IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 March 2024, accepted 24 May 2024, date of publication 31 May 2024, date of current version 3 October 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3408062*

RESEARCH ARTICLE

# TextFocus: Assessing the Faithfulness of Feature Attribution Methods Explanations in Natural Language Processing

**ETTORE MARIOTTI[1], ANNA ARIAS-DUART[2,3], MICHELE CAFAGNA[4], ALBERT GATT[5], DARIO GARCIA-GASULLA[2,3], AND JOSE MARIA ALONSO-MORAL[1], (Member, IEEE)**

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), 15782 Santiago de Compostela, Spain
[2]Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain
[3]Universitat Politècnica de Catalunya (UPC)–BarcelonaTech, Les Corts, 08034 Barcelona, Spain
[4]University of Malta, MSD 2080 Msida, Malta
[5]Utrecht University, 3584 CS Utrecht, The Netherlands

Corresponding author: Ettore Mariotti (ettore.mariotti@usc.es)

This work was supported in part by NL4XAI Project funded by European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie Grant under Agreement 860621; in part by MCIN/AEI/10.13039/501100011033 and 11ESF Investing in Your Future under Grant PID2021-123152OB-C21; in part by MCIN/AEI/10.13039/501100011033 and the 11European Union NextGenerationEU/PRTR under Grant TED2021-130295B-C33; in part by the Galician Ministry of Culture, Education, Professional Training, and University (co-funded by European Regional Development Fund, ERDF/FEDER Program) under Grant ED431G2019/04 and Grant ED431C2022/19; and in part by European Union–Horizon 2020 Program under the Scheme 11INFRAIA-01-2018-2019–Integrating Activities for Advanced Communities 11SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics (http://www.sobigdata.eu) under Grant 871042.

**ABSTRACT** Among the existing eXplainable AI (XAI) approaches, Feature Attribution methods are a popular option due to their interpretable nature. However, each method leads to a different solution, thus introducing uncertainty regarding their reliability and coherence with respect to the underlying model. This work introduces *TextFocus*, a metric for evaluating the faithfulness of Feature Attribution methods for Natural Language Processing (NLP) tasks involving classification. To address the absence of ground truth explanations for such methods, we introduce the concept of *textual mosaics*. A mosaic is composed of a combination of sentences belonging to different classes, which provides an implicit ground truth for attribution. The accuracy of explanations can be then evaluated by comparing feature attribution scores with the known class labels in the mosaic. The performance of six feature attribution methods is systematically compared on three sentence classification tasks by using *TextFocus*, with Integrated Gradients being the best overall method in terms of faithfulness and computational requirements. The proposed methodology fills a gap in NLP evaluation, by providing an objective way to assess Feature Attribution methods while finding their optimal parameters.

**INDEX TERMS** Artificial intelligence (AI), explainable AI (XAI), trustworthy AI, explanation faithfulness, feature attribution, feature importance, natural language processing (NLP).

## I. INTRODUCTION

Many Artificial Intelligence (AI) systems today are still black boxes, meaning that their inner workings are poorly understood and difficult to explain. This lack of transparency and understanding can lead to distrust in these systems and their outputs, which is why eXplainable AI (XAI) has become an important research area [1], [2], [3].

The main goal in XAI is to make AI systems more interpretable and explainable, and in this context Feature Attribution methods provide the target audience with a prominent class of explanations. Feature Attribution methods

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

can help to measure the understandability of models [4], find biases [5], debug model behaviours and uncover shortcuts [6], etc. There are many Feature Attribution methods, and it is not trivial to compare them since there is no ground truth for the correct explanations, and thus no way of assessing whether a method is working correctly or not.

To address this issue, XAI evaluation metrics are necessary, yet confusion arises when distinguishing between *faithfulness* (i.e., accuracy of the explanation to the model's reasoning) and *plausibility* (i.e., how convincing the explanation is to humans) [7]. Unfortunately, there is no metric to quantitatively evaluate the *faithfulness* of XAI methods in Natural Language Processing (NLP) without relying on humans.

Judging the faithfulness of a model's decisions to the actual data is difficult for humans, as they often lack a comprehensive understanding of the model's intricate inner workings. Even if a human-provided explanation might seem reasonable, it may not faithfully reflect the model's actual processes. This means that any human evaluation of an explanation would be based on its *plausibility* rather than its *faithfulness*. This is a risky undertaking, as plausible but incorrect explanations can be used to mislead and manipulate behaviour [8]. On the other hand, an implausible explanation is still useful to diagnose model behaviour as long as it is faithful to that behaviour.

Before assessing the plausibility of an explanation, it is crucial to ensure the faithfulness of the underlying process. While human evaluation is important as humans are the end users, faithfulness must be the first priority. Additionally, using human feedback for evaluating explanations can be expensive, time-consuming, and difficult to reproduce consistently [9].

### A. MAIN CONTRIBUTIONS

In this paper, we propose a novel fully automatic data-driven approach to assessing ***faithfulness***. This approach is cost-effective, fast, and easily reproducible, making it a more efficient and reliable option for evaluating the faithfulness of XAI explanations. The focus lies in determining how well explanations generated by Feature Attribution methods align with known elements in the input that correspond to specific classes. In doing so, we also investigate how different choices of hyperparameters affect the faithfulness of these methods. The framework aids in both selecting optimal hyperparameters and evaluating the faithfulness of XAI methods, contributing to more transparent and trustworthy AI systems. The main contributions of this work are:

1) We introduce an automatic way of assessing the faithfulness of XAI methods specifically for NLP tasks.
2) We provide a comprehensive benchmark of six XAI methods, including different variants where applicable, across three distinct sentence classification tasks.

The rest of the manuscript is organized as follows. Section II outlines related work. Section III goes in depth with the proposal of the new metric. Section IV describes the experimental setting. Section V introduces some sanity checks. Section VI presents and discusses the experimental results. Section VII summarizes some limitations of the proposal. Finally, concluding remarks and future work are pointed out in Section VIII.

## II. RELATED WORK

Despite the common use of deep learning models in the Computer Vision (CV) and NLP fields, techniques for evaluating the faithfulness of XAI methods differ.

Let us first categorize the evaluation methods into two types: categorical and numerical evaluations. Categorical evaluations involve axioms that explainability methods must fulfil (*e.g.*, Completeness or Implementation Invariance Sensitivity [10]). Numerical evaluations enable instead the ranking of Feature Attribution methods according to a given desirable property. Randomization tests, for example, measure the difference between the explanation obtained with a randomized and non-randomized model [11], or the difference between the explanation of the correct class and the explanation for a random class [12].

Some other methods try to measure the faithfulness of an XAI method to the model's behaviour, by perturbing the input according to their explanation and then examining output variations [13], [14], [15], [16]. While in CV such perturbations often take the form of pixel manipulations, similar methods are used in NLP by manipulating (e.g., replacing, deleting or zeroing the embeddings of) tokens [17], [18], [19], [20]. However, these modifications quickly generate instances that are mostly outside of the original data distribution. In some cases, they result in infelicitous strings (*e.g.*, a sentence without a verb).

In order to avoid the Out-of-Distribution (OOD) problem, a variation is proposed by [21], where instead of deleting the tokens corresponding to OOD words or zeroing their embeddings, they are sampled from a distribution inferred by a Masked Language Model, *e.g.*, BERT [22]. However, since these models only expect a part of the input to be masked, the re-sampling is performed on up to 20% of the total number of tokens.

Rather than changing the original instances in the dataset, Zhang et al. [23] proposed, in the CV domain, a ''Pointing game'' where a human labeler identifies a specific region of the image and measures how much the explanation matches the annotation. This technique was later adapted to the NLP domain [24], [25]. While this technique overcomes the lack of ground truth, this assumption may not always align with the model's prediction process. Consequently, this approach is more suited for evaluating plausibility (*i.e.,* how much it resonates with human intuition) rather than the faithfulness of the explanation (*i.e.,* how much it is true to the actual computation). Indeed, it has been shown that models can rely on patterns or artifacts in the data that humans could miss. For example Ribeiro et al. [26] noted that an image classifier

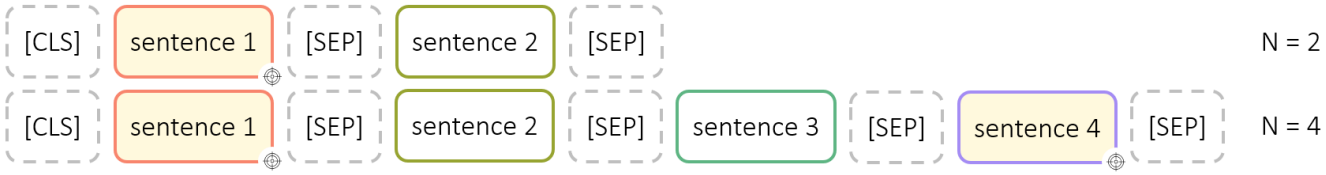



**FIGURE 1.** Example of the structure of two *text mosaics*. The first row corresponds to a mosaic of length two, that is, composed of two sentences. The second row corresponds to a mosaic of length four. In both cases, the target class's sentences are highlighted in yellow.

trained on distinguishing husky vs wolf was not focusing on the region corresponding to the animal, but rather on the snow present in the background of the picture. In this case a pointing game would have assigned a very low score to a faithful explainer (*i.e.,* one that would point to the snow as an evidence). In order to overcome this limitation, the approach by Arias-Duart et al. [27] relaxed the assumption of the pointing game with a construction that by design excluded the human involvement in the 'pointing' definition. To do this, they created an *image mosaic*: a composite grid of images where half belong to a target class and the other half to various other classes. These images are randomly selected, ensuring a diverse representation, yet they are not resized. The primary assumption is that when such a mosaic contains images from both target and non-target classes, the model's positive attributions – meaning the degree to which the model attributes parts of the image to the target class – should predominantly appear on the images that actually belong to the target class. This is a more relaxed assumption of the pointing game, as it is not stating precisely a 'segmentation mask' of the ideal explanation, it is just stating that the evidence should fall on the specific parts of the mosaics (which we know by construction, rather than by human annotation). By constructing this ground truth, the faithfulness of Feature Attribution methods can be measured by analysing the attributions' location.

As we will see in the next section, we draw inspiration from the previous work and propose a novel metric (*TextFocus*) to assess the faithfulness of Feature Attribution methods for NLP.

## III. METHODOLOGY

In this section we describe how the so-called Textual Mosaics (see Section III-A) are used to compute the *TextFocus* score (see Section III-B).

### A. TEXTUAL MOSAICS

We focus on text classification tasks. Thus, we have a list of sentences, with each one having a single label (*e.g.*, negative or positive, offensive or non-offensive). Using a subset of these sentences, we build what we call *textual mosaics*.

A *textual mosaic* is a string composed of $N$ sentences, with half belonging to the target class A, and the other half to other classes. The target class is the one that the Feature Attribution method is requested to explain. For each mosaic, we first randomly sample $N/2$ sentences from the target class, then randomly select $N/2$ sentences from the other classes. The sentences in the textual mosaic are then arranged in random order. Each sentence is separated by a `[SEP]` token, with the mosaic beginning with a `[CLS]` token and ending with a `[SEP]` token. We chose to use this construction to align with the input expected by many widely-used pretrained language models. Figure 1 illustrates two examples of *text mosaic* configurations; in the first row, one sentence corresponds to the target class (sentence 1) and the other to a non-target class (sentence 2). The second row shows a *text mosaic* of length four, with two sentences of the target class (sentences 1 and 4) and two of a non-target class (sentences 2 and 3).

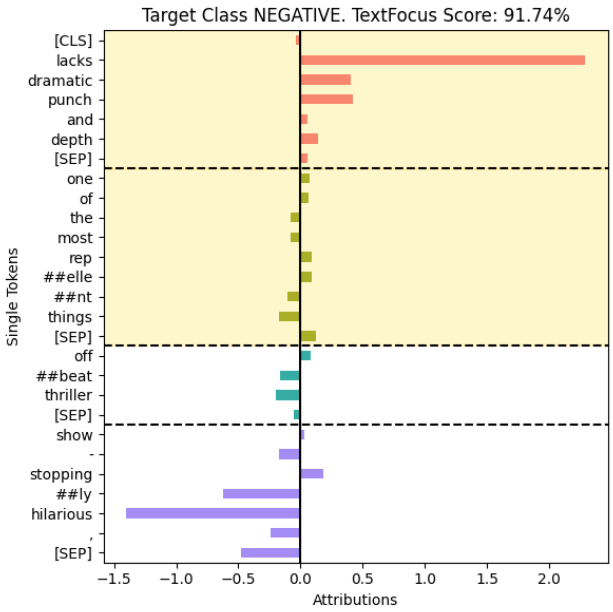



**FIGURE 2.** Attributions for a textual mosaic of size 4, with sentences in the target class highlighted in yellow. Sentences are divided by a dashed line, and attributions are color-coded by sentence. Attributions on the right favour the target class (in this case the negative class) and attributions on the left go against the target class.

An actual example of a mosaic of size $N = 4$, from a sentiment analysis task, is shown in Figure 2. The two target class sentences are highlighted in yellow (see the top part of the picture). Each token is displayed in order from top to bottom and each sentence is separated from the next by a dashed line. Bars are used to represent token attributions:

positive attributions (which favour the target class) are on the right side, while negative attributions (which go against the target class) are on the left. Finally, each token is also assigned a distinct colour, indicating the sentence to which it belongs.

The construction of the textual mosaic gives us the possibility of making the following **key assumptions**, which this work builds upon:

- Evidence for the target class should lie somewhere in the target class sentences, but not on the others.
- Evidence against the target class should be found in sentences not belonging to the target class.

When constructing the mosaics, we deliberately use validation data for which the model's prediction aligns with the true label. By doing so, we can be more confident that the model has indeed detected some form of 'evidence' supporting its classification. This focused approach serves to minimize confounding factors, thereby enabling a more accurate evaluation of the capabilities of XAI methods to identify and localise this 'evidence'.

This is done with the intention of having a more clean measurement of faithfulness, in virtue of dropping those data points where the model gets confused and might find evidence for other classes. This behaviour would jeopardise our measurement of faithfulness, as the evidence the model finds might be not related to the label provided by the dataset.

### B. A QUANTITATIVE METRIC: TEXTFOCUS

We need a quantitative metric to measure how accurately an XAI method is able to correctly assign positive attributions to features of a sample of the target class, and negative attributions to features of samples in the non-target classes.

One key distinction between *textual mosaics* and *image mosaics* is that the former has a variable number of tokens, unlike the latter, which typically has a fixed number of pixels. To prevent our metric from being affected by sentence length, we normalize the attribution for each sentence by its length. In addition, when computing *textual mosaics*, we disregard the effect of special tokens such as `[SEP]` and `[CLS]`, because they are structural placeholders in the mosaic.

More formally, let $S_i$ be the set of tokens in sentence $i$ of the mosaic, with $j$ being the token index not including special tokens. $\alpha_{i,j}$ is the attribution of token $j$ in sentence $i$ according to some XAI method. We define $M$ as the set of sentences of the whole mosaic, with $|M| = N$, and $C \subset M$ the set of sentences belonging to the target class.

A principled way of identifying True Positives ($TP$), False Positives ($FP$), True Negatives ($TN$), and False Negatives ($FN$) from mosaic constructions is provided by [28]. This allows us to adapt metrics from the classification literature to the evaluation of Feature Attribution methods. More formally:

$$TP := \sum_{i \in C} \frac{\sum_j \max(0, \alpha_{i,j})}{\|S_i\|} \tag{1}$$

$$FP := \sum_{i \in M \setminus C} \frac{\sum_j \max(0, \alpha_{i,j})}{\|S_i\|} \tag{2}$$

$$TN := \sum_{i \in M \setminus C} \frac{\sum_j \min(0, \alpha_{i,j})}{\|S_i\|} \tag{3}$$

$$FN := \sum_{i \in C} \frac{\sum_j \min(0, \alpha_{i,j})}{\|S_i\|} \tag{4}$$

The *TextFocus* score can then be computed as:

$$\textit{TextFocus}(M) := \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

This way, *TextFocus* gives us a measure of the overall accuracy of an XAI attribution method in NLP by calculating the proportion of the correctly attributed importance, considering both target and non-target class sentences, in relation to the total attributions of the entire mosaic.

Accordingly, we can assess the *faithfulness* of different attribution methods, compare their performance, and evaluate the impact of different hyperparameter choices on this score.

## IV. EXPERIMENTAL SETTING

As a proof of concept, we will use *TextFocus* for evaluating six different Feature Attribution methods (see Section IV-A). To do so, we will consider the datasets and models which are briefly introduced in Section IV-B.

### A. FEATURE ATTRIBUTION EXPLANATION TECHNIQUES

We implemented the six evaluated methods using Captum.[1] Some of them require computing the gradients with respect to the input. However, since NLP models take discrete, non-differentiable tokens as input, instead of computing the gradients with respect to the input, we compute the gradients with respect to the token embeddings. The evaluated methods are the following:

- **Gradient**: This method uses the gradients of the inputs computed with respect to the target class, as Feature Attribution map. The technique was first used to explain predictions in CV by [29]. We aggregate attributions on a token embedding using the mean, allowing for both positive and negative attributions, in contrast to the L2 or L1 norms employed in other contexts [30], [31].
- **Gradient X Activation**: This method consists in multiplying the result of the gradient by the input activation [32].
- **Integrated Gradients (IG)**: This method computes the integral of gradients along the straight line joining a baseline $x'$ and the input $x$ in the input space [10]. The approximation of the integral is made by summing up all the steps from the baseline $x'$ to input $x$. This technique is based on Aumann-Shapley values, a game-theoretic approach to cost-sharing found in the economics literature [33].

[1] https://captum.ai/

- **DeepLIFT**: [34] assigns relevance with respect to a baseline by backpropagating a relevance score via the Rescale Rule. We use the gradient formulation proposed by [35].
- **Gradient SHAP**: This method adds Gaussian noise to each input sample multiple times and each time selects a random point along the path between the baseline and the input. It computes the gradient of outputs with respect to those selected points, and the final SHAP values represent the expected value of the gradients multiplied by the difference between the inputs and baselines. SHAP values are approximated under the assumption that input features are independent and that the explanation model is linear between the inputs and the given baselines. However, these assumptions can be challenged when this method is applied to language tokens. The implementation adopted here is inspired by the work of [36] and follows the implementation of the SHAP package.[2]
- **LIME** [26]: This method produces Local Interpretable Model-agnostic Explanations. To do so, it generates perturbed instances of the original dataset by deleting random tokens from the instance to be explained. These new instances and their predictions are used to train a linear model which approximates locally the original model. Importantly, when fitting the linear model, each sample is weighted in proportion to its cosine similarity to the original instance, as computed using the model's embedding. The coefficients of the linear model comprise the final attribution scores.

Some of these methods (i.e., IG, DeepLIFT, and Gradient SHAP) fulfill the Sum-to-Zero property (also called *Completeness* in [10] and [35]), which states that the attributions of the method $\alpha_j$ have to sum up to the difference between the value of the function $f$ of the input $x$ and the *baseline*, namely:

$$f(x) - f(baseline) = \sum_j \alpha_j \tag{6}$$

This property can be used to assess a necessary condition for the convergence of the method, i.e., computing the absolute convergence error $\delta$ as follows:

$$\delta := (f[input] - f[baseline]) - \sum_j \alpha_j \tag{7}$$

And then the relative percentage error is:

$$ErrorPerc := \frac{\delta}{(f[input] - f[baseline])} * 100 \tag{8}$$

Accordingly, to ensure accuracy and fair comparison between methods, we iteratively increase the number of steps/samples until the *ErrorPerc* is under a fixed threshold (in our experiment 1%). It is important to note that a small $\delta$ does not guarantee a correct solution; it is a necessary, yet not sufficient, condition to be met.

In addition, it is worth noting that some of these methods require baselines for their computation. Ideally, the baseline should represent a reference input whose "signal" is absent, so that it can be compared to the input being analysed. While in CV those baselines are usually black images (where all the pixel values set to 0), in NLP a zero input embedding is meaningless. Therefore, in this domain, different types of special tokens (*i.e.*, `[PAD]`, `[MASK]`, `[UNK]`, `[SEP]` or `[CLS]`) can be used as baselines.

Since the choice of the baseline can affect the output and performance of the Feature Attribution method [6], we will evaluate the performance of each method according to the considered baselines. Notice that we will delve deeper into this issue later in Section VI-A.

### B. DATASETS AND MODELS

In the experiments, we will concentrate on the task of text classification. We will use three models based on DistilBERT [37]. Each model is fine-tuned for the following datasets:

- **Standfod Stentiment (SST-2)** is a binary sentiment classification dataset containing parts of movie reviews (with a mean length of about 20 tokens) excerpted from `rottentomatoes.com` and labelled on Amazon Mechanical Turk [38]. The SST-2 dataset is composed of 11,855 sentences labeled as positive or negative.
- **Internet Movie Database (IMDB)** [39] is also a binary sentiment classification dataset of highly popular movie reviews collected from `imdb.com`. The dataset consists of a total of 50k sentences, half of which are for training and half for testing. The length of the IMDB sentences is much greater than the SST-2 sentences (mean length of about 300 tokens).
- **Emotion** [40] is a multi-class classification dataset of 20k English tweets scraped from Twitter and labeled with six basic emotions: anger, fear, joy, love, sadness and surprise.

DistilBERT models fine-tuned for each task are found in the official repository of HuggingFace[3,4,5]. The SST-2 model reaches an accuracy of 98.9% on the test set, while the IMDB model reaches an accuracy of 92.8%. On Emotion, the model has an accuracy of 92.5%.

For the SST-2 and Emotion datasets we use mosaics with $N = 4$, similar to [27], since larger sizes would result in higher computational complexity and memory demands. However, for the IMDB mosaics, we use $N = 2$. This is due to the input size limitation of DistilBERT, which is of 512 tokens. In order to prevent truncating the sentence (and thus potentially losing information), when constructing

[2]https://github.com/shap/shap

[3]https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

[4]https://huggingface.co/lvwerra/distilbert-imdb

[5]https://huggingface.co/sabre-code/distilbert-base-uncased-finetuned-emotion

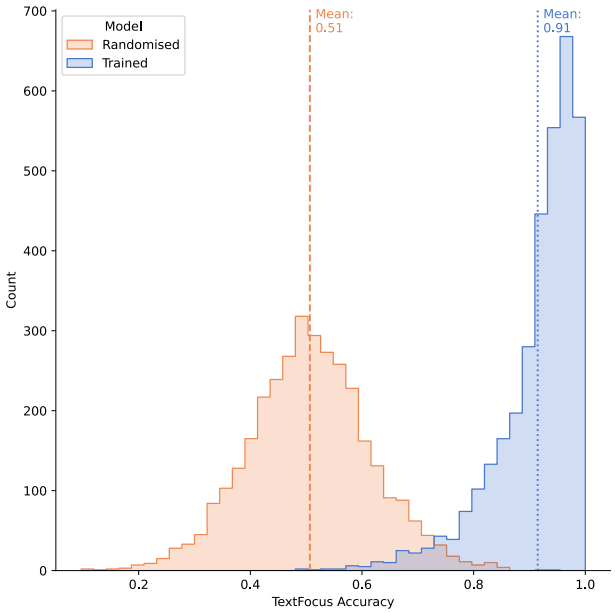



**FIGURE 3.** Histogram of *TextFocus* scores obtained with IG using a trained model (blue) and a randomized model (orange) on 2k *textual mosaics* on SST-2.

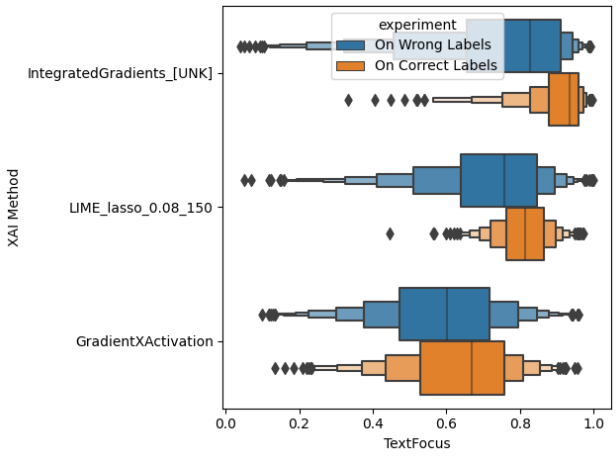



**FIGURE 4.** The impact of evaluating *TextFocus* on the data predicted correctly vs incorrectly. On the data points where the model agrees with the label we have a more clean measurement of *TextFocus*. Data extracted from analysing 1600 mosaics on the Emotion dataset.

*textual mosaics* from IMDB we select sentences with lengths of less than 256 tokens.

## V. SANITY CHECKS

We performed several sanity checks to assess the overall robustness of the mosaic design for *TextFocus*.

### A. RANDOMIZATION TEST

First of all, we analyse the behaviour of *TextFocus* when applied to a random model. With a randomly initialised model, we expect an XAI method to compute token attributions which are unrelated to the labels, since the model has not been trained to make associations between input tokens and output labels. So, we would expect attributions from the random model to be distributed randomly.

We computed the *TextFocus* distribution on 2k text mosaics from SST-2 with the Integrated Gradients (IG) method and an `[UNK]` token as a baseline. Figure 3 shows the evaluation of two models: randomized (orange; centre of the picture) and non-randomized (blue; right-hand side of the picture). The latter corresponds to the one introduced in Section IV-B.

The random model produces a normal distribution centered around 0.51. The results suggest that, as expected, the model relies on patterns that are not localised according to the data labels. On the other hand, the average *TextFocus* score of the trained model is 0.91, which indicates that the Feature Attribution method is able to successfully assign the majority of attributions to the correct tokens within mosaic sentences.

### B. CORRECT LABELS VERSUS WRONG LABELS

In the construction of the textual mosaic we select those data points whose prediction agrees with the provided label. In order to check what is the effect of this step we set up a simple experiment on the Emotion dataset where we do the opposite: we evaluate *TextFocus* on those validation points whose prediction is different from the dataset label. As illustrated in Figure 4, the scores are significantly high, indicating that there is some correct evidence found by the XAI methods in the mistaken predictions. More importantly, we can see that the distribution of *TextFocus* on the correct sentences (i.e., where the model agrees with the dataset label) is higher than the distribution of *TextFocus* on the wrong sentences (i.e., where the model disagrees with the dataset label). The variance on the correct labels is also smaller, indicating a more precise measurement.

### C. ASSESSMENT OF THE IMPACT OF MOSAIC STRUCTURE

One natural question that arises in the construction of textual mosaics is whether the way we build these mosaics influences the explanation of the model classifications, particularly when compared to analysing the sentences individually. We explored this using mosaics of length $n$, made by repeating the same sentence $n$ times.

We study the order of the tokens in the mosaic as a possible confounding factor. In particular, if the position $k$ of the token in the sentence $j$, affects its feature attribution $a_{jk}$, then we should observe a different attribution $a_{j'k}$ for the same token in a different position.

To validate this, we introduce Sentence Attribution Stability (SAS) and rank each token within a mosaic sentence by its attribution. Comparisons are made using Rank Biased Overlap (RBO), akin to the approach in [41]. The details about SAS definition and implementation are given with Algorithm 1.

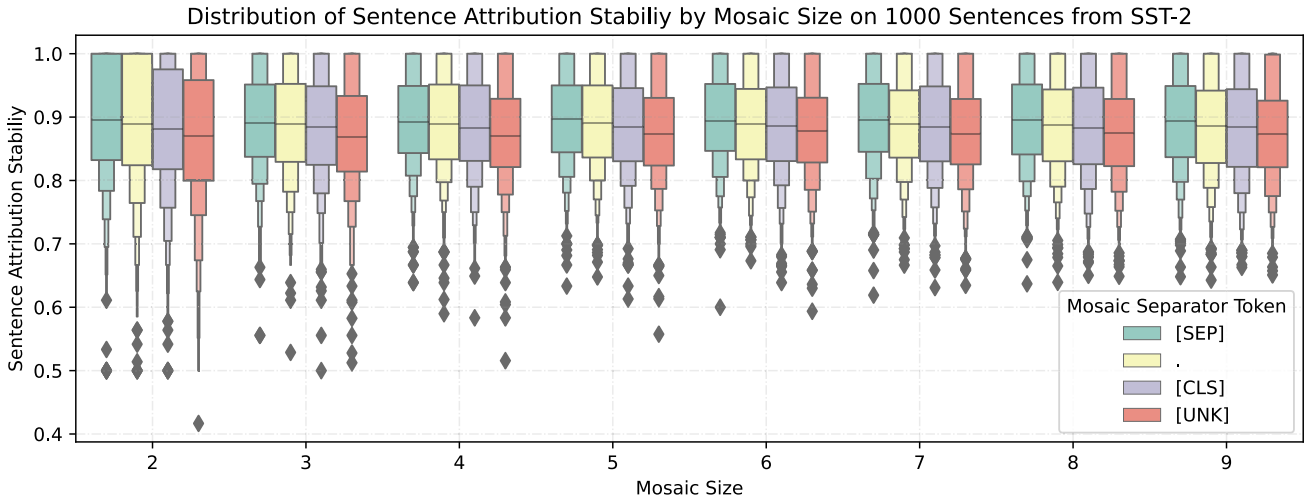



**FIGURE 5.** Distribution of SAS scores across mosaic sizes and separator tokens. Higher scores indicate better attribution stability. The consistently high average underscores the robustness of our methodology, irrespective of mosaic design choices.

**Algorithm 1** Sentence Attribution Stability (SAS)

```
Initialize an empty list: sorted_sentences = []
Initialize an empty list: distances = []
for each (sentence, attribution) in (sentences, attributions) do
    Replace tokens in the sentence with unique numeric identifiers
    Create tuples (numeric_token, attribution)
    Sort tuples by attribution in descending order
    Append sorted numeric tokens to sorted_sentences
end for
Define Pairs as all unique combinations of indices (i, j) for sorted_sentences
for each unique pair (i, j) in Pairs do
    rbo=RBO(sorted_sentences[i], sorted_sentences[j])
    Append rbo to distances
end for
SAS = 1/(n 2) × ∑ distances
```

The idea behind SAS is to measure how the importance of tokens fluctuates when the same sentence appears in different positions of the mosaic. The algorithm sorts the tokens in each sentence by their importance scores and then compares these sorted lists. If the importance rankings of tokens change noticeably depending on their mosaic position, this variation will be captured by the SAS score.

We perform this experiment using 1,000 SST-2 sentences, DistilBERT and Integrated Gradients with an `[UNK]` baseline. As illustrated in Figure 5, our sanity check is supported by the notably high distribution of SAS scores (around 0.9) that remains consistent across different mosaic sizes. The reduced variance with increasing mosaic size is likely attributable to the increased stability from more pairwise comparisons. Intriguingly, the mosaic's structure remains robust irrespective of the mosaic separator token chosen, though there's a higher score for the `[SEP]` token.

## VI. RESULTS

In this section, we discuss the results obtained from the evaluation of the different explainability methods. First, we study the role of the choice of baselines on the methods' performance (see Section VI-A). Then, we use *TextFocus* to assess the Feature Attribution methods, shedding light on the most faithful XAI methods and discussing the trade-off between the performance and execution time (see Section VI-B). For the sake of reproducibility, and in accordance with guidelines for developing responsible AI, we share the code to replicate these experiments.[6]

### A. THE IMPACT OF THE BASELINE CHOICE

In this experiment, we analysed the effect on the performance of the Feature Attribution methods, depending on the tokens chosen to construct baselines. To conduct this evaluation, we use the three explainability methods requiring baselines (*i.e.*, IG, DeepLIFT, and Gradient SHAP). A good baseline should represent the absence of a signal. BERT-like models [22] provide a natural choice for a baseline using special tokens such as `[MASK]`, `[PAD]`, or `[UNK]`. For the sake of completeness, we also tried `[CLS]` and `[SEP]`. Figure 6 shows the results of this experiment for each dataset: Emotion (on the left), SST-2 (on the center), IMDB (on the right). Each colour corresponds to a different Feature Attribution method.

Interestingly, IG performs quite well across all the baselines. However, one can observe a higher mean and smaller standard deviation of the scores for the `[UNK]` token.

[6]The experiments were run on a A100 Nvidia GPU. The data and code needed for reproducing the experiments are available at https://gitlab.nl4xai.eu/ettore.mariotti/TextFocus

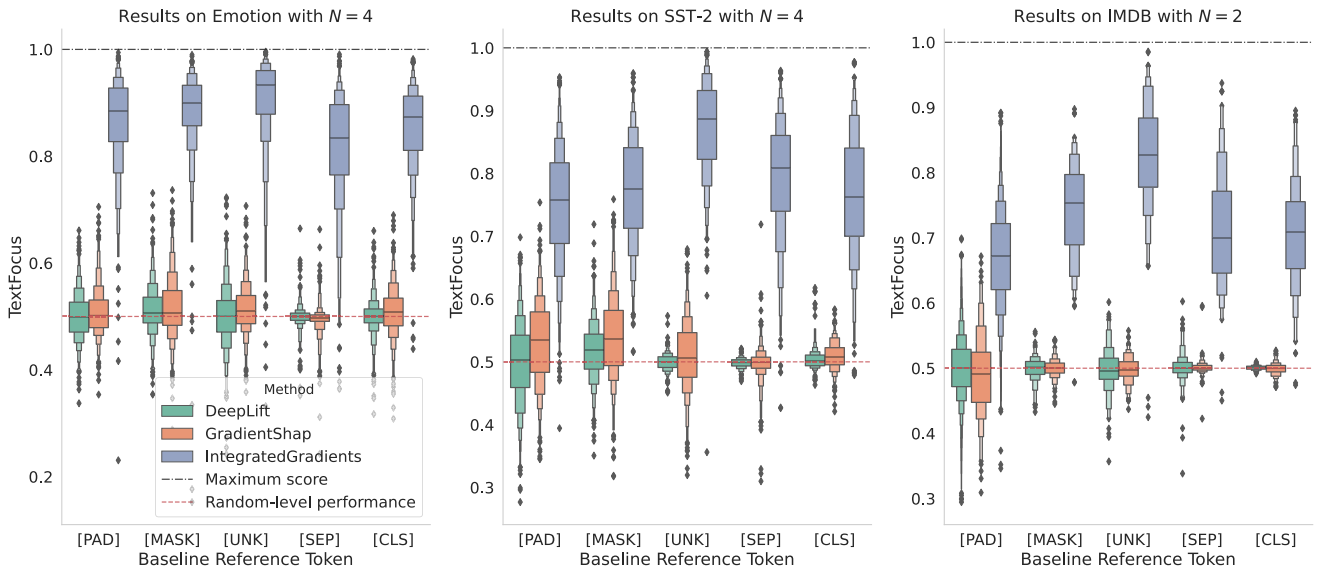



**FIGURE 6.** ***TextFocus*** **distribution for the three methods requiring baselines (*i.e.*, Gradient SHAP, IG and DeepLIFT). Each method is depicted with a box-plot of a different colour. On the x-axis the different special tokens used as baselines.**

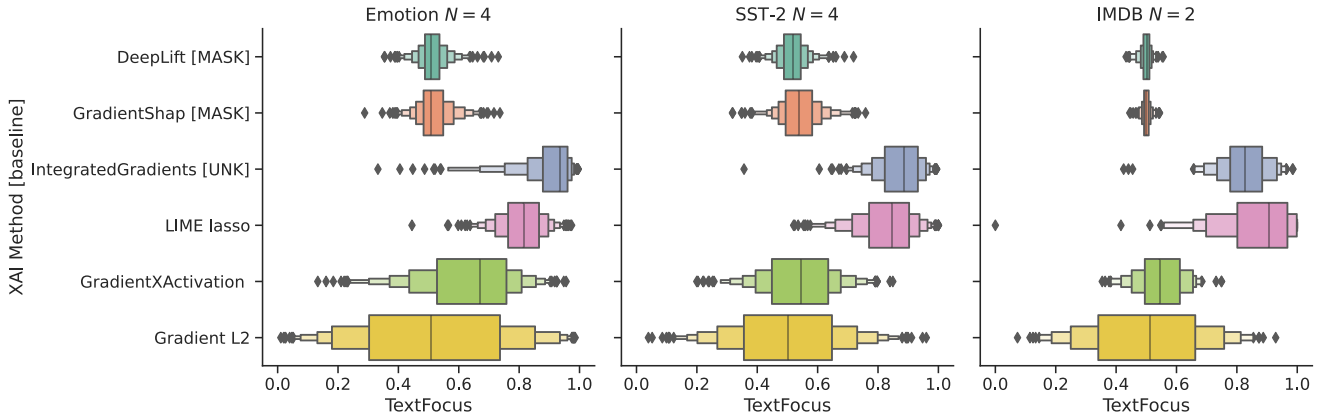



**FIGURE 7.** ***TextFocus*** **score distributions for the seven methods benchmarked, obtained from the evaluations of 1202 mosaics on Emotion (left side), 1746 mosaics on SST-2 (center) and 7014 mosaics on IMDB (right side). The baselines used for the methods requiring them are specified in square brackets. The similarity between the results (*e.g.*, the rank is preserved) additionally suggests that choosing $N = 2$ or $N = 4$ does not significantly impact the results.**

On the other hand, the results of our study suggest that both DeepLIFT and Gradient SHAP often exhibit behaviour that appears random, making it difficult to identify a baseline that performs better than the others. For our next experiments, we have chosen to use the `[MASK]` token as the baseline for both DeepLIFT and Gradient SHAP, as it yielded the highest scores on average.

### B. BENCHMARKING FEATURE ATTRIBUTION METHODS

We benchmark the six Feature Attribution methods introduced in Section IV-A: Gradient, Gradient X Activation, IG, DeepLIFT, Gradient SHAP, and LIME. For each method requiring a baseline, we pick the special token obtaining the best results on average in Section VI-A, that is, the `[UNK]` token for IG, `[MASK]` for Gradient SHAP and DeepLIFT. We run this evaluation on both DistilBERT models introduced in Section IV-B. For each mosaic, we compute *TextFocus* for all target classes. We analyse a total of 1,746 mosaics of size $N = 4$ for the SST-2 dataset, 1,202 mosaics of size $N = 4$ for the emotion dataset and 7,014 mosaics of size $N = 2$ for the IMDB dataset (see Section IV-B for further details). Those results are shown in Figure 7.

IG, using the `[UNK]` token as a baseline, gets consistently good results in all the experiments: obtaining a high mean *TextFocus* of 0.918 in the SST-2 case, a mean *TextFocus* of 0.830 in the IMDB experiment, and a mean of 0.943 in the Emotion experiment. The second best method is LIME, getting a mean *TextFocus* 0.902 in the SST-2 dataset, 0.852 in the IMDB case, and 0.845 in the Emotion dataset. On the other hand, Gradient SHAP, DeepLift, Gradient X Activation and Gradient obtained scores that are very similar to chance behaviour, with a mean accuracy around 0.5 in all the

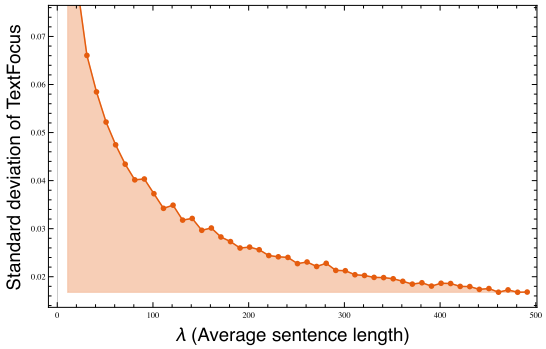



**FIGURE 8.** Standard Deviation of *TextFocus* scores over 1000 synthetic mosaics, as a function of λ (the average sentence length).

configurations in which we tested them. Therefore, IG and LIME turn up as methods able to assign more than 50% of the attribution to the correct sentences of the mosaic. According to these experiments, the *TextFocus* results are consistent across all datasets. It's worth noting that LIME's reliance on random token sampling and lasso regularization often leads to reproducibility issues. When applying LIME with the same model and the same input data in different runs, inconsistent attributions are common. This inconsistency arises because computational constraints limit the random sampling, and lasso regularization can cause different tokens to be deemed important in different runs, minimizing the importance of others.

It is worth noting that Figure 7 shows how some methods that behave in a random-like manner (e.g., DeepLIFT, Gradient SHAP, Gradient X Activation, and Gradient L2) exhibit smaller variance on the IMDB dataset compared to the Emotion or SST-2 datasets. To investigate this further, we generated artificial mosaics with $N = 4$, where the length of each sentence is modeled using a Poisson distribution with mean λ. Assuming that the attributions from a random model follow a normal distribution centered at 0, we computed *TextFocus* for these synthetic attributions. After calculating *TextFocus* for these artificial attributions, we observed that the final *TextFocus* definition followed a Normal-like distribution centered at 0.5, with a standard deviation inversely proportional to λ (see Figure 8). Thus, we can conclude that longer sentence lengths lead to reduced variance.

## VII. LIMITATIONS

*TextFocus* scores may be affected by two main factors: *shared evidence* when a non-target-class label in the same mosaic contains evidence of the target class (imagine a movie review that is overall positive but critical in some parts), and *missing evidence* when no evidence of the target class is found in the sub-mosaics labelled with it (e.g., mislabeled sentences). These issues may stem from the data's inherent properties (*e.g.*, spurious correlations). While this may cause inconveniences when using *TextFocus* to measure the Feature Attribution methods faithfulness, it can be an effective method for understanding how and why a model is confused. Similarly to how mosaics together with the *Focus* metric can uncover bias in vision models [5], future work could also use textual mosaics and *TextFocus* to discover unknown biases in NLP tasks.

We also performed an experiment to analyse whether the interaction between sentences of different classes within a mosaic affects the Feature Attributions, which may lead us to misleading conclusions. We swap the order of sentences within a mosaic without observing significant changes in the result. Even so, to mitigate this possible artifact, for every mosaic we randomize the position of the target class sentences.

## VIII. CONCLUSION AND FUTURE WORK

We introduced and tested *TextFocus* as a new score to assess the faithfulness of Feature Attribution methods in NLP. Of the six XAI methods evaluated, Gradient, Gradient SHAP, DeepLIFT and Gradient X Activation show a behaviour close to random.

IG with `[UNK]` baseline is the method achieving the best performance according to *TextFocus*. LIME also provides reliable explanations. Yet an inherent characteristic of the LIME method that should be taken into account is that despite obtaining reliable explanations, it is not deterministic. That is, depending on the tokens that are removed during the computation, the explanation may be slightly different. This feature may not be a disadvantage for some applications, but could be undesirable for others (*e.g.*, applications requiring high reproducibility).

An interesting avenue for future work would be to include *TextFocus* in the benchmarks provided by the paper and GitHub repository of [42]. By incorporating *TextFocus*, the diversity of evaluation methods could be enhanced, providing a more comprehensive understanding of the performance characteristics of different XAI approaches.

We also plan to extend *TextFocus* to encoder-decoder and decoder-only generative models. We also believe that our findings could be extrapolated to more complex challenges such as Language Generation or Question & Answering.

## REFERENCES

[1] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019.

[2] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, "DARPA's explainable AI (XAI) program: A retrospective," *Appl. AI Lett.*, vol. 2, no. 4, p. e61, Dec. 2021.

[3] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101805.

[4] E. Mariotti, J. M. Alonso-Moral, and A. Gatt, "Measuring model understandability by means of Shapley additive explanations," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Padua, Italy, Jul. 2022, pp. 1–8.

[5] A. Arias-Duart, F. Pares, V. Gimenez-Abalos, and D. Garcia-Gasulla, "Focus and bias: Will it blend?" in *Artificial Intelligence Research and Development*. IOS Press, 2022, pp. 325–334.

[6] J. Bastings, S. Ebert, P. Zablotskaia, A. Sandholm, and K. Filippova, "'Will you find these shortcuts?' A protocol for evaluating the faithfulness of input salience methods for text classification," 2021, *arXiv:2111.07367*.

[7] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4198–4205.

[8] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, "Do feature attribution methods correctly attribute features?" in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 9623–9633.

[9] A. Belz, C. Thomson, E. Reiter, and S. Mille, "Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP," in *Proc. Findings Assoc. Comput. Linguistics*, Toronto, ONT, Canada, 2023, pp. 3676–3687.

[10] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[11] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–11.

[12] L. Sixt, M. Granz, and T. Landgraf, "When explanations lie: Why many modified bp attributions fail," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 9046–9057.

[13] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[14] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3016–3022.

[15] D. A. Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.

[16] L. Rieger and L. K. Hansen, "IROF: A low resource evaluation metric for explanation methods," in *Proc. ICLR*, 2020, pp. 1–11.

[17] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," 2016, *arXiv:1612.08220*.

[18] V. Prabhakaran, B. Hutchinson, and M. Mitchell, "Perturbation sensitivity analysis to detect unintended model biases," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5740–5745.

[19] G. Chrysostomou and N. Aletras, "Improving the faithfulness of attention-based explanations with task-specific information for text classification," 2021, *arXiv:2105.02657*.

[20] S. Serrano and N. A. Smith, "Is attention interpretable?" 2019, *arXiv:1906.03731*.

[21] S. Kim, J. Yi, E. Kim, and S. Yoon, "Interpretation of NLP models through input marginalization," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 3154–3167.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MI, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[23] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, Oct. 2018.

[24] N. Poerner, H. Schütze, and B. Roth, "Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 340–350.

[25] J. DeYoung, S. Jain, N. Fatema Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "ERASER: A benchmark to evaluate rationalized NLP models," 2019, *arXiv:1911.03429*.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144.

[27] A. Arias-Duart, F. Parés, D. Garcia-Gasulla, and V. Giménez-Ábalos, "Focus! Rating XAI methods and finding biases," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2022, pp. 1–8.

[28] A. Arias-Duart, E. Mariotti, D. Garcia-Gasulla, and J. M. Alonso-Moral, "A confusion matrix for evaluating feature attribution methods," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3708–3713.

[29] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[30] L. Arras, A. Osman, K.-R. Müller, and W. Samek, "Evaluating recurrent neural network explanations," in *Proc. ACL Workshop BlackboxNLP, Analyzing Interpreting Neural Netw.*, Florence, Italy, 2019, pp. 113–126.

[31] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," 2017, *arXiv:1712.09913*.

[32] M. Denil, A. Demiraj, and N. de Freitas, "Extraction of salient sentences from labelled documents," 2014, *arXiv:1412.6815*.

[33] R. J. Aumann and L. S. Shapley, *Values of Non-Atomic Games*. Princeton, NJ, USA: Princeton Univ. Press, 1974.

[34] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2017, *arXiv:1704.02685*.

[35] M. Ancona, E. Ceolini, C. Oztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," Feb. 2018, *arXiv:1711.06104*.

[36] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[37] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[38] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empir. Methods Nat. Lang. Process.* Washington, DC, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642.

[39] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Language Technologies*, 2011, pp. 142–150.

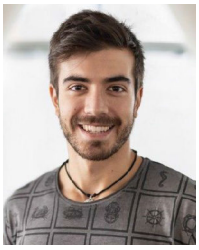
[40] E. Saravia, H.-C.-T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 3687–3697.

[41] A. Sarica, A. Quattrone, and A. Quattrone, "Introducing the rank-biased overlap as similarity measure for feature importance in explainable machine learning: A case study on Parkinson's disease," in *Proc. Int. Conf. Brain Informat.* Springer, 2022, pp. 129–139.

[42] X. Li, M. Du, J. Chen, Y. Chai, H. Lakkaraju, and H. Xiong, "$M^4$: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, Dec. 2023, pp. 1630–1643.

**ETTORE MARIOTTI** is currently pursuing the Ph.D. degree with Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS). His research, deeply rooted in the field of explainable artificial intelligence (XAI), encompasses a holistic approach to explaining machine learning models. His work spans the development and application of post-hoc explanation methods for black box models, alongside the creation of inherently interpretable white box' models.

**ANNA ARIAS-DUART** received the bachelor's degree in telecommunications technology and services engineering from Universitat Politècnica de València (UPV), in 2015, the double Diploma degree from UPV and Télécom ParisTech, Paris, in 2018, and the Ph.D. degree in artificial intelligence from Universitat Politècnica de Catalunya within the Industrial Doctorate Program in collaboration with SEAT, S.A. Her research primarily focuses on explainability and bias detection in artificial intelligence.



**MICHELE CAFAGNA** has a background in computer science and machine learning. He has been working in generative language model applied to journalism. He is currently a Ph.D. Fellow with the Institute of Linguistics and Language Technologies, University of Malta, and a Early Stage Researcher with the Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) Project. His primary research interests include generative multimodal models with a focus on model transparency and semantic grounding and explainablity.

**ALBERT GATT** received the Ph.D. degree in computing science from the University of Aberdeen, U.K., in 2007. He is currently a Professor of natural language generation (NLG) with Utrecht University, where he is a member with the NLP Group, Department of Information and Computing Sciences. His research focuses on NLG and on the interface between vision and language in deep neural networks. He has also conducted extensive research on evaluation, benchmarking, and explainability, especially for multimodal and generative models.

**DARIO GARCIA-GASULLA** received the Ph.D. degree in graph mining, in 2015. He worked in knowledge representation, logic inference, and reasoning. He has been doing AI research, since 2008. As a Postdoctoral Researcher, he has lead multiple research lines on deep learning, particularly in the field of interpretability, transfer learning, medical image, and foundation models. He is currently acts as a Lecturer with UPC–BarcelonaTech, while leading the HPAI research group at BSC.

**JOSE MARIA ALONSO-MORAL** (Member, IEEE) received the Ph.D. degree in telecommunication engineering from UPM, Spain, in 2007. He is currently an Associate Professor with CiTIUS-USC and a Coordinator of the H2020-MSCA-ITN-2019 NL4XAI Project. He is the Vice-Chair of the Task Force on Explainable Fuzzy Systems in the IEEE Computational Intelligence Society and an Associate Editor of *IEEE Computational Intelligence Magazine* and IEEE TRANSACTIONS ON FUZZY SYSTEMS.