mariappan, Ambikapathy Aladiyan, Hassan Haes Alhelou “Real-time implementation of adaptive neural

backstepping controller for battery-less solar-powered PMDC motor"IET Power Electronics, vol 16, issue 1, 128-144, 2023.

[13] Shiva Pujan Jaiswal, Vikas Singh Bhadoria, Ranjeeta Singh, Vivek Shrivastava, A Ambikapathy "Case Study on Modernization of a Micro-Grid and Its Performance Analysis Employing Solar PV Units " Energy Harvesting, Chapman and Hall/CRC, 81-104

[14] Aggarwal, A., Mittal, R., Gupta, S., Mittal, A. "Internet of things driven perceived value co-creation in smart cities of the future: a PLS-SEM based predictive model." Journal of Computational and Theoretical Nanoscience 16.9 (2019): 4053-4058.

[15] Singh, J.P., Chand, P.K., Mittal, A., Aggarwal, A. "High-performance work system and organizational citizenship behaviour at the shop floor." Benchmarking: An International Journal 27.4 (2020): 1369-1398.

[16] Shalender, Kumar, and Rajesh Kumar Yadav. "Strategic flexibility, manager personality, and firm performance: The case of Indian Automobile Industry." Global Journal of Flexible Systems Management 20 (2019): 77-90.

[17] Goyal, J., Singh, M., Singh, R., Aggarwal, A. "Efficiency and technology gaps in Indian banking sector: Application of meta-frontier directional distance function DEA approach." The Journal of finance and data science 5.3 (2019): 156-172.

[18] Sood, K., Dhanaraj, R.K., Balamurugan, B., Grima, S., Uma Maheshwari, R. eds. Big Data: A game changer for insurance industry. Emerald Publishing Limited, 2022.

[19] Rajan, S., and L. Joseph. "An adaptable optimal network topology model for efficient data centre design in storage area networks." International Journal on Recent and Innovation Trends in Computing and Communication 11.1 (2023): 43-50.

[20] Kumar, Sanjai, D. Chitradevi, and Sanju Rajan. "Stock price prediction using deep learning LSTM (long short-term memory)." 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2022.

[21] Kaliraja, C., D. Chitradevi, and Anju Rajan. "Predictive Analytics of Road Accidents Using Machine Learning." 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2022.

[22] Mahaveerakannan, R., et al. "An IoT based forest fire detection system using integration of cat swarm with LSTM model." Computer Communications 211 (2023): 37-45.

[23] Eliyas, Sherin, Angeline Benitta, and Sathish Kumar. "Kidney Stone Prediction Using Neural Network Classifier." 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2022.

[24] Eliyas, S., & Ranjana, P. (2024). A Potent View on the Effects of E-Learning. International Journal of Grid and High Performance Computing (IJGHPC), 16(1), 1-10.

[25] Eliyas, S., & Ranjana, P. (2022, December). Improving Continuous Intention in E-Learning through the use of a Combined Gamification Model. In 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET) (pp. 1-4). IEEE.

[26] Venkatachalam Janani and Chandrabose Shanthi. " Optimizing Region Detection in Enhanced Infrared Images Using Deep Learning.." Revue d'Intelligence Artificielle , Volume 37,No 4, (2023).

[27] Janani V; Dinakaran M. " Infrared image enhancement techniques-A review" Second International Conference on Current Trends In Engineering and Technology-ICCTET 2014, Pages 167-173,IEEE, 2014.

[28] Janani V,Shanthi C. " Human-Animal Conflict Analysis and Management-A Critical Survey." 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART). Pages 1003-1007, IEEE, 2022.

[29] H. Xu, "Enhancing Recommender Systems with NLP-based Biased Singular Value Decomposition," 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS), Chengdu, China, 2023, pp. 808-811, doi: 10.1109/ISCTIS58954.2023.10213075.

[30] B. Lalitha, S. Kadiyam, R. V. Kalidindi, S. M. Vemparala, K. Yarlagadda and S. V. Chekuri, "Applicant Screening System Using NLP," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 379-383, doi: 10.1109/ICIDCA56705.2023.10099953.

[31] A. Ferrari, L. Zhao and W. Alhoshan, "NLP for Requirements Engineering: Tasks, Techniques, Tools, and Technologies," 2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), Madrid, ES, 2021, pp. 322-323, doi: 10.1109/ICSE-Companion52605.2021.00137.

[32] R. Krasniqi and H. Do, "Generalizability of NLP-based Models for Modern Software Development Cross-Domain Environments," 2023 IEEE/ACM 2nd International Workshop on Natural Language-Based Software Engineering (NLBSE), Melbourne, Australia, 2023, pp. 11-13, doi: 10.1109/NLBSE59153.2023.00009.

[33] A. Lertpiya et al., "A Preliminary Study on Fundamental Thai NLP Tasks for User-generated Web Content," 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 2018, pp. 1-8, doi: 10.1109/iSAI-NLP.2018.8692946.

[34] K. Kaushik, S. A. Yadav, V. Chauhan and A. Rana, "An Approach for Implementing Comprehensive Reconnaissance for Bug Bounty Hunters," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 189-193, doi: 10.1109/IC3I56241.2022.10072942.

[35] R. Anuradha, S. B, A. Nagpal, P. Chaturvedi, R. Kalra and A. A. Alwan, "Deep Learning for Anomaly Detection in Large-Scale Industrial Data," 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gautam Buddha Nagar, India, 2023, pp. 1551-1556, doi: 10.1109/UPCON59197.2023.10434613.

[36] P. K. Kushwaha, A. Rana, S. Srivastava, A. Saifi, A. Tavish and P. Chaturvedi, "Employee Absenteeism Prediction Using Machine Learning," 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gautam Buddha Nagar, India, 2023, pp. 116-121, doi: 10.1109/UPCON59197.2023.10434342.

[37] V. Jadeja, A. L. N. Rao, A. Srivastava, S. Singh, P. Chaturvedi and G. Bhardwaj, "Convolutional Neural Networks: A Comprehensive Review of Architectures and Application," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, 2023, pp. 460-467, doi: 10.1109/IC3I59117.2023.10397695.

[38] S. Gupta, S. Thakur and A. Gupta, "Optimized Feature Selection Approach for Smartphone Based Diabetic Retinopathy Detection," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, pp. 350-355, doi: 10.1109/ICIPTM54933.2022.9754021.