

EvalxNLP: A Framework for Benchmarking Post-Hoc Explainability Methods on NLP Models

Mahdi Dhaini*, Kafaite Zahra Hussain, Efstratios Zaradoukas and Gjergji Kasneci

Technical University of Munich, School of Computation, Information and Technology, Department of Computer Science,
Boltzmannstr. 3, Garching, 85748, Germany

Abstract

As Natural Language Processing (NLP) models continue to evolve and become integral to high-stakes applications, ensuring their interpretability remains a critical challenge. Given the growing variety of explainability methods and diverse stakeholder requirements, frameworks that help stakeholders select appropriate explanations tailored to their specific use cases are increasingly important. To address this need, we introduce *EvalxNLP*, a Python framework for benchmarking state-of-the-art feature attribution methods for transformer-based NLP models. *EvalxNLP* integrates eight widely recognized explainability techniques from the Explainable AI (XAI) literature, enabling users to generate and evaluate explanations based on key properties such as faithfulness, plausibility, and complexity. Our framework also provides interactive, LLM-based textual explanations, facilitating user understanding of the generated explanations and evaluation outcomes. Human evaluation results indicate high user satisfaction with *EvalxNLP*, suggesting it is a promising framework for benchmarking explanation methods across diverse user groups. By offering a user-friendly and extensible platform, *EvalxNLP* aims at democratizing explainability tools and supporting the systematic comparison and advancement of XAI techniques in NLP.

Keywords

Natural Language Processing, Explainable AI, Feature Attribution, Evaluation, Large Language Models

1. Introduction

Although significant progress has been made in the field of NLP, transformer-based models often operate as *black boxes*, making it challenging to interpret their decision-making processes. In high-stakes domains like medical diagnosis and financial decision-making, transparency is essential for trust and accountability. Despite the rapid development of explainability methods, there remains a lack of standardized evaluation frameworks, particularly for NLP, where text data is inherently unstructured and context-dependent. Existing explainability assessments vary widely, spanning qualitative user studies and quantitative metrics like faithfulness and plausibility, yet no universal consensus exists on the most effective approach.

To address these challenges, we introduce *EvalxNLP*, a benchmarking framework for evaluating post-hoc explainability methods in text classification tasks. *EvalxNLP* supports multiple explanation techniques, and assesses them across key properties such as faithfulness, plausibility, and complexity. In addition, *EvalxNLP* integrates LLM-based natural language explanations to facilitate the users' understanding of the generated explanations and evaluations. We also conduct a user-based study to evaluate the usability and user satisfaction with the framework. Our framework provides a systematic, user-friendly platform that aims to democratize access to explainability tools, enabling both researchers and practitioners to compare and refine XAI techniques for NLP applications. By offering a unified and reproducible evaluation methodology, *EvalxNLP* advances the field of explainability, promoting more transparent and trustworthy AI systems.

Late-breaking work, Demos and Doctoral Consortium, colocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

*Corresponding author.

✉ mahdi.dhaini@tum.de (M. Dhaini); ge95nut@mytum.de (K. Z. Hussain); efstratios.zaradoukas@tum.de (E. Zaradoukas); gjergji.kasneci@tum.de (G. Kasneci)

ORCID 0000-0002-7831-3141 (M. Dhaini); 0009-0002-1358-2487 (K. Z. Hussain); 0009-0004-5762-3986 (E. Zaradoukas); 0000-0002-3123-7268 (G. Kasneci)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 1

Overview of the built-in explanation methods, metrics, and datasets in EvalxNLP

Method Category	Explainers	Property	Metric	Dataset	Classification Task
Gradient-based	Saliency	Faithfulness	Soft suff	MovieReviews	Sentiment
	GradientxInput		Softcomp	HateXplain	Hate Speech
Perturbation-based	IntegratedGradients	Faithfulness	FAD N-AUC	e-SNLI	NLI
	DeepLIFT		AUC-TP		
	LIME		IOU-F1		
	Guided BackProp	Plausibility	Token-F1		
	SHAPIQ		AUPRC		
	SHAP	Complexity	Complexity Sparseness		

2. Related Work

Most existing explainability evaluation frameworks are designed for general-purpose applications, meaning that they include explainability methods used in image or tabular applications. As a result, most of them lack dedicated support for text-based models. OpenXAI [1], BEExAI [2], and Quantus [3] are prime examples of supporting multiple data modalities but not including text-specific evaluation metrics. Performance assessment is typically based on generic criteria, without tailored adaptations for NLP tasks. Frameworks such as Inseq [4] support sequence generation models but lack built-in evaluation metrics. XAI-Bench [5] evaluates explainability methods using synthetic data, which may not fully capture real-world text applications [6]. While these frameworks provide partial solutions, they do not offer a comprehensive suite for text explainability. ferret [7] facilitates explainability evaluation for NLP models but supports only five feature-attribution methods and six evaluation metrics. Among existing XAI libraries, it is the only one with adequate text-specific explainability features, to be considered as an NLP-specialized explainability framework. But, it relies on some metrics that have been shown to be inaccurate especially for measuring faithfulness (as explained in section 3.3.1); Captum [8] provides implementations for 22 attribution methods but is limited to two evaluation metrics (Infidelity and Sensitivity) and lacks built-in benchmarking. Similarly, AIX360 [9] supports text data but evaluates only two properties, faithfulness and monotonicity, without systematic benchmarking capabilities. M4 [10] evaluates faithfulness for image and text modalities but does not assess plausibility or complexity. While these frameworks address various aspects of explainability, among existing XAI libraries, only ferret offers a complete set of key features [7]: multiple explainability methods, Transformers-readiness (built with close integration into the Hugging Face (HF) *transformers* library), evaluation APIs, explainable datasets (i.e., those with human-annotated rationales), and built-in visualization. In addition to these features, our tool also extends functionality by incorporating recent explainability methods, recent metrics for evaluating explanation properties and providing an LLM-based module that generates natural language explanations to enhance user understanding. It consolidates capabilities scattered across multiple frameworks, offering a robust suite for benchmarking, evaluation, and qualitative explanations. By seamlessly integrating these features into one comprehensive framework, like *EvalxNLP*, we support practitioners in benchmarking Ph-FA explanation methods.

3. The Framework

The *EvalxNLP* framework builds on the top of four main components described below. For the technical implementation details, we refer the reader to the documentation provided in the repository. We release the code, tutorials, and documentation in the following repository¹.

¹<https://github.com/dmah10/EvalxNLP>

3.1. Explainers

One goal of *EvalxNLP* is to enable users to generate diverse explanations through multiple explainability methods. The explainer component integrates eight widely recognized explainability methods from the XAI literature, specifically focusing on post-hoc feature attribution (Ph-FA) methods. *EvalxNLP* incorporates two categories of post-hoc methods: gradient-based and perturbation-based approaches. Gradient-based methods compute feature importance by leveraging gradients of the model’s output with respect to its input features. They efficiently utilize backpropagation, making them well-suited for deep learning models. *EvalxNLP* integrates five key methods, including Saliency (also called Gradients)[11], which calculates raw gradients to highlight important inputs; Gradient×Input [12], which scales gradients by input values for enhanced clarity; Integrated Gradients [13], which averages gradients along a path from a baseline to the input; DeepLift [14], which attributes differences in activations to individual inputs for more stable attributions; and Guided BackProp [15], which filters negative gradients to highlight only positive contributions. *EvalxNLP* implements these methods using Captum and ferret, providing a diverse set of efficient and interpretable attribution techniques. Perturbation-based methods integrated into the toolbox include the widely used LIME [16] and SHAP [17] methods, as well as the recently introduced SHAP with interactions method (SHAP-I) [18], which augments traditional Shapley values by incorporating feature interactions [19], a notable contribution over existing frameworks. The rationale behind providing a comprehensive range of Ph-FA methods is twofold: (1) to offer users access to a diverse set of explanations from established as well as novel methods, enabling a holistic assessment and selection of explanations tailored to specific use cases; and (2) to facilitate benchmarking, comparative analyses, and evaluations of these methods based on selected evaluation criteria

We implement the explanation methods by building on top of their original implementations (for LIME and SHAP) and existing open-source libraries for the remaining methods. Specifically, we use the original implementation for LIME. For SHAP, we utilize Partition SHAP, a variant that optimizes Shapley value computation by exploiting feature independence. For SHAP-I, we extend the implementation provided by the shapiq package [20], while gradient-based methods are implemented using the Captum library [8]. Our approach of building upon established open-source libraries aims to facilitate and support the expansion and development of open-source XAI libraries.

3.2. LLM explanations

Feature attribution explanations, especially those involving many features, can be difficult for lay users to interpret [21]. To address this, our tool integrates an LLM-based module that automatically generates natural language explanations to help users interpret both (1) the importance scores from various explanation methods and (2) the evaluation metric scores. These textual explanations enhance the comprehensibility of Ph-FA outputs, particularly for non-experts, and support more informed decision-making by providing textual explanations for the evaluation metric scores. Additionally, combining visual heatmaps with textual explanations offers a more accessible and holistic view of the model’s decision-making process. To prevent the unfaithful textual explanations for model decisions by LLMs [21], we don’t ask the LLM to provide its own explanations of the model decisions but only use LLM solely to verbalize the outputs of explanation methods such as importance scores into natural language to make them more comprehensible. Figure 1 presents an example of an explanation heatmap (Figure 1a) and LLM-generated explanation for the SHAP scores (Figure 1b) for a misclassified instance in the MovieReviews dataset [22] that is misclassified by XLM-RoBERTa-base [23]. As shown in Figure 1b, the LLM provides comprehensible textual explanation to ease understanding the scores by SHAP.

The LLM is integrated into the framework via an API, enabling seamless generation of textual explanations on demand. For LLM API support, our demo uses the Together AI² API, with Llama-3.3-70B-Instruct-Turbo as the default model for generating explanations. Users can switch models or providers and modify the LLM instructions as needed.

²<https://www.together.ai/>

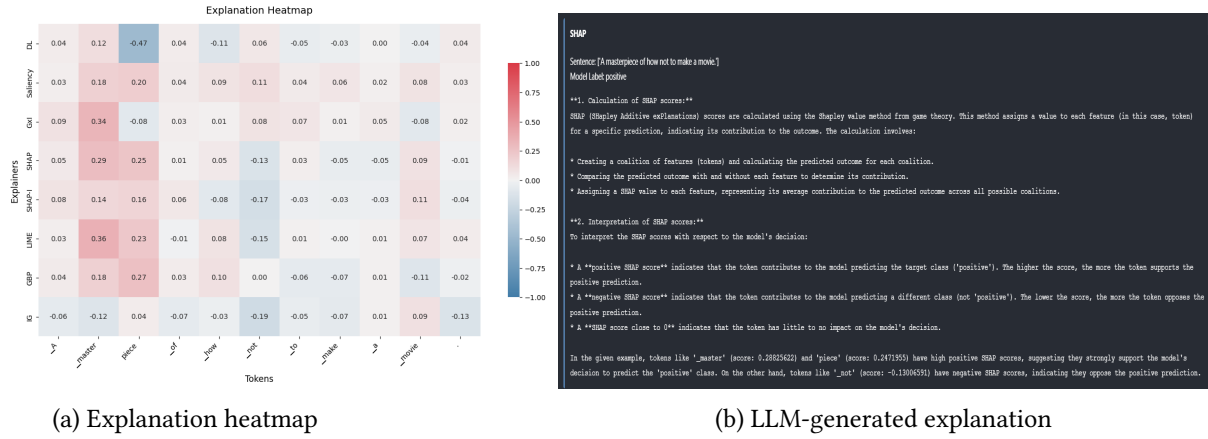


Figure 1: (Zoom in for a better view): (a) heatmap of the importance scores generated by the explainers for the single misclassified instance *A masterpiece of how not to make a movie* and (b) LLM-generated textual explanation for the scores by SHAP

3.3. Metrics

To evaluate explainability methods within our NLP framework, we use a comprehensive set of metrics covering three key properties: faithfulness, plausibility, and complexity. These ensure that explanations align with model reasoning while remaining interpretable and concise for users. By integrating diverse metrics, the framework supports a rigorous, holistic assessment that balances model fidelity, human interpretability, and explanation brevity. For mathematical details, we refer readers to the original papers introducing these metrics. (↓)/(↑) indicates that lower/higher values are better for a given metric.

3.3.1. Faithfulness

Faithfulness measures how well the generated explanations reflect the true behavior of the model. Compared to other frameworks that use sufficiency (suff) and comprehensiveness (comp) based on complete token removal, an approach shown to produce inaccurate faithfulness measurements [24], we employ soft suff and soft comp, which have proven more accurate in measuring faithfulness [25]. **Soft suff** [26] ↓ : Evaluates how well the most important tokens can retain the model’s prediction when other tokens are softly perturbed. It assumes that retaining more elements of important tokens should preserve the model’s output, while dropping less important tokens should have minimal impact. **Soft comp** [26] ↑ : Measures how much the model’s prediction changes when important tokens are softly perturbed using Bernoulli mask. It assumes that heavily perturbing important tokens should significantly affect the model’s output, indicating their importance to the prediction. **Feature Attribution Dropping (FAD) Curve and Normalized Area Under Curve (N-AUC)** [27] ↓ : Measures the impact of dropping the most salient tokens on model performance, with the steepness of the FAD curve indicating the method’s faithfulness. The N-AUC quantifies this steepness, where a lower score reflects better alignment of the attribution method with the model’s true feature importance. **Area Under the Threshold-Performance Curve (AUC-TP)** [28] ↓ : AUC-TP evaluates the faithfulness of saliency explanations by progressively masking the most important tokens (based on their saliency scores) and measuring the drop in the model’s performance. This AUC-TP value provides a single metric summarizing how significantly the model relies on the highlighted tokens, with lower values indicating better faithful explanations.

3.3.2. Plausibility

Plausibility measures how well the generated explanations align with human intuition. The following metrics are incorporated: **IOU-F1 Score** [29] ↑ : Computes the Intersection over Union (IoU) between predicted and ground-truth rationales, considering a (partial) match if the overlap is above 50% where these matches are used to calculate F1 scores. **Token-Level F1 Score** [29] ↑ : Measures alignment by

calculating the F1-score between predicted and human-annotated rationales at the token level. **Area Under Precision-Recall Curve (AUPRC)** [29] \uparrow : Evaluates plausibility by comparing the saliency scores of tokens with ground-truth rationale masks, computing the area under the precision-recall curve.

3.3.3. Complexity

Complexity evaluates how concise and interpretable the explanations are. Sparse explanations that highlight only a few important features are preferred. We use the following metrics: **Complexity** [30] \downarrow : Measures how evenly importance scores are distributed using Shannon entropy. Higher values indicate more complex explanations, while lower ones suggest concise attributions. **Sparseness** [31] \uparrow : Computes the sparsity of attributions using the Gini index, where higher scores indicate more concentrated importance on a few features.

3.4. Datasets

EvalxNLP is a framework for text classification that combines robust evaluation and explanation capabilities. It supports tasks like Sentiment Analysis, Hate Speech Detection, and Natural Language Inference (NLI), using rationale-annotated datasets that highlight key text segments. These human-provided rationales enable assessment of how well model explanations align with human reasoning. Currently, *EvalxNLP* includes a representative dataset for each of the aforementioned tasks and supports these three datasets by default, while also allowing users to extend it with additional classification datasets: **MovieReviews**: Designed for Sentiment Analysis, this dataset consists of 1,000 positive and 1,000 negative movie reviews, each annotated with phrase-level human rationales that justify the sentiment label. **HateXplain** [32]: Used for Hate Speech Detection, this dataset comprises 20,000 posts from Gab and Twitter, annotated with one of three labels: hate speech, offensive, or normal. **e-SNLI** [33]: A dataset for Natural Language Inference containing 549,367 examples, split into training, validation, and test sets. Each example includes a premise and a hypothesis labeled as entailment, contradiction, or neutral.

4. Case Study

To show the usability of *EvalxNLP* on real-world datasets, we present a case study showcasing how the tool can be used for benchmarking explainers for a sentiment analysis task on the Movie Reviews that include rationales [29] using an XLM-RoBERTa-base that is fine-tuned for sentiment analysis. *EvalxNLP* enables users to generate explanations and benchmark explainers either for single instances (local explanations) and across multiple instances (a subset or an entire dataset). This functionality depends on the user’s intention, such as identifying the best explanation method with respect to specific metrics and properties, either for individual sentences or aggregated across datasets. In this case study, we demonstrate how *EvalxNLP* benchmarks explainers using the full Movie Reviews dataset by aggregating evaluation metrics across all instances. Figure 2 presents the results indicating that DL achieves the highest overall faithfulness scores, particularly on soft metrics, suggesting it produces the most faithful explanations for this dataset. IG performs best on complexity metrics, indicating its explanations are simpler and easier to understand, while SHAP outperforms other methods on plausibility metrics, showing that its explanations align closely with human intuition. As expected, no single explanation method excels in all properties, confirming that practitioners must select methods based primarily on the evaluation property most relevant to their specific use case.

5. Human Evaluation

We also conducted a user-based study involving 20 participants to assess the usability of the tool. The participants were provided with instructions on how to run the tool and try its different functionalities.

Evaluator	Soft Comprehensiveness ↑	Soft Sufficiency ↓	AUTPC ↓	FAD ↓	Complexity ↓	Sparseness ↑	IOU F1 score ↑	Token F1 score ↑	AUPRC ↑
Explainer									
DL	0.094966	0.017995	0.510050	0.882353	0.011549	0.492214	0.186554	0.295325	0.247292
GBP	0.096274	0.013233	0.503015	0.905462	0.011551	0.492123	0.188023	0.297449	0.249667
GxI	0.119364	0.014157	0.496985	0.939076	0.011465	0.507764	0.184515	0.292689	0.250127
IG	0.115986	0.009149	0.507538	0.890756	0.011334	0.514918	0.185454	0.294009	0.244667
LIME	0.088791	0.002159	0.487940	0.920168	0.011820	0.423071	0.185320	0.293817	0.248880
SHAP	0.118418	0.047314	0.515075	0.947479	0.011738	0.425501	0.205684	0.315798	0.290448
SHAP-I	0.184629	0.091084	0.491457	0.936975	0.012039	0.285083	0.187693	0.295962	0.247983
Saliency	0.250609	0.001482	0.501005	0.955882	0.012106	0.292470	0.185667	0.294503	0.255862

Figure 2: Evaluation metrics results from EvalxNLP for the explainers on the *MovieReviews* dataset with best scores in **bold**. Darker colors refer to better values for each metric, and bold refers to the best value per metric.

The user study first collects demographic data, including participants’ profession, NLP experience, and prior exposure to benchmarking tools. Then, they were asked to evaluate the system based on some criteria using a 5-point Likert scale to measure usability and satisfaction, where 1 and 5 mean the worst and best values, respectively, for each criterion (For brevity, we don’t present the questions here and refer to Figure 3). Figure 3 presents the results of the human evaluation where these results were collected prior to the integration of the LLM component, which was later added to enhance the understanding of explanations generated by the various explainers. Another round of human evaluation is planned to assess the effectiveness of adding the LLM-component.

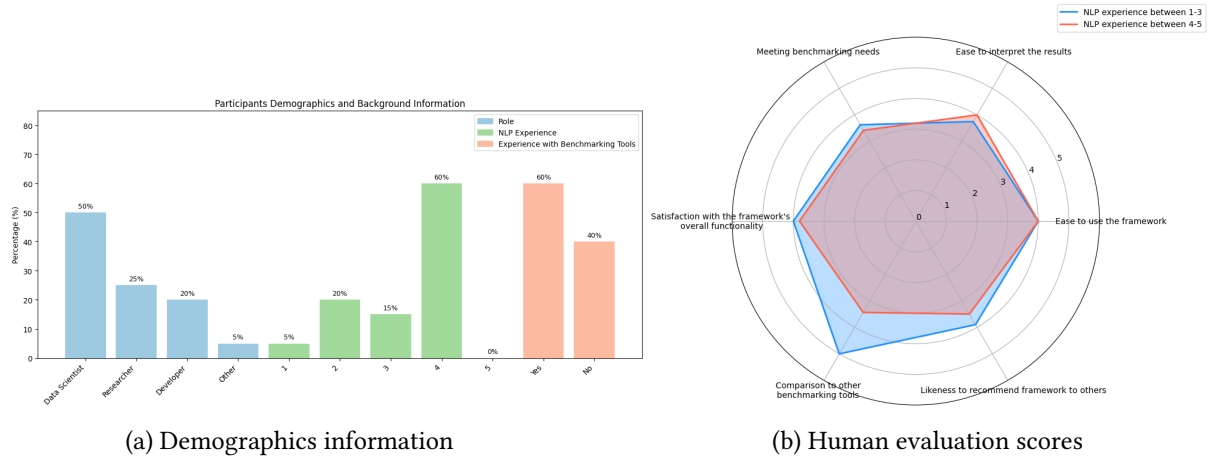


Figure 3: Zoom in for a better view) (a) Bar chart for the demographics and background information of the participants in the human evaluation. (b) Radar chart presenting the results of the human evaluation. Responses on framework comparison were collected only from participants who confirmed prior experience using other tools (2 participants for lower NLP experience and 9 participants for higher NLP experience).

Based on results in Figure 3b, The overall results are promising, with scores consistently above 3 out of 5 across all criteria for both participant groups. Particularly, the framework is easy to use, and all users, especially those with greater NLP experience, find the results easy to interpret. However, for the remaining criteria, participants with less NLP experience provided higher ratings compared to their more experienced counterparts. This indicates there remains room for improvement, particularly in enhancing the framework’s ability to meet the benchmarking needs of more experienced users.

6. Conclusion

Given the increasing number of available explainability methods and the diverse requirements stakeholders may have, there is a rising need for continued contributions to existing and new frameworks that support stakeholders in obtaining and selecting appropriate explanations tailored to their specific use cases. To address this gap, we introduce *EvalxNLP*, a novel Python framework designed to bench-

mark state-of-the-art Ph-FA explainability methods for transformer-based NLP models, particularly targeting classification tasks. The framework enables users to generate and evaluate explanations at the single-instance level and across entire real-world datasets across different metrics for three main explainability properties: faithfulness, complexity, and plausibility. *EvalxNLP* is targeted for use by various stakeholders, including laypeople, developers, and researchers, depending on their goals and also where certain properties could be more critical for specific users. For instance, developers can employ *EvalxNLP* to debug models and compare Ph-FA methods, prioritizing faithfulness metrics relevant to their needs. Our framework is developed to be easily extensible by the research community. **Limitations** of the framework include focusing on classification tasks and utilizing feature-attribution explainability methods. **Future directions** include expanding the supported methods and metrics, such as integrating recent non-feature attribution techniques like [34] and robustness metrics such as sensitivity [35]. We also plan to incorporate users feedback to refine explanation quality, and to extend the framework to generate premise–conclusion rules from Ph-FA methods to enhance explainability[36].

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions. This research has been supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS23069 Software Campus 3.0 (TU München).

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

- [1] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, Openxai: Towards a transparent evaluation of model explanations, *NeurIPS* (2022).
- [2] S. Sithakoul, S. Meftah, C. Feutry, Beexai: Benchmark to evaluate explainable ai, in: *WC on Explainable AI*, 2024.
- [3] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M. M.-C. Höhne, Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond, *JMLR* (2023).
- [4] G. Sarti, N. Feldhus, L. Sickert, O. van der Wal, Inseq: An interpretability toolkit for sequence generation models, in: *Proc. of ACL (System Demonstrations)*, 2023.
- [5] Y. Liu, S. Khandagale, C. White, W. Neiswanger, Synthetic benchmarks for scientific research in explainable machine learning, in: *NeurIPS Datasets and Benchmarks Track*, 2021.
- [6] L. Faber, A. K. Moghaddam, R. Wattenhofer, When comparing to ground truth is wrong: On evaluating gnn explanation methods, in: *Proc. of ACM SIGKDD*, 2021.
- [7] G. Attanasio, E. Pastor, C. Di Bonaventura, D. Nozza, ferret: a framework for benchmarking explainers on transformers, in: *Proc. of EACL (System Demonstrations)*, 2023.
- [8] V. Miglani, A. Yang, A. Markosyan, D. Garcia-Olano, N. Kokhlikyan, Using captum to explain generative language models, in: *Proc. of NLP-OSS Workshop at ACL*, 2023.
- [9] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al., Ai explainability 360: Impact and design, in: *Proc. of AAAI*, 2022.
- [10] X. Li, M. Du, J. Chen, Y. Chai, H. Lakkaraju, H. Xiong, M4: A unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models, *NeurIPS* (2023).
- [11] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, *arXiv preprint arXiv:1312.6034* (2013).
- [12] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: *Proc. of ICML*, 2017.
- [13] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: *Proc. of ICML*, 2017.

- [14] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: Proc. of ICML, 2017.
- [15] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806 (2014).
- [16] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proc. of ACM SIGKDD, 2016.
- [17] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, NeurIPS (2017).
- [18] M. Muschalik, H. Baniecki, F. Fumagalli, P. Kolpaczki, B. Hammer, E. Hüllermeier, shapiq: Shapley interactions for machine learning, NeurIPS (2024).
- [19] F. Fumagalli, M. Muschalik, P. Kolpaczki, E. Hüllermeier, B. Hammer, Shap-iq: Unified approximation of any-order shapley interactions, NeurIPS (2023).
- [20] M. Muschalik, H. Baniecki, F. Fumagalli, P. Kolpaczki, B. Hammer, E. Hüllermeier, shapiq: Shapley interactions for machine learning, in: NeurIPS Datasets and Benchmarks Track, 2024.
- [21] N. Feldhus, L. Hennig, M. D. Nasert, C. Ebert, R. Schwarzenberg, S. Möller, Saliency map verbalization: Comparing feature importance representations from model-free and instruction-based methods, in: Proc. of NLRSE Workshop at ACL, 2023.
- [22] O. Zaidan, J. Eisner, Modeling annotators: A generative approach to learning from annotator rationales, in: Proc. of EMNLP, 2008.
- [23] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proc. of LREC, 2022.
- [24] Z. Zhao, G. Chrysostomou, K. Bontcheva, N. Aletras, On the impact of temporal concept drift on model explanations, in: Findings of ACL: EMNLP, 2022.
- [25] Z. Zhao, N. Aletras, Incorporating attribution importance for improving faithfulness metrics, in: Proc. of ACL (Long Papers), 2023.
- [26] Z. Zhao, N. Aletras, Incorporating attribution importance for improving faithfulness metrics, in: Proc. of ACL (Long Papers), 2023.
- [27] H. Ngai, F. Rudzicz, Doctor XAvIer: Explainable diagnosis on physician-patient dialogues and XAI evaluation, in: Proc. of Workshop on Biomedical Language Processing, 2022.
- [28] P. Atanasova, A diagnostic study of explainability techniques for text classification, in: Accountable and Explainable Methods for Complex Reasoning over Text, 2024.
- [29] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proc. of ACL, 2020.
- [30] U. Bhatt, A. Weller, J. M. F. Moura, Evaluating and aggregating feature-based model explanations, Proc. of IJCAI (2021).
- [31] P. Chalasani, J. Chen, A. R. Chowdhury, X. Wu, S. Jha, Concise explanations of neural networks using adversarial training, in: Proc. of ICML, 2020.
- [32] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proc. of AAAI, 2021.
- [33] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, NeurIPS (2018).
- [34] T. Leemann, A. Fastowski, F. Pfeiffer, G. Kasneci, Attention mechanisms don't learn additive models: Rethinking feature importance for transformers, TMLR (2025).
- [35] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, P. K. Ravikumar, On the (in)fidelity and sensitivity of explanations, in: NeurIPS, 2019.
- [36] L. Rizzo, D. Verda, S. Berretta, L. Longo, A novel integration of data-driven rule generation and computational argumentation for enhanced explainable ai, Machine Learning and Knowledge Extraction 6 (2024) 2049–2073.