Flexible Instance-Specific Rationalization of NLP Models

George Chrysostomou, Nikolaos Aletras

Department of Computer Science, University of Sheffield
gchrysostomou1@sheffield.ac.uk, n.aletras@sheffield.ac.uk

Abstract

Recent research on model interpretability in natural language processing extensively uses feature scoring methods for identifying which parts of the input are the most important for a model to make a prediction (i.e. explanation or rationale). However, previous research has shown that there is no clear best scoring method across various text classification tasks while practitioners typically have to make several other ad-hoc choices regarding the length and the type of the rationale (e.g. short or long, contiguous or not). Inspired by this, we propose a simple yet effective and flexible method that allows selecting optimally for each data instance: (1) a feature scoring method; (2) the length; and (3) the type of the rationale. Our method is inspired by input erasure approaches to interpretability which assume that the most faithful rationale for a prediction should be the one with the highest difference between the model's output distribution using the full text and the text after removing the rationale as input respectively. Evaluation on four standard text classification datasets shows that our proposed method provides more faithful, comprehensive and highly sufficient explanations compared to using a fixed feature scoring method, rationale length and type. More importantly, we demonstrate that a practitioner is not required to make any ad-hoc choices in order to extract faithful rationales using our approach.[1]

1 Introduction

Large pre-trained transformer-based language models such as BERT (Devlin et al. 2019; Bommasani et al. 2021), currently dominate performance across language understanding benchmarks (Wang et al. 2019). These developments have opened up new challenges on how to extract faithful explanations (i.e. rationales[2]), which accurately represent the true reasons behind a model's prediction when adapted to downstream tasks (Jacovi and Goldberg 2020).

Recent studies use feature scoring (i.e. attribution) methods such as gradient and attention-based scores (Arras et al. 2016; Sundararajan, Taly, and Yan 2017; Jain and Wallace 2019; Chrysostomou and Aletras 2021b) to identify important (i.e. salient) segments of the input to subsequently extract them as rationales (Jain et al. 2020; Treviso and Martins 2020). However, a single feature scoring method is typically applied across the whole dataset (i.e. globally). This might not be optimal for individual instances resulting into less faithful explanations (Jacovi and Goldberg 2020; Atanasova et al. 2020). Additionally, rationales are usually extracted using a pre-defined fixed length (i.e. the ratio of a rationale compared to the full input sequence) and type (i.e. top k terms or contiguous) globally. We hypothesize that using a fixed length or type for different instances could result into shorter (i.e. not sufficient for explaining a model's prediction) or longer than needed rationales reducing rationale faithfulness, whilst finding the explanation length is an open problem (Zhang et al. 2021). Moreover to extract rationales, practitioners are currently required to make assumptions for the rationale parameters (i.e. feature scoring method, length and type), whilst different choice of parameters might substantially affect the faithfulness of the rationales.

In this paper, we propose a simple yet effective method that operates at instance-level and mitigates the a priori selection of a specific (1) feature scoring method; (2) length and (3) type when extracting faithful rationales. Our proposed method is flexible and allows the automatic selection of some of these instance-specific parameters or all. Inspired by erasure methods, it functions by computing differences between a model's output distributions obtained using the full input sequence and the input without the rationale respectively. We base this on the assumption that by removing important tokens from the sequence, we should observe large divergences in the model's predicted distribution (Nguyen 2018; Serrano and Smith 2019; DeYoung et al. 2020) resulting into more faithful rationales (Atanasova et al. 2020; Chen and Ji 2020). The contributions of our work are thus as follows:

- To the best of our knowledge, we are the first to propose a method for instance-specific faithful rationale extraction;

- We empirically demonstrate that rationales extracted with instance-specific flexible feature scoring method, length and type using our proposed method are more comprehensive than rationales with fixed, pre-defined parameters;

- We show that our method results in consistently highly sufficient rationales, mitigating the variability in faith-

---

[1]Code for experiments available at: https://github.com/GChrysostomou/instance-specific-rationale

[2]We use these terms interchangeably throughout the paper.

fulness of different feature scoring methods across datasets when used globally, i.e. the same for all instances (Atanasova et al. 2020).

2 Background and Related Work

Rationale Extraction

Given a trained model $\mathcal{M}$, an input $\mathbf{x} = [x_1, \ldots, x_T]$ and a predicted distribution over classes $\mathcal{Y}$, rationale extraction methods seek to identify the most important subset $\mathcal{R} \in \mathbf{x}$ of the input for explaining the model's prediction.

There are two common approaches for extracting rationales. The first consists of two modules jointly trained on an end-task, e.g. sentiment analysis (Lei, Barzilay, and Jaakkola 2016; Bastings, Aziz, and Titov 2019). The first module extracts the rationale (i.e. typically by learning to select which inputs should be masked) and the second module is trained using only the rationale. The second approach consists of using feature scoring (or attribution) methods (i.e. salience metrics) to first identify important parts of the input and then extract the rationales from $\mathcal{M}$ (Jain et al. 2020; Treviso and Martins 2020; DeYoung et al. 2020). A limitation of the first approach is that the models are hard to train compared to the latter and often do not reach high accuracy (Jain et al. 2020). Regarding the latter approach, a limitation is that the same feature scoring method is applied to all instances in a given dataset, irrespective of whether a feature scoring method is not the best for a particular instance (Atanasova et al. 2020; Jacovi and Goldberg 2020) while finding a suitable explanation length is an open problem (Zhang et al. 2021).

Computing Input Importance

Feature scoring methods Ω compute input importance scores $\boldsymbol{\omega}$ for each token in the sequence $\mathbf{x}$, such that $\boldsymbol{\omega} = \Omega(\mathcal{M}, \mathbf{x}, \mathcal{Y})$. High scores indicate that the associated tokens contributed more towards a model's prediction. Subsequently, $\mathcal{R}$ is extracted by selecting the K highest scored tokens (or K-gram for contiguous) in a sequence (DeYoung et al. 2020; Jain et al. 2020).

A common approach to computing $\boldsymbol{\omega}$ is by calculating the gradients of the prediction with respect to the input (Kindermans et al. 2016; Li et al. 2016; Arras et al. 2016; Sundararajan, Taly, and Yan 2017; Bastings and Filippova 2020). Jain et al. (2020) use attention weights to attribute token importance for rationale extraction, while Treviso and Martins (2020) propose sparse attention. Li et al. (2016) compute input importance scores by measuring the difference in a model's prediction between keeping and omitting each token, with Kim et al. (2020) also suggesting input marginalization as an alternative to token omission. Another way is using sparse linear meta-models that are easier to interpret (Ribeiro, Singh, and Guestrin 2016). Atanasova et al. (2020) however show that sparse linear meta-models are not as faithful as gradient-based approaches for interpreting large language models.

Evaluating Rationale Faithfulness

Having extracted $\mathcal{R}$, we typically need to evaluate how faithful that explanation is for a model's prediction. Several studies evaluate the faithfulness of explanations by training a separate classifier on an end-task using only the rationales as input (Jain et al. 2020; Treviso and Martins 2020). These classifiers are inherently faithful, as they are trained only on the rationales (Jain et al. 2020). Other studies compare the ability of different feature scoring methods to identify important tokens by using word erasure, i.e. masking (Samek et al. 2017; Serrano and Smith 2019; Atanasova et al. 2020; Chen and Ji 2020; DeYoung et al. 2020; Zhang et al. 2021; Chrysostomou and Aletras 2021a). The intuition is that by removing the most important tokens, it should result in a larger difference in the output probabilities, compared to removing a less important token which will also lead to drops in classification accuracy (Robnik-Šikonja and Kononenko 2008; Nguyen 2018; Atanasova et al. 2020). DeYoung et al. (2020) use erasure to evaluate the comprehensiveness and sufficiency of rationales. Carton, Rathore, and Tan (2020) suggest normalizing these metrics using the predictions of the model with a baseline input, to allow for a fairer comparison across models and datasets.

3 Instance-Specific Rationale Extraction

Our aim is to address the *"one-size-fits-all"* ad-hoc approach of previous work on rationale extraction with feature scoring methods that typically extracts rationales using the same feature scoring method, length and type across all instances in a dataset. Inspired by word erasure approaches (Nguyen 2018; Serrano and Smith 2019; DeYoung et al. 2020) we mask the tokens that constitute a rationale and record the difference δ in a model's output distribution by using the full text and the reduced input. Our main assumption is that a sufficiently faithful rationale is the one that will result into the largest δ (Atanasova et al. 2020; Chen and Ji 2020; DeYoung et al. 2020). Following this assumption, we can extract rationales by selecting for each instance a specific (1) *feature scoring method*; (2) *length*; and (3) *type*.[3]

Instance-level Feature Scoring Selection

Given a set of M feature scoring methods $\{\Omega_1, \ldots, \Omega_M\}$, we extract a rationale $\mathcal{R}$ as follows:

1. For each Ω_i in the set we compute input importance scores $\boldsymbol{\omega}_i = \Omega_i(\mathcal{M}, \mathbf{x}, \mathcal{Y})$;

2. We subsequently select the K highest scored tokens (TOPK) or the highest K-gram (CONTIGUOUS) to form a rationale $\mathcal{R}_i$, where K is the rationale length;

3. For each rationale we compute the difference δ_i, between the reference model output (using full text input) and the model output having masked the rationale, such that:

$$\delta_i = \Delta(\mathcal{Y}, \mathcal{Y}_i^m) = \Delta(\mathcal{M}(\mathbf{x}), \mathcal{M}(\mathbf{x}_{\backslash \mathcal{R}_i}))$$

---

[3]Similar to Jain et al. (2020), we consider two rationale types: (a) TOPK tokens ranked by a feature scoring method, treating each word in the input sequence independently; and (b) CONTIGUOUS span of input tokens of length K with the highest overall score computed by a feature scoring method.

where Δ is the function used to compute the difference between the two outputs;

4. We select the rationale $\mathcal{R}$ with the highest difference $\delta_{max} = max(\{\delta_1, \ldots, \delta_i, \ldots, \delta_M\})$.

For computing δ, we experiment with the following divergence metrics (Δ): (a) Kullback-Leibler (KL); (b) Jensen-Shannon divergence (JSD); (c) Perplexity (PERP.) and (d) Predicted Class Probability (CLASSDIFF).[4]

Instance-level Rationale Length Selection

For computing at instance-level the rationale length k and extracting the rationale R using a single feature scoring method Ω, we propose the following steps:

1. Given Ω, we first compute input importance scores $\boldsymbol{\omega} = \Omega(\mathcal{M}, \mathbf{x}, \mathcal{Y})$;

2. We then iterate over the sequence such that $k = range(1, N)$, where N is the fixed, pre-defined rationale length and k the possible rationale length at the current iteration. We set N as the upper bound rationale length for our approach to make results comparable with fixed length rationales.

3. At each iteration we begin by masking the top k tokens (as indicated by ω) to form a candidate rationale $\mathcal{R}_k$. When using TOPK we mask the k highest scored tokens, whilst with CONTIGUOUS we mask the highest scored k-gram;

4. We compute the difference δ_k between the reference model output $\mathcal{Y}$ and the model output having masked the candidate rationale $\mathcal{Y}_k^m = \mathcal{M}(\mathbf{x}_{\backslash \mathcal{R}_k})$;

5. We record every δ until $k = N$ and extract the rationale $\mathcal{R}$ with the highest difference $\delta_{max} = max(\{\delta_1, \ldots, \delta_k, \ldots, \delta_N\})$, where k at δ_{max} is the computed rationale length.[5]

Instance-level Rationale Type Selection

In a similar way to selecting a feature scoring method, our approach can also be used to select between different rationale types (i.e. CONTIGUOUS or TOPK) for each instance in the dataset.

Finally, our approach is flexible and can be easily modified to support selecting any of these parameters while keeping the rest fixed (i.e. feature scoring method, rationale length and rationale type) or by selecting any combination of them. An important benefit of our approach is that we extract rationales with different settings for each instance rather than using uniform settings globally (i.e. across the whole dataset), which we empirically demonstrate to be beneficial for faithfulness (§5).

---

[4]We describe the metrics in detail in App. B.

[5]We also experimented with early stopping, whereby the difference between δ_k and the δ_{max} until k are under a specified threshold, however this resulted in reduced performance (we included a more thorough analysis in the App. D).

Data	$\|W\|$	C	Splits Train/Dev/Test	F1	N
SST	18	2	6,920 / 872 / 1,821	90.1 ± 0.2	20%
AG	36	4	102,000 / 18,000 / 7,600	93.5 ± 0.2	20%
Ev.Inf.	363	3	5,789 / 684 / 720	83.0 ± 1.6	10%
M.RC	305	2	24,029 / 3,214 / 4,848	73.2 ± 1.7	20%

Table 1: Dataset statistics including average words at instance ($\|W\|$), number of classes (**C**), data splits, F1 macro performance and the fixed, pre-defined rationale ratio across all instances (N).

4   Experimental Setup

Tasks

For our experiments we use the following datasets (details in Table 1):

- **SST:** Binary sentiment classification without neutral sentences (Socher et al. 2013).

- **AG:** News articles categorized in Science, Sports, Business, and World topics (Corso, Gulli, and Romani 2005).

- **Evidence Inference (EV.INF.):** Abstract-only biomedical articles describing randomized controlled trials. The task is to infer the relationship between a given intervention and comparator with respect to an outcome (Lehman et al. 2019).

- **MultiRC (M.RC):** A reading comprehension task with questions having multiple correct answers that depend on information from multiple sentences (Khashabi et al. 2018). Following DeYoung et al. (2020) and Jain et al. (2020), we convert this to a binary classification task where each rationale/question/answer triplet forms an instance and each candidate answer is labeled as True/False

Models

Similar to Jain et al. (2020), we use BERT (Devlin et al. 2019) for SST and AG); SCIBERT (Beltagy, Lo, and Cohan 2019) for EV.INF. and ROBERTA (Liu et al. 2019) for M.RC. See App. A for hyperparameters.

Feature Scoring Methods

We use a random baseline and six other feature scoring methods (to compute input importance scores) similar to Jain et al. (2020) and Serrano and Smith (2019).

- **Random (RAND):** Random allocation of token importance.

- **Attention (α):** Token importance corresponding to normalized attention scores (Jain et al. 2020).

- **Scaled Attention ($\alpha\nabla\alpha$):** Scales the attention scores α_i with their corresponding gradients $\nabla\alpha_i = \frac{\partial \hat{y}}{\partial \alpha_i}$ (Serrano and Smith 2019) .

- **InputXGrad ($\mathbf{x}\nabla\mathbf{x}$):** Attributes input importance by multiplying the gradient of the input by the input with respect to the predicted class, where $\nabla x_i = \frac{\partial \hat{y}}{\partial x_i}$ (Kindermans et al. 2016; Atanasova et al. 2020) .

- **Integrated Gradients (IG):** Ranking words by computing the integral of the gradients taken along a straight path from a baseline input (zero embedding vector) to the original input (Sundararajan, Taly, and Yan 2017).

- **DeepLift:** Ranking words according to the difference between the activation of each neuron to a reference activation (Shrikumar, Greenside, and Kundaje 2017).

- **LIME:** Ranking words by learning an interpretable model locally around the prediction (Ribeiro, Singh, and Guestrin 2016).

Evaluating Explanation Faithfulness

- **F1 macro:** Similar to Arras et al. (2017) we measure the F1 macro performance of model $\mathcal{M}$ when masking the rationale in the original input ($\mathbf{x}_{\setminus \mathcal{R}}$). A key difference in our approach is that we use the predicted labels of the model with full input as gold labels, as we are interested in the faithfulness of explanations for the predictions of the model. Larger drops in F1 scores indicate that the extracted rationale is more faithful.[6]

- **Normalized Sufficiency (NormSuff):** We measure the degree to which the extracted rationales are sufficient for a model to make a prediction (DeYoung et al. 2020). Similar to Carton, Rathore, and Tan (2020) we bind sufficiency between 0 and 1 and use the reverse difference so that higher is better. We modify this metric and measure the normalized sufficiency (Carton, Rathore, and Tan 2020) such that:

$$\text{Suff}(\mathbf{x}, \hat{y}, \mathcal{R}) = 1 - max(0, p(\hat{y}|\mathbf{x}) - p(\hat{y}|\mathcal{R}))$$

$$\text{NormSuff}(\mathbf{x}, \hat{y}, \mathcal{R}) = \frac{\text{Suff}(\mathbf{x}, \hat{y}, \mathcal{R}) - \text{Suff}(\mathbf{x}, \hat{y}, 0)}{1 - \text{Suff}(\mathbf{x}, \hat{y}, 0)}$$
(1)

where $\text{Suff}(\mathbf{x}, \hat{y}, 0)$ is the sufficiency of a baseline input (zeroed out sequence) and $\hat{y}$ the model predicted class using the full text $\mathbf{x}$ as input, such that $\hat{y} = \arg\max(\mathcal{Y})$.

- **Normalized Comprehensiveness (NormComp):** We measure the extent to which a rationale is needed for a prediction (DeYoung et al. 2020). For an explanation to be highly comprehensive, the model's prediction when masking the rationale should have a high difference between the model's prediction with full text. Similarly to Carton, Rathore, and Tan (2020) we bind this metric between 0 and 1 and normalize it. We compute it by:

$$\text{Comp}(\mathbf{x}, \hat{y}, \mathcal{R}) = max(0, p(\hat{y}|\mathbf{x}) - p(\hat{y}|\mathbf{x}_{\setminus \mathcal{R}}))$$

$$\text{NormComp}(\mathbf{x}, \hat{y}, \mathcal{R}) = \frac{\text{Comp}(\mathbf{x}, \hat{y}, \mathcal{R})}{1 - \text{Suff}(\mathbf{x}, \hat{y}, 0)}$$
(2)

We do not conduct human experiments to evaluate explanation faithfulness since that is only relevant to explanation plausibility (i.e. how understandable by humans a rationale is (Jacovi and Goldberg 2020)) and in practice faithfulness

---

[6]We also conducted experiments using the dataset gold labels with results being comparable. We include all results in App. G.

and plausibility do not correlate (Atanasova et al. 2020). Finally we do not compare with select-then-predict methods (Lei, Barzilay, and Jaakkola 2016; Jain et al. 2020), as we are interested in faithfully explaining the model $\mathcal{M}$ and not forming inherently faithful classifiers.

Performance-Time Trade-off

Input erasure approaches typically require N forward passes to compute a rationale length (see §3) when removing one token at a time. Similar to Nguyen (2018); Atanasova et al. (2020), we expedite this process when selecting a rationale length by "skipping" every $X\%$ of tokens. For our work, we use a 2% skip rate which led to a seven-fold reduction in the time required to compute rationales for datasets comprising of long sequences, such as MRc and EvInf, with comparable performance in faithfulness to the slower process of removing one token at a time. We include in App. E the performance/skip-rate trade-off when (1) we do not use a skip-rate; (2) at 2% and (3) at 5%.

5 Results

Selecting Instance-specific Feature Scoring

Figure 1 compares the faithfulness of extracted rationales when using our proposed method for selecting an instance-specific feature scoring method (OURS) and our baselines, that use a single fixed pre-defined feature scoring method globally (i.e. across all instances in a dataset). We measure faithfulness using F1 macro (lower is better), mean Norm-Suff and mean NormComp (higher is better respectively). For clarity we show results using the TOPK rationale type, with results for CONTIGUOUS included in App. G.[7]

Overall, results demonstrate that rationales extracted with our proposed approach are highly sufficient and comprehensive. In fact, our approach results in more sufficient rationales against all single feature scoring methods in AG and is comparable with the best NormSuff scores in the remainder of the datasets. This suggests that even when rationales with our proposed method are not the most sufficient, they are consistently highly sufficient (i.e. rationales extracted with our approach are significantly more sufficient than fixed, pre-defined feature scoring methods in 18 out of 24 test cases). Compared to our six baselines, the rationales extracted with our approach are significantly more comprehensive across all four datasets (Wilcoxon Rank Sum, $p < .05$). Additionally, the larger drops in F1 macro performance demonstrate that rationales extracted with our proposed approach are more necessary for a model to make a prediction compared to a globally used, pre-defined feature scoring approach.

Our results strengthen the hypothesis that whilst some feature scoring methods are better than others globally, they might not be optimal for all instances in a dataset (Jacovi and Goldberg 2020) and our approach helps mitigate that. Similar to (Atanasova et al. 2020), we observe that the faithfulness performance of single feature scoring methods varies

---

[7]Also for clarity, all results presented in this work are using JSD for Δ. The other divergence functions performed comparably and we include a comparison between them in App. F.

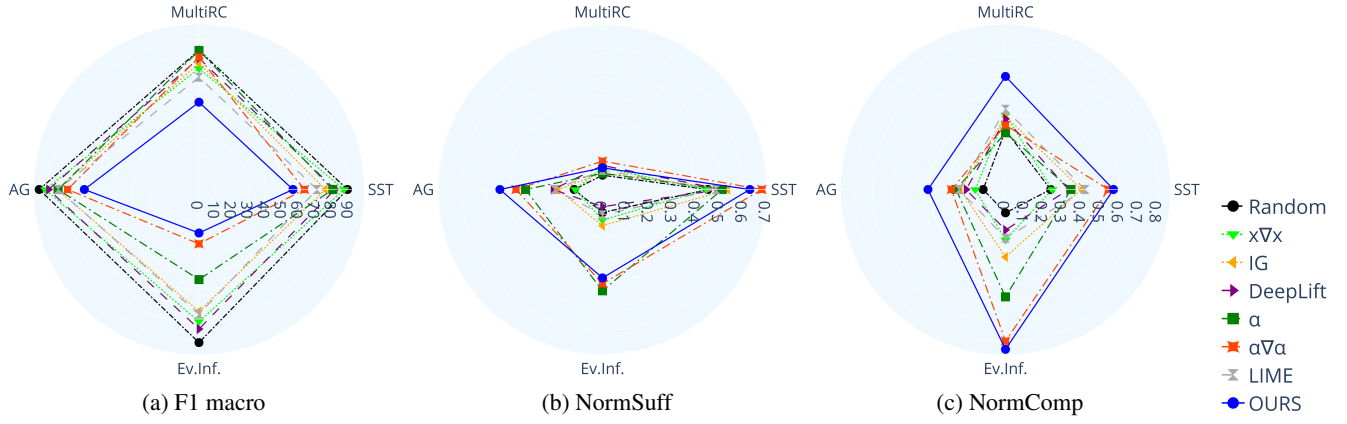(a) F1 macro   (b) NormSuff   (c) NormComp

Figure 1: F1 macro (lower is better), mean NormSuff (higher is better) and mean NormComp (higher is better), when using any single feature scoring method across all instances in a dataset and our proposed method of selecting a feature scoring method for each instance (OURS) for TOPK rationale types.

across datasets. For example LIME returns more comprehensive rationales than $\alpha\nabla\alpha$ in MultiRC, however is outperformed by the latter in SST. By returning consistently highly comprehensive and sufficient rationales, our propose method helps reducing the variability in faithfulness performance observed when using any single feature scoring method across datasets.

Selecting Instance-specific Rationale Length

Table 2 shows the Relative Improvement (R.I.) ratio in mean NormSuff and NormComp (>1.0 is better) between rationales extracted using a fixed pre-defined length (see N in Table 1) and rationales extracted using our method with instance-specific length across feature scoring methods and datasets. For brevity the detailed results, including F1 macro where we make similar observations to comprehensiveness, are included in App. G. Overall, rationales extracted using our approach are on average shorter than fixed length rationales. Specifically, rationale length drops from 20% to 16% on average in SST, AG; from 20% to 15% in M.Rc and from 10% to 7% in Ev.Inf..[8]

NormSuff scores indicate that our shorter on average rationales are overall less but comparably sufficient with longer, fixed-length rationales. For example with SST rationales with instance-specific length are 0.9-1.0 times less sufficient that rationale with pre-defined length. We find this particularly evident in datasets such as M.Rc and Ev.Inf., where our rationales are on average 4-5% shorter (approximately 15 tokens shorter on average for α in M.Rc) but still retain comparable sufficiency, while in some cases improving it (e.g. 1.2 R.I. in Ev.Inf. with DeepLift).

We also note that rationales extracted with instance-specific length are more comprehensive in most cases, despite being shorter on average compared to fixed-length rationales. For example in Ev.Inf., CONTIGUOUS rationales with I.G. are 1.4 times more comprehensive when we se-

---

[8]We include the computed rationale length results in App. C and the same evaluation under $2 \times N$ in App. G.

	FEAT	NormSuff				NormComp			
		SST	M.Rc	AG	Ev.Inf.	SST	M.Rc	AG	Ev.Inf.
TOPK	**DeepLift**	0.9	0.8	0.8	1.1	0.8	1.1	1.0	1.0
	LIME	1.0	0.7	0.9	0.9	0.9	1.1	1.0	1.0
	α	0.9	0.9	0.7	0.8	0.8	1.1	0.9	1.2
	$\alpha\nabla\alpha$	0.9	0.9	0.8	0.9	1.0	1.1	0.9	1.0
	IG	0.9	0.9	0.8	0.9	0.9	1.1	1.0	1.1
	$x\nabla x$	1.0	0.8	0.7	0.8	0.9	1.1	0.9	1.2
CONTIGUOUS	**DeepLift**	0.9	0.9	0.8	1.2	0.9	1.1	1.3	1.5
	LIME	0.9	0.7	0.8	0.9	1.0	1.1	1.2	1.3
	α	0.9	0.9	0.7	0.9	0.7	1.1	1.0	1.2
	$\alpha\nabla\alpha$	0.9	0.8	0.8	0.9	1.0	1.1	1.1	1.1
	IG	0.9	0.8	0.8	1.0	1.0	1.2	1.2	1.4
	$x\nabla x$	0.9	0.8	0.7	1.0	1.0	1.1	1.0	1.3

Table 2: Relative Improvement (R.I.) ratios for mean NormSuff and mean NormComp between fixed length rationales (see N in Table 1) extracted using our method and rationales with instance-specific length (>1.0 is better).

lect their length at instance-level. Results also indicate that using our proposed method benefits more CONTIGUOUS rationales compared to TOPK for comprehensiveness, leading to increased R.I. in the majority of cases. Overall, findings support our initial hypothesis that in certain cases a rationale with longer than needed length might contain unnecessary information and adversely impact its comprehensiveness.

Selecting Instance-specific Feature Scoring, Length and Type

Table 3 shows mean NormSuff and NormComp scores when using our proposed method to select at instance-level (I-L) a combination of: (1) the feature scoring method (FEAT); (2) the rationale length (LEN); and (3) the rationale type (TYPE). For comparison, we also show scores of the best performing fixed (FIX) feature scoring function, rationale type and length (see Figure 1). For brevity we include results with F1 macro in App. G, observing similar trends.

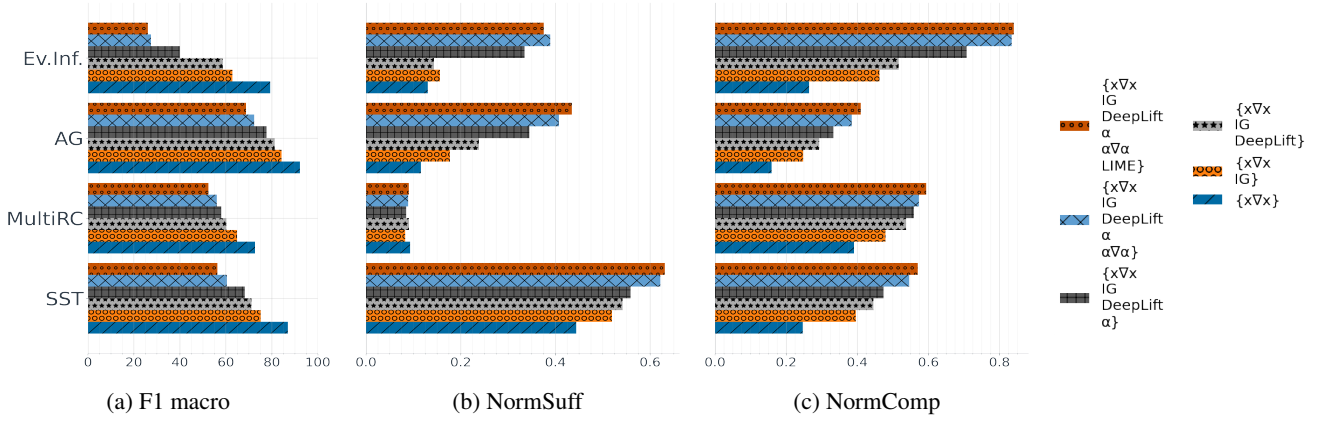We first observe that the highest NormSuff scores across

Figure 2: F1 macro (lower is better), mean NormSuff and mean NormComp (higher is better), when extracting rationales with our approach given decreasing numbers of feature scoring methods.

			NormSuff				NormComp			
TYPE	LEN	FEAT	SST	M.Rc	AG	Ev.Inf.	SST	M.Rc	AG	Ev.Inf.
TOPK	FIX	FIX	.68	**.12**	.37	.43	.54	.42	.28	.80
	I-L	FIX	.61	.11	.30	.37	.52	.46	.27	.82
	FIX	I-L	.63	.09	**.44**	.38	**.57**	.59	**.41**	.84
	I-L	I-L	.59	.07	.38	.36	.55	.62	.39	.86
CONT.	FIX	FIX	**.71**	.07	.41	**.85**	.46	.47	.17	.55
	I-L	FIX	.63	.06	.33	.78	.47	.54	.19	.62
	FIX	I-L	.67	.07	.42	.82	.46	.60	.22	.59
	I-L	I-L	.61	.05	.33	.76	.48	.65	.24	.67
I-L	I-L	I-L	.60	.06	.39	.49	**.57**	**.69**	**.41**	**.88**

Table 3: Mean NormSuff and NormComp scores when we select at instance-level (I-L) a combination of the: (1) rationale length (LEN); (2) feature scoring method (FEAT.); and (3) rationale type (TYPE). {TYPE}-FIX-FIX and {TYPE}-I-L-FIX values are from the highest scoring feature scoring method (see Figure 1). **Bold** values denote the highest performing combination in column-wise (higher is better).

three datasets (SST, MRc, EvInf), are from the best performing fixed scoring method with fixed length and rationale type. Additionally, the best performing combination of our proposed approach for sufficiency is when we only select the feature scoring method keeping the length and type fixed. This combination results in the highest NormSuff scores in AG (.44 with TOPK type compared to .42, which is the second best with CONTIGUOUS) and competitive NormSuff scores with the highest scoring combination (e.g. .82 in Ev.Inf. and CONTIGUOUS compared to .85). We assume that using combinations which include instance-specific lengths do not perform as well for sufficiency due to the shorter rationale length, which we have previously shown to partially degrade rationale sufficiency.

Finally, our results demonstrate that we obtain highly comprehensive rationales when selecting at instance level all parameters (FEAT. + LEN + TYPE) using our approach. In fact, this results in higher NormComp scores compared to any other setting combination across all datasets. For example in M.Rc., selecting all parameters results in a Norm-

Comp score of .69 which is .22 units higher than the rationales extracted with fixed feature scoring method and length and type. This highlights the efficacy of our approach in extracting highly comprehensive rationales, without requiring strong a priori assumptions about rationale parameters.

Ablation Study

We finally perform an ablation study to examine the behavior and effectiveness of our approach by sequentially removing one feature scoring method at a time to measure changes in F1 macro, NormSuff and NormComp. The intuition is that we should observe drops in faithfulness scores when removing feature attribution methods for our approach to be effective (i.e. we should extract more faithful rationales when having more feature scoring options to choose from). Figure 2 shows the results.

We first observe that removing one feature scoring method at a time results in increases in F1 macro (lower is better) and drops in NormComp scores (higher is better). This demonstrate that the faithfulness of the rationales extracted with our approach deteriorates as the number of feature scoring methods becomes smaller highlighting the efficacy of our proposed approach. For example, in Ev.Inf. by removing $\alpha\nabla\alpha$ results in a drop of .14 in mean NormComp (.84 when including $\alpha\nabla\alpha$ compared to .70 without it). On the other hand, we also observe that our method can still benefit from feature scoring methods that achieve low NormComp scores when used standalone, resulting in improvements in comprehensiveness and drops in F1 macro (e.g. α in SST). This indicates that our approach steadily improves rationale faithfulness for model's predictions given a larger pool of available feature scoring methods.

Results show a deterioration in NormSuff scores as the number of feature scoring methods becomes smaller, showing that our method results in more sufficient rationales when presented with a larger list of available feature scoring methods in the majority of the datasets. We hypothesize that this is not true for MultiRC due to the already low NormSuff scores of the rationales (e.g. no more than 0.12). By using all six feature scoring methods, our approach produces highly

Example 1	**Data.:AG Id: test_4614**

[FIXED-LEN + α]: ... game last `Friday night will stand , the CFL announced yesterday. While` a review ...
[I-L-LEN + α (Ours)]: ... game last Friday night will stand , `the CFL announced yesterday`. While a review ...
[Predicted Topic || True Topic]: Decreased significantly || Decreased significantly

Example 2	**Data.:Ev.INF. Id: 3162205_2**

[FIXED-LEN + $\alpha\nabla\alpha$]: ... computed tomography (3D - CT `) scans . ABSTRACT.RESULTS : The control sides treated with an autograft showed significantly better Lenke scores than the study sides treated with $\beta$ - CPP at 3 and 6 months postoperatively , but there was no difference between the two sides at 12 months` . The fusion ..
[I-L-LEN + $\alpha\nabla\alpha$ (Ours)]: ... computed tomography (3D - CT) scans `. ABSTRACT.RESULTS : The control sides treated with an autograft showed significantly better Lenke scores` than the study sides treated with β - CPP at 3 ...
[Predicted Relationship || True Relationship]: Increased significantly || No significant difference

Example 3	**Data.:SST Id: test_694**

[FIXED-LEN + α]: ... Frontal `is the antidote for Soderbergh` fans who think he s gone too commercial ...
[I-L-LEN + I-L-FEAT (Ours)]: ... Frontal is the antidote for Soderbergh fans who think he `s gone too commercial` ...
[Predicted Sentiment || True Sentiment]: Negative || Positive

Example 4	**Data.:SST Id: test_1039**

[FIXED-LEN + α]: It 's just `incredibly` dull.
[I-L-LEN + I-L-FEAT (Ours)]: It 's just `incredibly dull`.
[Predicted Sentiment || True Sentiment]: Negative || Negative

Table 4: Examples when using our approach (Ours) to select at instance-level (I-L) a combination of the: (1) rationale length (LEN); (2) feature scoring method (FEAT) against our baseline of fixed-length rationales from a fixed feature scoring method.

sufficient rationales and is comparable to the set achieved the highest sufficiency. For example in Ev.Inf. using all feature scoring methods results to a NormSuff score of approximately .38 compared to the highest scoring feature scoring set (all except LIME) and the lowest scoring ($\mathbf{x}\nabla\mathbf{x}$) which achieved .39 and .15 respectively.

We also tested different combinations of feature scoring methods with similar observations (see App. G). Finally, we experimented with doubling the upper bound of the rationale length (from N to $2 \times N$) for both fixed length rationales and our proposed approach. Our approach still yielded more comprehensive rationales compared to the fixed-length ones that were also highly sufficient (results included in App. G).

6 Qualitative Analysis

Table 4 shows examples of the qualitative comparison between our approach (Ours) for selecting at instance-level (I-L) a combination of the: (1) rationale length (LEN); (2) feature scoring method (FEAT against our baseline of fixed-length rationales from a fixed feature scoring method.

Concise rationales: Example 1 presents an instance from AG. Our approach extracts a rationale that is six tokens shorter than the one with fixed length while also achieving a higher NormComp score. However, the fixed length rationale scores higher in NormSuff. We can assume from this that sufficiency positively correlates with rationale length.

Error analysis: Our assumption is that if a model makes a wrong prediction, we should be able to extract the rationale that better demonstrates what led to a wrong prediction. Example 2 shows an instance from Ev.INF., where the model has wrongly predicted that "Lenke scores at 12 months" have 'increased significantly' instead of the correct 'no significant difference'. Surprisingly, both rationales recorded

maximum scores (1.0) in NormSuff and NormComp. We observe that the correct answer is included in the fixed length rationale, however the model made a wrong prediction. On the contrary, our rationale highlights something directly related to its prediction.

Example 3 presents an instance from SST, where the fixed-length rationale and the instance-specific rationale (ours) attend at different sections of the text. Our rationale scored lower for NormSuff, however we observe that it aligns more closely with the predicted sentiment.

When using a fixed pre-defined length is not sufficient: Example 4 presents a different scenario, where the fixed-length rationale for SST is at 20% whilst the upper bound N for our rationale is at 40%. The intuition is that in certain cases a fixed rationale length might not be sufficient for all instances to explain a prediction. We argue that our approach highlighted something more informative for the task ("incredibly dull" compared to "incredibly"), due to removing the restriction of a pre-defined fixed length.

7 Conclusions

We have proposed a simple yet effective approach for selecting at instance-level (1) feature scoring method; (2) length; and (3) type of the rationale. We empirically demonstrated that rationales extracted with our approach are significantly more comprehensive and highly sufficient, while being shorter compared to rationales extracted with a fixed feature scoring method, length and type. Finally, we consider our work an important step towards instance-level faithful rationalization while finding the most sufficient rationale, an interesting direction for future work.

Acknowledgments

References

Arras, L.; Horn, F.; Montavon, G.; Müller, K.-R.; and Samek, W. 2016. Explaining Predictions of Non-Linear Classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, 1–7. Berlin, Germany: Association for Computational Linguistics.

Arras, L.; Montavon, G.; Müller, K.-R.; and Samek, W. 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 159–168. Copenhagen, Denmark: Association for Computational Linguistics.

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020. A Diagnostic Study of Explainability Techniques for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3256–3274. Online: Association for Computational Linguistics.

Bastings, J.; Aziz, W.; and Titov, I. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2963–2977. Florence, Italy: Association for Computational Linguistics.

Bastings, J.; and Filippova, K. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 149–155. Online: Association for Computational Linguistics.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. Hong Kong, China: Association for Computational Linguistics.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

Carton, S.; Rathore, A.; and Tan, C. 2020. Evaluating and Characterizing Human Rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9294–9307. Online: Association for Computational Linguistics.

Chen, H.; and Ji, Y. 2020. Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4236–4251. Online: Association for Computational Linguistics.

Chrysostomou, G.; and Aletras, N. 2021a. Enjoy the Salience: Towards Better Transformer-based Faithful Explanations with Word Salience. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8189–8200. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Chrysostomou, G.; and Aletras, N. 2021b. Improving the Faithfulness of Attention-based Explanations with Task-specific Information for Text Classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 477–488. Online: Association for Computational Linguistics.

Corso, G. M. D.; Gulli, A.; and Romani, F. 2005. Ranking a stream of news. In Ellis, A.; and Hagino, T., eds., *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, 97–106. ACM.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458. Online: Association for Computational Linguistics.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. Online: Association for Computational Linguistics.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3543–3556. Minneapolis, Minnesota: Association for Computational Linguistics.

Jain, S.; Wiegreffe, S.; Pinter, Y.; and Wallace, B. C. 2020. Learning to Faithfully Rationalize by Construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4459–4473. Online: Association for Computational Linguistics.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 252–262. New Orleans, Louisiana: Association for Computational Linguistics.

Kim, S.; Yi, J.; Kim, E.; and Yoon, S. 2020. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3154–3167. Online: Association for Computational Linguistics.

Kindermans, P.-J.; Schütt, K.; Müller, K.-R.; and Dähne, S. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*.

Lehman, E.; DeYoung, J.; Barzilay, R.; and Wallace, B. C. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3705–3717. Minneapolis, Minnesota: Association for Computational Linguistics.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117. Austin, Texas: Association for Computational Linguistics.

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2016. Visualizing and Understanding Neural Models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 681–691. San Diego, California: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Nguyen, D. 2018. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1069–1078. New Orleans, Louisiana: Association for Computational Linguistics.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Krishnapuram, B.; Shah, M.; Smola, A. J.; Aggarwal, C. C.; Shen, D.; and Rastogi, R., eds., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144. ACM.

Robnik-Šikonja, M.; and Kononenko, I. 2008. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5): 589–600.

Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; and Müller, K. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11): 2660–2673.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Florence, Italy: Association for Computational Linguistics.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3145–3153. PMLR.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 3319–3328. PMLR.

Treviso, M.; and Martins, A. F. T. 2020. The Explanation Game: Towards Prediction Explainability through Sparse Communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 107–118. Online: Association for Computational Linguistics.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Wiegreffe, S.; and Pinter, Y. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 11–20. Hong Kong, China: Association for Computational Linguistics.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Zhang, W.; Huang, Z.; Zhu, Y.; Ye, G.; Cui, X.; and Zhang, F. 2021. On Sample Based Explanation Methods for NLP: Faithfulness, Efficiency and Semantic Evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5399–5411. Online: Association for Computational Linguistics.