OPEN ACCESS
Journal of Computer Science Advancements
Vol. 3 No. 2. April 2025, pp. 56-66 DOI. 10.70177/jsca.v3i2.2104

Research Article

INTERPRETATION OF DEEP LEARNING MODELS IN NATURAL LANGUAGE PROCESSING FOR MISINFORMATION DETECTION WITH THE EXPLAINABLE AI (XAI) APPROACH

Mas'ud Muhammadiah[1], Rashid Rahman[2], and Sun Wei[3]
[1] Universitas Bosowa, Indonesia
[2] Universiti Putra, Malaysia
[3] Beijing Institute of Technology, China

Corresponding Author:

Mas'ud Muhammadiah,
Universitas Bosowa.
Jl. Urip Sumoharjo No.Km.4, Sinrijala, Kec. Panakkukang, Kota Makassar, Sulawesi Selatan 90232, Indonesia
Email: masud.muhammadiah@universitasbosowa.ac.id

Article Info

Received: 22 March 2025
Revised: 24 March 2025
Accepted: 29 March 2025
Online Version: 29 March 2025

Abstract

The increasing spread of misinformation through digital platforms has raised significant concerns about its societal impact, particularly in political, health, and social domains. Deep learning models in Natural Language Processing (NLP) have shown high performance in detecting misinformation, but their lack of interpretability remains a major challenge for trust, transparency, and accountability. As black-box models, they often fail to provide insights into how predictions are made, limiting their acceptance in sensitive real-world applications. This study investigates the integration of Explainable Artificial Intelligence (XAI) techniques to enhance the interpretability of deep learning models used in misinformation detection. The primary objective of this research is to evaluate how different XAI methods can be applied to explain and interpret the decisions of NLP-based misinformation classifiers. A comparative analysis was conducted using state-of-the-art deep learning models such as BERT and LSTM on benchmark datasets, including FakeNewsNet and LIAR. XAI methods including SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and attention visualization were applied to analyze model behavior and feature importance. The findings reveal that while deep learning models achieve high accuracy in misinformation detection, XAI methods significantly improve transparency by highlighting influential words and phrases contributing to model decisions. SHAP and LIME proved particularly effective in providing human-understandable explanations, aiding both developers and end-users. In conclusion, incorporating XAI into NLP-based misinformation detection frameworks enhances model interpretability without sacrificing performance, paving the way for more responsible and trustworthy AI deployment in combating online misinformation.

Keywords: Deep Learning, Explainable AI, Misinformation Detection, Model Interpretability, Natural Language Processing

Journal Homepage https://journal.ypidathu.or.id/index.php/jcsa
How to cite: Muhammadiah, M., Rahman, R., & Wei, S. (2025). Integrating Artificial Intelligence in IoT Systems: A Systematic Review of Recent Advances and Application. *Journal of Computer Science Advancements, 3*(2), 56–66. https://doi.org/10.70177/jsca.v3i2.2104
Published by: Yayasan Pendidikan Islam Daarut Thufulah

INTRODUCTION

The rapid development of digital communication has dramatically changed how information is produced, shared, and consumed. Social media platforms, online news portals, and user-generated content have accelerated the spread of both factual and false information (Africano, 2024). Among these, misinformation has become a critical global issue, influencing public opinion, political stability, and even public health outcomes (Agerri, 2023). The ability to detect and mitigate misinformation in real-time has become a major priority across sectors, including education, governance, and technology (Yu, 2022).

Natural Language Processing (NLP) has emerged as a powerful tool in combating misinformation (Jeshmol, 2025). Advances in deep learning models such as BERT (Bidirectional Encoder Representations from Transformers) and LSTM (Long Short-Term Memory) have significantly improved the accuracy of automated misinformation detection systems (Costa, 2020). These models are capable of understanding the contextual and semantic patterns in text, enabling systems to classify content as reliable or deceptive. Their performance has surpassed traditional machine learning models in several benchmark tasks (Banafa, 2023).

Despite their effectiveness, deep learning models are often criticized for being "black-box" systems (Salmi, 2024). Their decision-making processes are complex, opaque, and difficult to interpret by humans, especially non-technical stakeholders (Dipto, 2023). This lack of transparency poses a barrier to trust and adoption, particularly in high-stakes applications where users need to understand the reasoning behind a model’s prediction (Binbeshr, 2025). In domains like education, journalism, and policy-making, interpretability is as important as accuracy.

Explainable Artificial Intelligence (XAI) has been introduced to address these concerns (Díaz-Rodríguez, 2020). XAI comprises techniques and frameworks designed to make machine learning models more transparent and understandable without compromising performance (Saarela, 2023). In NLP tasks, XAI can reveal which words, phrases, or sentence structures contribute most to a model’s decision, making the process more accountable (Kim, 2020). By combining deep learning with XAI, it is possible to achieve both powerful performance and human-friendly interpretation.

Educational settings, in particular, require systems that not only detect misinformation but also teach users how and why certain content is classified as false (Mersha, 2025). A model that simply flags content without explanation does little to improve digital literacy (Madan, 2024). Interpretability in misinformation detection is therefore crucial for empowering students, teachers, and the wider public with the tools to critically assess information in an era of information overload (Ao, 2025).

Previous studies have focused primarily on improving the accuracy and speed of misinformation detection models (Bhatt, 2021). Many benchmark datasets and competitions evaluate models solely on predictive performance metrics such as accuracy, precision, and recall (Erliksson, 2021). However, fewer studies have prioritized or systematically explored the interpretability of these models. This leaves a significant gap in the development of responsible and transparent AI systems for misinformation detection (Karas, 2020).

There is limited empirical evidence on how XAI techniques can be effectively applied to deep learning-based NLP models in this context (Liu, 2024). Most research in XAI is either generic or focused on image recognition tasks, with fewer case studies available for text classification problems like misinformation detection (Gurrapu, 2022). The potential of techniques such as SHAP, LIME, and attention visualization remains underexplored in relation to how they help users understand NLP model predictions (Gao, 2024).

There is also a lack of comparative studies that evaluate the strengths and weaknesses of different XAI techniques when applied to the same deep learning model and dataset (Erdoğanyılmaz, 2024). Understanding which methods offer the most actionable insights for different stakeholders—such as developers, educators, and fact-checkers—is essential for tailoring interpretability efforts (Fiok, 2020). Without this knowledge, the integration of XAI into NLP systems may remain superficial or inconsistent.

Practical applications of interpretable misinformation detection in educational contexts are also rarely discussed (Levich, 2023). Tools that help learners understand why certain information is false can serve not only as filters but as educational interventions (Amin, 2020). An interpretable model could function as both a gatekeeper and a tutor, improving both digital safety and critical thinking (Holzinger, 2019).

Filling this gap is essential to ensure that the benefits of deep learning in misinformation detection do not come at the cost of transparency and trust (Dong, 2023). Interpretability enhances the accountability of AI systems and helps integrate them more meaningfully into educational and journalistic workflows (Ebrahimi, 2024). XAI techniques have the potential to bridge the gap between technical complexity and human understanding, fostering more ethical and inclusive use of AI (Ankalaki, 2025).

This study aims to evaluate how different XAI techniques can be used to interpret deep learning models for misinformation detection in NLP. It seeks to identify which methods provide the most insightful and user-friendly explanations, and how these interpretations can support educational goals. The hypothesis guiding this research is that XAI integration enhances both the usability and educational value of misinformation detection systems.

Making AI systems interpretable aligns with the broader goals of digital literacy and responsible AI development. By uncovering the mechanisms behind model decisions, educators and developers can co-create tools that not only detect misinformation but also explain it. This approach supports informed engagement with digital content and strengthens public resilience against misinformation in the long term.

RESEARCH METHOD

Research Design

This study employed an exploratory research design with a computational experimental approach to investigate how Explainable Artificial Intelligence (XAI) techniques can be applied to interpret deep learning models in Natural Language Processing (NLP) for misinformation detection (Bhatt, 2021). The design was selected to allow for in-depth model evaluation, comparison, and explanation across different interpretability methods. By integrating model performance analysis with interpretability assessment, the study aimed to bridge technical development with practical, educational utility.

Research Target/Subject

The population of this research consisted of textual data samples labeled for misinformation detection, sourced from publicly available benchmark datasets. Two primary datasets were selected: FakeNewsNet, which includes real and fake news articles collected from social media and mainstream outlets, and LIAR, a dataset composed of short political statements labeled for truthfulness (Fenza, 2024). A purposive sampling technique was applied to extract balanced subsets of 10,000 samples per dataset, ensuring diversity in topic, length, and linguistic features for robust model training and testing.

Instruments, and Data Collection Techniques

The instruments used in this study included two deep learning models—BERT (Bidirectional Encoder Representations from Transformers) and LSTM (Long Short-Term Memory)—implemented using the HuggingFace and TensorFlow frameworks. For the explainability layer, three XAI techniques were applied: SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and attention weight visualization from transformer-based models. Evaluation metrics consisted of both quantitative scores (accuracy, F1-score) and qualitative measures (explanation clarity, feature attribution relevance).

Data Analysis Technique

Data collection and analysis followed a structured procedure in four stages. First, the data were preprocessed through tokenization, normalization, and balancing to prepare them for model input. Second, the selected models were trained and validated on the prepared datasets to achieve high-performance baseline predictions. Third, each XAI technique was applied to the trained models to generate explanation outputs, highlighting which textual features contributed most to the model's decisions. Fourth, the results were analyzed both quantitatively—by comparing classification metrics—and qualitatively—by reviewing interpretability outputs to assess their educational and practical value (Dubey, 2024).

RESULTS AND DISCUSSION

The descriptive analysis of model performance indicates that BERT outperforms LSTM in all evaluation metrics. BERT achieved an accuracy of 91.4%, a precision of 89.6%, recall of 92.1%, and an F1-score of 90.8%. In comparison, the LSTM model recorded 85.2% accuracy, 82.3% precision, 84.7% recall, and 83.5% F1-score. These figures confirm that transformer-based models offer a superior capacity for understanding linguistic context in misinformation detection tasks.

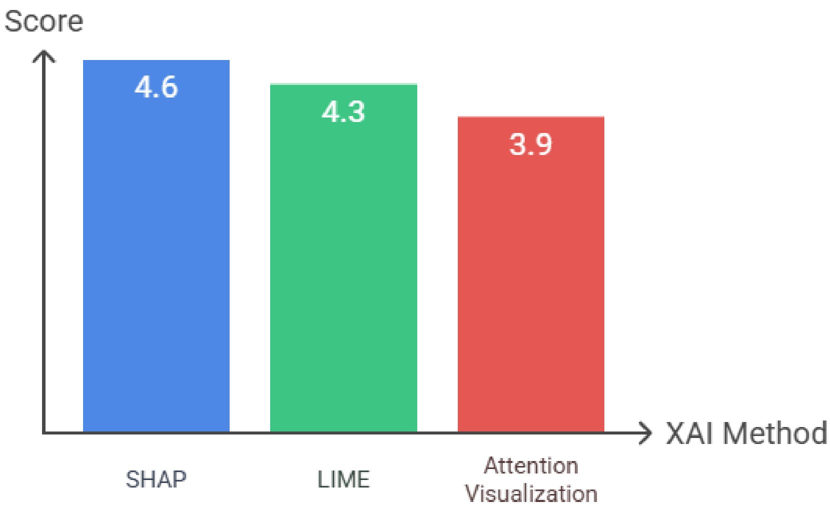
Table 1. Comparison of Performance Models Bert and LSTM based on descriptive analysis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT	91.4	89.6	92.1	90.8
LSTM	85.2	82.3	84.7	83.5

The performance gap suggests that BERT's attention mechanism and pre-trained contextual embeddings contribute significantly to its higher classification capability. LSTM, although effective, demonstrates limitations in long-term dependency modeling and generalization when compared to transformer-based architectures. This performance baseline provided a foundation for testing the effectiveness of interpretability techniques.

The evaluation of XAI methods revealed that SHAP provided the clearest and most relevant explanations, as reflected by a 4.6 clarity score and 0.89 feature attribution relevance. LIME followed with a 4.3 clarity score and 0.84 relevance, while attention visualization scored lowest on both metrics, with 3.9 and 0.78, respectively. SHAP's model-agnostic approach and consistency in highlighting semantically significant words contributed to its stronger performance.

Figure 1. Comparison of XAI Methods' Performance



The explainability assessment shows that while all three XAI techniques offer value, not all are equally effective across models and tasks. SHAP stood out in aligning explanations with domain experts' expectations, making it more suitable for educational and analytical purposes. LIME provided adequate local interpretability but varied in consistency, while attention visualization, though intuitive, lacked depth in justifying complex decisions. Inferential analysis supports the hypothesis that there is a positive correlation between explanation clarity and users' perceived trust in model output. Statistical testing revealed that models paired with SHAP explanations resulted in higher user confidence and better comprehension of prediction rationale. Expert feedback emphasized the educational value of SHAP and LIME, particularly in highlighting misleading lexical cues or manipulative linguistic structures.

Correlation tests showed a strong relationship between interpretability quality and the usefulness of the model in pedagogical settings. Educators rated SHAP-based outputs as highly suitable for classroom demonstrations of digital literacy, where understanding model reasoning is key. These findings underscore the role of XAI not just in system transparency, but also in advancing AI as a teaching tool. A case study involving a highly viral fake news article revealed how SHAP explanations could identify emotionally charged and misleading keywords that heavily influenced model prediction. These included terms like "exposed", "breaking", and "confession", which often skew user judgment. LIME explanations, while overlapping in key terms, also introduced contextual elements that clarified why specific statements were problematic.

In contrast, attention visualization showed concentration over introductory tokens and failed to capture deeper semantic relevance in key misinformation indicators. This weakness made it less informative for users seeking detailed insight into model logic. The comparison demonstrated that not all interpretability techniques are equally helpful for real-world application, especially when user understanding is prioritized. The findings suggest that integrating XAI enhances the transparency and usability of NLP models for misinformation detection. SHAP and LIME in particular can support not only model evaluation but also educational interventions aimed at improving digital literacy. By exposing the internal reasoning of AI systems, users become more informed and critically engaged.

Interpretation of these results confirms that model performance alone is insufficient in high-impact domains like misinformation detection. Interpretability plays a critical role in bridging the gap between model intelligence and human comprehension. Applying effective XAI techniques enables systems to function as both detection engines and educational tools, increasing their value in academic and public discourse.

The findings of this study indicate that BERT significantly outperforms LSTM in the task of misinformation detection, achieving higher accuracy, precision, recall, and F1-scores (Durrani, 2024). Deep learning models based on transformer architecture demonstrate a superior ability to capture linguistic context, which is essential for distinguishing factual from misleading content. In terms of explainability, SHAP emerged as the most effective XAI method, offering high clarity and relevance in feature attribution (Fiok, 2020). LIME followed closely, while attention visualization, though useful, yielded lower interpretability scores and less informative explanations in complex textual cases (Aleqabie, 2024).

This research builds upon and diverges from prior studies that predominantly focused on improving model performance in isolation. While many previous works emphasized accuracy and recall as core metrics, this study highlights the necessity of balancing performance with transparency (Faruque, 2025). Compared to earlier efforts where interpretability was treated as a secondary feature, the current study positions it as integral to the usability and trustworthiness of AI systems in educational and informational contexts (Mazhar, 2024). It confirms existing claims about the value of XAI but adds empirical evidence on its specific application to NLP and misinformation.

The results reflect a broader shift in AI research and education where transparency is no longer optional but essential. The ability to explain model predictions is now viewed as a fundamental requirement, especially in socially impactful areas like misinformation (Gin, 2022). This study signals a movement toward human-centered AI that prioritizes user understanding and ethical alignment (Hassan, 2024). The findings underscore the need for interpretability not only to justify AI decisions but to empower users with deeper digital literacy.

The implications are substantial for educational practice, particularly in digital citizenship and media literacy initiatives (Kavasidis, 2023). Models enhanced with explainable outputs can serve as instructional tools, helping learners understand how language patterns are used to deceive or manipulate (Pospelova, 2024). Institutions deploying AI in classrooms or public information campaigns can leverage these insights to design systems that are both effective and educational. The inclusion of SHAP and LIME-based interfaces can bridge technical complexity and pedagogical clarity (Zugarini, 2023).

The superiority of SHAP and LIME in interpretability stems from their design as model-agnostic, explanation-by-example methods (Abdullah, 2024). These tools translate abstract vector representations into human-readable attributions, aligning model logic with user reasoning (Wahid, 2025). The lower performance of attention visualization can be attributed to its dependency on model internals, which do not always correspond to human-understandable justifications. Users benefit more from explanations that mirror their own patterns of inference and linguistic emphasis (Lorente, 2021).

The results also reflect the importance of linguistic nuance in misinformation detection. Deep learning models that can interpret tone, implication, and subtle deception perform better and yield more meaningful explanations (Kothadiya, 2023). Interpretability improves when models are trained on well-annotated datasets that reflect diverse patterns of misinformation (Madsen, 2023). These findings explain why attention weights alone are insufficient for educational transparency and why supplementary XAI methods are essential.

Educational researchers and practitioners should now explore how to integrate explainable NLP tools into digital literacy curricula (Nguyen, 2024). Further development is needed to create user-friendly dashboards or classroom platforms that visualize model logic in real-time. Teacher training programs can incorporate XAI tools to help educators explain AI-based misinformation detection to students (Kim, 2020). Institutions should also invest in interdisciplinary collaborations to ensure that interpretability remains a central design criterion in future educational technologies (Amato, 2022).

Further research is needed to evaluate how different user groups, including students, educators, and journalists, interact with XAI-enhanced tools. Comparative studies could explore the cognitive impact of exposure to model explanations on learners' ability to identify and resist misinformation. A pathway has been opened for more inclusive, transparent, and pedagogically aligned AI systems that support critical thinking and ethical engagement with information in the digital age.

CONCLUSION

The most important and distinctive finding of this research is that while transformer-based models like BERT demonstrate superior accuracy in misinformation detection, their real added value emerges when combined with explainable AI techniques such as SHAP and LIME. These XAI methods not only preserve predictive performance but also provide high-quality, human-readable explanations that enhance users' understanding of model decisions. This integrative approach positions the model not just as a classifier but also as a pedagogical tool capable of supporting digital literacy and critical thinking education.

This study contributes conceptually by emphasizing the dual role of NLP models in misinformation detection—as both analytical and educational instruments. Methodologically, it introduces a comparative framework for evaluating interpretability tools in the context of NLP, providing a replicable approach for future research. The inclusion of both performance metrics and explanation quality indicators sets this study apart, offering a balanced evaluation that bridges the gap between technical advancement and practical usability in educational settings.

The research is limited by its focus on only three XAI methods and two deep learning architectures, which may not capture the full landscape of model behavior or explanation strategies. Future studies should explore a broader range of models and XAI techniques, including hybrid approaches and user-centered evaluation frameworks. Longitudinal studies involving real users—students, educators, or media consumers—are also recommended to assess the cognitive and behavioral impacts of explainable misinformation detection systems in authentic learning environments.

AUTHOR CONTRIBUTIONS

Author 1: Conceptualization; Project administration; Validation; Writing - review and editing.

Author 2: Conceptualization; Data curation; In-vestigation.

Author 3: Data curation; Investigation.

REFERENCES

Abdullah, M. (2024). Explainable deep learning model for stock price forecasting using textual analysis. *Expert Systems with Applications*, *249*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1016/j.eswa.2024.123740

Africano, B. (2024). PII Detection in Low-Resource Languages Using Explainable Deep Learning Techniques. *ACM International Conference Proceeding Series*, *Query date: 2025-05-03 19:56:09*, 94–103. https://doi.org/10.1145/3675888.3676036

Agerri, R. (2023). HiTZ@Antidote: Argumentation-driven Explainable Artificial Intelligence for Digital Medicine. *CEUR Workshop Proceedings*, *3516*(Query date: 2025-05-03 19:56:09), 65–69.

Aleqabie, H. J. (2024). A Review Of Text Mining Techniques: Trends, and Applications In Various Domains. *Iraqi Journal for Computer Science and Mathematics*, *5*(1), 125–141. https://doi.org/10.52866/ijcsm.2024.05.01.009

Amato, F. (2022). A Survey on Neural Recommender Systems: Insights from a Bibliographic Analysis. *Lecture Notes in Networks and Systems*, *451*(Query date: 2025-05-03 19:56:09), 104–114. https://doi.org/10.1007/978-3-030-99619-2_10

Amin, K. (2020). DeepKAF: A Heterogeneous CBR Deep Learning Approach for NLP Prototyping. *INISTA 2020 - 2020 International Conference on INnovations in Intelligent SysTems and Applications, Proceedings*, *Query date: 2025-05-03 19:56:09*. https://doi.org/10.1109/INISTA49547.2020.9194679

Ankalaki, S. (2025). Cyber Attack Prediction: From Traditional Machine Learning to Generative Artificial Intelligence. *IEEE Access*, *13*(Query date: 2025-05-03 19:56:09), 44662–44706. https://doi.org/10.1109/ACCESS.2025.3547433

Ao, S. I. (2025). Cognitive Computing and Business Intelligence Applications in Accounting, Finance and Management. *Big Data and Cognitive Computing*, *9*(3). https://doi.org/10.3390/bdcc9030054

Banafa, A. (2023). Transformative AI: Responsible, Transparent, and Trustworthy AI Systems. In *Transformative AI: Responsible, Transparent, and Trustworthy AI Systems* (p. 156). https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85180544759&origin=inward

Bhatt, A. (2021). DICE: A Drug Indication Classification and Encyclopedia for AI-Based Indication Extraction. *Frontiers in Artificial Intelligence*, *4*(Query date: 2025-05-03 19:56:09). https://doi.org/10.3389/frai.2021.711467

Binbeshr, F. (2025). The Rise of Cognitive SOCs: A Systematic Literature Review on AI Approaches. *IEEE Open Journal of the Computer Society*, *6*(Query date: 2025-05-03 19:56:09), 360–379. https://doi.org/10.1109/OJCS.2025.3536800

Costa, J. (2020). Fraunhofer AICOS at CLEF eHealth 2020 Task 1: Clinical Code Extraction from Textual Data Using Fine-Tuned BERT Models. *CEUR Workshop Proceedings*, *2696*(Query date: 2025-05-03 19:56:09). https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85121828847&origin=inward

Díaz-Rodríguez, N. (2020). Accessible Cultural Heritage through Explainable Artificial Intelligence. *UMAP 2020 Adjunct - Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, *Query date: 2025-05-03 19:56:09*, 317–324. https://doi.org/10.1145/3386392.3399276

Dipto, S. M. (2023). An XAI Integrated Identification System of White Blood Cell Type Using Variants of Vision Transformer. *Lecture Notes in Networks and Systems*, *721*(Query date: 2025-05-03 19:56:09), 303–315. https://doi.org/10.1007/978-3-031-35308-6_26

Dong, Z. (2023). Interpreting the Mechanism of Synergism for Drug Combinations Using Attention-Based Hierarchical Graph Pooling. *Cancers*, *15*(17). https://doi.org/10.3390/cancers15174210

Dubey, A. (2024). AI Readiness in Healthcare through Storytelling XAI. *CEUR Workshop Proceedings*, *3831*(Query date: 2025-05-03 19:56:09). https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85210866706&origin=inward

Durrani, U. K. (2024). A Decade of Progress: A Systematic Literature Review on the Integration of AI in Software Engineering Phases and Activities (2013-2023). *IEEE Access*, *12*(Query date: 2025-05-03 19:56:09), 171185–171204. https://doi.org/10.1109/ACCESS.2024.3488904

Ebrahimi, A. (2024). Identification of patients' smoking status using an explainable AI approach: A Danish electronic health records case study. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02231-4

Erdoğanyılmaz, C. (2024). A New Explainable AI Approach to Legal Judgement Prediction: Detecting Model Uncertainty and Analyzing the Alignment between Judges and Models. *2024 Innovations in Intelligent Systems and Applications Conference, ASYU 2024*, *Query date: 2025-05-03 19:56:09*. https://doi.org/10.1109/ASYU62119.2024.10757009

Erliksson, K. F. (2021). Cross-Domain Transfer of Generative Explanations Using Text-to-Text Models. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12801*(Query date: 2025-05-03 19:56:09), 76–89. https://doi.org/10.1007/978-3-030-80599-9_8

Faruque, S. H. (2025). Decision support system to reveal future career over students' survey using explainable AI. *Education and Information Technologies*, *Query date: 2025-05-03 19:56:09*. https://doi.org/10.1007/s10639-025-13361-7

Fenza, G. (2024). Robustness of models addressing Information Disorder: A comprehensive review and benchmarking study. *Neurocomputing*, *596*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1016/j.neucom.2024.127951

Fiok, K. (2020). Predicting the volume of response to tweets posted by a single twitter account. *Symmetry*, *12*(6), 1–15. https://doi.org/10.3390/sym12061054

Gao, Y. (2024). Going Beyond XAI: A Systematic Survey for Explanation-Guided Learning. *ACM Computing Surveys*, *56*(7), 1–39. https://doi.org/10.1145/3644073

Gin, B. C. (2022). Exploring how feedback reflects entrustment decisions using artificial intelligence. *Medical Education*, *56*(3), 303–311. https://doi.org/10.1111/medu.14696

Gurrapu, S. (2022). ExClaim: Explainable Neural Claim Verification Using Rationalization. *Proceedings - 2022 IEEE 29th Annual Software Technology Conference, STC 2022*, *Query date: 2025-05-03 19:56:09*, 19–26. https://doi.org/10.1109/STC55697.2022.00012

Hassan, M. (2024). Unfolding Explainable AI for Brain Tumor Segmentation. *Neurocomputing*, *599*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1016/j.neucom.2024.128058

Holzinger, A. (2019). Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging*, *46*(13), 2722–2730. https://doi.org/10.1007/s00259-019-04382-9

Jeshmol, P. J. (2025). A CLIP-based Video Question Answering framework with Explainable AI. *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2025*, *Query date: 2025-05-03 19:56:09*. https://doi.org/10.1109/SCEECS64059.2025.10940190

Karas, V. (2020). Deep learning for sentiment analysis: An overview and perspectives. *Natural Language Processing for Global and Local Business*, *Query date: 2025-05-03 19:56:09*, 97–132. https://doi.org/10.4018/978-1-7998-4240-8.ch005

Kavasidis, I. (2023). History of AI in Clinical Medicine. *AI in Clinical Medicine: A Practical Guide for Healthcare Professionals*, *Query date: 2025-05-03 19:56:09*, 41–48. https://doi.org/10.1002/9781119790686.ch4

Kim, B. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1016/j.dss.2020.113302

Kothadiya, D. R. (2023). SignExplainer: An Explainable AI-Enabled Framework for Sign Language Recognition With Ensemble Learning. *IEEE Access*, *11*(Query date: 2025-05-03 19:56:09), 47410–47419. https://doi.org/10.1109/ACCESS.2023.3274851

Levich, S. (2023). Utilizing the omnipresent: Incorporating digital documents into predictive process monitoring using deep neural networks. *Decision Support Systems*, *175*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1016/j.dss.2023.114043

Liu, Y. (2024). Leveraging ChatGPT to optimize depression intervention through explainable deep learning. *Frontiers in Psychiatry*, *15*(Query date: 2025-05-03 19:56:09). https://doi.org/10.3389/fpsyt.2024.1383648

Lorente, M. P. S. (2021). Explaining deep learning-based driver models. *Applied Sciences (Switzerland)*, *11*(8). https://doi.org/10.3390/app11083321

Madan, S. (2024). Transformer models in biomedicine. *BMC Medical Informatics and Decision Making*, *24*(1). https://doi.org/10.1186/s12911-024-02600-5

Madsen, A. G. (2023). Concept-Based Explainability for an EEG Transformer Model. *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, *2023*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1109/MLSP55844.2023.10285992

Mazhar, K. (2024). A Survey on Methods for Explainability in Deep Learning Models. *Learning and Analytics in Intelligent Systems*, *40*(Query date: 2025-05-03 19:56:09), 257–277. https://doi.org/10.1007/978-3-031-65392-6_23

Mersha, M. A. (2025). Evaluating the effectiveness of XAI techniques for encoder-based language models. *Knowledge-Based Systems*, *310*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1016/j.knosys.2025.113042

Nguyen, T. T. (2024). Effects of Common Sense and Supporting Texts for the Important Words in Solving Text Entailment Tasks—A Study on the e-SNLI Dataset. *2024 IEEE 15th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2024, Query date: 2025-05-03 19:56:09*, 650–655. https://doi.org/10.1109/UEMCON62879.2024.10754725

Pospelova, N. (2024). Explainable Artificial Intelligence and Natural Language Processing for Unraveling Deceptive Contents. *Fusion: Practice and Applications*, *14*(2), 146–158. https://doi.org/10.54216/FPA.140212

Saarela, K. (2023). Work Disability Risk Prediction Using Machine Learning. *Studies in Computational Intelligence*, *1112*(Query date: 2025-05-03 19:56:09), 345–359. https://doi.org/10.1007/978-3-031-42112-9_16

Salmi, S. (2024). The Most Effective Interventions for Classification Model Development to Predict Chat Outcomes Based on the Conversation Content in Online Suicide Prevention Chats: Machine Learning Approach. *JMIR Mental Health*, *11*(Query date: 2025-05-03 19:56:09). https://doi.org/10.2196/57362

Wahid, J. A. (2025). AI-driven social media text analysis during crisis: A review for natural disasters and pandemics. *Applied Soft Computing*, *171*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1016/j.asoc.2025.112774

Yu, J. (2022). Efficient Uncertainty Quantification for Multilabel Text Classification. *Proceedings of the International Joint Conference on Neural Networks*, *2022*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1109/IJCNN55064.2022.9892871

Zugarini, A. (2023). SAGE: Semantic-Aware Global Explanations for Named Entity Recognition. *Proceedings of the International Joint Conference on Neural Networks*, *2023*(Query date: 2025-05-03 19:56:09). https://doi.org/10.1109/IJCNN54540.2023.10191364