

The Quarrel of Local Post-hoc Explainers for Moral Values Classification in Natural Language Processing

Agiollo, Andrea; Cavalcante Siebert, Luciano; Murukannaiah, Pradeep Kumar; Omicini, Andrea

DOI

[10.1007/978-3-031-40878-6_6](https://doi.org/10.1007/978-3-031-40878-6_6)

Publication date

2023

Document Version

Final published version

Published in

Explainable and Transparent AI and Multi-Agent Systems - 5th International Workshop, EXTRAAMAS 2023, Revised Selected Papers

Citation (APA)

Agiollo, A., Cavalcante Siebert, L., Murukannaiah, P. K., & Omicini, A. (2023). The Quarrel of Local Post-hoc Explainers for Moral Values Classification in Natural Language Processing. In D. Calvaresi, A. Najjar, A. Omicini, R. Carli, G. Ciatto, R. Aydogan, Y. Mualla, & K. Främling (Eds.), *Explainable and Transparent AI and Multi-Agent Systems - 5th International Workshop, EXTRAAMAS 2023, Revised Selected Papers* (pp. 97-115). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 14127 LNAI). Springer. https://doi.org/10.1007/978-3-031-40878-6_6

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

<https://www.openaccess.nl/en/you-share-we-take-care>

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.



The Quarrel of Local Post-hoc Explainers for Moral Values Classification in Natural Language Processing

Andrea Agiollo¹(✉) , Luciano Cavalcante Siebert² ,
Pradeep Kumar Murukannaiah² , and Andrea Omicini¹

¹ Dipartimento di Informatica – Scienza e Ingegneria (DISI), Alma Mater Studiorum—Università di Bologna, Cesena, Italy

{andrea.agiollo, andrea.omicini}@unibo.it

² Delft University of Technology, Delft, The Netherlands

{L.CavalcanteSiebert, P.K.Murukannaiah}@tudelft.nl

Abstract. Although popular and effective, *large language models* (LLM) are characterised by a performance vs. transparency trade-off that hinders their applicability to sensitive scenarios. This is the main reason behind many approaches focusing on local post-hoc explanations recently proposed by the XAI community. However, to the best of our knowledge, a thorough comparison among available explainability techniques is currently missing, mainly for the lack of a general metric to measure their benefits. We compare state-of-the-art local post-hoc explanation mechanisms for models trained over moral value classification tasks based on a measure of correlation. By relying on a novel framework for comparing global impact scores, our experiments show how most local post-hoc explainers are loosely correlated, and highlight huge discrepancies in their results—their “quarrel” about explanations. Finally, we compare the impact scores distribution obtained from each local post-hoc explainer with human-made dictionaries, and point out that there is no correlation between explanation outputs and the concepts humans consider as salient.

Keywords: Natural Language Processing · Moral Values Classification · eXplainable Artificial Intelligence · Local Post-hoc Explanations

1 Introduction

Large Language Models (LLMs) represent the de-facto solution for dealing with complex Natural Language Processing (NLP) tasks such as sentiment analysis [45], question [19], and many others [34]. The ever-increasing popularity of such data-driven approaches is widely caused by their performance improvements against their counterparts. Indeed, Neural Network (NN) based approaches have shown uncanny performance over different NLP tasks such as grammar acceptability of a sentence [43] and text translation [40]. However, following the quest

for higher performance, research efforts gave birth to ever more complex NN architectures such as BERT [15], GPT [12], and T5 [35].

Although being powerful and, empirically, reliable, LLM suffer from a performance vs. transparency trade-off [11, 47]. Indeed, LLM are *black-box* models, as they rely on the optimisation of their numerical sub-symbolical components, which are mostly unreadable by humans. The *black-box* nature of LLMs hinder their applicability to some scenarios where transparency represents a fundamental requirement, e.g., NLP for medical analysis [29, 39], etc. Therefore, there exists the need to identify relevant mechanisms capable of opening such LLM black-boxes and diagnose their reasoning process, and presenting it in a human-understandable fashion. Towards this aim, a few different explainability approaches, focusing mostly on *Local Post-hoc explainer (LPE)* mechanisms, have been recently proposed. An LPE represents a popular solution to explain the reasoning process by highlighting how different portions of the input sample impact differently the produced output, by assigning a relevance score to each input component. These approaches apply to single instances of input sample, thus being *local*, and to optimised LLM—thus being *post-hoc*.

Despite a broad variety of LPE approaches, the state of the art lacks a fair comparison among them. A common trend for proposals of novel explanation mechanisms is to highlight its advantages through a set of tailored experiments. This hinders comparison fairness, making it very difficult to identify the best approach for obtaining explanations of NLP models or even to know if such a best approach exists. This is why we present a framework for comparing several well-known LPE mechanisms for text classification in NLP. Aiming at obtaining comparison fairness, we rely on aggregating the local explanations obtained by each local post-hoc explainer into a set of global impact scores. Such scores identify the set of concepts that best describe the underlying NLP model from the perspective of each LPE. These concepts, along with their aggregated impact scores, are then compared for each LPE against other LPE counterparts. The comparison between the aggregated global impact scores rather than the single explanations is justified by the locality of LPE approaches. Indeed, it is reasonable for local explanations of different LPEs to differ somehow, depending on the approach design, therefore making it complex to compare the quality of two LPEs over the same sample. However, it is also expected for the aggregated global impacts to be aligned between different LPE as they are applied to the same NN, which leverages the same set of relevant concepts for its inference. Therefore, when comparing the aggregated impact scores of different LPE, we expect them to be correlated—at least up to a certain extent.

We perform our comparison between LPE explanations across the social domains available in the Moral Foundation Twitter Corpus (MFTC) [20]. MFTC represents an example of a challenging task, as it is proposed to tackle moral values classification. Moral values are inherently subjective to human readers, therefore introducing possible disagreement inside annotations and making the overall optimisation pipeline sensitive to small changes. Moreover, identifying moral values represents a sensitive task, as it requires a deep and safe understanding of

complex concepts such as harm and fairness. Consequently, we believe MFTC to represent a suitable option for analysing the behaviour of LPEs over different scenarios. Moreover, relying on MFTC enables a comparison between extracted relevant concepts and a set of humanly tailored impact scores, namely Moral Foundations Dictionary (MFD) [21]. Therefore, allowing us to study the extent of correlation between LPE-extracted concepts and humanly salient concepts. Surprisingly, our experiments show how there are setups where the explanations of different LPEs are far from being correlated, highlighting how explanation quality is highly dependent on the chosen eXplainable Artificial Intelligence (xAI) approach and the respective scenario at hand. There are huge discrepancies in the results of different state-of-the-art local explainers, each of which identifies a set of relevant concepts that largely differs from the others—at least in terms of relative impact scores. Therefore, we stress the need for identifying a robust approach to compare the quality of explanations and the approaches for their extraction. Moreover, the comparison between the distribution of LPEs’ impact scores and the set of human-tailored impact scores shows how there exists almost always no correlation between salient concepts extracted from the NN model and concepts relevant for humans. The obtained results highlight the fragility of xAI approaches for NLP, caused mainly by the complexity of large NN models, their inclination to the extreme fitting of data—with no regard for concept meaning—and the lack of sound techniques for comparing xAI mechanisms.

2 Background

2.1 Explanation Mechanisms in NLP

The set of explanations extraction mechanisms available in the xAI community are often categorized along two main aspects [2, 17]: (i) *local* against *global* explanations, and (ii) *self-explaining* against *post-hoc* approaches. In the former context, *local* identifies the set of explainability approaches that given a single input, i.e., sample or sentence, produce an explanation of the reasoning process followed by the NN model to output its prediction for the given input [32]. In contrast, *global* explanations aim at expressing the reasoning process of the NN model as a whole [18, 22]. Given the complexity of the NN models leveraged for tackling most NLP tasks, it is worth noticing how there is a significant lack of *global* explainability systems, whereas a variety of *local* xAI approaches are available [31, 37].

About the latter aspect, we define *post-hoc* as those set of explainability approaches which apply to an already optimized black-box model for which it is required to obtain some sort of insight [33]. Therefore, a *post-hoc* approach requires additional operations to be performed after that the model outputs its predictions [14]. Conversely, inherently explainable, i.e., *self-explaining*, mechanisms aim at building a predictor having a transparent reasoning process by design, e.g., CART [30]. Therefore, a self-explaining approach can be seen as generating the explanation along with its prediction, using the information emitted by the model as a result of the process of making that prediction [14].

In this paper, we focus on *local post-hoc* explanation approaches applied to NLP. Here, it represents a popular solution to explain the reasoning process by highlighting how different portions, i.e., words, of the input sample impact differently the produced output, by assigning a relevance score to each input component. The relevance score is then highlighted using some saliency map to ease the visualisation of the obtained explanation. Therefore, it is also common for local post-hoc explanations to be referred to as *saliency* approaches, as they aim at highlighting salient components.

2.2 Moral Foundation Twitter Corpus Dataset

In our experiments, we select the MFTC dataset as the target classification task. The MFTC dataset is composed of 35,108 tweets – sentences –, which can be considered as a collection of seven different datasets. Each split of MFTC corresponds to a different context. Here, tweets corresponding to the dataset samples are collected following a certain event or target. As an example, tweets belonging to the Black Lives Matter (BLM) split were collected during the period of Black Lives Matter protests in the US. The list of all MFTC subjects is the following: *(i)* All Lives Matter (ALM), *(ii)* BLM, *(iii)* Baltimore protests (BLT), *(iv)* hate speech and offensive language (DAV), *(v)* presidential election (ELE), *(vi)* MeToo movement (MT), *(vii)* hurricane Sandy (SND). In our experiments we also considered training and testing the NN model over the totality of MFTC tweets. This was done to analyse the LPEs behaviour over an unbiased task, as the average morality of each MFTC split is influenced by the corresponding collection event.

Each tweet in MFTC is labelled, following the same moral theory, with one or more of the following 11 moral values: *(i)* care/harm, *(ii)* fairness/cheating, *(iii)* loyalty/betrayal, *(iv)* authority/subversion, *(v)* purity/degradation, *(vi)* non-moral. Ten of the 11 available moral values are obtained as a moral concept and its opposite expression—e.g., fairness refers to the act of supporting fairness and equality, while cheating refers to the act of refraining from cheating or exploiting others. Given morality subjectivity, each tweet is labelled by multiple annotators, and the final moral labels are obtained via majority voting.

Finally, similar to previous works [28, 36], we preprocess the tweets before using them as input samples for our LLM training. We preprocess the tweets by removing URLs, emails, usernames and mentions, as well as correcting common spelling mistakes and converting emojis to their respective lemmas using the Ekphrasis package¹ and the Python Emoji package², respectively.

3 Methodology

In this section, we present our methodology for comparing LPE mechanisms. We first propose an overview of the proposed approach in Sect. 3.1. Subsequently, the set of LPE mechanisms adopted in our experiments are presented in Sect. 3.2,

¹ <https://github.com/cbaziotis/ekphrasis>.

² <https://pypi.org/project/emoji/>.

and the aggregation approaches leveraged to obtain global impact scores from LPE outputs are described in Sect. 3.3. Finally, in Sect. 3.4 we present the metrics used to identify the correlation between LPEs.

3.1 Overview

Given the complexity of measuring different LPE approaches over single local explanations, we here consider measuring how much LPEs correlate with each other over a set of fixed samples. The underlying assumption of our framework is that various LPE techniques aim at explaining the same NN model used for prediction. Therefore, while explanations may differ over local samples, it is reasonable to assume that reliable LPEs when applied over a vast set of samples—sentences or set of sentences—should converge to similar (correlated) results. Indeed, the underlying LLM considers being relevant for its inference always the same set of concepts—lemmas. A lack of correlation between different LPE mechanisms would hint that there exists a conflict between the set of concepts that each explanation mechanisms consider as relevant for the LLM, thus making at least one, if not all, of the explanations unreliable.

Being interested in analysing the correlation between a set of LPEs over the same pool of samples, we first define ϵ_{NN} as a LPE technique applied to a NN model at hand. Being local, ϵ_{NN} is applied to the single input sample $\mathbf{x}_i$, producing as output one impact score for each component (token) of the input sample l_k . Throughout the remainder of the paper, we consider l_k to be the lemmas corresponding to the input components. Mathematically, we define the output impact score for a single token or its corresponding lemma as $j(l_k, \epsilon_{NN}(\mathbf{x}_i))$. Depending on the given ϵ_{NN} , the corresponding impact score j may be associated with a single label – i.e., moral value –, making j a scalar value, or with a set of labels, making j a vector—one scalar value for each label. To enable comparing different LPE, we define the aggregated impact scores of a LPE mechanism over a NN model and a set of samples $\mathcal{S}$ as $\epsilon_{NN}(\mathcal{S})$. In our framework we obtain $\epsilon_{NN}(\mathcal{S})$ aggregating $\epsilon_{NN}(\mathbf{x}_i)$ for each $\mathbf{x}_i \in \mathcal{S}$ using an aggregation operation $\mathcal{A}$, mathematically:

$$\epsilon_{NN}(\mathcal{S}) = \mathcal{A}(\{\epsilon_{NN}(\mathbf{x}_i) \text{ for each } \mathbf{x}_i \in \mathcal{S}\}). \quad (1)$$

Defining a correlation metric $\mathcal{C}$, we obtain from Eq. (1) the following for describing the correlation between two LPE techniques:

$$\mathcal{C}(\epsilon_{NN}(\mathcal{S}), \epsilon'_{NN}(\mathcal{S})) = \mathcal{C}(\mathcal{A}(\{\epsilon_{NN}(\mathbf{x}_i) \text{ for each } \mathbf{x}_i \in \mathcal{S}\}), \mathcal{A}(\{\epsilon'_{NN}(\mathbf{x}_i) \text{ for each } \mathbf{x}_i \in \mathcal{S}\})) \quad (2)$$

where ϵ_{NN} and ϵ'_{NN} are two LPE techniques applied to the same NN model.

3.2 Local Post-hoc Explanations

In our framework, we consider seven different LPE approaches for extracting local explanations $j(l_k, \epsilon_{NN}(\mathbf{x}_i))$ from an input sentence $\mathbf{x}_i$ and the trained

LLM—identified as NN . The seven LPEs are selected in order to represent as faithfully as possible the state-of-the-art of xAI approaches in NLP. Subsequently, we briefly describe each of the seven selected LPE. However, a detailed analysis of these LPEs is out of the scope of this paper and we refer interested readers to [14, 32, 38].

Gradient Sensitivity Analysis. The Gradient Sensitivity analysis (GS) probably represents the simplest approach for assigning relevance scores to input components. GS relies on computing gradients over inputs components as $\frac{\delta f_c(\mathbf{x}_i)}{\delta \mathbf{x}_{i,k}}$, which represents the derivative of the output with respect to the k^{th} component of $\mathbf{x}_i$. Following this approach local impact scores of an input component can be thus defined as:

$$j(l_k, \epsilon_{NN}(\mathbf{x}_i)) = \frac{\delta f_{\tau_m}(\mathbf{x}_i)}{\delta \mathbf{x}_{i,k}}, \quad (3)$$

where $f_{\tau_m}(\mathbf{x}_i)$ represents the predicted probability distribution of an input sequence $\mathbf{x}_i$ over a target class τ_m . While simple, GS has been shown to be an effective approach for understanding approximate input components relevance. However, this approach suffers from a variety of drawbacks, mainly linked with its inability to define negative contributions of input components for a specific prediction—i.e., negative impact scores.

Gradient $\times$ Input Aiming at addressing few of the limitations affecting GS, the Gradient $\times$ Input (GI) approach defines the relevance scores assignment as GS multiplied – element-wise – with $\mathbf{x}_{i,k}$ [25]. Therefore, mathematically speaking, GI impact scores are defined as:

$$j(l_k, \epsilon_{NN}(\mathbf{x}_i)) = \mathbf{x}_{i,k} \cdot \frac{\delta f_{\tau_m}(\mathbf{x}_i)}{\delta \mathbf{x}_{i,k}}, \quad (4)$$

where notation follows the one of Eq. (3). Being very similar to GS, GI inherits most of its limitations.

Layer-Wise Relevance Propagation. Building on top of gradient-based relevance scores mechanisms – such as GS and GI –, Layer-wise Relevance Propagation (LRP) proposes a novel mechanism relying on conservation of relevance scores across the layers of the NN at hand. Indeed, LRP relies on the following assumptions: (i) NN can be decomposed into several layers of computation; (ii) there exists a relevance score $R_d^{(l)}$ for each dimension $\mathbf{z}_d^{(l)}$ of the vector $\mathbf{z}^{(l)}$ obtained as the output of the l^{th} layer of the NN; and (iii) the total relevance scores across dimensions should propagate through all layers of the NN model, mathematically:

$$f(\mathbf{x}) = \sum_{d \in L} R_d^{(L)} = \sum_{d \in L-1} R_d^{(L-1)} = \dots = \sum_{d \in 1} R_d^{(1)}, \quad (5)$$

where, $f(\mathbf{x})$ represents the predicted probability distribution of an input sequence $\mathbf{x}$, and L the number of layers of the NN at hand. Moreover, LRP defines a propagation rule for obtaining $R_d^{(l)}$ from $R^{(l+1)}$. However, the derivation of such propagation rule is out of the scope of this paper and thus we refer interested readers to [8, 10]. In our experiments, we consider as impact scores the relevance scores of the input layer, namely $j(l_k, \epsilon_{NN}(\mathbf{x}_i)) = R_d^{(1)}$.

Layer-Wise Attention Tracing. Since LLMs rely heavily on self-attention mechanisms [42], recent efforts propose to identify input components relevance scores analysing solely the relevance scores of attentions heads of LLM models, introducing Layer-wise Attention Tracing (LAT) [1, 44]. Building on top of LRP, LAT propose to redistribute the inner relevance scores $R^{(l)}$ across dimensions using solely self-attention weights. Therefore, LAT defines a custom redistribution rule as:

$$R_i^{(l)} = \sum_{k \text{ s.t. } i \text{ is input for neuron } k} \sum_h \mathbf{a}^{(h)} R_{k,h}^{(l+1)}, \quad (6)$$

where, h corresponds to the attention head index, while $\mathbf{a}^{(h)}$ are the corresponding learnt weights of the attention head. Similarly to LRP, we here consider as impact scores the relevance scores of the input layer, namely $j(l_k, \epsilon_{NN}(\mathbf{x}_i)) = R^{(1)}$.

Integrated Gradient. Motivated by the shortcomings of previously proposed gradient-based relevance score attribution mechanisms – such as GS and GI –, Sundararajan et al. [41] propose a novel Integrated Gradient approach. The proposed approach aims at explaining the input sample components relevance by integrating the gradient along some trajectory of the input space, which links some baseline value $\mathbf{x}'_i$ to the sample under examination $\mathbf{x}_i$. Therefore, the relevance score of the input k^{th} component of the input sample $\mathbf{x}_i$ is obtained following

$$j(l_k, \epsilon_{NN}(\mathbf{x}_i)) = (\mathbf{x}_{i,k} - \mathbf{x}'_{i,k}) \cdot \int_{a=0}^1 \frac{\delta f(\mathbf{x}'_i + t \cdot (\mathbf{x}_i - \mathbf{x}'_i))}{\delta \mathbf{x}_{i,k}} dt, \quad (7)$$

where $\mathbf{x}_{i,k}$ represents the k^{th} component of the input sample $\mathbf{x}_i$. By integrating the gradient along an input space trajectory, the authors aim at addressing the locality issue of gradient information. In our experiments we refer to the Integrated Gradient approach as HESS, as for its implementation we rely on the integrated hessian library available for hugging face models³.

SHAP. SHapley Additive exPlanations (SHAP) relies on Shapley values to identify the contribution of each component of the input sample toward the final prediction distribution. The Shapley value concept derives from game theory,

³ https://github.com/suinleelab/path_explain.

where it represents a solution for a cooperative game, found assigning a distribution of a total surplus generated by the players coalition. SHAP computes the impact of an input component as its marginal contribution toward a label τ_m , computed deleting the component from the input and evaluating the output discrepancy. Firstly defined for explaining simple NN models [31], in our experiments we leverage the extension of SHAP supporting transformer models such as BERT [26], available in the SHAP python library⁴.

LIME. Similarly to SHAP, Local Interpretable Model-agnostic Explanations (LIME) relies on input sample perturbation to identify its relevant components. Here, the predictions of the NN at hand are explained via learning an explainable surrogate model [37]. More in detail, to obtain its explanations LIME constructs a set of samples from the perturbation of the input observation under examination. The constructed samples are considered to be close to the observation to be explained from a geometric perspective, thus considering small perturbation of the input. The explainable surrogate model is then trained over the constructed set of samples, obtaining the corresponding local explanation. Given an input sentence, we here consider obtaining its perturbed version via words – or tokens – removal and words substitution. In our experiments, we rely on the already available LIME python library⁵.

3.3 Aggregating Local Explanations

Once local explanations of the NN model are obtained for each input sentence – i.e., tweet –, we aggregate them to obtain a global list of concept impact scores. Before aggregating the local impact scores, we convert the words composing local explanations into their corresponding lemmas – i.e., concepts – to avoid issues when aggregating different words expressing the same concept—e.g., hate and hateful. As there exists no bullet-proof solution for aggregating different impact scores, we adopt four different approaches in our experiments, namely:

- *Sum.* A simple summation operation is leveraged to obtain the aggregated score for each lemma. While simple this aggregation approach is effective when dealing with additive impact scores such as SHAP values. However, it suffers from lemma frequency issues, as it tends to overestimate frequent lemmas having average low impact scores. Global impact scores are here defined as $J(l_k, \epsilon_{NN}) = \sum_{i=1}^N j(l_k, \epsilon_{NN}(\mathbf{x}_i))$. Therefore, we here define $\mathcal{A}$ as

$$\mathcal{A}(\{\epsilon_{NN}(\mathbf{x}_i) \text{ for each } \mathbf{x}_i \in \mathcal{S}\}) = \left\{ \sum_{i=1}^N j(l_k, \epsilon_{NN}(\mathbf{x}_i)) \text{ for each } l_k \in \mathcal{S} \right\}. \quad (8)$$

⁴ <https://github.com/slundberg/shap>.

⁵ <https://github.com/marcotcr/lime>.

- *Absolute sum.* We here consider summing the absolute values of the local impact scores – rather than their true values – to increase the awareness of global impact scores towards lemmas having both high positive and high negative impact over some sentences. Mathematically, we obtain aggregated scores as $J(l_k, \epsilon_{NN}) = \sum_{i=1}^N |j(l_k, \epsilon_{NN}(\mathbf{x}_i))|$.

$$\mathcal{A}(\{\epsilon_{NN}(\mathbf{x}_i) \text{ for each } \mathbf{x}_i \in \mathcal{S}\}) = \left\{ \sum_{i=1}^N |j(l_k, \epsilon_{NN}(\mathbf{x}_i))| \text{ for each } l_k \in \mathcal{S} \right\}. \quad (9)$$

- *Average.* Similar to the sum operation, we here consider obtaining aggregated scores averaging local impact scores, thus avoiding possible overshooting issues arising when dealing with very frequent lemmas. Mathematically, we define $J(l_k, \epsilon_{NN}) = \frac{1}{N} \cdot \sum_{i=1}^N j(l_k, \epsilon_{NN}(\mathbf{x}_i))$.

$$\mathcal{A}(\{\epsilon_{NN}(\mathbf{x}_i) \text{ for each } \mathbf{x}_i \in \mathcal{S}\}) = \left\{ \frac{1}{N} \cdot \sum_{i=1}^N j(l_k, \epsilon_{NN}(\mathbf{x}_i)) \text{ for each } l_k \in \mathcal{S} \right\}. \quad (10)$$

- *Absolute average.* Similarly to absolute sum, we here consider to average absolute values of local impact scores for better-managing lemmas having a skewed impact as well as tackling frequency issues. Global impact scores are here defined as $J(l_k, \epsilon_{NN}) = \frac{1}{N} \cdot \sum_{i=1}^N |j(l_k, \epsilon_{NN}(\mathbf{x}_i))|$.

$$\mathcal{A}(\{\epsilon_{NN}(\mathbf{x}_i) \text{ for each } \mathbf{x}_i \in \mathcal{S}\}) = \left\{ \frac{1}{N} \cdot \sum_{i=1}^N |j(l_k, \epsilon_{NN}(\mathbf{x}_i))| \text{ for each } l_k \in \mathcal{S} \right\}. \quad (11)$$

Being aware that the selection of the aggregation mechanism may influence the correlation between different LPEs, in our experiments we analyse LPEs correlation over the same aggregation scheme. Moreover, we also consider analysing how aggregation impacts the impact scores correlation over the same LPE, highlighting how leveraging the absolute value of impact score is highly similar to adopting its true value—see Sect. 4.3.

3.4 Comparing Explanations

Each aggregated global explanation J depends on a corresponding label τ_m – i.e., moral value – since LPEs produce either a scalar impact value for a single τ_m or a vector of impact scores for each τ_m . Therefore, recalling Sect. 4.3, we can define the set of aggregated global scores depending on the label they refer to as following:

$$\mathcal{J}_{\tau_m}(\epsilon_{NN}, \mathcal{S}) = \{J(l_k, \epsilon_{NN}) | \tau_m \text{ for each } l_k \in \mathcal{S}\}. \quad (12)$$

$\mathcal{J}_{\tau_m}(\epsilon_{NN}, \mathcal{S})$ represents a distribution of impact scores over the set of lemmas – i.e., concepts – available in the samples set for a specific label. To compare the distributions of impact scores extracted using two LPEs – i.e., $\mathcal{J}_{\tau_m}(\epsilon_{NN}, \mathcal{S})$ and $\mathcal{J}_{\tau_m}(\epsilon'_{NN}, \mathcal{S})$ – we use Pearson correlation, which is defined as the ratio between the covariance of two variables and the product of their standard deviations, and

it measures their level of linear correlation. The selected correlation metric is applied to the normalised impact scores. Indeed, different LPEs produce impact scores which may differ relevantly in terms of their magnitude. Normalising the impact scores, we map impact scores to a fixed interval, allowing for a direct comparison of $\mathcal{J}_{\tau_m}$ over different ϵ_{NN} . Mathematically, we refer to the normalised global impact scores as $\|\mathcal{J}_{\tau_m}\|$. Therefore, we define the correlation score between two sets of global impact scores for a single label as:

$$\rho(\|\mathcal{J}_{\tau_m}(\epsilon_{NN}, \mathcal{S})\|, \|\mathcal{J}_{\tau_m}(\epsilon'_{NN}, \mathcal{S})\|) = \rho(\|\{J(l_k, \epsilon_{NN}) | \tau_m \text{ for each } l_k \in \mathcal{S}\}\|, \|\{J(l_k, \epsilon_{NN}) | \tau_m \text{ for each } l_k \in \mathcal{S}\}\|) \quad (13)$$

where ρ refers to the Pearson correlation used to compare couples of $\mathcal{J}_{\tau_m}(\epsilon_{NN}, \mathcal{S})$. Throughout our analysis we experimented with similar correlation metrics, such as Spearman correlation and simple vector distance – similarly to [27] –, obtaining similar results. Therefore, to avoid redundancy we here show only the Pearson correlation results. Throughout our experiments, we consider a simple *min-max* normalisation process, scaling the scores to the range $[0, 1]$.

As our aim is to obtain a measure of similarity between LPEs applied over the same set of samples, we can average the correlation scores ρ obtained for each label τ_m over the set of labels $\mathcal{T}$. Therefore, we mathematically define the correlation score of two LPEs, putting together Eqs. (2), (12) and (13) as:

$$\mathcal{C}(\epsilon_{NN}(\mathcal{S}), \epsilon'_{NN}(\mathcal{S})) = \frac{1}{M} \cdot \sum_{m=1}^M \rho(\|\mathcal{J}_{\tau_m}(\epsilon_{NN}, \mathcal{S})\|, \|\mathcal{J}_{\tau_m}(\epsilon'_{NN}, \mathcal{S})\|) \quad (14)$$

where M is the total number of labels, i.e., moral principles, belonging to $\mathcal{T}$.

4 Experiments

In this section, we present the setup and results of our experiments. We present the model training details and its obtained performance in Sect. 4.1. We then focus on the comparison between the available LPEs, showing the correlation between their explanations in Sect. 4.2. Section 4.3 analyses how correlation scores are affected by the selected aggregation mechanism $\mathcal{A}$. Finally, in Sect. 4.4 we analyse the extent to which LPEs explanations are aligned with human notions of moral values.

4.1 Model Training

We follow state-of-the-art approaches for dealing with morality classification task [9, 24]. Thus, we treat the morality classification problem as a multi-class multi-label classification task, leveraging BERT as the LLM to be optimised [15]. We define one NN model for each MFTC split and optimise its parameters over the 70% of tweets, leaving the remaining 30% for testing purposes. However, conversely from recent approaches, we here do not rely on the *sequential training*

paradigm, but rather train each model solely on the MFTC split at hand. Indeed, in our experiments, we do not aim at obtaining strong transferability between domains, but rather we focus on analysing LPEs behaviour.

We leverage the pre-trained *bert-base-uncased* model – available in the Hugging Face python library⁶ – as the starting point of our training process. Each model is trained for 3 epochs using a standard binary cross entropy loss [46], a learning rate of 5×10^{-5} , a batch size of 16 and a maximum sequence length of 64. We keep track of the macro F1-score for each model to identify its performance over the test samples. Table 1 shows the performance of the trained BERT model.

Table 1. BERT performance over MFTC datasets.

	ALM	BLM	BLT	DAV	ELE	MT	SND	TOT
F_1 score	63.04%	82.59%	64.51%	88.12%	63.14%	52.16%	56.85%	69.10%

4.2 Are Local Post-hoc Explainers Aligned?

We analyse the extent to which different LPEs are aligned in their process of identifying impactful concepts for the underlying NN model. With this aim, we train a BERT model over a specific dataset (following the approach described in Sect. 4.1) and compute the pairwise correlation $\mathcal{C}(\epsilon_{NN}(\mathcal{S}), \epsilon'_{NN}(\mathcal{S}))$ (as described in Sect. 3) for each pair of LPE in the selected set. To avoid issues caused by model overfitting over the training set, which would render explanations unreliable, we apply each ϵ_{NN} over the test set of the selected dataset.

Using the pairwise correlation values we construct the correlation matrices shown in Figs. 1 and 2, which highlight how there exist a very weak correlation score between most LPEs over different datasets. Here, it is interesting to notice how, there exists few specific couples or clusters of LPE which highly correlate with each other. For example, GS, GI and LRP show moderate to high correlation score, mainly due to their reliance on computing the gradient of the prediction to identify impactful concepts. However, this is not the case for all LPE couples relying on similar approaches. For example, GI and gradient integration – HESS in the matrices – show little to no correlation, although they both are gradient-based approach for producing local explanations. Similarly, SHAP and LIME show no correlation even if they both rely on input perturbation and are considered the state-of-the-art.

Figures 1 and 2 highlight how the vast majority of LPE pairs show very small to no correlation at all, exposing how there exists a disagreement between the selected approaches. This finding represents a fundamental result of our study, as it highlights how there is no accordance between LPE even when they are applied

⁶ <https://github.com/huggingface>.

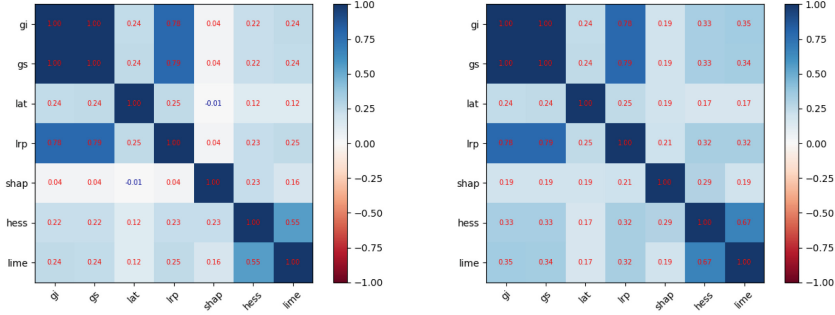


Fig. 1. $\mathcal{C}(\epsilon_{NN}(S), \epsilon'_{NN}(S))$ using average aggregation (left) and absolute average aggregation (right) as $\mathcal{A}$ over the BLM dataset.

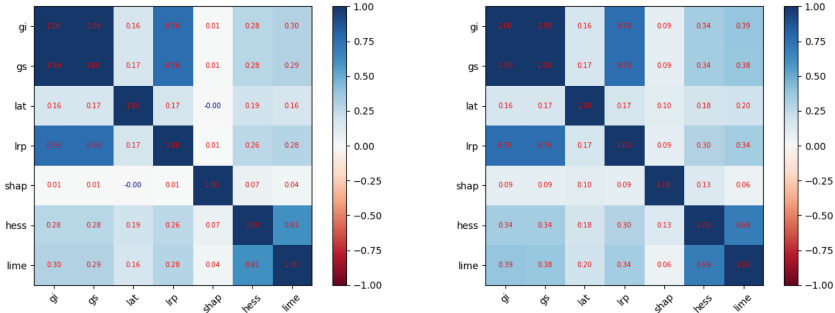


Fig. 2. $\mathcal{C}(\epsilon_{NN}(S), \epsilon'_{NN}(S))$ using average aggregation (left) and absolute average aggregation (right) as $\mathcal{A}$ over the ELE dataset.

to the same model and dataset. The reason behind such large discrepancies among LPE might be various, but mostly bear down to the following:

- Few of the LPE considered in the literature do not represent reliable solutions for identifying the reasoning principles of LLMs.
- Each of the uncorrelated LPEs highlight a different set or subset of reasoning principles of the underlying model.

Therefore, our results show how it is also complex to identify a set of fair and reliable metrics to spot the best LPE or even reliable LPEs, as they seem to gather uncorrelated explanations. Similar results to the ones shown in Figs. 1 and 2 are obtained for all dataset splits and are made available at <https://tinyurl.com/QU4RR3L>.

4.3 How Does Impact Scores Aggregation Affect Correlation?

Since our LPE correlation metric is dependent on $\mathcal{A}$, we here analyse how the selection of different aggregation strategies impacts the correlation between

LPEs. To understand the impact of $\mathcal{A}$ on $\mathcal{C}$, we plot the correlation matrices for a single dataset, varying the aggregation approach, thus obtaining the four correlation matrices shown in Fig. 3.

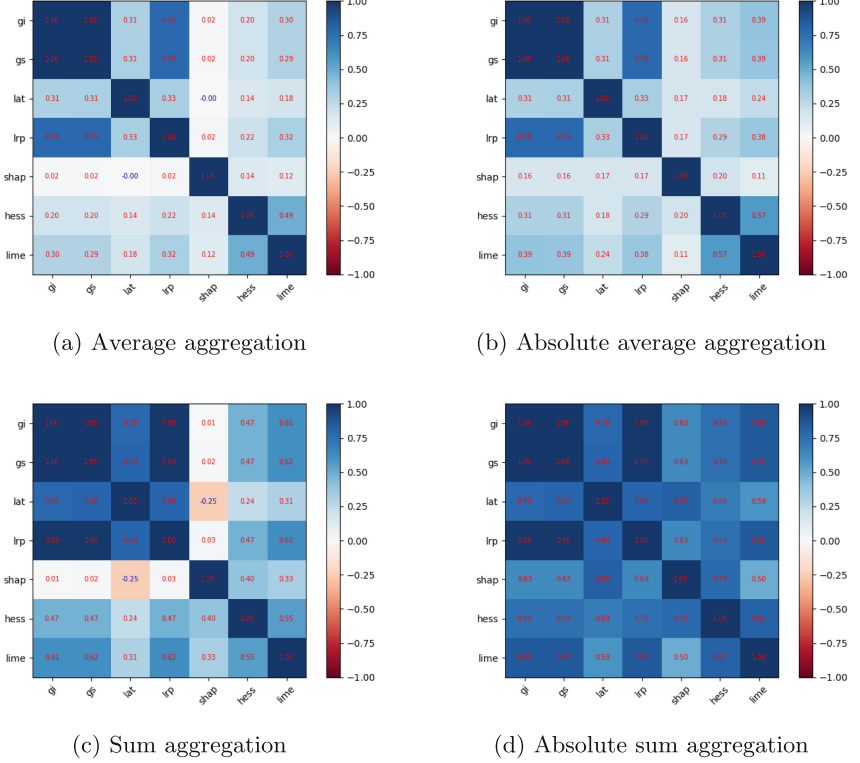


Fig. 3. $\mathcal{C}(\epsilon_{NN}(S), \epsilon'_{NN}(S))$ using different aggregations over the ALM dataset.

From Figs. 3c and 3d it is possible to notice how there exists a strong correlation between different LPEs. This results seems to be in contrast with the results found in Sect. 4.2. However, the reason behind the strong correlation achieved when relying on summation aggregation is not caused by the actual correlation between explanations, but rather on the susceptibility of summation to tokens frequency. Indeed, since the summation aggregation approaches do not take into account the occurrence frequency of lemmas in $\mathcal{S}$, they tend to overestimate the relevance of popular concepts. Intuitively, using this aggregations, a rather impactless lemma appearing 5000 times would obtain a global impact higher than a very impactful lemma appearing only 10 times. These results highlight the importance of relying on average based aggregation approaches when considering to construct global explanations from the LPE outputs.

Figure 3 also highlights how leveraging the absolute value of LPEs incurs in higher correlation scores. The reason behind such a phenomenon is to be found in the impact scores distributions. Indeed, while true local impact scores are distributed over the set of real numbers $\mathbb{R}$, computing the absolute value of local impacts j shifts their distribution to $\mathbb{R}^+$, shrinking possible differences between positive and negative scores. Moreover, it is also true that LPE outputs rely much more heavily on scoring positive contributions using positive impact scores, and tends to give less focus to negative impact scores. Therefore, it is generally true that the output of LPEs is unbalanced towards positive impact scores, making negative impact scores mostly negligible.

4.4 Are Local Post-hoc Explainers Aligned with Human Values?

As our experiments show the huge variability in the response by available state-of-the-art LPE approaches, we check whether there exists at least one LPE that is aligned with human interpretation of values. To do so, we compare the set of global impact scores $\mathcal{J}$ extracted by each LPE against two sets of lemmas which are considered to be relevant for humans. The set of humanly-relevant lemmas, along with their impact scores are obtained from the MFD and the extended Moral Foundations Dictionary (eMFD). The MFD is a dictionary of relevant lemmas for the set of moral values belonging to MFTC. Such a dictionary is generated manually by picking relevant words from a large list of words for each foundation value [21]. Meanwhile, eMFD represents an extension of MFD constructed from text annotations generated by a large sample of human coders.

Similar to the comparison of Sect. 4.2, we rely on Pearson correlation, measuring the correlation coefficient $\mathcal{C}$ between each LPE and MFD or eMFD, treating MFD as if it was a distribution of relevant concepts. Figure 4 shows the results for our study over the BLT dataset for different aggregation mechanisms.

Alarmingly, the results show how there exists no positive correlation between any of the LPE approaches and both MFD and eMFD. Although it is possible that the trained model learns relevant concepts that are specific to the target domain – i.e., BLT in Fig. 4 – it is concerning how strongly uncorrelated LPE and human interpretation of values are. Indeed, while BERT may focus on a few specific concepts which are not human-like, it is assumed and proven to be effective in learning human-like concepts over the majority of NLP tasks. Especially if we consider our BERT model to be only fine-tuned on the target domain, it is very unreasonable to assume these results to be caused by BERT learning concepts that are not aligned with human values. Rather, it is fairly reasonable to deduce that the considered LPEs are far from being completely aligned to the real reasoning process of the underlying BERT model, thus incurring in such high discrepancy with human-labeled moral values.

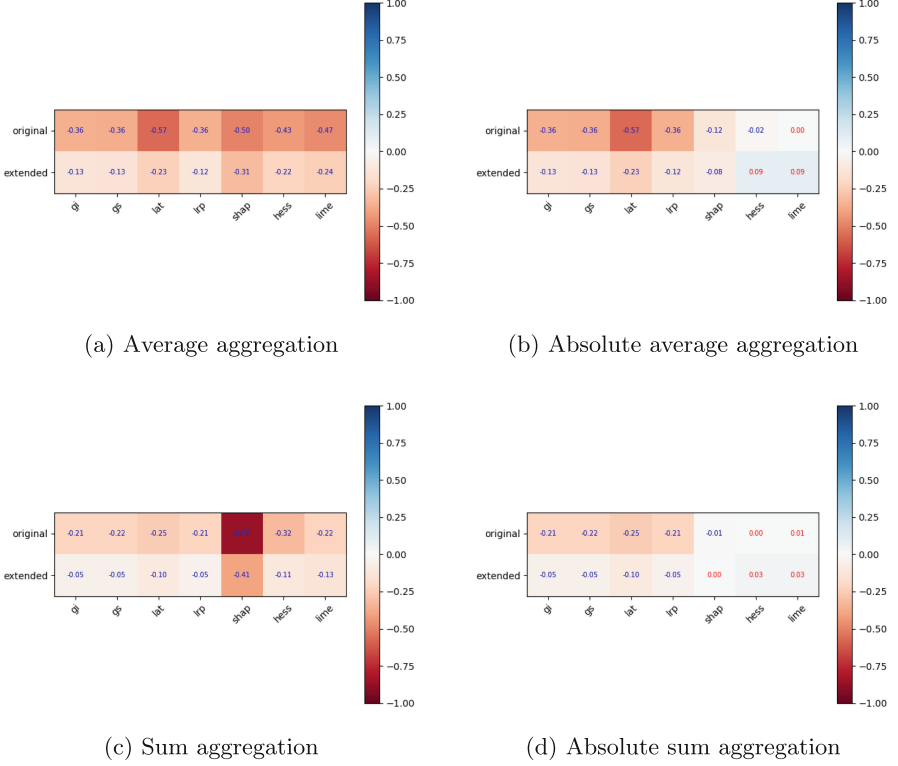


Fig. 4. $\mathcal{C}(\epsilon_{NN}(\mathcal{S}), MFD)$ using different aggregations over the BLT dataset.

5 Conclusion and Future Work

We propose a new approach for the comparison among state-of-the-art local post-hoc explanation mechanisms, aiming at identifying the extent to which their extracted explanations correlate. We rely on a novel framework for extracting and comparing global impact scores from local explanations obtained from LPEs, and apply such a framework over the MFTC dataset. Our experiments show how most LPEs explanations are far from being mutually correlated when LPEs are applied over a large set of input samples. These results highlight what we called the “quarrel” among state-of-the-art local explainers, apparently caused by each of them focusing on a different set or subset of relevant concepts, or imposing a different distribution on top of them. Further, we compare the impact scores distribution obtained from each LPEs with a set of human-made dictionaries. Our experiments alarmingly show how there exists no correlation between LPE outputs and the concepts considered to be salient by humans. Therefore, our experiments highlight the current fragility of xAI approaches for NLP.

Our proposal is a solid starting point for the exploration of the reliability and soundness of xAI approaches in NLP. In our future work, we aim at investigating

more in-depth the issue of robustness of LPE approaches, adding novel LPEs to our comparison such as [16], and aiming at identifying if it is possible to rely on them to build a surrogate of the model from a global perspective. Moreover, we also consider as a promising research line the possibility of building on top of LPE approaches so as to obtain reliable global explanations of the underlying NN model. Finally, in the future we aim at extending the in-depth analysis of LPEs to domains different from NLP, such as computer vision [4, 5, 13], graph processing [6, 23], and neuro-symbolic approaches [3, 7].

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4190–4197. Association for Computational Linguistics, July 2020. <https://doi.org/10.18653/v1/2020.acl-main.385>
2. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018). <https://doi.org/10.1109/ACCESS.2018.2870052>
3. Agiollo, A., Ciatto, G., Omicini, A.: Graph neural networks as the copula mundi between logic and machine learning: a roadmap. In: Calegari, R., Ciatto, G., Denti, E., Omicini, A., Sartor, G. (eds.) WOA 2021–22nd Workshop “From Objects to Agents”. CEUR Workshop Proceedings, vol. 2963, pp. 98–115. Sun SITE Central Europe, RWTH Aachen University, October 2021. <http://ceur-ws.org/Vol-2963/paper18.pdf>, 22nd Workshop “From Objects to Agents” (WOA 2021), Bologna, Italy, 1–3 September 2021. Proceedings
4. Agiollo, A., Ciatto, G., Omicini, A.: Shallow2Deep: restraining neural networks opacity through neural architecture search. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) Explainable and Transparent AI and Multi-agent Systems. Third International Workshop, EXTRAAMAS 2021. LNCS, vol. 12688, pp. 63–82. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82017-6_5
5. Agiollo, A., Omicini, A.: Load classification: a case study for applying neural networks in hyper-constrained embedded devices. *Appl. Sci.* **11**(24) (2021). <https://doi.org/10.3390/app112411957>, <https://www.mdpi.com/2076-3417/11/24/11957>, Special Issue “Artificial Intelligence and Data Engineering in Engineering Applications”
6. Agiollo, A., Omicini, A.: GNN2GNN: graph neural networks to generate neural networks. In: Cussens, J., Zhang, K. (eds.) Uncertainty in Artificial Intelligence. Proceedings of Machine Learning Research, vol. 180, pp. 32–42. ML Research Press, August 2022. <https://proceedings.mlr.press/v180/agiollo22a.html>, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1–5 August 2022, Eindhoven, The Netherlands
7. Agiollo, A., Rafanelli, A., Omicini, A.: Towards quality-of-service metrics for symbolic knowledge injection. In: Ferrando, A., Mascardi, V. (eds.) WOA 2022–23rd Workshop “From Objects to Agents”, CEUR Workshop Proceedings, vol. 3261, pp. 30–47. Sun SITE Central Europe, RWTH Aachen University, November 2022. <http://ceur-ws.org/Vol-3261/paper3.pdf>
8. Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.R., Wolf, L.: XAI for transformers: better explanations through conservative propagation. In: International Conference on Machine Learning, pp. 435–451. PMLR (2022). <https://proceedings.mlr.press/v162/ali22a.html>

9. Alshomary, M., Baff, R.E., Gurcke, T., Wachsmuth, H.: The moral debater: a study on the computational generation of morally framed arguments. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8782–8797. Association for Computational Linguistics, Dublin, Ireland, May 2022. <https://doi.org/10.18653/v1/2022.acl-long.601>
10. 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), e0130140 (2015). <https://doi.org/10.1371/journal.pone.0130140>
11. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021). <https://doi.org/10.1145/3442188.3445922>
12. Brown, T., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020). <https://dl.acm.org/doi/abs/10.5555/3495724.3495883>
13. Buhrmester, V., Münch, D., Arens, M.: Analysis of explainers of black box deep neural networks for computer vision: a survey. *Mach. Learn. Knowl. Extr.* **3**(4), 966–989 (2021). <https://doi.org/10.3390/make3040048>
14. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A survey of the state of explainable AI for natural language processing. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp. 447–459. Association for Computational Linguistics, Suzhou, China, December 2020. <https://aclanthology.org/2020.aacl-main.46>
15. 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 Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN, USA, June 2019. <https://doi.org/10.18653/v1/N19-1423>
16. Främling, K., Westberg, M., Jullum, M., Madhikermi, M., Malhi, A.: Comparison of contextual importance and utility with LIME and Shapley values. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) Explainable and Transparent AI and Multi-agent Systems - Third International Workshop, EXTRAAMAS 2021. LNCS, vol. 12688, pp. 39–54. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82017-6_3
17. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **51**(5), 1–42 (2018). <https://doi.org/10.1145/3236009>
18. Hailesilassie, T.: Rule extraction algorithm for deep neural networks: a review. *Int. J. Comput. Sci. Inf. Secur.* **14**(7), 376–381 (2016). https://www.academia.edu/28181177/Rule_Extraction_Algorithm_for_Deep_Neural_Networks_A_Review
19. Hao, T., Li, X., He, Y., Wang, F.L., Qu, Y.: Recent progress in leveraging deep learning methods for question answering. *Neural Comput. Appl.* **34**(4), 2765–2783 (2022). <https://doi.org/10.1007/s00521-021-06748-3>
20. Hoover, J., et al.: Moral foundations Twitter corpus: a collection of 35k tweets annotated for moral sentiment. *Soc. Psychol. Pers. Sci.* **11**(8), 1057–1071 (2020). <https://doi.org/10.1177/194855061987662>

21. Hopp, F.R., Fisher, J.T., Cornell, D., Huskey, R., Weber, R.: The extended moral foundations dictionary (eMFD): development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behav. Res. Methods* **53**, 232–246 (2021). <https://doi.org/10.3758/s13428-020-01433-0>
22. Ibrahim, M., Louie, M., Modarres, C., Paisley, J.: Global explanations of neural networks: mapping the landscape of predictions. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 279–287 (2019). <https://doi.org/10.1145/3306618.3314230>
23. Jaume, G., et al.: Quantifying explainers of graph neural networks in computational pathology. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021*, pp. 8106–8116. Computer Vision Foundation/IEEE (2021). <https://doi.org/10.1109/CVPR46437.2021.00801>
24. Kiesel, J., Alshomary, M., Handke, N., Cai, X., Wachsmuth, H., Stein, B.: Identifying the human values behind arguments. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers)*, pp. 4459–4471 (2022). <https://doi.org/10.18653/v1/2022.acl-long.306>
25. Kindermans, P.J., et al.: The (un)reliability of saliency methods. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L., Müller, K.R. (eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_14
26. Kokalj, E., Škrlić, B., Lavrač, N., Pollak, S., Robnik-Šikonja, M.: BERT meets Shapley: extending SHAP explanations to transformer-based classifiers. In: *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pp. 16–21 (2021)
27. Liscio, E., et al.: What does a text classifier learn about morality? An explainable method for cross-domain comparison of moral rhetoric. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 1–12, Toronto (2023, to appear)
28. Liscio, E., Dondera, A., Geadau, A., Jonker, C., Murukanniah, P.: Cross-domain classification of moral values. In: *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2727–2745. Association for Computational Linguistics, Seattle, United States, July 2022. <https://doi.org/10.18653/v1/2022.findings-naacl.209>
29. Liu, G., et al.: Medical-VLBERT: medical visual language BERT for COVID-19 CT report generation with alternate learning. *IEEE Trans. Neural Netw. Learn. Syst.* **32**(9), 3786–3797 (2021). <https://doi.org/10.1109/TNNLS.2021.3099165>
30. Loh, W.Y.: Fifty years of classification and regression trees. *Int. Stat. Rev.* **82**(3), 329–348 (2014). <https://doi.org/10.1111/insr.12016>
31. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
32. Luo, S., Ivison, H., Han, C., Poon, J.: Local interpretations for explainable natural language processing: a survey. *arXiv preprint arXiv:2103.11072* (2021)
33. Madsen, A., Reddy, S., Chandar, S.: Post-hoc interpretability for neural NLP: a survey. *ACM Comput. Surv.* **55**(8), 1–42 (2022). <https://doi.org/10.1145/3546577>
34. Otter, D.W., Medina, J.R., Kalita, J.K.: A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **32**(2), 604–624 (2021). <https://doi.org/10.1109/TNNLS.2020.2979670>

35. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**(1), 5485–5551 (2020). <https://dl.acm.org/doi/abs/10.5555/3455716.3455856>
36. Ramachandran, D., Parvathi, R.: Analysis of Twitter specific preprocessing technique for tweets. *Procedia Comput. Sci.* **165**, 245–251 (2019). <https://doi.org/10.1016/j.procs.2020.01.083>
37. 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 (2016). <https://doi.org/10.18653/v1/N16-3020>
38. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE* **109**(3), 247–278 (2021). <https://doi.org/10.1109/JPROC.2021.3060483>
39. Sheikhalishahi, S., et al.: Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med. Inform.* **7**(2), e12239 (2019). <https://doi.org/10.2196/12239>
40. Stahlberg, F.: Neural machine translation: a review. *J. Artif. Intell. Res.* **69**, 343–418 (2020). <https://doi.org/10.1613/jair.1.12007>
41. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) *International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 70, pp. 3319–3328. PMLR, August 2017. <https://proceedings.mlr.press/v70/sundararajan17a.html>
42. Tay, Y., Bahri, D., Metzler, D., Juan, D.C., Zhao, Z., Zheng, C.: Synthesizer: rethinking self-attention for transformer models. In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 139, pp. 10183–10192. PMLR, July 2021. <https://proceedings.mlr.press/v139/tay21a.html>
43. Warstadt, A., Singh, A., Bowman, S.R.: Neural network acceptability judgments. *Trans. Assoc. Comput. Linguist.* **7**, 625–641 (2019). <https://doi.org/10.1162/tacl.a.00290>
44. Wu, Z., Nguyen, T.S., Ong, D.C.: Structured self-attention weights encode semantics in sentiment analysis. In: *Proceedings of the Third Blackbox NLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 255–264. Association for Computational Linguistics (2020). <https://doi.org/10.18653/v1/2020.blackboxnlp-1.24>
45. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. *Wiley Interdisc. Rev. Data Mining Knowl. Discov.* **8**(4) (2018). <https://doi.org/10.1002/widm.1253>
46. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc. (2018). https://proceedings.neurips.cc/paper_files/paper/2018/file/f2925f97bc13ad2852a7a551802feea0-Paper.pdf
47. Zini, J.E., Awad, M.: On the explainability of natural language processing deep models. *ACM Comput. Surv.* **55**(5), 1–31 (2022). <https://doi.org/10.1145/3529755>