Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Università Ca'Foscari Venezia

Crossing the Divide: Designing Layers of Explainability

Zangari Alessandro, Matteo Marcuzzo, Matteo Rizzo, Andrea Albarelli, Andrea Gasparetto

iNEST Spoke 6 “Tourism, Culture and Creative Industries”
RT: 1 Sub. RT: 3

Abstract
In the era of deep learning, the opaque nature of sophisticated models often stands at odds with the growing demand for transparency and explainability in Artificial Intelligence. This paper introduces a novel approach to text classification that emphasises explainability without significantly compromising performance. We propose a modular framework to distil and aggregate information in a manner conducive to human interpretation. Our methodology's core is that features extracted at the finest granularity are inherently explainable and reliable. Compared with explanation methods based on word-level importance, this layered aggregation of low-level features allows us to trace a clearer decision trail of the model’s decision-making process. Our results demonstrate that this approach yields effective explanations with a marginal reduction in accuracy, presenting a compelling trade-off for applications where understandability is paramount.

Text

Introduction

Our paper delves into the challenge of enhancing explainability in Artificial Intelligence (AI), particularly focusing on text classification within Natural Language Processing (NLP). The introduction outlines the growing demand for transparency and understandability in AI, especially when deploying sophisticated deep learning (DL) models. These models often prioritise performance over the ability to explain their decisions, making the explanations more of an afterthought than an integral part of the development process. DL-based approaches, unlike traditional machine learning (ML), can learn complex, high-order features, making it difficult for humans to comprehend their meaning or origin. This lack of transparency raises concerns about these models' reliability and potential biases, particularly in high-stakes environments where legal regulations demand explainability. The paper proposes a novel framework that balances manual and automatic feature extraction to minimise performance loss while enhancing explainability. This framework uses black-box models for extracting low-level textual features, followed by a traditional classifier for task-solving.

The paper addresses the current state of explainable NLP, which is still an active research topic and lacks standardisation. Various explanation methods are discussed, including local and global explanations, specifically focusing on model-agnostic methods like LIME and SHAP (Lundberg and Lee, 2017; Ribeiro et al., 2016). However, the reliability of these methods has been questioned, and the need for standardised terminology and evaluation practices in explainable AI (XAI) is emphasised (Adebayo et al., 2018; Garreau and Mardaoui, 2021).

The proposed approach suggests using DL models for extracting granular features, which are more likely to be trusted and understood. We argue that an explainable model does not necessarily need to explain how low-level features are extracted but should focus on the final layer of decision-making. By combining DL and traditional feature extraction methods, the approach aims to provide effective explanations while retaining the effectiveness of neural models used as feature extractors.

For our experiments, we use two NLP tasks to demonstrate the effectiveness of the explanation strategy. We compare various classifiers and utilise SHAP for evaluating feature significance in model reasoning. The paper also discusses the importance of faithfulness in explanations, measuring how closely an explanation aligns with the model's reasoning process. Finally, we conducted user studies to validate the proposed approach. The results show positive user feedback towards the novel explanation strategy, although we note the limitations due to the limited sample size and high variance in the data. We conclude by highlighting the potential of our approach in providing more consistent and reliable explanations and suggest future research directions to explore further and validate the strategy. The document presents a comprehensive study on enhancing explainability in NLP through a novel approach

that combines DL and traditional ML methods. Our focus on balancing performance with explainability addresses a critical challenge in AI, particularly in applications where understanding the reasoning behind decisions is crucial.

Related work

This section of the paper discusses advancements in the field of explainable NLP, specifically focusing on methods for making ML models more explainable and transparent, especially those in NLP. This section is critical as it sets the stage for our proposed approach by highlighting existing strategies and their limitations.

One of the key themes in this section is the distinction between traditional ML methods and DL models. Traditional ML relies on manually crafted features, which, while straightforward, require intensive labour and time to create (Gasparetto et al., 2022; Li et al., 2022). In contrast, DL models, particularly those used in NLP, autonomously extract complex, high-order features, making them more task-specific but less transparent. We emphasise the challenge of understanding these models due to layers of compression and non-linear transformations that obscure their decision-making processes. Indeed, this enhanced expressiveness of features extracted with DL comes with a significant drawback. The encoding process severs any discernible connection to the input features or raw data, making it exceedingly difficult for humans to understand their meaning or trace their origins.

The paper categorises explainable AI (XAI) methods into two types: local and global explanations. Local explanations focus on individual observations, while global explanations aim to elucidate the model's overall behaviour (Linardatos et al., 2020; Madsen et al., 2022; Zini and Awad, 2022). Notable approaches in local explanations include LIME and SHAP (Lundberg and Lee, 2017; Ribeiro et al., 2016). LIME uses input sample perturbations and logistic regression to assign feature importance, while SHAP, considered a gold standard in many cases, applies Shapley values from game theory as measures of the contribution of input features to a specific output. However, the reliability of these methods has been questioned, prompting a need for more robust solutions. However, other common explanation strategies suffer from even more criticism. Many of these strategies utilise saliency in the form of gradients and attention scores for explanations (Adebayo et al., 2018; Bastings et al., 2022). Gradient-based methods like (Sundararajan et al., 2017) assess output changes due to minor input variations, but their effectiveness is challenged by instances where significant features yield zero gradients (Nielsen et al., 2022). Attention-based methods try to correlate these scores to important input components, though their reliability as feature importance indicators has also been disputed (Bibal et al., 2022; Serrano and Smith, 2019). Counterfactual explanations, derived from adversarial examples and contrastive learning (Linardatos et al., 2020), represent another approach, noted for their human suitability (Miller, 2019). However, their application and the faithfulness of their explanations have likewise been questioned (Hoedt et al., 2023; Madsen et al., 2022).

We also discuss evaluating explanation quality, focusing on *plausibility* and *faithfulness*. Plausibility refers to how realistic an explanation appears to a human observer, often assessed through user studies. On the other hand, faithfulness measures how much a model relies on the elements of an explanation for decision-making (Rizzo et al., 2023; Saranya and Subhashini, 2023). In this study, we discuss some metrics that have been proposed for evaluating faithfulness, highlighting the need for standardised terminology and evaluation practices in XAI (Falis et al., 2021).

Despite the growth in explainable NLP, we note that the pace of developing new NLP methods surpasses the research into their explainability. This gap is exacerbated by a lack of consensus on fundamental concepts within the field, underscoring the urgent need for standardised terminology and evaluation practices in XAI to foster coherent and practical advancements.

In summary, this section provides a comprehensive overview of the current state of explainable NLP. It highlights the challenges in making DL models, particularly in NLP, transparent and explainable. The section underscores the need for more effective local explanation methods, the importance of evaluating explanation quality, and the critical need for standardised practices in XAI research (Miller, 2019; Rawal et al., 2022).

Approach and evaluations

This paper presents a novel approach to text classification that prioritises explainability without significantly compromising performance. This approach introduces a modular framework for distilling and aggregating information in a way conducive to human understanding. It is based on the premise that features extracted at the finest granularity are inherently explainable and reliable, offering a clearer decision trail of the model's decision-making process. The results demonstrate effective explanations with only a marginal reduction in accuracy, presenting a compelling trade-off for applications where understandability is paramount.

Significant advancements in NLP have been observed in recent years due to the widespread use of Large Language Models (LLMs). Despite their popularity, these models' lack of transparency raises concerns about their reliability and potential biases (Bender et al., 2021; Durán and Jongsma, 2021). As the scale of LLMs (and DNNs, in general) is increasing at a far higher pace than the speed at which explainability methods are being developed, it has become evident that different approaches might be warranted, especially whenever human comprehension is a key factor in the system. At the same time, it is hard to discard DL methods entirely because of their outstanding performance. Thus, we propose and design a process that balances manual and automatic feature extraction. The key idea of this approach is that DL models may yet be used in an explainable fashion by utilising them to extract finely-grained features.

At the core of this idea stands the argument that, at some point, we must ultimately rely on knowledge that may or may not be precise (Adadi and Berrada, 2018; Glanville,

1982). Indeed, every model must exercise faith in the data it is built on at one point or another. For instance, a model trained on sensor data has to believe the hardware performing such readings is itself working correctly and that the resulting data reflects reality. Likewise, we might utilise DL approaches to extract granular features for which we can afford a certain degree of trust (Durán and Jongsma, 2021; McCoy et al., 2022), and then utilise these features to make more complex decisions. This approach also relies on the concept that an explainable model does not necessarily have to explain how low-level features are extracted, but rather that explainability is a useful tool at the "last layer" of a decision system. By utilising a model built on such features, we can create effective explanations (as they pertain to concrete textual properties) without entirely discarding the effectiveness of neural approaches used as feature extractors.

We validate our approach through experiments on two NLP tasks, demonstrating the effectiveness of the explanation strategy with a user study and analysing its fidelity to model decisions. The feature extraction and pre-processing for our proposed approach detail the extraction of various features, including those leveraging pre-trained LMs and traditional statistical features. We prune unimportant features by doing a feature selection using Recursive Feature Elimination (RFE) and the selected features are scaled. The experimental settings present the datasets, methods used in the paper, and technical details. The datasets include the IMDb dataset for binary sentiment classification and the "Call Me Sexist But" (CMSB) dataset for binary sexism detection. The *IMDb* consists of 50000 movie reviews from the IMDb website, used for binary classification. Reviews are labelled as positive (score ≥ 7) or negative (score ≤ 4), excluding neutral ones. It is evenly split with 25000 reviews for training and testing (Maas et al., 2011). The *CMSB* dataset contains over 13000 texts from various sources, including tweets and surveys, for binary sexism detection (D. M. Samory, 2021; M. Samory, 2021). It features both original and minimally altered adversarial texts to shift sexist content to non-sexist. The paper compares the effectiveness of various classifiers and a Transformer model fine-tuned on the dataset. Our explanation approach focuses on local explainability, using SHAP for feature evaluation and an eXplainable User Interface (XUI) for communicating explanations to users.

The proposed approach involves using black-box models for extracting low-level textual features, followed by a traditional classifier for task-solving. This method aims to minimise performance loss while enhancing explainability. The paper validates this approach through experiments on two NLP tasks, demonstrating the effectiveness of the explanation strategy with a user study and analysing its fidelity to model decisions. We utilise two metrics proposed by previous research to evaluate the faithfulness of explanation strategies, specifically measuring *comprehensiveness* and *sufficiency* (DeYoung et al., 2020; Jacovi and Goldberg, 2020). Both give some insight into the alignment between the explanation and the actual reasoning process of the model. Indeed, faithfulness is solely dependent on the model and the generated explanation, quantifying their alignment.

The concept of *faithfulness* is a critical term in the field of model explanation. In general, it refers to how closely an explanation reflects the actual reasoning process of the model (Carvalho et al., 2019; Chan et al., 2022). This concept is crucial because it helps us understand whether the explanation provided by the model is a true representation of its internal workings or not. As mentioned, faithfulness does not rely on external factors or user interpretation, making it a robust and reliable metric for evaluating the quality of explanations. Our two metrics of choice, *comprehensiveness* and *sufficiency* provide a holistic view of the faithfulness of the explanations. *Comprehensiveness* suggests that an interpretation is faithful if the important features are highly representative of the entire input. If these features are removed, there should be a significant change in the model's confidence. This means that the important features carry substantial weight in the model's decision-making process. On the other hand, *sufficiency* complements comprehensiveness. It assumes that if a percentage of the non-critical features are masked out, the model's confidence in the original decision should remain relatively stable. This indicates that the non-critical features do not significantly impact the model's decision. We use these metrics as an objective measure to compare our explanation strategies. The strategy that achieves a higher comprehensiveness score and a lower sufficiency score is considered the most faithful. This balance ensures that the explanation is both comprehensive and sufficient.

For our proposed classifier, features are eliminated by assigning them the median value of the training set. This method minimises their informativeness by confining them to a commonly occurring range of values. This approach ensures that the eliminated features do not skew the model's decision. With the pre-trained language model, words in the input sentence are masked using a special sequence. This sequence is employed by the tokenizer for unknown symbols. This method ensures that the masked words do not influence the model's decision. In our experiments, we use a set of predefined values for comprehensiveness and sufficiency. The evaluation is carried out on the test set of the two datasets, which consist of approximately 3500 examples each. This large sample size ensures the robustness and reliability of our evaluation.

A user study was also conducted to validate the approach regarding faithfulness and plausibility. The study involved 21 participants who compared explanations of both kinds. In summary, our proposed approach to text classification tasks emphasises explainability by utilising DL and non-DL extractors to create a knowledge base of fine-grained features. While not as strong as LM-based approaches in terms of performance metrics, the methods still perform well while being more faithful, and preliminary results on a user study of reduced sample size showcase positive user feedback toward the explanation strategy.

Results and Discussion

The paper focuses on a novel approach to text classification in the context of AI, emphasising explainability without significant performance compromise. The

methodology involves a modular framework designed to distil and aggregate information conducive to human interpretation. It is centred on the premise that features extracted at the finest granularity are inherently explainable and reliable. The approach is validated through experiments on two NLP tasks, demonstrating the effectiveness of the explanation strategy with a user study and analysing its fidelity to model decisions.

Our research highlights that while traditional ML relies on manually crafted features, DL-based approaches learn complex, high-order features, which are more task-specific but less transparent. This creates a gap between the model's internal representations and human comprehension, especially concerning high-dimensional feature extraction and optimisation in DL. The paper proposes a design process that strategically utilises DL and traditional methods to tackle smaller, more defined sub-problems. This involves using black-box models for extracting low-level textual features, followed by a traditional classifier for task-solving, aiming to enhance explainability with minimal performance loss. The paper also addresses the challenges in evaluating explanation quality, focusing on faithfulness and plausibility. Faithfulness measures the reliance of a model on the explanation's elements for decision-making, whereas plausibility assesses the realism of an explanation to a human observer, often through user studies. The study uses metrics for faithfulness and conducts user studies to validate the approach.

Quantitatively, while the language model-based approach shows better performance, the proposed method is competitive, with only a slight decrease in accuracy and F1 score but better explainability. The faithfulness analysis shows that the proposed explanation method aligns more with the XGBoost model than the RoBERTa-base counterpart, indicating greater faithfulness. However, the explanation strategy does not fully capture the interactions between features, which is a limitation.

The proposed high-level feature-based explanation strategy in user studies showed a preference, although the limited user pool prevents definitive conclusions. The results also suggest that our selected top features may sometimes be too complex for the user to reconstruct the model decision based on them. This underscores the importance of careful consideration in both the interpretation and subsequent application of these findings. While the observed trends are promising, they cannot be reliably generalised to a broader population. The high variance in the data points to a wide range of user experiences, which could be attributed to diverse user backgrounds, differing levels of engagement with the method, or other external factors not controlled in the study. The study highlights the need for further research with a larger and more diverse sample to validate the strategy on a larger scale. Despite a slight performance decrease measured by the F1 score, the approach provides more consistent and reliable explanations and suggests a degree of generalizability to new datasets.

Conclusion

The paper presents a novel approach to text classification tasks emphasising explainability without significantly compromising performance. We propose a layered method, combining DL and non-DL extractors, to create a base of fine-grained features for classifiers. This approach offers a clearer understanding of the decision logic of models, as these features are inherently more abstract.

Performance-wise, while not as strong as approaches based on modern language models, this method still competes effectively. A notable aspect is its more consistent and reliable explanations. The feature extractors, not fine-tuned for specific datasets, suggest potential generalizability to new datasets. The approach's modularity allows customisation and bias control by adding or removing feature extractors.

User studies indicate a preference for this high-level feature-based explanation strategy, although limited user pool size prevents definitive conclusions. In terms of faithfulness – the degree to which an explanation mirrors the model's true reasoning process – the method outperforms the pre-trained language model on the comprehensiveness metric. However, it shows higher sensitivity to top feature removal, indicating a reliance on feature combinations. In contrast, the pre-trained model, benefiting from extensive internal knowledge from pre-training, is more stable in explanations.

We acknowledge a 3-5% decrease in performance (measured by F1 score) compared to the state-of-the-art models but highlight the benefits of more reliable explanations. We suggest that significant improvements are possible by fine-tuning feature extraction for specific tasks. Future research could develop a comprehensive library of general feature extractors applicable to any dataset and explore the method's application to a broader range of datasets and tasks.

References

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18. Curran Associates Inc., Red Hook, NY, USA, pp. 9525–9536.

Bastings, J., Ebert, S., Zablotskaia, P., Sandholm, A., Filippova, K., 2022. "Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 976–991. https://doi.org/10.18653/v1/2022.emnlp-main.64

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? , in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21. Association for Computing Machinery, New York, NY, USA, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., Watrin, P., 2022. Is Attention Explanation? An Introduction to the Debate, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp. 3889–3900. https://doi.org/10.18653/v1/2022.acl-long.269

Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. Electron. Basel 8, 832.

Chan, C.S., Kong, H., Guanqing, L., 2022. A Comparative Study of Faithfulness Metrics for Model Interpretability Methods, in: Muresan, S., Nakov, P., Villavicencio, A. (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp. 5029–5038. https://doi.org/10.18653/v1/2022.acl-long.345

DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C., 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models, in: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 4443–4458. https://doi.org/10.18653/v1/2020.acl-main.408

Durán, J.M., Jongsma, K.R., 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J. Med. Ethics 47, 329–335. https://doi.org/10.1136/medethics-2020-106820

Falis, M., Dong, H., Birch, A., Alex, B., 2021. CoPHE: A Count-Preserving Hierarchical Evaluation Metric in Large-Scale Multi-Label Text Classification, in: Moens, M.-F., Huang, X., Specia, L., Yih, S.W. (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. Association for Computational Linguistics, pp. 907–912. https://doi.org/10.18653/v1/2021.emnlp-main.69

Garreau, D., Mardaoui, D., 2021. What does LIME really see in images?, in:

Proceedings of the 38th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 3620–3629.

Gasparetto, A., Marcuzzo, M., Zangari, A., Albarelli, A., 2022. A Survey on Text Classification Algorithms: From Text to Predictions. Information 13, 83. https://doi.org/10.3390/info13020083

Glanville, R., 1982. Inside every white box there are two black boxes trying to get out. Behav. Sci. 27, 1–11. https://doi.org/10.1002/bs.3830270102

Hoedt, K., Praher, V., Flexer, A., Widmer, G., 2023. Constructing adversarial examples to investigate the plausibility of explanations in deep audio and image classifiers. Neural Comput. Appl. 35, 10011–10029. https://doi.org/10.1007/s00521-022-07918-7

Jacovi, A., Goldberg, Y., 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?, in: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S., He, L., 2022. A Survey on Text Classification: From Traditional to Deep Learning. ACM Trans. Intell. Syst. Technol. 13, 31:1-31:41. https://doi.org/10.1145/3495162

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 23, 18. https://doi.org/10.3390/e23010018

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C., 2011. Learning Word Vectors for Sentiment Analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA, pp. 142–150.

Madsen, A., Reddy, S., Chandar, S., 2022. Post-hoc Interpretability for Neural NLP: A Survey. ACM Comput. Surv. 55, 155:1-155:42. https://doi.org/10.1145/3546577

McCoy, L.G., Brenna, C.T.A., Chen, S.S., Vold, K., Das, S., 2022. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J. Clin. Epidemiol. 142, 252–257. https://doi.org/10.1016/j.jclinepi.2021.11.001

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Nielsen, I.E., Dera, D., Rasool, G., Ramachandran, R.P., Bouaynaya, N.C., 2022. Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks. IEEE Signal Process. Mag. 39, 73–84. https://doi.org/10.1109/MSP.2022.3142719

Rawal, A., McCoy, J., Rawat, D.B., Sadler, B.M., Amant, R.St., 2022. Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives. IEEE Trans. Artif. Intell. 3, 852–866. https://doi.org/10.1109/TAI.2021.3133846

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining

the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

Rizzo, M., Veneri, A., Albarelli, A., Lucchese, C., Nobile, M., Conati, C., 2023. A Theoretical Framework for AI Models Explainability with Application in Biomedicine, in: 2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). Presented at the 2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–9. https://doi.org/10.1109/CIBCB56990.2023.10264877

Samory, D.M., 2021. The "Call Me Sexist But" Dataset [WWW Document]. GESIS Blog. URL https://blog.gesis.org/the-call-me-sexist-but-dataset/ (accessed 1.3.24).

Samory, M., 2021. The "Call me sexist but" Dataset (CMSB). https://doi.org/10.7802/2251

Saranya, A., Subhashini, R., 2023. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decis. Anal. J. 7, 100230. https://doi.org/10.1016/j.dajour.2023.100230

Serrano, S., Smith, N.A., 2019. Is Attention Interpretable?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 2931–2951.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17. JMLR.org, Sydney, NSW, Australia, pp. 3319–3328.

Zini, J.E., Awad, M., 2022. On the Explainability of Natural Language Processing Deep Models. ACM Comput. Surv. 55, 103:1-103:31. https://doi.org/10.1145/3529755