On the Adaptability of Attention-Based Interpretability in Different Transformer Architectures for Multi-Class Classification Tasks*

Sofia Katsaki$^{\dagger 1 [0009-0001-2270-408X]}$, Christos Aivazidis$^{\dagger 2 [XXXX-YYYY-ZZZZ-AAAA]}$, Nikolaos Mylonas$^{1 [0000-0002-5733-543X]}$, Ioannis Mollas$^{1 [0000-0002-7765-7903]}$, and Grigorios Tsoumakas$^{1 [0000-0002-7879-669X]}$

[1] School of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54632, Greece {skatsaks,myloniko,iamollas,greg}@csd.auth.gr
[2] School of Mathematics, Aristotle University of Thessaloniki, Thessaloniki 54632, Greece
{aivazidis}@math.auth.gr

Abstract. Transformers are widely recognized as leading models for NLP tasks due to their attention-based architecture. However, their complexity and numerous parameters hinder the understanding of their decision-making processes, restricting their use in high-risk domains where accurate explanations are crucial. To overcome this challenge, a technique named Optimus was introduced recently. Optimus provides an adaptive selection of head, layer, and matrix operations, to provide feature importance based interpretations for transformers. This work extends Optimus, adapting to two new transformer models, as well as the new task of multi-class classification, while also optimizing the time response of the technique. Experiments showed that the performance of Optimus remains consistent through different encoder-based transformer models and classification tasks.

Keywords: Interpretable Machine Learning · Local Interpretability · Transformers · Attention · Text Classification

1 Introduction

Transformers [31], widely recognized as the leading models for Natural Language Processing (NLP) tasks [25], employ an attention-based architecture to

[†] These authors contributed equally to this work

[*] The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 514)

effectively model intricate relationships among words and sentences. Their adaptability enables their application in diverse downstream tasks, facilitated by task-specific layers. As a result, transformers have emerged as the state-of-the-art approach for numerous NLP problems, including text classification, as examined in this study.

Despite their impressive performance, Transformer models are known for their lack of interpretability [29]. This is because of their complex architecture and numerous parameters, which make it difficult to understand how they arrive at their decisions. As a result, their use in high-risk domains [15,12], where accurate explanations are crucial for human lives and economic considerations, is limited. However, considering the significant potential of Transformers in these domains, there is a need for interpretability techniques specifically designed for Transformers.

To facilitate the utilization of transformer models across diverse applications, the need for interpretability techniques arises. While several model-agnostic techniques, such as Local Interpretable Model-agnostic Explanations (LIME) [26] or SHAP [19], can be applied to transformers, there are also model-specific techniques, including Layer-Relevance Propagation (LRP) [2,8], Integrated Gradients (IG) [30], and attention information, such as Rollout [1] and AttExplainer [24]. Regarding the latter, attention-based interpretations have been heavily criticized lately [13,4], in a recent work of us, we introduce Optimus, a transformer-specific family of local interpretation techniques, improving a faithfulness-based metric through adaptive selection of head, layer, and matrix operations [23].

While Optimus has shown to be a powerful interpretability technique for binary and multi-label classification tasks using BERT [9] and DistilBERT [28] transformers, this study explores the application of this technique on two different models, RoBERTa [18] and ALBERT [16] for multi-class classification. In addition to this adaptation, a series of experiments are conducted to demonstrate the effectiveness of Optimus in these models for the selected task, using interpretability metrics as benchmarks. We also explore an optimization procedure that we applied to enhance the speed of Optimus, and finally, we compare the time response of Optimus with its competitor, IG.

The remaining sections are organized as follows: Section 2 provides background concepts and discusses studies on transformer interpretability, including Optimus. Section 3 outlines the methodology for selecting and testing transformer architectures with Optimus, along with the corresponding modifications. The experiments are showcased in Section 4, while Section 5 concludes the study and highlights potential future directions.

2 Background and Related Work

In this section, we delve into various subjects that are relevant to our work and also examine previous research that is related to this paper. Initially, we present the concept of interpretability and highlight two widely recognized techniques associated with it. Next, we examine the metrics employed to evaluate the effectiveness of interpretability techniques. Following that, we explore the interpretability aspect of Transformers and the specific technique that forms the foundation of this study.

2.1 Interpretability

Interpretability refers to the ability of extracting reasoning behind a model's decision. It is an important aspect of machine learning and artificial intelligence in general, as it allows us to gain valuable insights about the inner workings of models that would otherwise be considered *black boxes.* These insights can then help us explain the decisions of models, which is paramount in critical applications such as healthcare, finance and autonomous systems. [7,3,14]. Interpretability techniques can be split into different categories based on their applicability on machine learning models and the scope of their provided interpretations [32].

Regarding the first categorization, we can find two distinct types of interpretability techniques, model-agnostic and model-specific. The former refers to techniques that can be applied indifferently on any type of machine learning model, and usually make use of only the model's decisions to provide explanations. The latter, includes techniques that utilize the unique characteristics of the machine learning model they try to explain, and are therefore applicable only on a specific type of model or a family of similar models. Concerning the scope of the provided interpretations, we can distinguish two types of methods, local and global. Local techniques provide interpretations for a specific instance at a time, while global methods provide an overview of the entire model's decision process.

Two of the most well-known interpretability techniques are LIME [26] IG [30]. LIME, as the name suggests, is a model-agnostic local method, which tries to create a simple interpretable model, using the predictions of the black box model, that is to be interpreted. To achieve this, LIME perturbs the input data around the instance of interest and generates a new dataset used to train the interpretable model. IG on the other hand is a neural-network specific, local interpretability technique, that uses the gradients of the model. Specifically, it calculates the gradients of the model's output with respect to the input features, and by integrating these gradients, it assigns an importance scores to each feature.

Metrics To determine the most suitable interpretability technique for a given task, it is crucial to conduct thorough evaluations [17]. While performing user-oriented experiments with individuals involved in the specific task is an ideal approach, it may not always be practical or feasible in many scenarios.

As an alternative, quantitative metrics that can be tested in a controlled lab environment are highly valuable and widely utilized. One effective method to assess the performance of interpretability techniques is by comparing the generated interpretations with a ground truth interpretation, also referred to as a rationale. Common metrics used to evaluate the accuracy of interpretations against rationales include AUPRC (Area Under the Precision-Recall Curve) and F1 token score [10]. However, it should be noted that ground-truth interpretations are not always available for all datasets.

In evaluating interpretability techniques, unsupervised metrics are valuable for assessing properties such as robustness, comprehensibility, and faithfulness. Robustness measures the stability of a technique by evaluating the degree of change in interpretations when instances are slightly modified [21]. Comprehensibility quantifies the percentage of non-zero weights in an interpretation, aiding end-user understanding [27]. Faithfulness evaluation metrics, including the popular faithfulness score, emulate user behavior to assess interpretation validity [11]. Truthfulness evaluates interpretations by iteratively removing tokens and analyzing the model's response [22]. These metrics provide objective insights without relying on human input, contributing to the evaluation of interpretability techniques.

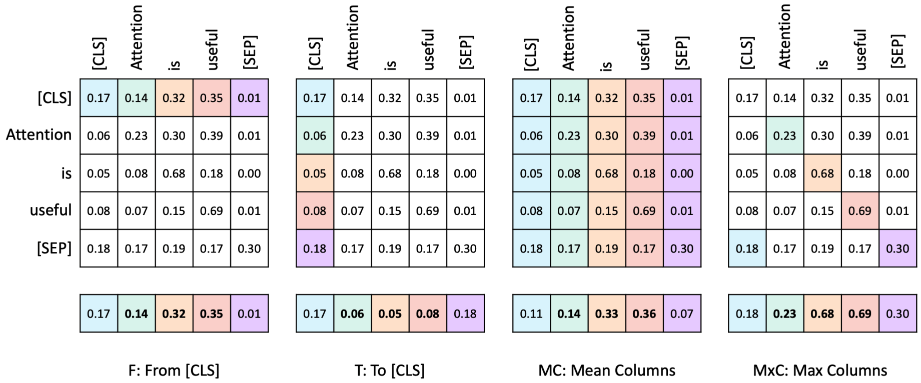
2.2 Transformers interpretability

Transformer specific interpretability techniques are scarce in the literature. Model agnostic techniques such as LIME or neural specific ones like IG can be used to provide interpretations for the decisions of Transformers. These techniques, however, do not make use of the unique attention-based architecture of those models and therefore their interpretations are not always sufficient.

A recent study introduced the Optimus family of techniques, for transformer-specific interpretability. These techniques make use of the attention matrices, that are readily available during the model's inference. These matrices exist in each attention head of each attention layer of the transformer model, and are combined in such a way to provide the best interpretation for each specific instance, batch of instances and label, depending on the specific technique of the Optimus family (Optimus Prime, Optimus Batch, Optimus Label respectively).

The combination procedure of Optimus makes use of different operations among attention matrices found in the literature, as well as a newly proposed operation the selection of specific head or layer. The best combination of operations on the head, layer, and matrix layer are selected in a way that optimizes an unsupervised faithfulness-based metric. Matrix level concerns how the interpretation is extracted, after aggregating the attention matrices from each head and layer of the Transformer. The resulting matrix from this aggregation has a size of $S \times S$, where S denotes the length of the input sequence. On the other hand, feature importance interpretations should be vectors of size S. Therefore, a procedure is needed to extract such vector from the final attention matrices. The operations studied in the original work are shown in Figure 1. Additionally, a baseline attention setup is considered in this work, making use of the most common operations found in the literature.

To aid with the interpretability process, the previous study also introduces a faithfulness-based metric called Ranked Faithful Truthfulness (RFT), which is applicable in feature importance interpretations. This metric, evaluates all parts



F: From [CLS]

	[CLS]	Attention	is	useful	[SEP]
[CLS]	0.17	0.14	0.32	0.35	0.01
Attention	0.06	0.23	0.30	0.39	0.01
is	0.05	0.08	0.68	0.18	0.00
useful	0.08	0.07	0.15	0.69	0.01
[SEP]	0.18	0.17	0.19	0.17	0.30
	0.17	**0.14**	**0.32**	**0.35**	0.01

T: To [CLS]

	[CLS]	Attention	is	useful	[SEP]
[CLS]	0.17	0.14	0.32	0.35	0.01
Attention	0.06	0.23	0.30	0.39	0.01
is	0.05	0.08	0.68	0.18	0.00
useful	0.08	0.07	0.15	0.69	0.01
[SEP]	0.18	0.17	0.19	0.17	0.30
	0.17	**0.06**	**0.05**	**0.08**	0.18

MC: Mean Columns

	[CLS]	Attention	is	useful	[SEP]
[CLS]	0.17	0.14	0.32	0.35	0.01
Attention	0.06	0.23	0.30	0.39	0.01
is	0.05	0.08	0.68	0.18	0.00
useful	0.08	0.07	0.15	0.69	0.01
[SEP]	0.18	0.17	0.19	0.17	0.30
	0.11	**0.14**	**0.33**	**0.36**	0.07

MxC: Max Columns

	[CLS]	Attention	is	useful	[SEP]
[CLS]	0.17	0.14	0.32	0.35	0.01
Attention	0.06	0.23	0.30	0.39	0.01
is	0.05	0.08	0.68	0.18	0.00
useful	0.08	0.07	0.15	0.69	0.01
[SEP]	0.18	0.17	0.19	0.17	0.30
	0.18	**0.23**	**0.68**	**0.69**	0.30

Fig. 1. Head, layer and matrix operations

of the interpretation, while also utilizing the impact each token had towards the decision during the evaluation process. A penalty is further applied on each token based on the ranking of its importance towards the decision.

A recent work introduced an attention-based interpretability method called CLS-A, which averages the attention matrices on the heads of the last layer and keeps the [CLS] token as the interpretation [5]. This combination of operations, which is one of the many studied in Optimus was found to provide interpretations of quality close to those of state-of-the-art methods, including LIME, in a user-oriented experiment.

3 Our Technique

In this work, our objectives are twofold: a) to extend Optimus to different transformer architectures, and b) to apply it to multi-class classification tasks. In addition to our primary objectives, we also focused on optimizing the time response of Optimus to make it more efficient. The following sections detail the specific steps we took to achieve these objectives.

3.1 Transformer architectures

To expand the application of Optimus to additional Transformer models, we initially had to identify certain key properties required for compatibility. Specifically, we sought Transformer models capable of performing sequence classification tasks, being encoder-based, and having readily accessible attention matrices for each head and layer. These criteria were essential to ensure the feasibility of integrating Optimus with the selected Transformer models.

The transformer being able to perform sequence classification was the main requirement, since Optimus is employed for text classification interpretability. Additionally, since in downstream tasks like sequence classification, a common practice is to employ only encoder-based models, which properly represent the terms of an input sequence and reveal token associations and dependencies, we chose for transformers used in our analysis to be encoder-based. Finally, the attention matrices being easily accessible is due to them being the primary source from which Optimus derives its interpretations.

Another restriction we imposed was regarding the pooling strategy employed during the sequence classification fine-tuning process of the Transformer models. Specifically, we only considered models that utilized either the embedding of the [CLS] token or the entire input sequence by averaging the token embeddings. These criteria led to the exclusion of models like GPT-2, which rely on different pooling strategies, that didn't meet our requirements.

With the above in mind, across the multiple models we reviewed, we explored two new Transformers namely, RoBERTa and ALBERT, which both fill our criteria. It is worth noting that slight modifications had to be made to Optimus to make the technique consistent with each of those models. Specifically, the [UNK] token present in BERT and DistilBERT was exchanged with the ¡unk¿ token for both RoBERTa and ALBERT, as those models do not recognize [UNK]. Additionally, the tokenizer used for Optimus was modified to be compatible with the ones utilized by those models.

3.2 Multi-class classification - Downstream task

Initially, Optimus was designed for binary and multi-label classification tasks. However, this study goes beyond by not only introducing two new models to Optimus, but also exploring its applicability in multi-class classification tasks. Multi-class classification was not originally tested in Optimus, so in this work, we made adaptations to the implementation of the RFT metric to ensure its suitability for multi-class scenarios. Specifically, instead of only handling binary and multi-label classification, code for handling multi-class classification was also introduced. As a result, we propose a new variation within the Optimus family called Optimus Class technique, denoted as OC. This variant finds the best attention setup for each class of each examined instance.

3.3 Optimization actions

An issue with the previous implementation of Optimus was that when producing token level interpretations, the time response of the technique showed a steep increase. The main cause behind this phenomenon was that when Optimus searched for the best attention setup according to the unsupervised metric, the examined model was continuously queried to provide predictions about each of the perturbed instances produced by the faithfulness-based metric.

To address this issue, we employed twin models. One model was utilized to generate attention matrices and hidden states, facilitating the creation of combinations - candidate interpretations. The other model was exclusively used for making predictions, particularly in the context of the RFT metric, which demands multiple model audits. With these two models, we can obtain the predictions needed for Optimus to find the best attention setups, much faster, overall decreasing the required time response. Furthermore, we introduced additional implementation improvements to further enhance the performance of Optimus. These enhancements aimed to optimize efficiency, increase speed, and improve overall functionality of the technique.

4 Experiments

This section introduces our experimental setup, including the models employed, the datasets utilized, and the performance evaluation of the examined techniques under two distinct interpretability evaluation metrics. The code for our experiments is available in the GitHub repo of Optimus under the branch 'multiclass' [†].

Datasets For our experiments, we utilized two single-label datasets, each consisting of more than two classes, to ensure the evaluation of Optimus's performance on multi-class tasks. HateXplain (HX) [20] is a single-label dataset from the hate speech domain, containing posts collected from Twitter and Gab. The annotators classified the posts as hateful, offensive, normal, or undecided if the distribution of their votes were uniformly distributed amongst the 3 classes. The latter was not taken into account due to the ambiguity of the results. Additionally, we randomly selected 10,000 samples from the original dataset where we performed a 7,000-1,000-2,000 train-val-test random split to train the models and evaluate the results. The second dataset, ESNLI [6], is a single-label classification dataset for natural language understanding. Given two sentences, the premise and hypothesis, the objective is to determine their relationship: entailment, contradiction, or neutral. We selected the first 10,000 samples where we performed a 7,000-1,000-2,000 train-val-test random split to train the models and evaluate the results. Both datasets were examined on token-level, meaning that the interpretations provided by the interpretability techniques concern each token of the input sequence rather than whole sentences. A few statistics about the datasets can be found in Table 1.

Table 1. Key statistics for each dataset and models' performance on them. Information about mean size is presented in token. Performance is measured in terms of F_1 macro (%)

			Performance			
Dataset	Mean Size	Classes	BERT	DistilBERT	RoBERTa	ALBERT
HX	23.9	3	79.06	76.34	83.04	79.22
ESNLI	24.4	3	62.58	64.93	63.28	63.33

[†] https://tinyurl.com/4wjac6ap

4.1 Setup

The transformer models deployed in our experimentation on the multi-class sequence classification task, were the ones encountered in the initial publication of Optimus, namely BERT and DistilBERT, along with two additional models, that were ALBERT and RoBERTa. We selected the base implementation for all the transformers utilized. Both BERT and DistilBERT contained casing information, while ALBERT and RoBERTa did not. It should be mentioned that for ALBERT, we used the second version of the pre-trained model. Table 1 presents the F_1 macro score results obtained from fine-tuning the models on the two datasets.

4.2 Metrics

Through the course of our experimentation on the attention-based interpretability approach of Optimus, two metrics were utilized, with the view to assess the performance of the explanations provided by each one of the techniques. First and foremost, we employed the RFT metric, which factors both faithfulness and truthfulness, which was discussed in Section 2.2. The second metric deployed was AUPRC, introduced in Section 2.1. The inclusion of the latter was deemed as indispensable, since HX, as well as ESNLI, incorporate rationales delivered by human annotators. It is worth noting that for both of the examined metrics and datasets, the performance showcased is the average score of the metric across all test instances as evaluated on the predicted classes.

4.3 Results

Table 2 presents the RFT performance of Integrated Gradients (IG) along with the variations of Optimus. LIME was excluded from this set of experiments as the results of the original paper suggest that it is both time-consuming and low-performing. Similar to the original work, ‘B’ represents the Baseline attention setup, which includes averaging attention heads and layers, as well as using the “From [CLS]” strategy at the matrix level. ‘OB’ refers to Optimus Batch. The newly introduced variant, Optimus Class, is denoted as OC in the table. Similarly, Table 3 presents the interpretability techniques’ performance in terms of AUPRC.

As we can see from the results, Optimus provides competitive interpretations compared to IG even on the task of multi-class classification in terms of both metrics. This phenomenon remains the same across all different Transformer models examined. Specifically, the baseline attention setup achieves results close or even higher to IG, while the variants of OB and OC outperform IG in most cases (6 out of 8 for both in terms of RFT and 7 out of 8 for OB and 6 out of 8 for OC in terms of AUPRC).

Based on our experimental procedure, we showed that the quality of interpretations provided by Optimus is consistent with different encoder-based Transformers, showcasing the applicability of the technique for a wider range

Table 2. Performance of interpretability techniques in terms RFT when explaining several transformer models on different datasets. Best performance denoted with bold

	IG	B	OB	OC
ESNLI (BERT)	0.456	0.488	0.615	**0.876**
ESNLI (DistilBERT)	0.385	0.481	0.552	**0.706**
ESNLI (RoBERTa)	0.442	0.266	0.597	**0.876**
ESNLI (ALBERT)	0.259	0.612	0.664	**0.863**
HX (BERT)	**0.476**	0.337	0.371	0.458
HX (DistilBERT)	**0.467**	0.357	0.379	0.455
HX (RoBERTa)	0.350	0.350	0.355	**0.422**
HX (ALBERT)	0.314	0.408	0.433	**0.562**

Table 3. Performance of interpretability techniques in terms AUPRC when explaining several transformer models on different datasets. Best performance denoted with bold

	IG	B	OB	OC
ESNLI (BERT)	0.290	0.514	**0.614**	0.443
ESNLI (DistilBERT)	0.301	0.576	**0.651**	0.498
ESNLI (RoBERTa)	0.316	0.274	**0.593**	0.408
ESNLI (ALBERT)	0.337	0.602	**0.604**	0.438
HX (BERT)	0.508	0.488	**0.541**	0.500
HX (DistilBERT)	0.481	0.498	**0.531**	0.506
HX (RoBERTa)	0.477	**0.523**	0.514	0.489
HX (ALBERT)	**0.464**	0.408	0.422	0.413

of models. Additionally, attention-based interpretations were found to be of the same or even higher quality than those provided by state-of-the-art methods for the task of multi-class classification as well.

Time optimization By optimizing our technique and adopting a twin-model architecture, we successfully reduced the runtime of Optimus. Specifically, when tested on the ESNLI dataset using the BERT model, we achieved a reduction in time-response for a single instance from 3.35 (original time-response) to 2.70 seconds. This improvement renders Optimus suitable for online scenarios, where quick interpretation is essential.

Time comparison In addition, we conducted a time comparison between IG and our newly proposed variant OC. The results of this analysis can be found in Table 4, which demonstrates that despite the time optimization efforts in Optimus, IG still exhibits faster performance compared to our technique. Specifically, it is observed that OC is slower than IG across all examined datasets and models. This can be attributed to the time-consuming nature of the RFT metric, which is required for the interpretation production process in Optimus.

Table 4. Average time response (seconds) of the examined techniques across different models and datasets

	ESNLI				HX			
	BERT	DistilBERT	RoBERTa	ALBERT	BERT	DistilBERT	RoBERTa	ALBERT
IG	0.75	0.5	0.75	0.88	0.85	0.51	0.75	0.83
OC	2.70	1.67	3.17	3.08	3.08	1.75	3.42	3.51

5 Conclusions

In this work, we explored the adaptability of Optimus, a transformer-specific attention-based interpretability technique, on different encoder-based transformers and the task of multi-class classification. Results were found to be consistent with the original claims of the technique being applicable to additional encoder-based Transformer models, while also showcasing the competitiveness of attention-based interpretations in another popular downstream task, multi-class classification. Specifically, two different multi-class datasets were used, and two interpretability evaluation metrics were measured.

Our proposed variant, Optimus Class, tested on two new Transformer models, seems to be consistent with the results showcased in the original paper, outperforming IG in most cases for both metrics and datasets. Nevertheless, despite our time optimization process improving the overall time response of the technique, Optimus Class, is still slower than IG, when evaluated on token-level interpretations.

5.1 Limitations

This extension of Optimus inherits some limitations from the original technique, while also optimizing the process to some extent. By utilizing two separate models—one for predictions and one solely for attention—the technique seeks to mitigate some of the challenges. However, the inherent limitations, such as the time-consuming nature of operation selection and scalability issues, still persist to some degree.

5.2 Future directions

Our primary objectives for future analysis of Optimus will be to encompass an expanded scope of transformers, including those with an encoder-decoder-based or decoded-oriented architecture. We will also integrate a wider variety of NLP tasks and incorporate additional datasets. Another avenue of investigation would entail studying the cognitive systems of the end-users, to comprehend their perception and deductive reasoning processes, with an aim to provide more user-friendly IML APIs, which allow end-users to manually adjust the features and parameters of these IML systems. To achieve this, we would like to conduct a large-scale, extensive human-oriented user study.

References

1. Abnar, S., Zuidema, W.H.: Quantifying attention flow in transformers. CoRR **abs/2005.00928** (2020), `https://arxiv.org/abs/2005.00928`
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), 1–46 (07 2015). https://doi.org/10.1371/journal.pone.0130140, `https://doi.org/10.1371/journal.pone.0130140`
3. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R.: How to explain individual classification decisions. J. Mach. Learn. Res. **11**, 1803–1831 (aug 2010)
4. Bastings, J., Filippova, K.: The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In: BlackboxNLP@EMNLP. pp. 149–155. ACL, Online (2020)
5. Bhan, M., Achache, N., Legrand, V., Blangero, A., Chesneau, N.: Evaluating self-attention interpretability through human-grounded experimental protocol (2023)
6. Camburu, O.M., Rocktäschel, T., Lukasiewicz, T., Blunsom, P.: e-snli: Natural language inference with natural language explanations. Advances in Neural Information Processing Systems **31** (2018)
7. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1721–1730. KDD '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2783258.2788613, `https://doi.org/10.1145/2783258.2788613`
8. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 782–791 (June 2021)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. ACL, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
10. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: ERASER: A benchmark to evaluate rationalized NLP models. In: Proceedings of the 58th Annual Meeting of the ACL. pp. 4443–4458. ACL, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.408
11. Du, M., Liu, N., Yang, F., Ji, S., Hu, X.: On attribution of recurrent neural network predictions via additive decomposition. In: The World Wide Web Conference. pp. 383–393 (2019)
12. EU: Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence (ai act) and amending certain union legislative acts. EUR-Lex-52021PC0206 (2021), `https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence`
13. Jain, S., Wallace, B.C.: Attention is not explanation. In: NAACL-HLT. pp. 3543–3556. ACL, Minneapolis, Minnesota (2019)
14. Koopman, P., Wagner, M.: Challenges in autonomous vehicle testing and validation. SAE International Journal of Transportation Safety **4**(1), 15–24 (apr

2016). https://doi.org/https://doi.org/10.4271/2016-01-0128, https://doi.org/10.4271/2016-01-0128

15. Kuziemski, M., Misuraca, G.: Ai governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. Telecommunications Policy **44**(6), 101976 (2020). https://doi.org/https://doi.org/10.1016/j.telpol.2020.101976, https://www.sciencedirect.com/science/article/pii/S0308596120300689, artificial intelligence, economy and society
16. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=H1eA7AEtvS
17. Lertvittayakumjorn, P., Toni, F.: Human-grounded evaluations of explanation methods for text classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, November 3-7. pp. 5194–5204. ACL, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1523
18. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
19. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc., Long Beach, California (2017), shorturl.at/bdsT4
20. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: A benchmark dataset for explainable hate speech detection. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, February 2-9. pp. 14867–14875. AAAI Press, Online (2021), https://ojs.aaai.org/index.php/AAAI/article/view/17745
21. Melis, D.A., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. In: Advances in Neural Information Processing Systems. pp. 7775–7784. Montreal, Canada (2018)
22. Mollas, I., Bassiliades, N., Tsoumakas, G.: Truthful meta-explanations for local interpretability of machine learning models (2022)
23. Mylonas, N., Mollas, I., Tsoumakas, G.: An attention matrix for every decision: Faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification (2022), to appear in Data Mining & Knowledge Discovery, Springer, as part of the ECML PKDD 2023 Journal Track
24. Niu, R., Wei, Z., Wang, Y., Wang, Q.: Attexplainer: Explain transformer via attention by reinforcement learning. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 724–731. Vienna, Austria (7 2022). https://doi.org/10.24963/ijcai.2022/102
25. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: A survey. Science China Technological Sciences **63**(10), 1872–1897 (2020)
26. Ribeiro, M.T., Singh, S., Guestrin, C.: 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. pp. 1135–1144. ACM (2016)

27. Robnik-Sikonja, M., Bohanec, M.: Perturbation-based explanations of prediction models. In: Human and Machine Learning - Visible, Explainable, Trustworthy and Transparent, pp. 159–175. Springer, International (2018). https://doi.org/10.1007/978-3-319-90403-0_9
28. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In: NeurIPS EMC^2 Workshop (2019)
29. Schwenke, L., Atzmueller, M.: Show me what you're looking for: Visualizing abstracted transformer attention for enhancing their local interpretability on time series data. The International FLAIRS Conference Proceedings **34** (Apr 2021). https://doi.org/10.32473/flairs.v34i1.128399, https://journals.flvc.org/FLAIRS/article/view/128399
30. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 6-11 August. vol. 70, pp. 3319–3328. PMLR, Sydney, NSW, Australia (2017), http://proceedings.mlr.press/v70/sundararajan17a.html
31. 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 **30** (2017)
32. Zhang, Y., Tiňo, P., Leonardis, A., Tang, K.: A survey on neural network interpretability. IEEE Trans. Emerg. Top. Comput. Intell. **5**(5), 726–742 (2021). https://doi.org/10.1109/TETCI.2021.3100641, https://doi.org/10.1109/TETCI.2021.3100641