Journal of Theoretical and Applied Information Technology
15th February 2025. Vol.103. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EXPLAINABLE MACHINE LEARNING FOR TEXT CLASSIFICATION: A NOVEL APPROACH TO TRANSPARENCY AND INTERPRETABILITY

Dr. MOHAMMED ABDUL WAJEED[1], Dr. KHAJA MIZBAHUDDIN QUADRY[2], Dr. MOKSUD ALAM MALLIK[3], Dr. K. RAJESH KHANNA[4]

[1]Professor, Lords Institute of Engineering and Technology, Computer Science & Engineering Department, Hyderabad, Telangana, India
[2]Associate Professor, Lords Institute of Engineering and Technology, Computer Science & Engineering Department, Hyderabad, Telangana, India
[3]Associate Professor, Lords Institute of Engineering and Technology, Computer Science & Engineering-Data Science Department, Hyderabad, Telangana, India
[4]Assistant Professor, Vaagdevi College of Engineering, Computer Science & Engineering Department, Warangal, Telangana, India
E-mail: [1]drwajeed1@gmail.com

ABSTRACT

The increasing use of machine learning (ML) models for text classification has sparked debates about their interpretability and transparency. Despite their exceptional performance in applications like sentiment analysis, spam detection, and topic categorization, the often-obscure nature of ML models presents challenges for users seeking to understand the decision-making process. Explainable Machine Learning (XML) aims to tackle these issues by offering human-comprehensible explanations of model outputs, thereby building trust and enabling improved decision-making in critical sectors such as healthcare, legal, and finance. This study offers a thorough examination of XML techniques applied to text classification models, encompassing attention mechanisms, feature importance methods, and post-hoc interpretability frameworks like LIME and SHAP. Furthermore, we assess these techniques in terms of their capacity to enhance model transparency, weighing the trade-offs between interpretability and performance. The research also delves into the future trajectory of XML in text classification, including the incorporation of user-focused explanations and regulatory requirements for AI transparency. Our results indicate that explainability not only enhances model trustworthiness but also aids in identifying model biases and enhancing overall performance.

Keywords: *Five Machine learning, Text Classification, Explainable Machine Learning. LIME, SHAP etc.*

1. INTRODUCTION

In the field of natural language processing (NLP), text classification is a fundamental task that involves assigning predefined categories to textual information. This technology has widespread applications, ranging from spam email detection and social media sentiment analysis to legal document sorting [11]. Recent progress in machine learning, especially deep learning, has substantially improved the precision and speed of text classification systems. Nevertheless, many of these models, particularly intricate structures like deep neural networks, are often considered "black boxes," with decision-making processes that are not easily understood by humans.

As machine learning models increasingly play a role in crucial decision-making across various sectors including healthcare, finance, and law, the demand for interpretability and transparency has grown significantly. Regulatory measures such as the European Union's General Data Protection Regulation (GDPR) stress the importance of the "right to explanation," encouraging organizations to provide clarity on how their AI systems arrive at decisions. In response to these concerns, Explainable Machine Learning (XML) has emerged as a promising solution, aiming to make machine learning models more comprehensible and provide users with clear insights into their predictions.

The importance of explainability in machine learning for text classification is multifaceted.

Firstly, comprehending how a model processes and categorizes text can help identify and address biases, particularly in sensitive areas where training data may contain historical or social prejudices. For example, models trained on biased text collections may inadvertently perpetuate or even intensify discrimination, leading to unfair outcomes in areas such as credit approval or hiring processes. Secondly, explainability builds confidence in AI systems by enabling users to confirm that a model's predictions are consistent with logical reasoning or domain expertise. This transparency is crucial for gaining the trust of non-technical stakeholders who depend on these systems.

In recent times, numerous methods for explaining machine learning models have emerged. These approaches can be divided into two main groups: intrinsic and post-hoc techniques. Intrinsic methods focus on creating models that are inherently easy to interpret, such as decision trees or rule-based systems. While these offer transparencies, they often sacrifice performance in complex tasks. On the other hand, post-hoc techniques aim to elucidate the decisions of any model, including sophisticated neural networks, without modifying their structure. Some well-known post-hoc methods include Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and attention mechanisms in neural networks.

Although progress has been made, text classification poses unique challenges for explainability compared to other areas of machine learning. The interpretation of textual data is particularly complex due to its high dimensionality and the ambiguous nature of natural language. Words and phrases often have subtle meanings that can change based on context, making it challenging to precisely explain a model's classification decisions. Furthermore, researchers often face a dilemma between model accuracy and interpretability: highly interpretable models may underperform, while highly accurate models may be more difficult to explain.

This study aims to provide a thorough review of explainability techniques in text classification. We explore various methods, including attention mechanisms that emphasize important words or tokens, feature attribution techniques such as LIME and SHAP, and other model-agnostic approaches. We also examine the trade-offs between interpretability and performance, as well as the broader implications of explainability for ethical AI and bias mitigation. Lastly, we investigate future directions for XML in text classification, with a focus on enhancing the user-centricity of explanations and addressing emerging legal and ethical requirements.

The structure of this paper is as follows: Section 2 reviews related work on explainable machine learning and text classification. Section 3 outlines the various explainability methods used in text classification tasks. Section 4 presents an experimental evaluation of these methods. Section 5 discusses the results and their implications for model interpretability, and Section 6 concludes the paper with potential future directions.

2. RELATED WORK

In recent years, the field of explainable machine learning (XML) has received significant interest, primarily due to the increasing demand for transparency and reliability in artificial intelligence systems. This segment examines notable research in explainability for text classification, organizing the relevant literature into three categories: inherent explainability, post-hoc explanations, and domain-specific investigations.

2.1 Intrinsic Explainability

Sections Models designed to be inherently interpretable, offering transparency without additional explanatory techniques, are referred to as intrinsically explainable [8]. Traditionally, more straightforward models like decision trees, rule-based classifiers, and linear models have been preferred for their transparency. These models provide clarity on feature importance and decision-making processes, making them naturally easier to understand.

For example, traditional models such as decision trees [13][12] and logistic regression [14] enable users to follow the decision-making process by examining tree divisions or the importance assigned to various features. While these models have been successfully applied to text classification tasks, their performance often falls short of more sophisticated models, like deep neural networks, particularly when handling large-scale, high-dimensional datasets common in NLP.

Recent efforts have aimed to improve interpretability in deep learning models. For instance, attention mechanisms [15] in recurrent neural networks (RNNs) and transformers [3][4] have offered a partial solution by emphasizing the most significant tokens or sentences influencing the

model's decision. The attention mechanism has been widely employed in text classification to provide human-comprehensible explanations of which input text portions are driving the model's classification [5]. However, despite attention providing useful insights, some researchers contend that it may not always correspond to true causal importance, raising concerns about its reliability as an explanation.

2.2 Post-hoc Explainability

Techniques for post-hoc explainability seek to elucidate complex, opaque models without modifying their internal workings. These approaches are model-independent and can be utilized for any classifier, regardless of its intricacy, making them especially useful for interpreting deep learning models.

LIME (Local Interpretable Model-Agnostic Explanations) is one of the most commonly employed post-hoc methods [1]. It creates local approximations of a model's decision by altering input data and constructing simpler, interpretable models (like linear classifiers) to clarify specific predictions. LIME has been effectively used in text classification tasks to identify key words or phrases that significantly impact a model's prediction. However, it faces challenges in terms of scalability and consistency, as it may yield different explanations for slightly varied inputs.

Another notable post-hoc technique, SHAP (SHapley Additive exPlanations) [2] offers consistent and theoretically sound explanations rooted in game theory. SHAP values assign a prediction to each feature's contribution, ensuring equitable distribution of feature importance. In text classification, SHAP has been used to explain how individual words or n-grams influence the model's output, providing an interpretable and robust framework. While SHAP delivers global explanations that remain consistent across inputs, unlike LIME, it comes with higher computational costs, particularly when applied to large, high-dimensional datasets.

Additional post-hoc methods include gradient-based approaches such as Grad-CAM [16] and Layer-wise Relevance Propagation (LRP) . These techniques have been adapted from computer vision to NLP tasks, offering visualizations that show which text segments contribute most significantly to the final classification [9]. By utilizing model gradients or activations to trace back from a prediction to the input features that had the greatest impact, these approaches are particularly valuable for understanding the internal mechanisms of deep neural networks.

2.3 Application-Specific Explainability in Text Classification

Numerous research efforts have concentrated on implementing explainability techniques in specific text classification domains, including healthcare, law, and social media analysis, where interpretability is crucial [10][6].

In the medical field, interpretable models are gaining traction to enhance clinical decision support systems. utilized LIME and SHAP to elucidate how patient records and clinical notes influence diagnoses and treatment suggestions in medical text classification models [7]. This approach is vital in healthcare environments, as comprehending model reasoning can enhance physician confidence and minimize the risk of erroneous or prejudiced decisions.

Within the legal sphere, [17] introduced a framework for explicable legal judgment prediction, employing attention mechanisms and post-hoc interpretability tools to ensure model predictions align with legal rationale. Given the high-stakes nature of legal rulings, explainability is essential to validate model outputs in accordance with legal principles.

Explainability has also been instrumental in uncovering bias in social media text classification. For example, [18] employed SHAP to examine how hate speech detection models evaluate various words or phrases, exposing biases that result in disproportionate misclassification of certain demographic groups. By utilizing explainable models, these investigators were able to identify and address sources of bias in both the training data and model architecture.

2.4 Challenges and Limitations

Despite the development of various methods for explainable text classification, several issues persist. A key challenge is striking a balance between model performance and interpretability [19][20]. While straightforward, easily understood models often fail to deliver satisfactory results on

complex datasets, cutting-edge deep learning approaches typically lack transparency. Scientists continue to investigate ways to address this disparity, but discovering an ideal equilibrium remains an unresolved issue.

Furthermore, the comprehensibility of explanations is subjective and can differ among users. For instance, a technical specialist might find detailed explanations based on gradients valuable, whereas a layperson may prefer simpler, rule-based or visual explanations. Consequently, user-focused explainability, which involves customizing explanations for different audiences, has emerged as a growing area of research interest.

3. PROPOSED WORK

3.1 Proposed Approach: Context Aware Explanation Networks (CAEN)

We introduce a new framework called Context-Aware Explanation Networks (CAEN) to tackle the shortcomings of existing explainability methods in text classification. These shortcomings include inconsistent explanations, contextual deficiency, and challenges in balancing model accuracy with interpretability. CAEN's fundamental approach is to generate flexible, context-sensitive explanations that not only emphasize crucial elements influencing a model's decision but also take into account the broader textual context and user requirements.

CAEN combines built-in and post-analysis explainability techniques into a single framework, optimized for use with deep learning models in text classification. The system introduces two novel features:

Contextually-Aware Attention System: In contrast to conventional attention mechanisms that concentrate on single words or phrases, CAEN utilizes a context-sensitive attention system. This approach evaluates not just the significance of individual words, but also their semantic environment. As a result, the model can deliver more comprehensive and nuanced explanations by considering how word importance fluctuates based on the textual context.

Customizable Explanation Components: The second innovation involves developing explanation components that can be tailored to various user types. These components are produced by an auxiliary network, trained simultaneously with the classification model. This network learns to generate explanations customized for specific audiences, such as technical experts or non-specialists. The network modifies the detail level and structure of the explanation according to user preferences or needs, ensuring the model remains interpretable and accessible to a wide range of stakeholders.

3.2 Components of CAEN

Conventional attention mechanisms in text classification allocate significance scores to individual tokens, typically based on their direct impact on classification results. However, these methods fail to consider the intricate interplay between words and their surrounding context. In response, CAEN introduces an improvement that captures contextual relationships by utilizing context-aware embeddings.

These context-aware embeddings are generated using a blend of transformer-based encoders (such as BERT or RoBERTa) and sentence-level context windows that capture far-reaching dependencies among words. A dynamic attention layer then processes these embeddings, adjusting token importance scores based on both their individual significance and their relevance to the broader sentence or document context.

To illustrate, in a sentiment analysis task, the word "not" within the phrase "not happy" would receive high importance due to its negating effect on "happy." While traditional attention models might overlook the full interaction between these words, the context-aware attention mechanism ensures that their relationship is properly considered when generating explanations.

A significant hurdle in interpretable machine learning is accommodating the diverse interpretability requirements of different users. While technically proficient individuals may prefer in-depth, quantitative explanations, those without technical expertise might find simplified, visual representations more beneficial. To tackle this issue, CAEN incorporates an auxiliary explanation network that adapts to generate user-specific interpretations.

The explanation network is composed of three key components:

User Profiling Component: This element collects data about the user, including their technical knowledge, desired explanation depth, and favored

explanation style (such as visual, textual, or quantitative). It can be manually set up or deduced from user interactions over time.

Explanation Generation Component: This module tailors the model's explanations based on the user profile. For technically adept users, it might produce detailed, gradient-based explanations that illustrate how each word or phrase contributes to the model's decision. For non-technical users, it might offer high-level interpretations using plain language or graphical emphasis to highlight the most crucial parts of the text.

Evaluation and Feedback Component: This element facilitates ongoing enhancement of explanation quality. It gathers user feedback on provided explanations and adjusts future interpretations accordingly. This adaptive process ensures that explanations become increasingly aligned with user expectations over time.

3.3 CAEN Workflow

The CAEN process can be divided into these key stages:

Input Processing: The text is tokenized and fed into a transformer-based model, generating context-aware embeddings that capture both individual word significance and broader token relationships.

Contextual Attention: A specialized attention layer calculates importance scores for each token, taking into account both specific word relevance and the overall context of the sentence or document.

Text Categorization: The classification model utilizes the attention mechanism's outputs to perform tasks such as sentiment analysis or topic classification.

Explanation Creation: The explanation network leverages context-aware attention results and user profiles to produce customized explanations, ranging from token importance visualizations to comprehensive descriptions based on user preferences.

User Input and Enhancement: User interactions and feedback are collected post-model engagement, which are then used to improve future explanations, boosting both interpretability and user contentment.

3.4 Evaluation Matrix

To assess CAEN's effectiveness, we suggest the following metrics:

Interpretability Score: Determined through user studies where participants evaluate the clarity and utility of the provided explanations.

Classification Accuracy: The model's performance on standard text classification tasks, ensuring that explainability features do not compromise predictive capabilities.

Explanation Fidelity: The extent to which generated explanations accurately represent the model's internal decision process. This can be quantified using metrics like faithfulness (alignment between explanation and model predictions) and completeness (coverage of all relevant factors).

User Satisfaction: Feedback gathered from diverse user groups to gauge how well the explanations meet their expectations and requirements.

3.5 Potential Use Cases

Healthcare: CAEN can elucidate medical text classification models in patient diagnosis systems, offering context-sensitive explanations tailored to clinicians' expertise levels.

Legal: For legal document classification, the model can provide justifications for decisions based on contextual interpretations of legal terminology.

Content Moderation: On social media platforms, CAEN can deliver nuanced, context-driven explanations for decisions in hate speech or misinformation detection, ensuring both moderators and users comprehend the reasons behind flagged content.

3.6 Comparison with Existing Models

Correct CAEN demonstrates several benefits over conventional techniques such as LIME, SHAP, and attention mechanisms:

Contextual Understanding: The context-sensitive attention feature delivers more sophisticated and precise explanations by taking into account word relationships.

Personalized User Experience: The customized explanation components ensure that interpretations are both comprehensible and enlightening for various user demographics.

Flexibility: The system continually enhances its performance based on user input, resulting in increased responsiveness and adaptability over time.

4. EXPERIMENTAL SETUP

We assessed the efficacy of the newly developed Context-Aware Explanation Networks (CAEN) by performing experiments on three commonly utilized text classification datasets spanning various domains: sentiment analysis, topic

classification, and hate speech detection. This portion of the study outlines the experimental methodology, followed by a comprehensive examination of the outcomes. To evaluate CAEN, we employed the following datasets:

IMDB Movie Reviews (Sentiment Analysis): Dataset This collection comprises 50,000 strongly opinionated movie reviews. The objective is to categorize each review as either positive or negative. Sentiment analysis is a frequently encountered text classification challenge where comprehending sentiment-indicative words (such as "excellent" versus "terrible") is vital. Elucidating model predictions in this context helps unveil why specific words or phrases resulted in a particular sentiment classification.

AG News (Topic Classification): This compilation includes over 120,000 news articles sorted into four categories: World, Sports, Business, and Sci/Tech. Topic classification models often depend on field-specific terminology, and explanations assist users in recognizing which words (e.g., "government," "soccer") contributed to a specific classification.

Hate Speech Detection (Hate Speech and Offensive Language): This collection consists of 24,783 tweets labelled as hate speech, offensive language, or neither. In hate speech detection, explainability is crucial to ensure that models do not unintentionally reflect biases present in their training data. Clarifying why a model classified certain text as hate speech is essential for enhancing both model transparency and user confidence.

4.1 Experimental Setup

For each dataset, we utilized a BERT-based classifier, fine-tuned for specific text classification tasks. The CAEN framework was incorporated into this structure, consisting of these essential elements:
Primary Classifier: A BERT model, pre-trained and fine-tuned for each task. We opted for the bert-base-uncased version, which is commonly employed in text classification.
Context-Sensitive Attention System: Implemented above the BERT encoder outputs; this system modifies token importance based on their contextual relationships.
Personalized Explanation Network: We developed two explanation profiles per task: one for technically proficient users (offering detailed token-level explanations with attention weights) and another for non-technical users (providing sentence-level explanations with highlights).

4.2 Comparison Models

To assess CAEN's performance, we contrasted it with these baseline models:
BERT with Conventional Attention: BERT employing traditional token-level attention weights to indicate token significance for each classification.
LIME: A model-agnostic post-hoc explainability approach that generates local explanations by altering input text and elucidating the original model's predictions.
SHAP: A model-agnostic explainability technique founded on Shapley values, offering global and local feature attributions.

CAEN's effectiveness was measured using these metrics:
Accuracy: Overall classification precision for each task.
Interpretability Score: A qualitative measure derived from user studies where participants (both technical and non-technical) evaluated the clarity, utility, and relevance of explanations on a 1 to 5 scale.
Fidelity: The extent to which generated explanations accurately represent actual model behavior, assessed using metrics such as completeness and faithfulness.
User Satisfaction: Feedback from user studies evaluating how well the explanations met the requirements of both technical and non-technical users.

4.3 Experimental Result

Table 1: Classification Performance

Dataset	BERT	BERT+ LIME	BERT+ SHAP	CAEN
IMDB	91.4	90.7	90.9	92.3
AG News	93.6	93.1	93.3	94.0
Hate Speech	78.9	78.2	78.5	79.5

It is evident from table 1, the Context-Aware Attention Network (CAEN) exhibited a minor enhancement in classification accuracy across all datasets when compared to the baseline models. This indicates that incorporating context-aware attention did not negatively impact predictive performance and, in certain instances, actually improved it.

Alongside quantitative assessments, we performed user studies involving two distinct groups: individuals with technical expertise and those without technical backgrounds. These participants evaluated the explanations produced by each methodology.

From the table 2 in terms of interpretability and user satisfaction, CAEN demonstrated superior performance compared to LIME and SHAP. This was especially evident among users with technical backgrounds, who valued the comprehensive and context-sensitive explanations provided by CAEN. Additionally, users without technical expertise showed a preference for CAEN due to its simplified explanations at the sentence level, which were customized to suit their level of understanding.

Table 2: Interpretability and User Satisfaction

Dataset	LIME	SHAP	Technical CAEN	Non-Technical CAEN
IMDB	3.7	4.0	4.8	4.5
AG News	3.9	4.2	4.7	4.6
Hate Speech	3.4	3.9	4.6	4.4

Table 3: Explanation Fidelity

Dataset	LIME	SHAP	CAEN
IMDB	85.6	87.2	89.5
AG News	86.9	88.3	90.1
Hate Speech	78.4	79.7	82.3

The accuracy with which generated explanations reflect true model behavior is measured by explanation fidelity. Among all datasets, From the above table, CAEN exhibited the highest fidelity, suggesting that its context-aware attention more effectively captured the model's underlying decision-making process.

4.4 Discussion of Results

The enhanced classification accuracy seen with CAEN indicates that the context-aware attention mechanism not only improves explainability but also delivers superior feature representations for text classification. By emphasizing context-dependent word significance, CAEN identifies subtleties that conventional models may overlook, particularly in tasks where word meanings vary based on the surrounding text.

Interpretability and User Feedback

Accuracy and Performance: User studies demonstrated a marked preference for CAEN's context-aware explanations compared to those offered by LIME and SHAP. Technically proficient users appreciated the comprehensive explanations that emphasized important tokens within the context of entire sentences or documents, while non-technical users favored the simplified, high-level explanations. This flexibility in catering to different user types is a key strength of CAEN and addresses the increasing demand for user-oriented explainability in AI systems.

Fidelity and Completeness: Fidelity results show that CAEN's explanations more accurately represent the underlying model's predictions than those provided by LIME and SHAP. The context-aware attention mechanism is instrumental in this regard, ensuring that explanations mirror the actual decision-making process rather than offering a superficial approximation. This enhanced fidelity is especially crucial in high-stakes applications, such as hate speech detection and healthcare, where model accountability is essential.

5. CONCLUSION AND FUTURE WORK

The findings from the experiments show that CAEN enhances both the precision and understandability of text classification models. Through its provision of context-aware and user-specific explanations, CAEN represents a notable improvement over conventional post-hoc explainability techniques. These outcomes underscore the crucial role of context in comprehending model decisions and the significance of adaptable explanations tailored to various user types.

The Context-Aware Explanation Networks (CAEN) framework shows potential for enhancing explainability and classification accuracy. However, there are several opportunities to further improve its effectiveness, durability, and adaptability through future research and development. The following are crucial areas for upcoming investigations:

5.1 Enhancing User Profiling and Personalization

The existing CAEN system's user profiling component depends on predetermined or manually set user preferences to create customized explanations. Future research could explore more advanced, flexible user profiling techniques, such as:

Adaptive User Profiling: Utilizing machine learning to automatically deduce user preferences based on interaction patterns and feedback over time. This would allow CAEN to continuously adjust the depth and style of explanations in real-time, providing a more individualized and fluid user experience.

Multi-User Adaptation: Developing methods to produce explanations suitable for multiple users in collaborative settings (e.g., medical teams or legal professionals). These explanations should cater to the needs of various stakeholders, ensuring comprehensibility across different expertise levels simultaneously.

5.2 Expanding to Multimodal Data

CAEN is currently designed for text classification tasks. A potential expansion would involve incorporating multimodal data (e.g., combining text with images, audio, or video) to create explainable models for more complex tasks, such as medical diagnosis (merging patient records with medical images) or social media analysis (using both text and images). Future studies could concentrate on:

Multimodal Context-Aware Attention: Modifying the context-aware attention mechanism to process inputs from multiple modalities, ensuring that explanations consider the interplay between text and other data types.

Cross-Modal Explanations: Creating explanation layers that emphasize how different modalities (e.g., text and images) contribute to the overall decision-making process, offering comprehensive, transparent reasoning for predictions.

5.3 Integrating Causal Explanations

A limitation of current attention-based methods, including CAEN, is that they highlight important features (e.g., words or phrases) but do not provide clear causal reasoning behind decisions. Incorporating causal inference techniques into CAEN could enhance the quality of explanations by identifying causal relationships between input features and model predictions. Potential research areas include:

Causal Attention Mechanisms: Introducing causality-driven attention mechanisms that focus on causal relationships between text features, moving beyond correlations or associations.

Counterfactual Explanations: Enhancing CAEN to generate counterfactual explanations (e.g., "If this word was changed, the model would classify the text differently"), helping users better understand the decision-making process.

5.4 Decreasing Computational Overhead

One of the challenges with CAEN, particularly in using sophisticated mechanisms like context-aware embeddings and user-specific explanation layers, is the computational overhead. Future work could focus on:

Model Compression and Optimization: Developing more efficient methods for computing context-aware attention, potentially through techniques like model pruning, quantization, or distillation. This would enable CAEN to scale to large datasets and real-time applications without compromising explainability.

Lightweight Explanation Modules: Exploring ways to reduce the computational complexity of the user-tailored explanation layers, ensuring that explanations can be generated efficiently, especially for resource-constrained environments like mobile applications or edge devices.

Dynamic User Profiling: Utilizing machine learning algorithms to automatically deduce user preferences based on their interaction patterns and feedback over time. This approach would allow CAEN to continuously adjust the complexity and format of explanations in real-time, providing a more individualized and fluid user experience.

Multi-User Compatibility: Developing methods to produce explanations that cater to multiple users in collaborative settings (e.g., medical teams or legal professionals). These explanations should strike a balance between the requirements of various stakeholders, ensuring comprehensibility across different levels of expertise simultaneously.

REFERENCES:

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 1135–1144. 2016

[2] Lundberg, S. M., & Lee, S.-I. . A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 4765–4774, 2017

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. . Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008, 2017

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. . BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 4171–418 2019.
[5] Lei, T., Barzilay, R., & Jaakkola, T. . Rationalizing neural predictions. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 107–117 2016.
[6] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[7] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. Explaining explanations: An overview of interpretability of machine learning. IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89 2018
[8] Tjoa, E., & Guan, C. A survey on explainable artificial intelligence (XAI): Towards medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.. 2020
[9] Chen, H., Li, Z., Hu, B., & Chen, D. . Explaining neural network predictions on sentence pairs via learning word-group masks. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1595–1607 2020.
[10] Jacovi, A., & Goldberg, Y. . Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 4198–4205 2020
[11] M A Wajeed, A C S Rao, T K S Shavali, A Rasool Improving Text Classification in Federated Learning through Transfer Learning: A combined approach for enhanced convergence and personalization in Journal of Theoretical and Applied Information Technology Vol 102, No. 23 December 2024.
[12] Doshi-Velez, F., & Kim, B. . Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017
[13] Quinlan, J. Ross. "Improved use of continuous attributes in C4. 5." Journal of artificial intelligence research 4 (1996): 77-90.
[14] Hosmer Jr, David W., Stanley Lemeshow, and Susanne May. Applied survival analysis: regression modeling of time-to-event data. Vol. 618. John Wiley & Sons, 2008.
[15] Bahdanau, Dzmitry, et al. "Learning to understand goal specifications by modelling reward." arXiv preprint arXiv:1806.01946 (2018).
[16] Selvaraju, Ramprasaath R., et al. "Grad-CAM: visual explanations from deep networks via gradient-based localization." International journal of computer vision 128 (2020): 336-359.
[17] Zhong, Xiaoting, et al. "Explainable machine learning in materials science." npj computational materials 8.1 (2022): 204.
[18] Garg, Priya, M. K. Sharma, and Parteek Kumar. "Improving Hate Speech Classification Through Ensemble Learning and Explainable AI Techniques." Arabian Journal for Science and Engineering (2024): 1-14.
[19] Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.
[20] L. Chen, Y. Li, H. Zhang and S. Wei, "A novel explainable structure for text classification," 2022 European Conference on Natural Language Processing and Information Retrieval (ECNLPIR), Hangzhou, China, 2022, pp. 92-95, doi: 10.1109/ECNLPIR57021.2022.00028.