---

**Article:**
Chrysostomou, G. and Aletras, N. (Submitted: 2021) Variable instance-level explainability for text classification. arXiv. (Submitted)

---

# Variable Instance-Level Explainability for Text Classification

**George Chrysostomou** **Nikolaos Aletras**
Department of Computer Science, University of Sheffield
United Kingdom
{gchrysostomou1, n.aletras}@sheffield.ac.uk



## Abstract

Despite the high accuracy of pretrained transformer networks in text classification, a persisting issue is their significant complexity that makes them hard to interpret. Recent research has focused on developing feature scoring methods for identifying which parts of the input are most important for the model to make a particular prediction and use it as an explanation (i.e. rationale). A limitation of these approaches is that they assume that a particular feature scoring method should be used across all instances in a dataset using a predefined fixed length, which might not be optimal across all instances. To address this, we propose a method for extracting variable-length explanations using a set of different feature scoring methods at instance-level. Our method is inspired by word erasure approaches which assume that the most faithful rationale for a prediction should be the one with the highest divergence between the model's output distribution using the full text and the text after removing the rationale for a particular instance. Evaluation on four standard text classification datasets shows that our method consistently provides more faithful explanations compared to previous fixed-length and fixed-feature scoring methods for rationale extraction.[1]

## 1 Introduction

Large pre-trained transformer-based language models such as BERT (Devlin et al., 2019), currently dominate performance across language understanding benchmarks (Wang et al., 2018). These developments have opened up new challenges on how to extract faithful explanations (i.e. rationales[2]) that accurately represent the true reasons behind their predictions when adapted to downstream tasks (Jacovi and Goldberg, 2020).

Recent studies use feature scoring methods such as gradient and attention scores (Arras et al., 2016; Sundararajan et al., 2017; Jain and Wallace, 2019) to identify important (i.e. salient) segments of the input and extract them as rationales (Jain et al., 2020; Treviso and Martins, 2020). However, a limitation of these approaches is that they set an a priori fixed rationale length (i.e. the ratio of a rationale compared to the full input sequence) across instances. We hypothesize that a fixed ratio can either not suffice for explaining a model's prediction in certain instances in the dataset or in some cases provides a larger number of tokens than needed, thus reducing the faithfulness of a rationale. Additionally, these approaches extract rationales using a single feature scoring method across a dataset which might not be the best for every instance (Jacovi and Goldberg, 2020; Atanasova et al., 2020).

Motivated by these limitations, we propose a method for extracting the best variable-length rationale from a set of different feature scoring methods, for each instance in a dataset. We achieve this by computing differences between a model's output distributions obtained using the full input sequence and the input without the rationale respectively. Our method is based on the assumption that by removing important tokens from the sequence, we should observe large differences in the model's confidence for the correct class (Nguyen, 2018; Serrano and Smith, 2019) resulting into more faithful rationales (Atanasova et al., 2020; Chen and Ji, 2020).

The contributions of our paper are as follows:

- We propose a method for extracting variable-length rationales with variable feature scoring methods at each instance in a dataset,;
- We empirically demonstrate that rationales generated with our proposed approach, are on average shorter and more faithful compared

[1] Code for experiments will be released.
[2] We use these terms interchangeably throughout the paper.

to longer rationales from any single feature scoring method;

- We show that we are able to select the feature scoring methods that leads to more faithful rationales at instance level, from a set of feature scoring methods, irrespective of the rationale length.

## 2 Background and Related Work

### 2.1 Rationale Extraction

Given a model $\mathcal{M}$, an input $\mathbf{x} = [x_1, \ldots, x_T]$ and a prediction $\mathcal{Y}$ which represents a distribution over classes, rationale extraction methods seek to identify the most important (i.e. salient) 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 (Lei et al., 2016; Bastings et al., 2019; Chalkidis et al., 2021). 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 methods (i.e. salience metrics) to first identify important parts of the input and then extract the rationales from an already trained model (Jain et al., 2020; Treviso and Martins, 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 if a feature scoring method is not the best for a particular instance using a predefined fixed rationale length.

### 2.2 Computing Input Importance

Feature scoring methods $\Omega$ 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.

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 et al., 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 as indicators of input token importance. Li et al. (2016) compute input importance scores by measuring the difference in a model's prediction between keeping and omitting each token. Kim et al. (2020) also suggest input marginalisation as an alternative to token omission. Another way is to use sparse linear meta-models that are easier to interpret (Ribeiro et al., 2016; Lundberg and Lee, 2017). Atanasova et al. (2020) however show that sparse linear meta-models are not as faithful as gradient-based approaches.

### 2.3 Evaluating Explanation Faithfulness

Having extracted a rationale, 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). The classifiers are inherently faithful, as they are trained only on the rationales (Jain et al., 2020). Other studies conduct comparative analyses on the ability of feature scoring methods to identify important tokens by using erasure as their basis (Samek et al., 2017; Arras et al., 2017; Nguyen, 2018; Serrano and Smith, 2019; Atanasova et al., 2020; Vashishth et al., 2019; Grimsley et al., 2020; Chen and Ji, 2020). The intuition is that by removing the most important tokens, we expect to see a larger difference in the output probabilities, compared to removing a less important token leading to drops in classification accuracy (Robnik-Šikonja and Kononenko, 2008; Nguyen, 2018; Atanasova et al., 2020). Nguyen (2018) and Serrano and Smith (2019) for example, count the fraction of tokens needed to be removed to cause a prediction switch for $\mathcal{M}$ (i.e. decision flip). The lower the fraction of tokens needed for a decision flip, the more faithful the rationale by a feature scoring method is.

## 3 Methodology

Our aim is to address the *"one-size-fits-all"* approach of previous work on rationale extraction with feature scoring methods that typically extracts fixed-length rationales using the same feature scoring method across all instances in a dataset.

Inspired by word erasure approaches (Nguyen, 2018; Serrano and Smith, 2019), we remove sequentially the highest ranked tokens by a given feature scoring method from the sequence (reduced input) and add them to the rationale. We record the difference $\delta$ in a model's output distribution with

using the full text and the reduced input. Our main assumption is that a sufficiently faithful rationale is the one that will cause the largest $\delta$ (Atanasova et al., 2020; Chen and Ji, 2020). Following this assumption, we can extract rationales of variable (1) *length*; (2) *feature scoring method*; and (3) *type*.[3]

### 3.1 Extracting Variable-Length Rationales

Our method consists of the following steps for extracting variable length rationales using a single feature scoring method for a single input sequence (see Algorithm 1):

1. We first compute the reference output probability distribution passing the full input $\mathbf{x}$ through model $\mathcal{M}$, $\mathcal{Y} = \mathcal{M}(\mathbf{x})$;
2. We subsequently compute input importance scores $\boldsymbol{\omega} = \Omega(\mathcal{M}, \mathbf{x}, \mathcal{Y})$ for the entire input sequence using a feature scoring method $\Omega$;
3. We rank all input tokens $\mathbf{x}$ by their importance scores $\omega$ in decreasing order, such that $\mathbf{x}_{ranked} = \text{argsort}(\mathbf{x}, \boldsymbol{\omega})$. For CONTIGUOUS rationales, we first split the input sequence into $n$-grams and rank them by decreasing importance (arg sort$_{n-gram}(\mathbf{x}, \boldsymbol{\omega})$);
4. We then iterate through input tokens until we reach the max rationale length $N$ (upper bound). For TOPK rationales, we begin by retrieving the input tokens $\{1, ..., n\}$ from $\mathbf{x}_{ranked}$ at each step $n$ of the iteration. For CONTIGUOUS rationales, we extract the top $n$-gram;
5. We then mask the top-$n$ tokens or top $n$-gram, depending on the type of rationale, to obtain $\mathbf{x}_{masked}$. The masked input sequence $\mathbf{x}_{masked}$ is then passed from the model $\mathcal{M}$, to obtain the output distribution $\mathcal{Y}^m$;
6. We compute and record the divergence $\delta$ of distribution $\mathcal{Y}^m$ from $\mathcal{Y}$, such that $\delta = \Delta(\mathcal{Y}, \mathcal{Y}^m)$. For computing $\delta$ we experiment with the following divergence metrics ($\Delta$): (a) Kullback-Leibler (KL) ; (b) Jensen-Shannon divergence (JSD) ; (c) Perplexity (PERP.) and (d) Predicted Class Probability (CLASSDIFF).[4]

[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.

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

**Algorithm 1:** Compute Variable-Length Rationale $K$ for TOPK type

**Input:** $\mathbf{x}$ ; $\mathcal{M}$ ; $\Omega$ ; $N$
**Output:** $\mathcal{R}$, $\delta_{max}$
$\delta_{max} = 0$
/* output distribution for $\mathbf{x}$ */
$\mathcal{Y} = \mathcal{M}(\mathbf{x})$
/* compute feature importance */
$\boldsymbol{\omega} = \Omega(\mathcal{M}, \mathbf{x})$
/* rank input tokens by $\boldsymbol{\omega}$ */
$\mathbf{x}_{ranked} = \text{argsort}(\mathbf{x}, \boldsymbol{\omega})$;
$\mathbf{x}_{masked} = \mathbf{x}$
**for** $n \leftarrow 1$ *in* $N$ **do**
  $ind_{masked} = \mathbf{x}_{ranked}[n]$
  $\mathbf{x}_{masked}[ind_{masked}]$ = [MASK]
  /* output distribution for $\mathbf{x}_{masked}$ */
  $\mathcal{Y}^m = \mathcal{M}(\mathbf{x}_{masked})$;
  /* compute divergence $\delta$ between output distributions */
  $\delta = \Delta(\mathcal{Y}, \mathcal{Y}^m)$;
  **if** $\delta > \delta_{max}$ **then**
    $\delta_{max} = \delta$;
    $\mathcal{R} \leftarrow x_{masked}[ind_{masked}]$
  **end**
**end**

7. Finally, we extract the rationale $\mathcal{R}$ with length $K$ with $K \leq N$ at step $n$ where we recorded the highest divergence $\delta$.[5]

### 3.2 Instance-level Feature Scoring Selection

Rationales are often extracted and evaluated using a single feature scoring method across all instances in a dataset (Serrano and Smith, 2019; Nguyen, 2018; Treviso and Martins, 2020; Jain et al., 2020). However, Atanasova et al. (2020) demonstrated that there is no clear best feature scoring method across text classification tasks, which might also hold for different instances in the same dataset (Jacovi and Goldberg, 2020).

Given a set of different feature scoring methods $\{\Omega_1, \ldots, \Omega_k\}$, we first extract rationales $\boldsymbol{\mathcal{R}} = [\mathcal{R}_{\Omega_1}, \ldots, \mathcal{R}_{\Omega_k}]$ using Algorithm 1 and select the one with highest $\delta_{max}$ (hereby denoted by FEAT$_{max(\delta)}$).

### 3.3 Instance-level Rationale Type Selection

In a similar way, our approach can also be used to select between different rationale types (i.e. contiguous or TopK) for each instance in the dataset, hereby denoted by TYPE$_{max(\delta)}$. Finally, our approach is flexible and can be easily modified to support fixed-length rationales by directly computing

[5] We also experimented with using $\delta - \delta_{max} \leq threshold$ resulting into reduced performance (see Appx. D).

$\delta$ between the original input and the input without the fixed length rationale.

A benefit of our method compared to Jain et al. (2020) and Treviso and Martins (2020), is that we do not need to train separate classifiers over the rationales to evaluate the faithfulness of explanations. The primary reason behind this is that we are interested in finding whether a rationale is faithful for model $\mathcal{M}$, and not to form inherently faithful classifiers (Jain et al., 2020). Finally, another important benefit of our approach is that we evaluate rationales at instance level rather than globally (across a dataset).

## 4 Experimental Setup

### 4.1 Tasks

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

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

**AG:** News articles categorized in Science, Sports, Business, and World topics (Del Corso et al., 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 composed of questions with multiple correct answers that depend on information from multiple sentences (Khashabi et al., 2018). Similar to 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 or False.

### 4.2 Models

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

### 4.3 Feature Scoring Methods

We opt using a random baseline and four other feature scoring methods (to compute input importance scores) as in Jain et al. (2020) and Serrano and Smith (2019). All scores are normalized (sum up to 1).

| DATA | $\lvert W\rvert$ | C | SPLITS TRAIN/DEV/TEST | F1 |
|---|---|---|---|---|
| SST | 18 | 2 | 6,920 / 872 / 1,821 | $90.7 \pm 0.1$ |
| AG | 36 | 4 | 102,000 / 18,000 / 7,600 | $92.7 \pm 0.2$ |
| EV.INF. | 363 | 3 | 5,789 / 684 / 720 | $80.6 \pm 0.6$ |
| M.RC | 305 | 2 | 24,029 / 3,214 / 4,848 | $75.9 \pm 0.4$ |

Table 1: Dataset statistics including average words at instance ($|W|$), number of classes (**C**) and data splits.

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

**Attention ($\alpha$):** Token importance corresponding to normalized attention scores (Jain et al., 2020).

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

**InputXGrad (x$\nabla$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 to the original input, where the baseline is the zero embedding vector (Sundararajan et al., 2017).

### 4.4 Evaluating Explanation Faithfulness

**F1 macro:** F1 macro performance of model $\mathcal{M}$ when masking the rationale in the original input ($\mathbf{x}_{\setminus\mathcal{R}}$) similar to (Arras et al., 2017). Larger drops in F1 scores indicate that the extracted rationale is more faithful.

**Word Relevance:** Following Arras et al. (2017), we also mask input tokens one-by-one in decreasing order ranked by a feature scoring method, and record at each step the model's performance.

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

## 5 Results

Table 2 presents the predictive performance (macro F1) of each model $\mathcal{M}$ obtained by masking the rationale from the full input. Rationales are extracted using our proposed approaches consisting of: (1)

| | | SST | | AG | | EV.INF | | M.RC | | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| FULL INPUT | | 90.7 | | 92.7 | | 80.6 | | 75.9 | | 85.0 |
| | N | 20% | | 20% | | 10% | | 20% | | |
| LEN | FEAT SCORING | TOPK | CONT | TOPK | CONT | TOPK | CONT | TOPK | CONT | |
| FIXED (K=N) | FIXED-RAND | 84.3 | 84.8 | 90.6 | 90.6 | 76.3 | 78.4 | 60.0 | 46.0 | 76.2 |
| | FIXED-$\alpha$ | 69.7 | 74.5 | 77.3 | 87.4 | 46.0 | 59.0 | 36.8 | 45.3 | 62.0 |
| | FIXED-$\alpha\nabla\alpha$ | 72.1 | 76.1 | 79.9 | 87.3 | 36.0 | 59.3 | 38.6 | 43.3 | 61.6 |
| | FIXED-$\mathbf{x}\nabla\mathbf{x}$ | 83.5 | 83.7 | 88.5 | 88.7 | 71.7 | 74.4 | 46.6 | 45.1 | 72.8 |
| | FIXED-IG | 83.5 | 82.9 | 88.0 | 89.4 | 72.0 | 75.5 | 51.0 | 44.8 | 73.4 |
| OURS | | | | | | | | | | |
| FIXED (K=N) | VAR-FEAT | 61.8 | 67.4 | 70.9 | 83.5 | 28.3 | 53.7 | 37.5 | 40.7 | 55.5 |
| VAR (K≤N) | FIXED-RAND | 85.9 | 86.8 | 91.3 | 91.3 | 78.8 | 80.3 | 66.4 | 52.5 | 79.2 |
| | FIXED-$\alpha$ | 71.1 | 74.8 | 77.7 | 87.3 | 36.8 | 50.1 | 38.3 | 43.1 | 59.9 |
| | FIXED-$\alpha\nabla\alpha$ | 72.8 | 76.0 | 79.9 | 86.5 | 32.6 | 48.8 | 39.1 | 42.9 | 59.8 |
| | FIXED-$\mathbf{x}\nabla\mathbf{x}$ | 84.0 | 83.5 | 88.6 | 88.9 | 70.0 | 66.5 | 48.0 | 45.3 | 71.8 |
| | FIXED-IG | 84.0 | 83.0 | 88.2 | 89.5 | 71.0 | 73.7 | 51.1 | 47.0 | 73.4 |
| VAR (K≤N) | VAR-FEAT | 64.0 | 67.7 | 71.5 | 83.0 | 27.5 | 44.5 | 36.6 | 40.5 | 54.4 |
| | VAR-FEAT+TYPE | **59.9** | | **70.4** | | **26.0** | | **36.2** | | **48.1** |

Table 2: Macro F1 scores for measuring the faithfulness of explanations by masking the rationale ($\mathbf{x}_{\setminus\mathcal{R}}$) (lower is better). $K$ and $N$ denote the rationale length per instance and its upper bound respectively. FIXED and VAR indicate fixed and variable (i.e. different per instance) length, feature scoring method or type.

fixed length, fixed type and variable feature scoring method per instance; (2) variable length, fixed feature scoring method and fixed type; (3) variable length, variable feature scoring method and fixed type; and (4) variable length, variable feature scoring method and variable type (bottom part of the table). As a baseline, we use fixed length, fixed feature scoring methods (RAND, $\alpha$, $\alpha\nabla\alpha$, $\mathbf{x}\nabla\mathbf{x}$ and IG) and type (top-part of the table). For reference, we also present the predictive performance of $\mathcal{M}$ using the full input text.[6]

Results demonstrate that using a fixed feature scoring method can be sub-optimal, as performance is incosistent across rationale types and datasets. For example $\alpha$ produces lower F1 scores in SST with TOPK, but is outperformed by $\mathbf{x}\nabla\mathbf{x}$ and CONTIGUOUS rationales in M.RC. This shows that there is no single best feature scoring method across datasets and instances (Jacovi and Goldberg, 2020; Atanasova et al., 2020) and strengthens our assumption that selecting the best feature scoring method per instance can improve faithfulness. Varying the feature scoring method (VAR-FEAT) and keeping the rationale length and type fixed, consistently yields lower F1 scores compared to our baseline (F1 score of 55.5 on average compared to 61.6 with $\alpha\nabla\alpha$). This highlights the effectiveness of our approach in selecting a better feature scoring method for each instance in a dataset.

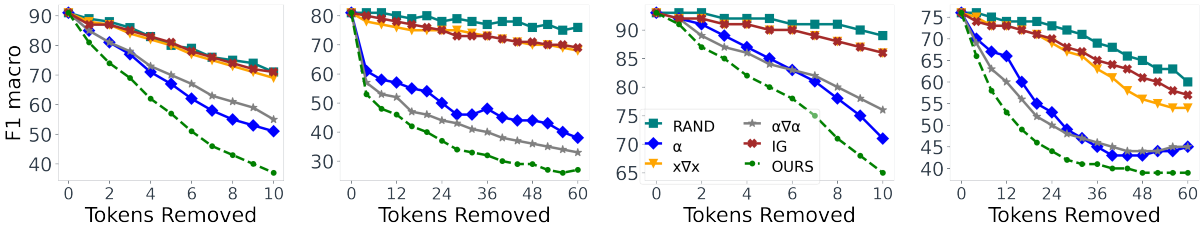
Varying the rationale length (VAR-LEN) and keeping fixed the feature scoring method and type across instance in a dataset improves faithfulness in some cases, but overall results are comparable. For example in SST with $\alpha$ varying the rationale length results in F1 scores of 71.1 compared to 69.7 with the fixed length. In short sequence datasets (SST, AG) this happens because the extracted rationales from our proposed approach are on average only a token shorter (approx. 4% shorter) than the fixed length rationales.[7] We argue that where performances are comparable, having more concise rationales helps explain better a model's prediction, by omitting unnecessary information (see §7).

In datasets where sequence lengths are longer (EV.INF. , M.RC) we record higher drops in F1 performance, despite of our rationales being shorter by 3-4% (approx. 14 tokens) on EV.INF. and 7-12% (approx. 30 tokens) with M.RC. For example in EV.INF. with $\alpha\nabla\alpha$ and TOPK rationale type, our shorter variable length rationales result in a drop of 2.4 points (F1 of 32.6 compared to 36.0). Results demonstrate that having longer rationales can sometimes be detrimental to explanation faithfulness, as they can contain context that is not supportive of a model's prediction (see §7).

Using both VAR-LEN and VAR-FEAT, results in consistently lower F1 macro scores compared to our baseline (54.4 on average compared to 61.6). Comparing though against fixed length rationales with VAR-FEAT, we can assume that it is not as

[6]For brevity, we present results using JSD to compute the divergence $\delta$ between the predictive distributions using the full input and the input with the masked rationale (see §3). We obtain similar performance using other divergence metrics (e.g. KL). For a full stack of results, see Appx. E.

[7]We include the computed rationale lengths with our approach in Appx. C.




(a) SST (b) EV.INF (c) AG (d) M.RC

Figure 1: F1 macro performance with removal of important tokens in decreasing order, for different feature scoring methods and OURS denoting that we select rationales from the best feature scoring method at each instance in the dataset.

effective in SST and AG. This is justified by the performances of the variable length rationales with a fixed feature scoring method in these datasets, which are comparable to fixed length rationales due to the similar number of tokens that comprise them as discussed previously.

We also observe that TOPK rationales result in lower F1 macro scores compared to CONTIGUOUS. We hypothesize that the model attends in most instances to sparse segments of the input rather than contiguous ones when making a prediction. Accounting for rationale type (VAR-LEN + VAR-FEAT + VAR-TYPE), we obtain more faithful rationales across all datasets (consistent drops in F1 performance). In SST for example we obtain F1 scores of 59.9 compared to 69.7, when comparing against fixed length rationales with TOPK and $\alpha$. In M.RC we hypothesize that F1 macro drops are lower, 36.2 compared to 36.8 with TOPK and $\alpha$, due to the baseline already providing competitive performance. Results confirm our hypothesis that flexible rationale extraction per instance is better than a one-size-fits-all fixed approach.

## 6 Quantitative Analysis

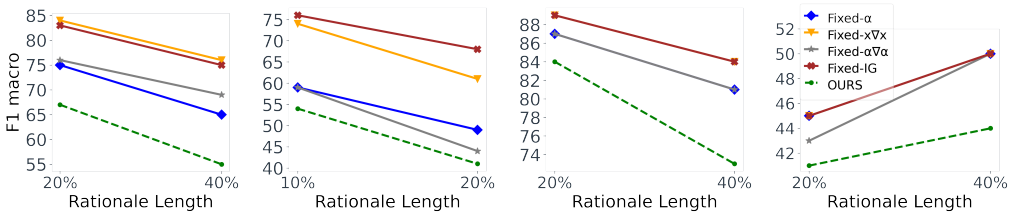
**Computational Efficiency:** Our approach of extracting variable-length rationales can impose a computational overhead, when computing $\delta$ at each token until we reach $N$. This entails $N$ forward passes, for each feature scoring method $\Omega$ (being similar to evaluating feature scoring methods by counting decision flips). We can improve efficiencies by simply using larger increments (more than one token) when calculating $\delta$ (Atanasova et al., 2020). For example we repeated experiments using 2% increments (such that we calculate $\delta$ every {2%, 4%, .. , $N$} of the sequence), resulting in F1 macro average performance of 54.6 with VAR-LEN and VAR-FEAT compared to 54.4 when computing $\delta$ at every token. A 2% increment in datasets such as M.RC, can reduce the number of forward passes and time by a factor of 6. We include results with a more thorough analysis in Appx. F.

**Variable Feature Scoring Method at Different Rationale Lengths:** In Figure 1 we use Word Relevance (Arras et al., 2017) to examine the effectiveness of our approach in selecting the best fixed-length rationale for each instance in a dataset, from a list of feature scoring methods with increasing rationale lengths. For our approach to be successful in extracting faithful rationales, we expect lower F1 macro performance than any single feature scoring method irrespective of the rationale length.

Results suggest that our approach successfully selects the most faithful rationale for each instance in the dataset, irrespective of the rationale length. For example in EV.INF., with just 30 tokens removed the best performing feature scoring method ($\alpha\nabla\alpha$) results in F1 macro scores of approximately 45% compared to 40% by selecting the best feature scoring method at each instance.

**Rationale Length and Faithfulness:** Similar to Jain et al. (2020), we hypothesize that the information a rationale holds and its length correlate. Figure 2 shows F1 macro scores for; (1) our baseline of fixed length (FIXED-LEN) rationales from fixed feature scoring methods (FIXED-FEAT) and (2) our variable length (VAR-LEN) rationales from the best feature scoring method at each instance (VAR-FEAT). For our approach to be successful, we expect lower F1 macro scores with a masked rationale (more faithful) compared to the baseline, irrespective of the upper bound $N$.[8]

[8]We also conduct this experiment with average information difference (Robnik-Šikonja and Kononenko, 2008) as an alternative to F1, observing similar outcomes. We include results in Appx. G.3




(a) SST (b) EV.INF (c) AG (d) M.RC

Figure 2: F1 macro scores for our proposed variable-length rationales from the best feature scoring method (**OURS**, VAR-LEN + VAR-FEAT) and our baseline of fixed length (FIXED-LEN), fixed feature scoring method (FIXED-FEAT) across different rationale ratios (for CONTIGUOUS rationale types). For our approach the upper bound $N$ is the same as the fixed rationale length $K$.

We first notice that our VAR-LEN + VAR-FEAT rationales are more faithful (lower F1 macro scores) compared to our baseline, across all datasets and rationale lengths. This suggests that our approach is successful in extracting the best variable-length rationales for each instance in the dataset, irrespective of the selected upper bound. What is particularly surprising is that with M.RC there is a decrease in information difference with longer rationale lengths, which is counter-intuitive. We hypothesize this is due to the rationales containing tokens that are not supportive of the predicted classes and as such reducing performance. We would expect that by increasing further the rationale length, that we would notice an increase in information difference. Nevertheless, our proposed approach (VAR-LEN + VAR-FEAT) still manages to yield rationales that result in larger drops in F1 scores and are as such more faithful.

**Increasing the Number of Feature Scoring Methods:** We now examine how the effectiveness of our approach for finding the best rationales at each instance, is affected by the number of feature scoring methods. Figure 3 shows F1 macro performance of our approach (VAR-LEN + VAR-FEAT) with increasing numbers of feature scoring methods. We expect that for our approach to be successful, F1 macro performance should degrade when masking the rationale with an increasing set of feature scoring methods (increased options for importance rankings per instance).

As expected by increasing our ranking choices, we are able to retrieve more faithful rationales per instance. This is reflected on the decreases across all datasets in F1 macro performance with increasing options of feature scoring methods. For example with AG, performance drops from 76.8 with 2 feature scoring methods to 71.5 with 4. This

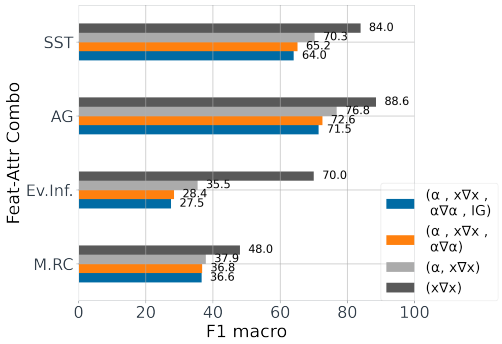



Figure 3: F1 macro performance of variable length, variable feature scoring method rationales (VAR-LEN + VAR-FEAT) with increased options of feature scoring methodes per instance, for TOPK type rationales.

highlights: (1) the effectiveness of our approach in selecting the feature scoring method that results to the more faithful rationale and (2) that there is no single best feature scoring method for all instances in a dataset (Atanasova et al., 2020).

## 7 Qualitative Analysis

Table 3 demonstrates examples of rationales extracted with our proposed approach (VAR-LEN + VAR-FEAT) and fixed length rationales from a variable feature scoring method (FIXED-LEN + VAR-FEAT), to perform a qualitative comparison between fixed and variable-length rationales.

**Concise rationales:** Examples 1 and 2 present examples from the datasets AG and EV.INF. respectively, where the model has predicted the correct label. In Example 1, our approach extracts a rationale that is 6 tokens shorter than the fixed-length one. We hypothesize that this is due to the model having sufficient information to predict a class with 6 fewer tokens and as such not requiring any additional information.

This is more evident in Example 2 where the

| | |
|---|---|
| **Example 1** | **Data.:AG Id: test_97** |
| **[FIXED-LEN + VAR-FEAT]:** ... in quarterly profit , raised its **full - year forecast and said it plans to enter the fast - growing Chinese market , sending its shares higher.** ... | |
| **[VAR-LEN + VAR-FEAT (Ours)]:** ... in quarterly profit , raised its full - year forecast and said **it plans to enter the fast - growing Chinese market , sending its shares higher.** ... | |
| **[Predicted Topic** \|\| **Actual Topic]:** Business \|\| Business | |
| **Example 2** | **Data.:EV.INF. Id: 4598102_3** |
| **[FIXED-LEN + VAR-FEAT]:** ... and week 8 ( **P = 0.007 ) compared with the control group . PANSS - negative scores in the aripiprazole group also decreased significantly at week 4 ( P = 0.005 ) and week 8 ( P < 0.001 ) compared** with the control group . ... | |
| **[VAR-LEN + VAR-FEAT (Ours)]:** ... and week 8 ( P = 0.007 ) compared with the control group . PANSS - negative scores in the aripiprazole group also **decreased significantly** at week 4 ( P = 0.005 ) and week 8 ( P < 0.001 ) compared with the control group . ... | |
| **[Intervention** \|\| **Comparator** \|\| **Outcome]:** Adjunctive aripiprazole \|\| No additional treatment \|\| PANSS-negative scores at week 4 and 8 | |
| **[Predicted relationship** \|\| **Actual relationship]:** Decreased significantly \|\| Decreased significantly | |
| **Example 3** | **Data.:EV.INF. Id: 3162205_2** |
| **[FIXED-LEN + VAR-FEAT]:** ... 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 .. | |
| **[VAR-LEN + VAR-FEAT (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 $\beta$ - CPP at 3 and 6 months postoperatively , but there was no difference between the two sides at 12 months . The fusion .. | |
| **[Intervention** \|\| **Comparator** \|\| **Outcome]:** Porous $\beta$-calcium pyrophosphate ($\beta$-CPP) plus autograft \|\| Autograft alone \|\| Lenke scores at 12 months | |
| **[Predicted relationship** \|\| **Actual relationship]:** Increased significantly \|\| No significant difference | |
| **Example 4** | **Data.:SST Id: test_1039** |
| **[FIXED-LEN + VAR-FEAT]:** It 's just **incredibly** dull. | |
| **[VAR-LEN + VAR-FEAT (Ours)]:** It 's just **incredibly dull**. | |
| **[Predicted Label** \|\| **Actual Label]:** Negative \|\| Negative | |

Table 3: True examples of extracted rationales with our proposed approach of variable-length (VAR-LEN) and fixed-length FIXED-FEAT rationales from the best feature scoring method (VAR-FEAT) at instance-level.

variable-length is formed of only 2 tokens, "decreased significantly", a phrase which directly correlates with the task of inferring relationships. Observing other examples from EV.INF., we have noticed that where a reported relationship is specifically stated in the input sequence, our approach generates shorter than average rationales. Since our approach is based on observing the model's behavior, we hypothesize that the model places too much importance on such phrases, often neglecting the surrounding context. Whilst here the model predicted correctly, our variable-length rationales can be more informative when a model does not.

**Error analysis:** Example 3 presents such a case from the dataset EV.INF., where the model has predicted falsely that "Lenke scores at 12 months" increased significantly instead of no significant difference. If we consider the fixed length rationale, we can observe that the correct answer is included in the extracted rationale, however our model predicted wrong. We argue that this restricts our understanding of the model's reasoning behind an incorrect prediction. On the contrary, our variable-length rationale highlights something directly related to its prediction. Albeit the wrong prediction, our approach provides a more concise rationale, making it is easier to interpret the model's reasoning.

**When a fixed 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 variable-length rationale is at 40%. The intuition is that in certain cases a fixed rationale length might not suffice for all instances to explain a prediction. We argue that our proposed approach highlighted something more informative for the task ("incredibly dull"), compared to the fixed length rationale ("incredibly"), due to removing the restriction of a pre-defined fixed length.

## 8 Conclusions

We propose a methodology for extracting variable length rationales from the best feature scoring method and rationale type at instance level. We achieve this by computing the difference in a model's prediction with full text and a reduced input resulting from rankings of feature scoring methods. We demonstrate that without defining a feature scoring method, a rationale type or a rationale length we can obtain consistently more faithful rationales. We show that our approach is flexible by decomposing it and showing consistent improvements in explanation faithfulness quantitatively and qualitatively. We consider applying our approach to other tasks, such as machine translation and summarization, an interesting direction for future work.

## References


Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 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*, pages 159–168.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Jasmijn Bastings and Katja Filippova. 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*, pages 149–155, Online. Association for Computational Linguistics.

Joost Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases.

Hanjie Chen and Yangfeng Ji. 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)*, pages 4236–4251, Online. Association for Computational Linguistics.

Gianna M Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, pages 97–106.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1780–1790, Marseille, France. European Language Resources Association.

Alon Jacovi and Yoav Goldberg. 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*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 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)*, pages 3543–3556.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 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)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3154–3167, Online. Association for Computational Linguistics.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 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)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 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*, pages 681–691.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4768–4777.

Dong Nguyen. 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)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

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

W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Marcos Treviso and André F. T. Martins. 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*, pages 107–118, Online. Association for Computational Linguistics.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *arXiv preprint arXiv:1909.11218*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 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)*, pages 11–20.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

## A Model Hyperparameters

Table 4 presents the hyper-parameters used to train the models across different datasets, along with F1 macro performance on the development set. Models where finetuned across 3 runs for 5 epochs. We implement our models using the Huggingface library (Wolf et al., 2019) and use default parameters of the ADAMW optimiser apart from the learning rates. We use a linear scheduler with 10% of the steps in the first epoch as warmup steps. Experiments are run on a single Nvidia Tesla V100 GPU.

## B Divergence metrics ($\Delta$)

To compute the value $\delta$ for how much $\mathcal{Y}^m$ differs from $\mathcal{Y}$, we consider four divergence measures $\Delta$, also previously used in literature (Robnik-Šikonja and Kononenko, 2008; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), together with a random baseline:

**Random Selection (RAND-SELECT)** : Randomly selecting the best feature scoring method at each instance.

**Kullback Leibler (KL)** : A non-symmetric divergence measure of how a particular distribution divergences from a reference distribution:

$$KL(\mathcal{Y}||\mathcal{Y}^*) = \mathcal{Y}(\log(\mathcal{Y} - \log(\mathcal{Y}^m)) \quad (1)$$

**Jensen-Shannon (JSD)** : A symmetric divergence metric based on the KL divergence of two distributions from their mean:

$$JSD(\mathcal{Y}||\mathcal{Y}^m) = \frac{1}{2}(KL(\mathcal{Y}||\mu) + \frac{1}{2}(KL(\mathcal{Y}^m||\mu)) \quad (2)$$

where $\mu$ is the average distribution of $\mathcal{Y}$ and $\mathcal{Y}^m$.

| DATASET | MODEL | $lr^m$ | $lr^c$ | F1 |
|---|---|---|---|---|
| SST | BERT-BASE | 1E-5 | 1E-4 | $90.7 \pm 0.2$ |
| AG | BERT-BASE | 1E-5 | 1E-4 | $93.3 \pm 0.0$ |
| EV.INF. | SCIBERT | 1E-5 | 1E-4 | $82.5 \pm 0.9$ |
| M.RC | ROBERTA-BASE | 1E-5 | 1E-4 | $76.3 \pm 0.2$ |

Table 4: Model and their hyper-parameters for each dataset, including learning rate for the model ($lr^m$) and the classifier layer ($lr^c$) and F1 macro scores on the development set across three runs.

| (%) | FEAT | SST | AG | EV.INF. | M.RC |
|---|---|---|---|---|---|
| TOPK | $\alpha$ | $16 \pm 6$ | $16 \pm 4$ | $6 \pm 3$ | $9 \pm 7$ |
| | $\mathbf{x}\nabla\mathbf{x}$ | $16 \pm 6$ | $16 \pm 4$ | $7 \pm 3$ | $13 \pm 7$ |
| | $\alpha\nabla\alpha$ | $16 \pm 5$ | $16 \pm 4$ | $6 \pm 3$ | $10 \pm 8$ |
| | IG | $16 \pm 5$ | $16 \pm 4$ | $7 \pm 3$ | $14 \pm 6$ |
| CONT | $\alpha$ | $15 \pm 6$ | $15 \pm 5$ | $6 \pm 3$ | $8 \pm 6$ |
| | $\mathbf{x}\nabla\mathbf{x}$ | $16 \pm 6$ | $15 \pm 5$ | $6 \pm 3$ | $10 \pm 7$ |
| | $\alpha\nabla\alpha$ | $16 \pm 6$ | $15 \pm 5$ | $6 \pm 3$ | $9 \pm 7$ |
| | IG | $16 \pm 6$ | $15 \pm 5$ | $6 \pm 3$ | $11 \pm 7$ |

Table 5: Average variable rationale lengths computed using JSD, across instances with standard deviation for TOPK and CONTIGUOUS rationale types.[9]

**Perplexity (PERP.)** : A measure of how well a model can predict a sample, where:

$$P(\mathcal{Y}||\mathcal{Y}^m) = \exp^{\mathcal{H}(\mathcal{Y},\mathcal{Y}^m)} \quad (3)$$

where we consider $\mathcal{Y}$ as the ground truth and $H(\mathcal{Y}||\mathcal{Y}^m)$ is the Cross Entropy Loss.

**Class Difference (CLASSDIFF)** : The direct difference between the predicted class probability from the model with full text ($\mathbf{x}$) and the same class probability with reduced text ($\mathbf{x}_{\setminus\mathcal{R}}$) :

$$P(\mathcal{Y}||\mathcal{Y}^m) = \mathcal{Y}[class] - \mathcal{Y}^m[class] \quad (4)$$

where $class = \arg\max(\mathcal{Y})$.

## C Computed variable lengths

In Table 5 we present the average computed rationale lengths using JSD across each feature scoring method, for each dataset and rationale type. To extract rationales we use an upper bound $N$ equal to the fixed rationale length $K$ (as indicated in Table 2 and as defined by (Jain et al., 2020)).

We first observe that rationales extracted with our proposed approach are on average shorter than the a priori set rationale ratio. In datasets such as SST and AG where we have short sequences on average (See Table 1), these differences are negligible, as a 4% decrease translates to approximately having a single token less in the rationale. Observing the standard deviations we also see that there are instances which exhaust the upper bound (20%) to form a rationale and also instances which require

[9] In SST the standard deviations exceed 20% due to a number of instances being shorter than 4 tokens. As such removing a single token is recorded as higher than 20%. Similarly the true recorded fixed ratio is actually higher than 20%

far less than the predefined ratio. This strengthens our initial hypothesis that certain instances might not require as many tokens to successfully explain a prediction.

In datasets with longer sequences lengths on average (EV.INF. and M.RC) such differences are more evident. For example with $\alpha$ and CONTIGUOUS rationales in M.RC, our proposed approach results in rationales which are on average 12% shorter than the fixed length rationales. This translates to approximately 37 less tokens to form a rationale. What is particularly interesting is that by observing the standard deviations for M.RC and EV.INF., we notice that the vast majority of instances does not exhaust the upper bound $N$ to form a rationale. We consider this particularly important for longer sequences, as often when acquiring an explanation for a model's prediction it is desirable to avoid a noisy interpretation of why a model predicted a particular class.

## D Alternative formulation of variable-length rationales

We also examined introducing a thresholding approach in early experimentation when extracting variable length rationales, whereby $\delta_{prev} - \delta_{max} < thresh$, similar to early stopping to avoid exhausting the upper bound $N$. Early results suggested that for datasets with shorter length sequences (SST and AG) this approach performed worse albeit comparably. In datasets with longer sequence lengths (EV.INF. and M.RC), the threshold approach performed poorly. On a closer inspection this was attributed to finding a threshold too early in the sequence thus not capturing all the necessary information. We experimented with JSD and the following thresholds : {1e-4, 1e-3, 1e-2}.

We considered using patience to avoid such naive thresholding, however the computation time would increase significantly. The reason being that we conduct experiments at instance-level, in batches. Introducing patience, would entail that we conduct it using a single instance at a time thus increasing computations by the batch size. As such we have not conducted this approach as it would be computationally expensive.

## E Comparing divergence metrics

We first compare the effectiveness of divergence metrics in computing a variable rationale length and selecting the best feature scoring method at instance level. Table 7 presents F1 macro scores for our proposed variable length rationales from the best feature scoring method at instance level (VARDIVMETRICS). For completion we also include a baseline, whereby we randomly select a feature scoring method as best at each instance (RANDSELECT).

Results suggest that the best performing divergence metrics (used to compute $\delta$) are JSD and KL with an average F1 macro of 55.2. JSD appears to be more consistent, as it manages to outperform the other divergence metrics in 5 out 8 instances, whilst being second in 2. As expected the random selection baseline, fails to perform comparably with any of the divergence metrics. As such, we presented results in the main body of our work with $\delta$ calculated using JSD for clarity, with the full stack of results in Appendix G.

## F Reducing Time Complexities

Selecting the best variable length rationale at each instance in a dataset can be computationally expensive when we compute $\delta$ for every token, being similar to counting decision flips (Nguyen, 2018; Serrano and Smith, 2019; Atanasova et al., 2020). This takes into consideration that we have to perform a forward pass for every token until we reach $N$ tokens, for each feature attribution approach $\Omega$. In the following segments we describe approaches to reduce computational times.

**Reducing search granularity:** Similar to Atanasova et al. (2020), we can reduce significantly computation times by reducing the granularity of our search. In our implementation (see Algo. 1) we describe masking each token or n-gram sequentially, which can be altered to skip tokens. For example, consider a sequence with 200 tokens and an upper-bound $N = 20\%$ and as such $N_t = 40$. Instead of computing $\delta$ for each token, we can compute it for every 5 tokens and as such reducing complexity by 5. Similarly we can compute $\delta$ at every 2% of the sequence until we reach $N$. For example for EV.INF., where $N$ =10% we compute $\delta$ at every {2%, 4%, ... 10%} thus keeping the forward passes constant across instances.

In Table 6, we present average F1 macro performance across datasets and rationale types, when computing $\delta$ at every 2% and 5% of tokens in a sequence using JSD. We compare against: (1) our best performing baseline FIXED-$\alpha\nabla\alpha$ (top of the

| LEN | FEAT SCORING | $\delta$@ | AVG. F1 |
|---|---|---|---|
| FIXED | FIXED-$\alpha\nabla\alpha$ | - | 61.6 |
| OURS | | | |
| FIXED | VAR-FEAT | - | 55.5 |
| VAR | VAR-FEAT | TOKEN | 54.4 |
| | | 2% | 54.6 |
| | | 5% | 55.0 |

Table 6: Average F1 macro scores (across datasets and rationale types) on the test set, with masked rationale ($\mathbf{x}_{\setminus\mathcal{R}}$) as model $\mathcal{M}$ input. $\delta$@ shows the increments at which we calculate $\delta$ (using JSD) when extracting a variable length rationale at each instance.

table) ; (2) fixed length rationales (FIXED-LEN) from the best feature scoring method at each instance (VAR-FEAT) and (3) our proposed approach of variable length (VAR-LEN) rationales from the best feature scoring method (VAR-FEAT with $\delta$ computed at each token ($\delta$@TOKEN), every 2% ($\delta$@2%) and every 5% ($\delta$@5%) until we reach $N$.

Results suggest that by computing $\delta$ at each token yields the most faithful rationales, with the incremental 2% approach resulting in similar performance. This is encouraging considering the computational time saved by calculating $\delta$ incrementaly. For example in M.RC, computing $\delta$ at each token until we reach $N$ requires on average approximately 60 forward passes (when $N$=20%), compared to 10 with the 2% increment.

**Combining feature scoring rankings:** We considered further reducing our computation time by merging importance scores from all feature scoring methods. The intuition is that we obtain a combined ranking and avoid selecting the best feature scoring method at each instance and computing a variable length for all feature scoring methods. We attempted this by averaging the normalized importance scores for each sequence from all the feature scoring methods, however as expected results where not comparable (63.2 average F1 macro compared to 54.4) with our proposed approach or even our best performing baseline.

# G Additional Results

## G.1 F1 model-macro

In Table 8 we also examine a variation of F1 macro, F1 model-macro, whereby the true labels are model's $\mathcal{M}$ predictions with full text ($\mathbf{x}$). The intuition is that a lower F1 model-macro score indicates that a rationale is more faithful as the model struggles to predict with a masked input $\mathbf{x}_{\setminus\mathcal{R}}$, what it had predicted with full text.

We again observe similar outcomes to Table 7 with F1 macro results. JSD performs better than the remainder of the metrics, with KL and CLASSDIFF being the closest competitors to JSD.

## G.2 F1 macro Full Results

In Table 7 we present results across the tested divergence metrics. As expected the baseline of random selection (RAND-SELECT) performs poorly, with the rest of divergence metrics performing comparably. JSD is better on average and performs the best in the majority of cases. In Tables 9, 10 and 11 we present the detailed results for the other divergence metrics under study.

## G.3 Average Information Difference

We also conduct experiments for comparing our baseline fixed-length rationales with a fixed feature attribution, with our variable-length rationales from the best feature scoring methods using Average Information Difference (Av.I.D). Av.I.D. This metric measures the information loss between the predicted class probability using the full input ($\mathbf{x}$) and the input with the rationale masked ($\mathbf{x}_{\setminus\mathcal{R}}$) (Robnik-Šikonja and Kononenko, 2008):

$$\text{Av.I.D} = \frac{1}{D}\sum^{D} \log_2 p(y|\mathbf{x}) - \log_2 p(y|\mathbf{x}_{\setminus\mathcal{R}}) \quad (5)$$

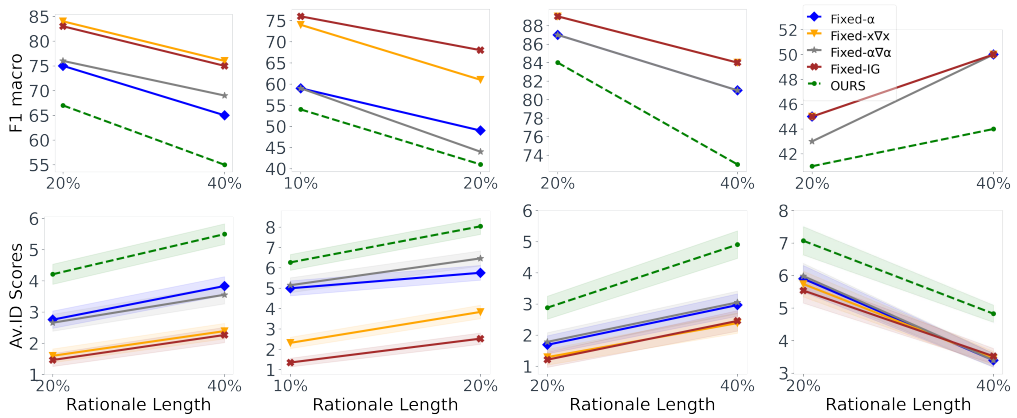
where, $D$ is the number of documents in a dataset, $p(y|\mathbf{x})$ the predicted class probability using the full input and $p(y|\mathbf{x}_{\setminus\mathcal{R}})$ the predicted class probability by masking the rationale. The intuition is that by masking the rationale in an instance (i.e. a model's explanation) we should observe high information loss. Therefore, the higher the Av.ID, the more faithful a rationale. We present results in Figure 4, observing similar performance between F1 macro and Av.I.D.

| | DivMetrics | SST | | AG | | Ev.Inf | | M.RC | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| K | N (Upper-bound) | 20% | | 20% | | 10% | | 20% | | |
| | Type | TopK | Cont | TopK | Cont | TopK | Cont | TopK | Cont | |
| | Rand-Select | 85.9 | 86.9 | 91.5 | 91.5 | 78.6 | 80.6 | 66.3 | 52.5 | 79.2 |
| | KL | **64.0** | **67.5** | 72.0 | 83.2 | **26.3** | 45.0 | 36.7 | **40.4** | **54.4** |
| Var ($K \leq$N) | JSD | **64.0** | 67.7 | **71.5** | **83.0** | 27.5 | **44.5** | **36.6** | 40.5 | **54.4** |
| | Perp. | 64.6 | 68.1 | 72.3 | 83.6 | 26.9 | 44.7 | 37.3 | 41.2 | 54.8 |
| | ClassDiff. | 64.5 | 68.0 | 72.1 | 83.3 | 26.4 | 45.0 | 36.7 | 40.5 | 54.6 |

Table 7: F1 macro scores on the test set, with masked rationale ($\mathbf{x}_{\setminus\mathcal{R}}$) as model $\mathcal{M}$ input. $K$ represents the rationale length, with Var denoting that our approach computes at each instance a rationale length. DivMetrics indicates the divergence metric ($\Delta$) used compute $\delta$, which from we select the best feature scoring method at instance level (see §3). We also include random selection (Rand-Select) as a baseline. **Bold** and underlined values indicate the best and second best respectively, across each dataset and rationale type (lower is better).

| | DivMetrics | SST | | AG | | Ev.Inf | | M.RC | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| K | N (Upper-bound) | 20% | | 20% | | 10% | | 20% | | |
| | Type | TopK | Cont | TopK | Cont | TopK | Cont | TopK | Cont | |
| | Rand-Select | 82.1 | 90.3 | 96.7 | 96.3 | 94.2 | 95.4 | 79.1 | 59.3 | 86.7 |
| | KL | **51.1** | 64.3 | 70.9 | 84.0 | 17.2 | 39.0 | 19.8 | **27.7** | 46.8 |
| Var (K$\leq$N) | JSD | 51.2 | **64.2** | **70.5** | **83.8** | 17.0 | **38.8** | **19.7** | **27.7** | **46.6** |
| | Perp. | 51.7 | 64.6 | 71.4 | 84.5 | 17.5 | **38.8** | 23.2 | 30.8 | 47.8 |
| | ClassDiff. | 51.4 | 64.7 | 71.0 | 84.1 | **16.8** | **38.8** | **19.7** | **27.7** | 46.8 |

Table 8: F1 model-macro scores on the test set (F1 model-macro with full text is 100%), with masked rationale ($\mathbf{x}_{\setminus\mathcal{R}}$) as model $\mathcal{M}$ input. $K$ represents the rationale length, with Var denoting that our approach computes at each instance a rationale length. DivMetrics indicates the divergence metric ($\Delta$) used compute $\delta$, which from we select the best feature scoring method at instance level (see §3). We also include random selection (Rand-Select) as a baseline. **Bold** and underlined values indicate the best and second best respectively, across each dataset and rationale type (lower is better).




(a) SST (b) Ev.Inf (c) AG (d) M.RC

Figure 4: F1 macro and Av.ID scores for our proposed variable-length rationales from the best feature scoring method (**OURS**, Var-Len + Var-Feat) and our baseline of fixed length (Fixed-Len), fixed feature scoring method (Fixed-Feat) across different rationale ratios (for Contiguous rationale types). For our approach the upper bound $N$ is the same as the fixed rationale length $K$.

| | | SST | | AG | | Ev.Inf | | M.RC | | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| Full Input | | 90.7 | | 92.7 | | 80.6 | | 75.9 | | 85.0 |
| | N | 20% | | 20% | | 10% | | 20% | | |
| Len | Feat Scoring | TopK | Cont | TopK | Cont | TopK | Cont | TopK | Cont | |
| Fixed (K=N) | Fixed-Rand | 84.3 | 84.8 | 90.6 | 90.6 | 76.3 | 78.4 | 60.0 | 46.0 | 76.2 |
| | Fixed-$\alpha$ | 69.7 | 74.5 | 77.3 | 87.4 | 46.0 | 59.0 | 36.8 | 45.3 | 62.0 |
| | Fixed-$\alpha\nabla\alpha$ | 72.1 | 76.1 | 79.9 | 87.3 | 36.0 | 59.3 | 38.6 | 43.3 | 61.6 |
| | Fixed-$\mathbf{x}\nabla\mathbf{x}$ | 83.5 | 83.7 | 88.5 | 88.7 | 71.7 | 74.4 | 46.6 | 45.1 | 72.8 |
| | Fixed-IG | 83.5 | 82.9 | 88.0 | 89.4 | 72.0 | 75.5 | 51.0 | 44.8 | 73.4 |
| Ours | | | | | | | | | | |
| Fixed | Var-Feat | **61.9** | **67.4** | **71.0** | 83.5 | 28.2 | 53.7 | 37.4 | 40.7 | 55.5 |
| Var (K≤N) | Fixed-Rand | 85.9 | 86.8 | 91.3 | 91.3 | 78.8 | 80.3 | 66.4 | 52.5 | 79.2 |
| | Fixed-$\alpha$ | 71.0 | 74.9 | 78.0 | 87.4 | 36.3 | 50.0 | 38.5 | 43.1 | 59.9 |
| | Fixed-$\mathbf{x}\nabla\mathbf{x}$ | 83.9 | 83.5 | 88.6 | 88.9 | 70.2 | 66.9 | 48.0 | 45.2 | 71.9 |
| | Fixed-$\alpha\nabla\alpha$ | 72.5 | 75.5 | 80.1 | 86.7 | 32.3 | 49.1 | 39.7 | 43.7 | 60.0 |
| | Fixed-IG | 83.9 | 83.2 | 88.2 | 89.6 | 71.3 | 73.7 | 51.0 | 47.1 | 73.5 |
| Var (K≤N) | Var-Feat | 64.0 | 67.5 | 72.0 | **83.2** | **26.3** | **45.0** | **36.7** | **40.4** | **54.4** |

Table 9: Macro F1 scores for measuring the faithfulness of explanations by masking the rationale ($\mathbf{x}_{\setminus\mathcal{R}}$) (lower is better). $K$ and $N$ denote the rationale length per instance and its upper bound respectively. Fixed and Var indicate fixed and variable (i.e. different per instance) length, feature scoring method or type. $\delta$ measured using KL.

| | | SST | | AG | | Ev.Inf | | M.RC | | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| Full Input | | 90.7 | | 92.7 | | 80.6 | | 75.9 | | 85.0 |
| | N | 20% | | 20% | | 10% | | 20% | | |
| Len | Feat Scoring | TopK | Cont | TopK | Cont | TopK | Cont | TopK | Cont | |
| Fixed (K=N) | Fixed-Rand | 84.3 | 84.8 | 90.6 | 90.6 | 76.3 | 78.4 | 60.0 | 46.0 | 76.2 |
| | Fixed-$\alpha$ | 69.7 | 74.5 | 77.3 | 87.4 | 46.0 | 59.0 | 36.8 | 45.3 | 62.0 |
| | Fixed-$\alpha\nabla\alpha$ | 72.1 | 76.1 | 79.9 | 87.3 | 36.0 | 59.3 | 38.6 | 43.3 | 61.6 |
| | Fixed-$\mathbf{x}\nabla\mathbf{x}$ | 83.5 | 83.7 | 88.5 | 88.7 | 71.7 | 74.4 | 46.6 | 45.1 | 72.8 |
| | Fixed-IG | 83.5 | 82.9 | 88.0 | 89.4 | 72.0 | 75.5 | 51.0 | 44.8 | 73.4 |
| Ours | | | | | | | | | | |
| Fixed (K=N) | Var-Feat | **61.8** | **67.3** | **71.0** | **83.6** | 28.5 | 53.8 | 37.5 | **40.7** | 55.5 |
| Var (K≤N) | Fixed-Rand | 85.9 | 86.8 | 91.3 | 91.3 | 78.8 | 80.3 | 66.4 | 52.5 | 79.2 |
| | Fixed-$\alpha$ | 72.5 | 76.0 | 78.5 | 87.6 | 38.9 | 51.3 | 43.6 | 45.6 | 61.8 |
| | Fixed-$\mathbf{x}\nabla\mathbf{x}$ | 84.3 | 83.9 | 88.6 | 89.0 | 69.5 | 68.1 | 49.5 | 46.9 | 72.5 |
| | Fixed- $\alpha\nabla\alpha$ | 73.4 | 76.7 | 80.4 | 86.9 | 33.6 | 50.0 | 41.4 | 44.5 | 60.9 |
| | Fixed-IG | 83.9 | 83.2 | 88.3 | 89.6 | 71.5 | 74.4 | 52.5 | 47.9 | 73.9 |
| Var (K≤N) | Var-Feat | 64.6 | 68.1 | 72.3 | **83.6** | **26.9** | **44.7** | **37.3** | 41.2 | **54.8** |

Table 10: Macro F1 scores for measuring the faithfulness of explanations by masking the rationale ($\mathbf{x}_{\setminus\mathcal{R}}$) (lower is better). $K$ and $N$ denote the rationale length per instance and its upper bound respectively. Fixed and Var indicate fixed and variable (i.e. different per instance) length, feature scoring method or type. $\delta$ measured using perplexity (Perp.).

| | | SST | | AG | | EV.INF | | M.RC | | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| FULL INPUT | | 90.7 | | 92.7 | | 80.6 | | 75.9 | | 85.0 |
| | N | 20% | | 20% | | 10% | | 20% | | |
| LEN | FEAT SCORING | TOPK | CONT | TOPK | CONT | TOPK | CONT | TOPK | CONT | |
| FIXED (K=N) | FIXED-RAND | 84.3 | 84.8 | 90.6 | 90.6 | 76.3 | 78.4 | 60.0 | 46.0 | 76.2 |
| | FIXED-$\alpha$ | 69.7 | 74.5 | 77.3 | 87.4 | 46.0 | 59.0 | 36.8 | 45.3 | 62.0 |
| | FIXED-$\alpha\nabla\alpha$ | 72.1 | 76.1 | 79.9 | 87.3 | 36.0 | 59.3 | 38.6 | 43.3 | 61.6 |
| | FIXED-$\mathbf{x}\nabla\mathbf{x}$ | 83.5 | 83.7 | 88.5 | 88.7 | 71.7 | 74.4 | 46.6 | 45.1 | 72.8 |
| | FIXED-IG | 83.5 | 82.9 | 88.0 | 89.4 | 72.0 | 75.5 | 51.0 | 44.8 | 73.4 |
| OURS | | | | | | | | | | |
| FIXED (K=N) | VAR-FEAT | **61.8** | **67.3** | **71.0** | 83.6 | 28.3 | 53.8 | 37.5 | 40.7 | 55.5 |
| VAR (K$\leq$N) | FIXED-RAND | 85.9 | 86.8 | 91.3 | 91.3 | 78.8 | 80.3 | 66.4 | 52.5 | 79.2 |
| | FIXED-$\alpha$ | 71.6 | 75.0 | 78.0 | 87.4 | 36.3 | 50.1 | 38.4 | 43.1 | 60.0 |
| | FIXED-$\mathbf{x}\nabla\mathbf{x}$ | 84.2 | 83.6 | 88.7 | 89.0 | 70.2 | 66.9 | 48.1 | 45.4 | 72.0 |
| | FIXED-$\alpha\nabla\alpha$ | 72.8 | 75.8 | 80.3 | 86.6 | 32.4 | 49.5 | 39.8 | 43.6 | 60.1 |
| | FIXED-IG | 84.1 | 83.2 | 88.3 | 89.6 | 71.0 | 73.2 | 51.2 | 47.2 | 73.5 |
| VAR (K$\leq$N) | VAR-FEAT | 64.5 | 68.0 | 72.1 | **83.3** | **26.4** | **45.0** | **36.7** | **40.5** | **54.6** |

Table 11: Macro F1 scores for measuring the faithfulness of explanations by masking the rationale ($\mathbf{x}_{\setminus\mathcal{R}}$) (lower is better). $K$ and $N$ denote the rationale length per instance and its upper bound respectively. FIXED and VAR indicate fixed and variable (i.e. different per instance) length, feature scoring method or type. $\delta$ measured using predicted class difference (CLASS-DIFF).