Exploring Practical Gaps in Using Cross Entropy to Implement Maximum Mutual Information Criterion for Rationalization

Wei Liu[◇] **Zhiying Deng**[†*] **Zhongyu Niu**[◇] **Jun Wang**[‡*]
Haozhao Wang[◇*] **Ruixuan Li**[◇*]

[◇]School of Computer Science and Technology, HUST, China
{idc_lw, zy_niu, hz_wang, rxli}@hust.edu.cn
[†]Faculty of Artificial Intelligence in Education, Central China Normal University, China
zhiyingdzy@gmail.com
[‡]iWudao Tech, China
jwang@iwudao.tech

Abstract

Rationalization is a framework that aims to build self-explanatory NLP models by extracting a subset of human-intelligible pieces of their inputting texts. It involves a cooperative game where a selector selects the most human-intelligible parts of the input as the rationale, followed by a predictor that makes predictions based on these selected rationales. Existing literature uses the cross-entropy between the model's predictions and the ground-truth labels to measure the informativeness of the selected rationales, guiding the selector to choose better ones. In this study, we first theoretically analyze the objective of rationalization by decomposing it into two parts: the model-agnostic informativeness of the rationale candidates and the predictor's degree of fit. We then provide various empirical evidence to support that, under this framework, the selector tends to sample from a limited small region, causing the predictor to overfit these localized areas. This results in a significant mismatch between the cross-entropy objective and the informativeness of the rationale candidates, leading to suboptimal solutions. To address this issue, we propose a simple yet effective method that introduces random vicinal[1] perturbations to the selected rationale candidates. This approach broadens the predictor's assessment to a vicinity around the selected rationale candidate. Compared to recent competitive methods, our method significantly improves rationale quality (by up to 6.6%) across six widely used classification datasets.

1 Introduction

With the success of deep learning, there are growing concerns over interpretability (Lipton, 2018). Ideally, the explanation should be both faithful (reflecting the model's actual behavior) and plausible (aligning with human understanding) (Jacovi and Goldberg, 2020; Chan et al., 2022). Post-hoc explanations, which are trained separately from the prediction process, may not faithfully represent an agent's decision, despite appearing plausible (Lipton, 2018). In contrast to post-hoc methods, ante-hoc (or self-explaining) techniques typically offer increased transparency (Lipton, 2018) and faithfulness (Yu et al., 2021), as the prediction is made based on the explanation itself. There is a stream of research that has exposed the unreliability of post-hoc explanations and called for self-explanatory methods (Rudin, 2019; Ghassemi et al., 2021; Ren et al., 2024).

In this study, our primary focus is on investigating a general model-agnostic self-explaining framework called Rationalizing Neural Predictions (RNP, also known as rationalization) (Lei et al., 2016). RNP utilizes a cooperative game involving a selector and a predictor. This game is designed with a focus on ''data-centric'' (i.e., it is to explain the connection between a text and the model-agnostic task label, rather than explaining the output of a specific model) feature importance: The selector first identifies the most informative part of the input, termed the rationale (in practice, rationale selection is achieved through masking unuseful tokens). Subsequently, the rationale is transmitted to the predictor to make predictions, as illustrated in Figure 1. The

[*] The corresponding authors.

[1] The term ''vicinal'' is borrowed from vicinal risk minimization (Chapelle et al., 2000); ''vicinal'' means *neighboring or adjacent*.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 577–594, 2025. https://doi.org/10.1162/tacl_a_00758
Action Editor: Richard Sproat. Submission batch: 12/2024; Revision batch: 2/2025; Published 6/2025.
© 2025 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

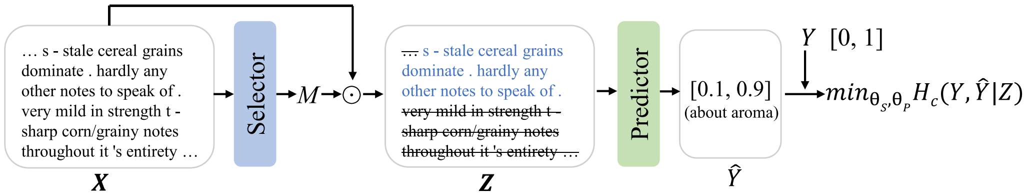

… s - stale cereal grains
dominate . hardly any
other notes to speak of .
very mild in strength t -
sharp corn/grainy notes
throughout it 's entirety …
X
Selector
M
⊙
~~…~~ s - stale cereal grains
dominate . hardly any
other notes to speak of .
~~very mild in strength t -~~
~~sharp corn/grainy notes~~
~~throughout it 's entirety …~~
Z
Predictor
[0.1, 0.9]
(about aroma)
$\hat{Y}$
Y [0, 1]
$min_{\theta_S,\theta_P} H_c(Y,\hat{Y}|Z)$

Figure 1: The standard rationalization framework RNP. The task is binary sentiment classification. $X, Z, \hat{Y}, Y$ represent the input, the selected rationale candidate, the prediction and the ground truth label; M is a sequence of binary masks. θ_S, θ_P are the parameters of the selector and the predictor; and H_c denotes the cross-entropy.

selector and predictor are trained cooperatively to maximize prediction accuracy. RNP and its variants have been one of the mainstreams to facilitate the interpretability of NLP models (Sha et al., 2021, 2023; Yue et al., 2023; Liu et al., 2023c; Storek et al., 2023; Zhang et al., 2023). And besides its use for interpretability, rationalization can also serve as a method for data cleaning, since the extracted (Z, Y) samples can act as a new dataset. Some recent studies find that a predictor trained with such a dataset can be more robust (Chen et al., 2022) and generalizable (Wu et al., 2022), since task-irrelevant, harmful information has been removed.

Despite its strength, the cooperative game of rationalization is difficult to train if the selector and the predictor are not well coordinated (Yu et al., 2021). In this paper, we begin by analyzing the objective of cooperative rationalization, decomposing the practical assessment of rationale candidates by the selector (i.e., $H_c(Y, \hat{Y}|Z)$) into two parts: the model-agnostic informativeness of the rationale candidates (i.e., $H(Y|Z)$), and the degree to which the predictor fits the conditional distribution $P(Y|Z)$ (i.e., $D_{KL}(P(Y|Z)||P(\hat{Y}|Z))$), as shown in §4.1. Initially, due to random initialization, the selector samples all rationale candidate regions evenly, and all rationale candidates are fairly fitted by the predictor, resulting in similar D_{KL}. Therefore, the selector tends to choose candidates with relatively lower $H(Y|Z)$. We then empirically observe that once the selector identifies some good (but not necessarily optimal) rationale candidates that have relatively low $H(Y|Z)$, it dramatically increases the probability of selecting these candidates and decreases the probability of selecting others, sampling within a very small area (in fact, even a single point, as shown in Empirical observation 1 of §4.2). This results in the predictor's perspective being limited to a small local area, only capable of reasonably fitting a narrow range of rationale candidates. Consequently, the optimal rationale, despite having the lowest $H(Y|Z)$, may be overlooked due to a high $D_{KL}(P(Y|Z)||P(\hat{Y}|Z))$ (please refer to Empirical observation 2 in §4.2).

To mitigate this issue, ensuring that the predictor equally fits any rationale candidates might be a viable solution. However, we also find the negative result that completely random sampling across all possible areas, to allow the predictor to fit all rationale candidates equally, renders the predictor ineffective due to an excessively low signal-to-noise ratio (see §4.3). As a result, we try to develop a compromise solution.

Given the continuity of textual semantics (''continuity'' does not refer to the adjacency of words, but rather to the gradual, smooth change in meaning when a small number of words are altered, rather than a sudden shift), an empirical conjecture is that if the selector identifies suboptimal rationale candidates with relatively low $H(Y|Z)$, then the optimal rationale is likely within its vicinity. Therefore, the predictor only needs to focus on fitting the rationale candidates within this vicinity. Inspired by the philosophy of variational autoencoders (VAE), we construct the vicinity by introducing random perturbations to the selected rationale candidates and use this new vicinity to train the predictor. The architecture of our method is shown in Figure 7, with Figure 8 showing its intuition. Through this approach, we significantly improved the rationale quality extracted by the standard RNP on six datasets from two widely used rationalization benchmarks, surpassing some recently published RNP variants.

Our contributions are summarized as follows: (1) **Problem Identification**: We formally analyze the objective of rationalization and identify the practical gap between the rationales' informativeness and the model's assessment of them, which is the primary contribution of this paper. (2)

Empirical Evidence: We provide various empirical evidence to demonstrate how this gap can affect the quality of extracted rationales. (3) **Simple Solution**: We introduce a simple yet effective method named **v**icinal **e**valuation **r**ationalization (VER) to mitigate this gap. Empirical results on several widely used benchmarks demonstrate the effectiveness of the proposed method. Additionally, our method is straightforward and does not require any additional auxiliary modules, preserving its potential to be combined with future methods.

2 Related Work

2.1 Cooperative Rationale Extraction

Rationalization is a general framework first proposed by Lei et al. (2016). By extracting rationales before making predictions, this framework can ensure the rationale's faithfulness to the model prediction, as the predictor is just trained on the selected rationales (Yu et al., 2021; Chang et al., 2020). Rationalization has been one of the mainstreams to facilitate the interpretability of NLP models (Yue et al., 2023; Storek et al., 2023; Zhang et al., 2023; Liu et al., 2023b, 2024a, 2025). Recently, there has also been some work attempting to extend it to the field of graph learning (Luo et al., 2020) and computer vision (Yuan et al., 2022). Apart from improving interpretability, recent work has also discovered that it can serve as a method of data cleaning, as training a predictor with the extracted rationales has been found to increase robustness (Chen et al., 2022) and generalization (Wu et al., 2022; Gui et al., 2023).

Improving the Training of Rationalization. The rationalization framework involves a cooperative framework between the selector and the predictor, which requires careful optimization to coordinate and is hard to train. Given this challenge, a number of research efforts focus on refining the optimization process to improve the rationalization. Bastings et al. (2019) replaced the Bernoulli sampling distributions with rectified Kumaraswamy distributions. Chang et al. (2019) introduced a adversarial game and produces both positive and negative rationales. Jain et al. (2020) disconnected the training regimes of the selector and predictor networks using saliency threshold. Chang et al. (2020) tried to learn the invariance in rationalization. Liu et al. (2023c) suggested that the selector and the predictor need to be assigned different learning rates. Yu et al. (2019) found that if not properly coordinated, the selector and the predictor may collude to use trivial patterns to make the correct prediction. To address this issue, 3PLAYER (Yu et al., 2019) tries to make the unselected parts incapable of predicting the label, so that informative parts are squeezed into the selected parts; DMR (Huang et al., 2021) tries to match the distribution between the selected rationales and the raw inputs; FR (Liu et al., 2022) shares the encoder between the selector and the predictor to make them regularize each other; MGR (Liu et al., 2023a) reduces the likelihood of collusion by training multiple selectors; NIR (Storek et al., 2023) constructs a large vocabulary based on the dataset and uses manually defined rules to determine which tokens in the rationale candidates selected by the selector are likely to be irrelevant to the prediction, and randomly replaces them with other meaningful words from the vocabulary. Inter_RAT (Yue et al., 2023) reduces training difficulties by blocking shortcuts. CR (Zhang et al., 2023) selects rationales by additionally calculating the necessity and sufficiency of each token. A2I (Liu et al., 2024b) attempts to reduce the influence of model-added spurious correlations through an attack-based instructor.

Our research can also be categorized as a solution to training difficulties, and the distinct contribution is a detailed analysis of a new potential cause of these training difficulties. Our practical method may seem similar to NIR at first glance, but the underlying concept is completely different. NIR stabilizes the training of the predictor by replacing tokens in the selected rationale candidates that are likely to lack specific semantics with tokens from the vocabulary that have rich semantics (which may not be present in the original input X). However, a fundamental flaw of NIR is that if the predictor scores the replaced rationale candidate highly, the feedback received by the selector would be to increase the probability of selecting the original rationale candidate before replacement (which is not necessarily good because the high score may come from the newly added tokens). We will take NIR as a baseline and also include some of the other latest methods.

2.2 Generative Explanation with LLMs

Generative explanation is a research line that is close but orthogonal to our research. With the great success of LLMs, a new research line for explanation is chain-of-thought. By generating (in contrast to selecting) intermediate reasoning steps before inferring the answer, the reasoning steps can be seen as a kind of explanation. The intriguing technique is called chain-of-thought (CoT) reasoning (Wei et al., 2022). However, LLMs sometimes exhibit unpredictable failure modes (Kıcıman et al., 2023) or hallucination reasoning (Ji et al., 2023), making this kind of generative explanation not trustworthy enough in some high-stakes scenarios. Also, some recent research finds that LLMs are not good at extractive tasks (Qin et al., 2023; Li et al., 2023; Ye et al., 2023).

The Potential Impact of Rationalization in the Era of LLMs. Compared to traditional ‘‘model-centric’’ XAI methods which solely focus on the model's learned information, ‘‘data-centric’’ approaches primarily aim to extract model-agnostic patterns inherent in the data. So, apart from improving interpretability, rationalization can serve as a method of data cleaning (Seiler, 2023).

Domain-specific large models often require supervised fine-tuning using domain-specific data. Uncleaned data may contain harmful information such as biases and stereotypes (Sun et al., 2024). Recent research suggests that training predictors with extracted rationales can remove irrelevant harmful information, enhancing robustness (Chen et al., 2022) and generalization (Wu et al., 2022; Gui et al., 2023). Considering that small models are sufficient for simple supervised tasks and are more flexible and cost-effective for training on single datasets (e.g., searching hyperparameters and adding auxiliary regularizers), using small models for rationalization on a single dataset and then using the extracted rationales for supervised fine-tuning might prevent large models from learning harmful information from new data. Additionally, shortening input texts can also reduce the memory required for fine-tuning (Guan et al., 2022). A recent study also finds that training a small model for data selection and producing a small subset is useful for fine-tuning LLMs (Xia et al., 2024).

3 Background of Rationalization

Notations. We consider the text classification task. We use $f_s(\cdot)$ and $f_p(\cdot)$ to represent the selector and the predictor, with θ_s, θ_p being their parameters, respectively. Upper case letters, such as X and Y, represent randoms variables. And lower case letters represent variable values. For the sake of notation brevity, we do not distinguish between vectors and scalars. $X = X_{1:l}$ represents the texts of length l and Y represents the classes in a dataset $\mathcal{D}$. Generally, the dataset $\mathcal{D}$ is considered representing the distribution of $P(Y|X)$ such that it can be approximated by a predictor which produces $P(\hat{Y}|X)$ (usually by minimizing the cross-entropy $H_c(Y, \hat{Y}|X)$):

$$H_c(Y, \hat{Y}|X) = H(Y|X) + D_{KL}(P(Y|X)||P(\hat{Y}|X)), \quad (1)$$

where we use H_c and H to distinguish between cross-entropy and entropy. When we minimize the cross-entropy, we are in fact minimizing $D_{KL}(P(Y|X)||P(\hat{Y}|X)$, and we will finally get $P(Y|X) = P(\hat{Y}|X)$ if $D_{KL}(P(Y|X)||P(\hat{Y}|X)) = 0$. This is just what a normal classification task does.

Then, we introduce the rationalization task. The overall target of the rationalization task is to find the evidence from X that best supports Y (not the model predicted $\hat{Y}$ in ante-hoc explanation). Based on this target, the intuition is that good rationales can be used to achieve high prediction accuracy. As a result, the practical way to identify rationales is to train a selector and a predictor that cooperatively maximize the prediction accuracy, as specified below.

For each $(x, y) \in \mathcal{D}$, the selector first outputs a sequence of binary mask $m = f_s(x) = [m_1, \cdots, m_l] \in \{0, 1\}^l$, where l is the sequence length.[2] Then, it forms the rationale candidate z by the element-wise product:

$$z = m \odot x = [m_1 x_1, \cdots, m_l x_l]. \quad (2)$$

To simplify the notation, we denote $f_s(X)$ by Z in the following sections, i.e., $f_s(X) = Z$ (here $f_s(X)$ represents the whole dataset's rationale set, so we use upper case Z to represent it). With the selector's sampling, we get a set of (Z, Y) samples $\mathcal{D}_z$, which is generally considered to be represent the distribution $P(Y|Z)$ and serve as the training data of the predictor. Then, standard

[2] In practice, the selector first outputs a Bernoulli distribution for each token and the mask for each token is independently sampled using gumbel-softmax. Using reinforcement learning (Covert et al., 2023) to enable gradient propagation is also a viable approach. However, this paper follows the most commonly used gumbel-softmax in the rationalization field for end-to-end training.

rationalization attempts to identify the rationale by minimizing the cross-entropy:

$$\min_{\theta_s,\theta_p} H_c(Y,\hat{Y}|Z),$$
$$s.t.,\ (X,Y)\sim\mathcal{D},\ Z=f_s(X),\ \hat{Y}=f_p(Z). \tag{3}$$

Compactness and Coherence Regularizer. To make the selected rationale human-intelligible, rationalization methods usually constrain the rationales by compact and coherent regularization terms. In this paper, we use the most widely used constraints proposed by Chang et al. (2020):

$$\Omega(m)=\lambda_1\left|\frac{||m||_1}{l}-\alpha\right|+\lambda_2\sum_{t=2}^{l}\big|m_t-m_{t-1}\big|. \tag{4}$$

The first term encourages that the percentage of the tokens being selected as rationales is close to a pre-defined level α. The second term encourages the rationales to be coherent.

4 Motivation

4.1 Linking Equation (3) to Mutual Information and Understanding the Potential Gaps

The theoretical support of Equation (3) for the selector to identify rationales is the maximum mutual information (MMI) criterion:

$$\begin{aligned} Z^* &= \arg\max_Z I(Y;Z)\\ &= \arg\max_Z (H(Y)-H(Y|Z))\\ &= \arg\min_Z H(Y|Z),\\ s.t.,\ & Z=f_s(X),\ (X,Y)\sim\mathcal{D}. \end{aligned} \tag{5}$$

Although theoretically supported by the MMI criterion, (5) is intractable in practice because the real distribution of $P(Y|Z)$ is not directly accessible. In practice, the common way is to use the cross-entropy $H_c(Y,\hat{Y}|Z)$ to approximate the entropy $H(Y|Z)$, and this is what Equation (3) does.

It then leads to two questions: Why can $H_c(Y,\hat{Y}|Z)$ be used to approximate $H(Y|Z)$, and what specific approximations have been assumed? Exploring these two questions will help us gain a more specific and detailed understanding of the effectiveness of Equation (3), going beyond vague intuitions. This can assist us in more accurately identifying potential issues in practice.

We rewrite (3) to decompose it into two parts:

$$H_c(Y,\hat{Y}|Z)=H(Y|Z)+D_{KL}(P(Y|Z)||P(\hat{Y}|Z)), \tag{6}$$

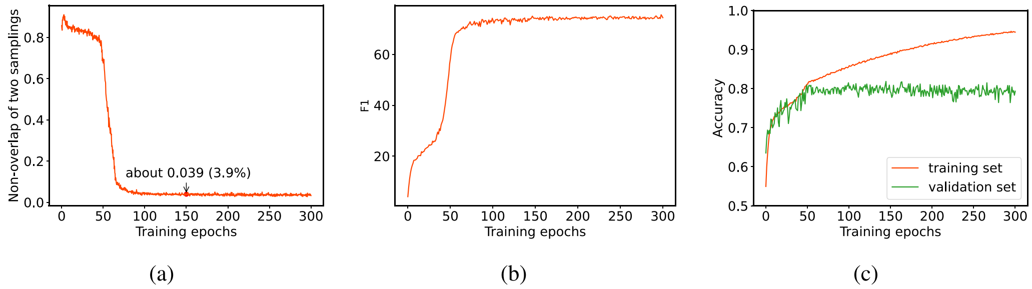
where $D_{KL}(P(Y|Z)||P(\hat{Y}|Z))$ represent the KL-divergence, indicating the error in approximating the real distribution $P(Y|Z)$ with the predictor produced $P(\hat{Y}|Z)$. Equation (6) consists of two parts: the model-agnostic informativeness of Z (i.e., $H(Y|Z)$), and the degree of the predictor fitting $P(Y|Z)$.

4.2 Empirical Observations/Evidence on the Practical Gap

Although the effectiveness of Equation (3) in identifying rationales is theoretically supported with the MMI criterion, the previous analysis in §4.1 tells us there is a gap between theory and practice. This section aims to provide empirical evidence to verify the existence of the theoretical gap.

Specifically, the selector gets the feedback about the importance of the selected rationale candidates through the predictor's output $H_c(Y,\hat{Y}|Z)$, which consists of two components: the model-agnostic informativeness of Z, and how well the predictor assesses Z (i.e., $D_{KL}(P(Y|Z)||P(\hat{Y}|Z))$), as is shown in Equation (6). Thus, the selector will be guided to choose the rationales that have both low $H(Y|Z)$ and low $D_{KL}(P(Y|Z)||P(\hat{Y}|Z))$, while the optimal rationales can only guarantee themselves with low $H(Y|Z)$. In practice, however, the predictor can only learn to assess rationales that have already been selected by the selector. If the selector's selection set is focused on a small local area, the assessment of the predictor on unsampled regions will be biased. The problem is that although the optimal rationales have the lowest $H(Y|Z)$, they are not guaranteed to have low $D_{KL}(P(Y|Z)||P(\hat{Y}|Z))$, resulting in obtaining only suboptimal rationales. In the following, we will show some empirical evidence that the optimal rationales are usually not properly valued and do not have the lowest cross-entropy under the RNP framework.

Empirical Observation 1. (the selector samples rationale candidates in a small area). Although the selector's choice already incorporates some randomness, we find that in practical training, when it identifies some relatively good rationale candidates, it quickly increases the probability



(a)

(b)

(c)

Figure 2: An experiment of the standard RNP on the Beer-Appearance dataset. We select about 20% (close to the sparsity of human-annotated rationales in this dataset) tokens from the raw input to be the rationale. (a): The non-overlap between two rationale candidates sampled by the two selectors that belong to two adjacent training iterations. (b): The rationale quality (overlap between model-selected rationales and human-annotated ones). (c): The prediction (classification) accuracy on the training (orange) and the validation set (green).

of selecting these potentially suboptimal rationale candidates to rapidly improve predictive accuracy. This causes its sampling to concentrate in a very small local area, potentially missing the optimal solution.

Figure 2 shows empirical evidence of it. Consider the current selector as f_{s_1} and it becomes f_{s_2} after a training iteration. We take a same batch of texts x, and the non-overlap between two rationales in Figure 2(a) is calculated as

$$m^{(1)} = f_{s_1}(x),\ m^{(2)} = f_{s_2}(x). \tag{7}$$

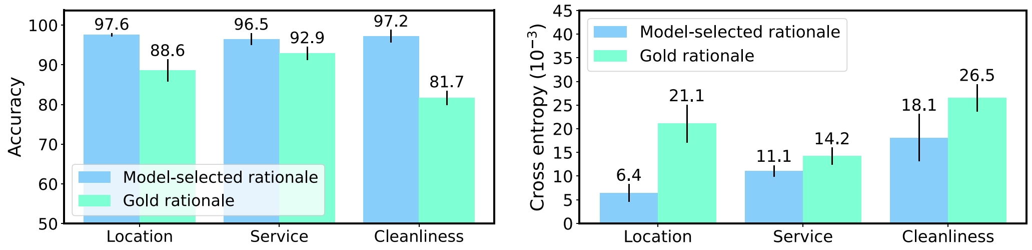
The non-overlap is calculated as $\frac{||m^{(1)}-m^{(2)}||_1}{||m^{(1)}||_1+||m^{(2)}||_1}$. We take the average of it within a mini-batch.

Initially, the selector explores a wide region, but after finding some relatively good rationale candidates, it rapidly increases the probability of selecting these candidates. Figure 2(a) shows that after about 100 training epochs, the difference of selected rationale candidates between two iterations is only about 4%. Considering that the rationale sparsity is about 20% and the maximum sequence length is 256, there are only about $4\% * 20\% * 256 \approx 2$ tokens that are changed between two training iterations (even fewer, as most texts are shorter than 256 tokens, see Table 4), which means that the selector's sampling is focused on an almost ''single point'' and the predictor sees very limited data. This causes the predictor's focus to quickly converge to a certain local area, reducing the likelihood that other regions are reasonably assessed. At this point, the quality of the rationale no longer grows (Figure 2(b)), while the training accuracy continues to rise at a high rate (Figure 2(c)). At this stage, the predictor is merely overfitting to some trivial patterns in this locality, as evidenced by the fact that we observed no increase in accuracy on the validation set (Figure 2(c)).

Empirical Observation 2. (the cross-entropy does not assess the unselected optimal rationale properly). In intuition, the gold human-annotated rationales are most informative and have the lowest $H(Y|Z)$. Thus, ideally (e.g., a predictor that works like a human expert such that the KL-divergence in Equation (6) is equal to 0 for all possible Z), using the human-annotated rationales can best predict Y. However, we empirically observe that sometimes the human-annotated rationales get lower prediction accuracy than the model-selected ones when sent to RNP's trained predictor.

Figure 3(a) shows the accuracy of RNP's trained predictor with model-selected and human-annotated rationales as input on three rationalization datasets. We observe that sometimes the human-annotated gold rationales get much lower accuracy than the model-selected ones. It's clear that gold rationales have lower $H(Y|Z)$. So, the reason of this phenomenon is that the predictor does not fit the gold rationale well and leads to high $D_{KL}(P(Y|Z)||P(\hat{Y}|Z))$ for it. This observation to some extent supports our hypothesis that the predictor does indeed overfit to the points selected by the selector while underfitting the potentially optimal rationales around them, leading to inaccurate assessment of using $H(Y,\hat{Y}|Z)$ to approximate $H(Y|Z)$.

Empirical Observation 3. (the predictor is overfitting to trivial patterns). In Empirical observation 1, we mentioned that after the selector



(a) (b)

Figure 3: Two kinds of (a) accuracy and (b) cross-entropy on three classification datasets from *HotelReviews* benchmark.‘‘model-selected rationale’’: the accuracy/cross-entropy of the prediction made by the predictor with the input being the model-selected rationales; ‘‘gold rationale’’: the accuracy/cross-entropy with the input being human-annotated gold rationales. Detailed setup is in Appendix A.3.

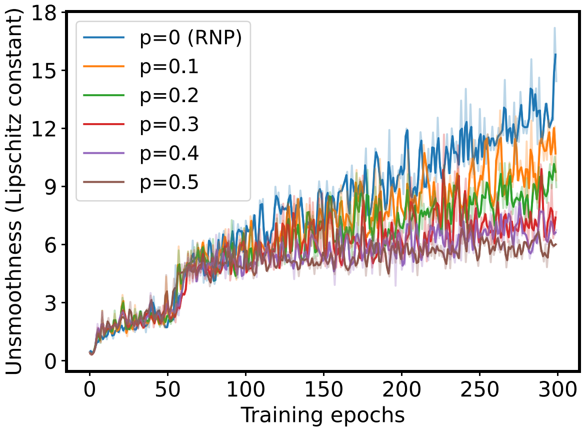


Figure 4: The average unsmoothness of the predictor. The details of the y-axis metric are in Appendix A.4. The dataset is one of the most widely used rationalization dataset Beer-Appearance.

confines its sampling to a very small area, the predictor overfits this small region while neglecting others. Now, we further provide another evidence of overfitting. Another indirect indicator of the extent to which the model overfits trivial patterns is the smoothness of the model (Virmaux and Scaman, 2018; Fazlyab et al., 2019). If a predictor learns many trivial patterns instead of genuinely meaningful semantic features, then its function surfaces typically exhibit non-smooth patterns such as steep steps or spikes, resulting in poor Lipschitz continuity (Liu et al., 2023c; Szegedy et al., 2014; Weng et al., 2018). Typically, Lipschitz continuity is approximated by the value of the variation of the output with respect to the input over the entire dataset (i.e., Lipschitz constant, lower is better). We follow the method used in Liu et al. (2023c) to compute the Lipschitz constant, which is given in Appendix A.4.

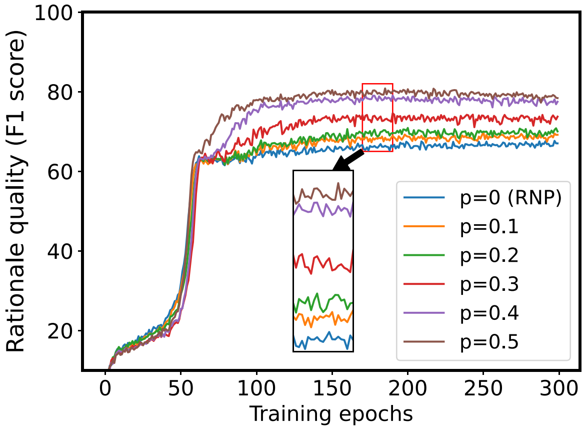


Figure 5: The rationale quality (F1 score) with different perturbation rate. The results are from the same experiments as Figure 4.

Figure 4 shows the unsmoothness of the predictor during training. ‘‘$p = 0$’’ refers to the vanilla RNP, and ‘‘$p \in [0.1, 0.5]$’’ refer to our method (which will be introduced in §4.3). Lower y value means a smoother model function. We see that the Lipschitz constant (unsmoothness) of RNP grows for a long time, while our method with $p = 0.5$ grows very slowly after about 70 epochs. Figure 5 shows the corresponding rationale quality. Figures 4 and 5 are from the same experiments, whose details are in Appendix A.4. Combining Figures 4 and 5, we further observe that the predictor’s overfitting significantly prevents the selector from finding good rationales.

4.3 A Simple Method with Vicinal Perturbation and Assessment

Empirical Observation 4. (negative results of equally assessing all possible rationale candidates). An intuitive way to address the issue

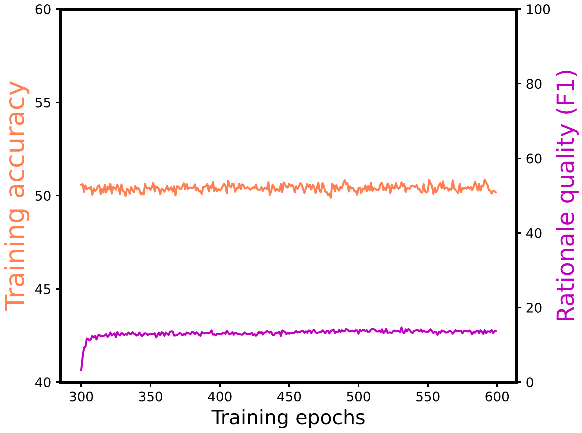


Figure 6: The training accuracy and rationale quality of a RNP model that first trains the predictor with randomly sampled $\mathcal{D}_z$ and then trains the selector to maximize the prediction accuracy.

is to decouple the training data of the predictor from the selector. If the predictor's training data ($\mathcal{D}_z$ as mentioned in §3) is sampled randomly rather than conditioned on the selector, then the rationale candidates in every regions will be able to be equally fitted, mitigating the impact of unbalanced evaluation on Eq. (6). In practice, however, random sampling may lead to a very low signal-to-noise ratio, preventing the predictor from effectively learning the semantics. For example, in a widely used rationalization dataset called Beer-Appearance, the average percentage of the ground-truth rationale is about 18.4%. This means that about 81.6% of the tokens in an input text act as noise, making this method intractable in practice. Figure 6 shows the results of such a method. In the first 300 epochs, we randomly sample a set of rationale candidates (with the sparsity constraint only) to form a $\mathcal{D}_z$ for each epoch to train the predictor. And in the latter 300 epochs, the predictor is fixed and we train the selector in the same way as RNP (please refer to Appendix A.6 for more details). Under this training approach, the predictor is expected to be capable of evaluating rationale candidates effectively in any regions. But Figure 6 does not show the results as we expect. In fact, the model does not work at all, due to the low signal-to-noise ratio received by the predictor.

Based on the issues mentioned above, we try to develop a compromise solution. Considering the continuity of textual semantics, an empirical conjecture is that if the selector identifies suboptimal rationale candidates with low $H(Y|Z)$, then the optimal rationales should be within their vicinity (note that continuity here does not refer to the adjacent positioning of words, but rather to the number of words that need to be changed). This conjecture is based on two considerations. First, in a random state (where the predictor has no bias toward any Z offset), all rationale candidates have similar $D_{KL}(P(Y|Z)||P(\hat{Y}|Z))$ values, but the optimal rationales have lower $H(Y|Z)$. Combining this with Equation (6), theoretically, the optimal rationales are more likely to be favored by the predictor at the beginning of training. Second, empirically, the vanilla RNP can select relatively good rationales in most cases.

Based on this assumption, the predictor only needs to evaluate the vicinity of the selector's selected region. Then, borrowing from the philosophy of VAE, we randomly perturb the selector's output to sample a vicinity. Specifically, we start by randomly flipping the elements in the selector's output m for which $m_i = 1$ with a probability of p (p is a hyperparameter), and then we count the number of flips. Subsequently, we select the same number of elements for which $m_j = 0$ and flip them to 1.

Formally, for the selector's output $m = [m_1, m_2, \cdots, m_l]$, we perturb it to be $m' = [m'_1, m'_2, \cdots, m'_l]$:

$$\begin{aligned} m'_i &= \mathbb{1}(\epsilon_i > p),\ s.t.,\ \epsilon_i \sim U(0,1),\ \forall i,\ m_i = 1, \\ m'_j &= \mathbb{1}(\epsilon_j \in \text{top}_k(\epsilon_J)),\ s.t.,\ \epsilon_j \sim U(0,1), \\ &\forall j,\ m_j = 0, \end{aligned} \tag{8}$$

where $k = ||m_i - m'_i||_1$ and ϵ_J is the collection of all ϵ_j. p is a hyperparameter, indicating the ratio of tokens in a rationale z that are replaced by the tokens in the raw input x but not seen by the predictor. It's worth noting that, similar to VAE, our approach doesn't repeatedly perturb a single data point; rather, full sampling of the vicinity is achieved by individually sampling multiple data points from a mini-batch. The perturbation occurs only during training. This method is somewhat similar to ϵ-*greedy* in reinforcement learning, but is not the same. ϵ-*greedy* involves complete random exploration with a probability of ϵ and the exploration is more akin to the method described as unsuccessful at the beginning of §4.3 (i.e., Empirical observation 4). In contrast, our approach focuses on exploration around local optima.

Figure 7 shows the comparison between the standard rationalization framework RNP and our

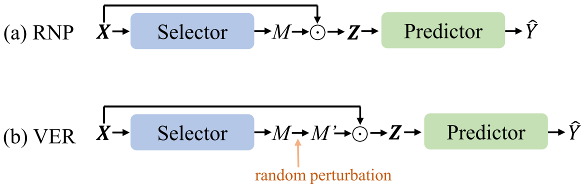


Figure 7: The comparison of the standard rationalization framework RNP and our method VER.

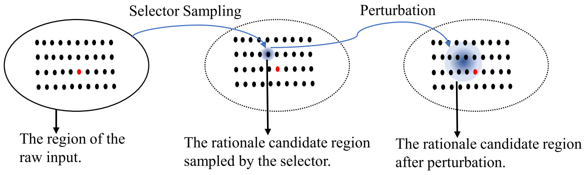

Selector Sampling
Perturbation
The region of the raw input.
The rationale candidate region sampled by the selector.
The rationale candidate region after perturbation.

Figure 8: The intuitive understanding of the perturbation. The left is the raw input X. The black dots represent rationale candidates, among which the red one is the optimal rationale. The blue region is the area that the selector has a high probability to sample.

method. Our approach is very simple and does not involve any adjustments to the model structure, making it scalable to those complex methods that rely on structural modifications. Figure 8 provides a toy example to show the intuition of our method. On the left is the original raw input X. The black dots represent rationale candidates, among which the red one is the optimal rationale. The blue region is the area that the selector has a high probability to sample. The region is quite small because the selector quickly becomes overconfident about certain suboptimal rationale candidates, leading to a concentration of sampling in a very narrow local area (see Empirical observation 1 in §4.2).

5 Experiments

5.1 Setup

Datasets. Although the process of rationale extraction is unsupervised, the rationalization task requires comparing the rationale quality extracted by different models. This necessitates that the test set includes ground-truth rationales, which imposes special requirements on the datasets. Following the conventional setup in the field of rationalization, we employ six text classification datasets from two widely used benchmarks.

The datasets are Beer-Appearance, Beer-Aroma, and Beer-Palate (collected from the BeerAdvocate [McAuley et al., 2012] benchmark); and Hotel-Location, Hotel-Service, and Hotel-Cleanliness (collected from the HotelReviews [Wang et al., 2010] benchmark). Among them, the beer-related datasets are most important and used by nearly all previous research in the field of rationalization. All of these datasets contain human-annotated ground-truth rationales on the test set (but not on the training set), making it convenient to compare different methods' performance fairly. Among them, the three beer-related datasets are most important and used by nearly all of previous research in the field of rationalization. The statistics of the datasets are in Appendix A.1.

To verify the generalizability of our VER, we also perform supplementary experiments on two different tasks. We use the MultiRC dataset for the reading comprehension task, and use the FEVER dataset for the fact extraction and verification task. These two datasets are taken from the ERASER benchmark (DeYoung et al., 2020).

Baselines. We compare with RNP (Lei et al., 2016), DMR (Huang et al., 2021), Inter_RAT (Yue et al., 2023), NIR (Storek et al., 2023), CR (Zhang et al., 2023), and A2I (Liu et al., 2024b). Among them, RNP represents the direct counterpart of our VER, and other methods stand for the latest literature. All of them have been discussed in §2. Aside from these rationalizations methods, we also compare with two popular post-hoc methods: LIME (Ribeiro et al., 2016) and Attention.[3]

Implementation Details. Each of the selector and predictor pairs contains an encoder (e.g., RNN/Transformer) followed by a linear layer. We use two types of encoders: GRUs (with 100-dimensional GloVe as the word embedding, following Inter_RAT, Table 1) and ''bert-base-uncased'' (following CR, Table 2). **In comparison with the baselines, we keep the perturbation rate of our VER as $p = 0.3$ for all the datasets. Then, we further show the influence of hyperparameter p separately** (Figure 9). For DMR, NIR, A2I, and our VER, considering they are all variants of the standard RNP, we first manually tune the hyperparameters for RNP, and then apply the hyperparameters to other methods.

[3] We refer to Lei et al. (2016) to implement the attention-based model. Specifically, the attention-based model calculates the normalized attention vector based on all input tokens, and then average pools the hidden layer states based on the attention weights, and uses the average vector for task prediction. Meanwhile, attention-based model selects top-k% tokens as rationale from the attention weights.

Datasets	Beer-Appearance					Beer-Aroma				
Methods	S	Acc	P	R	F1	S	Acc	P	R	F1
LIME	15.1 (0.0)	N/A	53.1 (0.1)	43.3 (0.1)	47.7 (0.1)	15.1 (0.0)	N/A	32.9 (0.1)	31.8 (0.1)	32.3 (0.1)
Attention	15.1 (0.0)	N/A	46.5 (3.6)	37.9 (2.9)	41.8 (3.2)	15.1 (0.0)	N/A	50.0 (3.4)	48.4 (3.3)	49.2 (3.3)
RNP	14.7 (0.7)	78.2 (3.3)	75.0 (0.5)	59.7 (3.1)	66.5 (2.1)	15.2 (1.0)	81.7 (2.4)	67.0 (12.1)	64.7 (8.8)	65.8 (10.4)
Inter_RAT	15.2 (1.1)	N/A	57.0 (5.3)	46.9 (2.3)	51.4 (3.2)	16.1 (0.7)	N/A	57.9 (2.4)	60.3 (2.5)	59.0 (2.1)
NIR	15.5 (0.4)	80.8 (1.2)	73.3 (1.3)	61.3 (2.2)	66.8 (1.7)	15.2 (0.2)	82.1 (7.1)	69.8 (6.0)	67.9 (5.3)	68.8 (5.6)
DMR	15.0 (1.1)	80.7 (2.9)	75.0 (0.8)	60.1 (1.7)	66.7 (1.2)	15.2 (1.3)	82.6 (1.3)	68.3 (3.9)	66.3 (5.6)	67.2 (4.5)
A2I	14.9 (0.3)	81.0 (1.2)	75.2 (0.9)	60.6 (1.7)	67.1 (1.3)	14.8 (0.1)	82.7 (2.3)	69.4 (2.5)	65.9 (2.6)	67.6 (2.5)
VER (ours)	14.4 (0.2)	81.5 (2.5)	**82.2** (1.8)	**64.1** (1.5)	**72.0** (1.5)	14.9 (0.4)	84.2 (2.1)	**71.8** (4.2)	**68.6** (2.2)	**70.0** (3.1)

Datasets	Beer-Palate					Hotel-Location				
Methods	S	Acc	P	R	F1	S	Acc	P	R	F1
LIME	15.1 (0.0)	N/A	5.1 (0.0)	6.2 (0.0)	5.6 (0.0)	15.2 (0.0)	N/A	12.1 (0.1)	21.6 (0.2)	15.5 (0.2)
Attention	15.1 (0.0)	N/A	35.5 (2.2)	43.1 (2.7)	38.9 (2.4)	15.2 (0.0)	N/A	14.8 (8.1)	26.5 (14.5)	19.0 (10.4)
RNP	15.4 (0.6)	75.0 (3.5)	46.3 (3.6)	57.4 (4.9)	51.2 (4.0)	14.7 (0.4)	97.6 (0.4)	41.4 (1.6)	72.0 (2.7)	52.6 (1.9)
Inter_RAT	15.3 (1.2)	N/A	42.9 (1.1)	52.8 (3.7)	47.3 (1.7)	15.1 (1.2)	N/A	28.8 (1.7)	50.8 (2.1)	36.7 (1.3)
NIR	14.3 (0.5)	82.4 (3.1)	49.8 (4.5)	57.4 (5.6)	53.3 (4.9)	13.9 (0.2)	97.3 (0.6)	42.2 (1.5)	69.3 (2.7)	52.5 (1.9)
DMR	15.2 (1.3)	81.9 (2.9)	47.1 (4.1)	57.5 (5.3)	52.0 (4.6)	14.8 (0.9)	97.2 (0.7)	41.6 (3.3)	72.7 (3.9)	52.9 (3.5)
A2I	15.3 (0.7)	81.3 (3.2)	50.9 (3.9)	62.0 (4.2)	55.8 (4.1)	15.0 (0.7)	97.5 (0.5)	41.7 (2.1)	73.1 (2.3)	53.1 (2.2)
VER (ours)	15.0 (0.2)	83.9 (2.1)	**54.8** (4.5)	**65.9** (5.0)	**59.9** (4.7)	14.6 (0.2)	97.8 (0.3)	**44.3** (0.4)	**76.3** (0.8)	**56.1** (0.4)

Datasets	Hotel-Service					Hotel-Cleanliness				
Methods	S	Acc	P	R	F1	S	Acc	P	R	F1
LIME	15.2 (0.0)	N/A	11.9 (0.1)	15.7 (0.2)	13.5 (0.2)	15.2 (0.0)	N/A	10.9 (0.1)	18.7 (0.1)	13.8 (0.1)
Attention	15.1 (0.0)	N/A	14.1 (3.4)	18.5 (4.5)	16.0 (3.9)	15.1 (0.0)	N/A	16.3 (3.6)	27.7 (5.7)	20.6 (4.5)
RNP	15.3 (0.3)	96.5 (1.5)	41.0 (1.5)	54.6 (1.1)	46.8 (1.4)	15.3 (0.2)	97.2 (1.6)	28.1 (0.7)	48.7 (1.1)	35.6 (0.8)
Inter_RAT	15.0 (0.8)	N/A	28.9 (1.1)	38.1 (1.9)	32.8 (1.1)	14.4 (1.1)	N/A	27.2 (2.1)	44.1 (2.4)	33.6 (2.1)
NIR	15.0 (0.3)	96.9 (0.4)	40.9 (1.5)	53.5 (1.2)	46.3 (1.4)	15.5 (0.4)	96.7 (0.9)	28.0 (0.6)	49.2 (1.0)	35.7 (0.6)
DMR	14.8 (0.7)	96.6 (2.0)	41.2 (2.1)	54.1 (3.1)	46.9 (2.6)	14.9 (0.5)	97.0 (0.5)	28.5 (1.1)	48.1 (1.7)	35.9 (1.3)
A2I	15.1 (0.4)	96.7 (0.6)	41.4 (1.7)	54.6 (1.3)	47.1 (1.5)	15.3 (0.3)	96.8 (1.0)	28.8 (0.6)	49.7 (1.1)	36.5 (0.7)
VER	14.9 (0.3)	97.1 (1.0)	**42.4** (0.9)	**55.3** (0.5)	**48.0** (0.6)	15.3 (0.2)	97.3 (0.4)	**30.3** (0.7)	**52.3** (0.7)	**38.3** (0.7)

Table 1: Results on standard benchmarks. Values in parentheses are the standard deviations. Notes: LIME and Attention belong to post-hoc methods and the predictor is not trained on the rationales, so we do not report the accuracy metric for them.

For Inter_RAT, since it has originally been implemented on the beer-related datasets, we apply its original hyperparameters but only adjust the sparsity regularizer in Equation (4). More details are in Appendix A.2.

Metrics. Since predictive accuracy is influenced by many other unknown factors besides the quality of the extracted rationale, it is not a good metric. So, following the previous literature of rationalization, we mainly focus on rationale quality, which is measured by the overlap between the human-annotated rationales and the model-selected tokens. (We note that this may not be a perfect metric as evaluating explanation quality is a complex problem [Pruthi et al., 2022]. Nevertheless, this metric has been widely adopted in previous research in this field and has been empirically validated to make sense.) The terms $P, R, F1$ denote precision, recall, and $F1$ score, respectively. Among them, P indicates how much useless information is removed from the raw text (the cleanliness of the extracted rationale), and R indicates how much useful information is extracted (comprehensiveness of the extracted rationale). These metrics are the most frequently used in rationalization. The term S represents the average sparsity of the selected rationales, that is, the percentage of selected tokens in relation to the full text. Since the sparsity of ground-truth rationales on beer- and hotel-related datasets is around $10\% \sim 20\%$, we adjust α in Equation (4) to make S be about 15% (since Equation (4) is only a soft constraint, it cannot strictly limit S to be exactly 15%). Acc stands for the predictive accuracy. For FEVER and MultiRC datasets, we follow NIR to set S to be about 20%. For the results with BERT encoder, we follow CR to make S be about 10%.

5.2 Results

Rationale Quality. Table 1 shows the comparisons of recent methods and our VER with perturbation $p = 0.3$. In terms of the rationale quality (F1 score), we outperform all of the

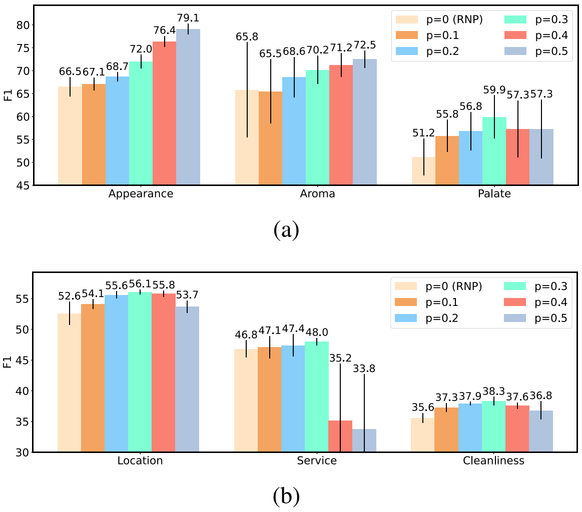
Methods	S	P	R	F1
RNP*	10.0 (n/a)	40.0 (1.4)	20.3 (1.9)	25.2 (1.7)
A2R*	10.0 (n/a)	55.0 (0.8)	25.8 (1.6)	34.3 (1.4)
INVRAT*	10.0 (n/a)	56.4 (2.5)	27.3 (1.2)	36.7 (2.1)
CR*	10.0 (n/a)	59.7 (1.9)	31.6 (1.6)	39.0 (1.5)
VER(ours)	10.8 (0.5)	**63.4** (8.7)	**37.6** (7.0)	**47.2** (7.9)

Table 2: Results with BERT encoder. The dataset is the most widely used Beer-Appearance. *: Results obtained from CR (Zhang et al., 2023).

baselines by a large margin. Specifically, the improvements are up to 6.6% (Beer-Palate dataset). We also improve the prediction accuracy to some extent on the three beer-related datasets. But the improvements on the hotel-related datasets are not significant. The possible reason is that the prediction accuracy of RNP on these datasets is already very high. We also follow CR to conduct a supplement experiment with BERT encoder on the most widely used Beer-Appearance dataset and compare with some other baselines that have been implemented with BERT, and the results are shown in Table 2. We still beat all the baselines.

Perturbation Rate. We show the results for p values of $[0, 0.1, 0.2, 0.3, 0.4, 0.5]$ in Figure 9 to demonstrate the robustness of our proposed method. In most cases, our method can improve the results of the standard RNP. On the one hand, a higher perturbation rate can bring the predictor a better evaluation of a larger region; on the other hand, an overhigh perturbation rate can also bring a very low signal-to-noise ratio, as mentioned at the beginning of §4.3. We also observe that, for the *Appearance* and *Aroma* datasets, $p = 0.5$ is best, while for other datasets, a lower $p = 0.3$ is best. This is because in the former two datasets, the gold rationales occupy a larger proportion of the raw input (see the dataset statistics in Table 4 of Appendix A.1), thus have a higher signal-to-noise ratio and are able to tolerate larger perturbations. This phenomenon is consistent with our findings at the beginning of §4.3 that too a high perturbation rate can fail due to the high signal-to-noise ratio (i.e., Empirical observation 4).

Quality of the Predictor's Assessment of Human-annotated (i.e., optimal) Rationales. Ideally, the human-annotated (i.e., optimal) rationale is most informative and should achieve high prediction accuracy when input to a good predictor. But as illustrated in §4.2, this is not always the case in practice, because the predictor could inaccurately assess the gold rationales. To further demonstrate that our VER can enhance the predictor's ability to assess rationales, we conducted additional experiments comparing the assessment of human-annotated gold annotated rationales by the RNP's predictor and the VER's predictor.



(a)

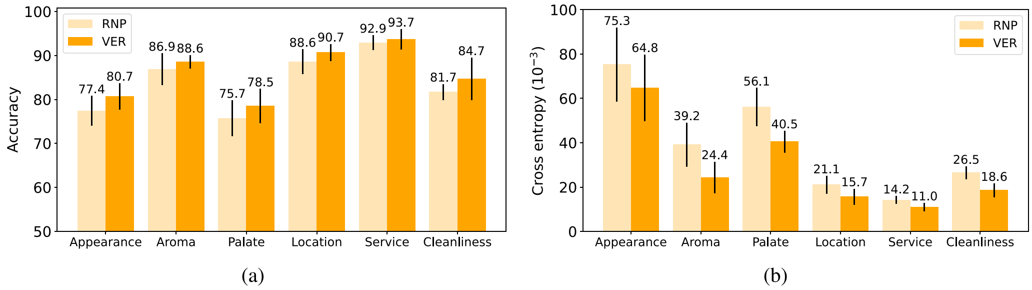
(b)

Figure 9: The influence of the perturbation rate p on (a) beer-related datasets and (b) hotel-related datasets.

Figure 10(a) and 10(b) shows the predictor's accuracy/cross-entropy with the inputs being human-annotated rationales (the details of the experimental settings are in Appendix A.5). We observe that our method can make the predictor assess the gold rationales more accurately, which allows the selector to receive more accurate feedback about the gold rationales, thereby correctly increasing the probability of selecting them.

Results on Different Tasks. To verify the effectiveness of our method on tasks other than classification, we also perform experiments on the reading comprehension task (using the MultiRC dataset) and the fact extraction and verification task (using the FEVER dataset). The datasets for the above two tasks are taken from the ERASER benchmark (DeYoung et al., 2020).

We mainly compare with the baselines that have already performed experiments on these datasets to avoid unfair comparisons resulted from hyperparameter choices. We follow the settings of NIR to use ‘‘bert-base-uncased’’ as the encoder of both the selector and the predictor (see Appendix A.2) and get the results of the baselines from its original



(a)

(b)

Figure 10: The quality of the predictor's assessment of human-annotated gold rationales. (a) Accuracy. (b) Cross-entropy. The x-axis represents different datasets.

Datasets / Methods	MultiRC					FEVER				
	S	Acc	P	R	F1	S	Acc	P	R	F1
A2R*	20.0 (n/a)	66.1 (1.9)	18.5 (1.6)	21.9 (2.2)	19.3 (1.8)	20.0 (n/a)	82.1 (3.2)	36.3 (0.6)	44.0 (0.3)	36.7 (0.5)
NIR*	20.0 (n/a)	66.4 (0.8)	22.6 (1.2)	**26.9** (1.8)	23.8 (1.4)	20.0 (n/a)	78.2 (1.9)	39.0 (2.5)	**47.2** (2.9)	39.5 (2.5)
VER (ours)	19.2 (1.2)	67.1 (2.3)	**24.5** (1.8)	23.8 (2.2)	**24.1** (2.0)	22.5 (2.2)	82.1 (2.6)	**40.1** (1.6)	40.8 (1.8)	**40.4** (1.7)

Table 3: Results with different tasks. *: Results obtained from the paper of NIR (Storek et al., 2023).

paper. The results are shown in Table 3. We see that our VER is still effective on these two tasks.

6 Conclusion and Limitations

In this paper, we first theoretically analyze the common approximation of the mutual information maximization criterion, specifically the minimization of cross-entropy, and its shortcomings when applied to the cooperative self-explanatory framework. We provide extensive empirical evidence to verify the existence and negative impact of this issue in practice. Subsequently, we propose a very simple method to mitigate this problem, achieving considerable improvements compared to existing methods. A potential limitation of our approach is that we only consider correlation, not causation, so a future direction may be to integrate our work with causal research studies.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under grant 2024YFC3307900; the National Natural Science Foundation of China under grants 62376103, 62302184, 62436003 and 62206102; Major Science and Technology Project of Hubei Province under grant 2024BAA008; Hubei Science and Technology Talent Service Project under grant 2024DJC078; and Ant Group through CCF-Ant Research Fund. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

We sincerely thank the reviewers and the editor for their valuable feedback and efforts during the review process. They have helped considerably in improving the quality of this paper. We thank Yuankai Zhang for his help in providing more baseline results during the revision.

References

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers*, pages 2963–2977. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1284

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022. UNIREX: A unified learning framework for language model rationale extraction. In *International Conference on Machine Learning, ICML 2022, 17–23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2867–2889. PMLR. https://doi.org/10.18653/v1/2022.bigscience-1.5

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2019. A game theoretic approach to class-wise selective rationalization. In *Advances in Neural Information Processing*

Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pages 10055–10065.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2020. Invariant rationalization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1448–1458. PMLR.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. 2000. Vicinal risk minimization. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 416–422. MIT Press.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10–15, 2022*, pages 3792–3805. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.naacl-main.278`

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J. White, and Su-In Lee. 2023. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, pages 6424–6447. PMLR.

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. `https://doi.org/10.18653/v1/2020.acl-main.408`

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. 2019. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, pages 11423–11434.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750. `https://doi.org/10.1016/S2589-7500(21)00208-9`, PubMed: 34711379

Yue Guan, Zhengyi Li, Jingwen Leng, Zhouhan Lin, and Minyi Guo. 2022. Transkimmer: Transformer learns to layer-wise skim. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22–27, 2022*, pages 7275–7286. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.acl-long.502`

Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. 2023. Joint learning of label and environment causal independence for graph out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36:3945–3978.

Yongfeng Huang, Yujun Chen, Yulun Du, and Zhilin Yang. 2021. Distribution matching for rationalization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2–9, 2021*, pages 13090–13097. AAAI Press. `https://doi.org/10.1609/aaai.v35i14.17547`

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, ACL 2020, Online, July 5–10, 2020*, pages 4198–4205. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.386`

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*, pages 4459–4473. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.409

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. https://doi.org/10.1145/3571730

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016*, pages 107–117. The Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1011

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57. https://doi.org/10.1145/3236386.3241340

Wei Liu, Zhiying Deng, Zhongyu Niu, Jun Wang, Haozhao Wang, Zhigang Zeng, and Ruixuan Li. 2025. Breaking free from mmi: A new frontier in rationalization by probing input utilization. In *The Thirteenth International Conference on Learning Representations*.

Wei Liu, Zhiying Deng, Zhongyu Niu, Jun Wang, Haozhao Wang, YuanKai Zhang, and Ruixuan Li. 2024a. Is the MMI criterion necessary for interpretability? Degenerating non-causal features to plain noise for self-rationalization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10–15, 2024*.

Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Xinyang Li, Yuankai Zhang, and Yang Qiu. 2023a. MGR: multi-generator based rationalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9–14, 2023*, pages 12771–12787. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.715

Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Chao Yue, and YuanKai Zhang. 2022. Fr: Folded rationalization with a unified encoder. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc.

Wei Liu, Jun Wang, Haozhao Wang, Ruixuan Li, Zhiying Deng, and YuanKai Zhang. 2024b. Attacking for inspection and instruction: Debiasing self-explaining text classification.

Wei Liu, Jun Wang, Haozhao Wang, Ruixuan Li, Zhiying Deng, Yuankai Zhang, and Yang Qiu. 2023b. D-separation for causal self-explanation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*.

Wei Liu, Jun Wang, Haozhao Wang, Ruixuan Li, Yang Qiu, Yuankai Zhang, Jie Han, and Yixiong Zou. 2023c. Decoupled rationalization with asymmetric learning rates: A flexible lipschitz restraint. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6–10, 2023*, pages 1535–1547. ACM. https://doi.org/10.1145/3580305.3599299

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems 33: Annual*

Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual.

Julian J. McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10–13, 2012*, pages 1020–1025. IEEE Computer Society. https://doi.org/10.1109/ICDM.2012.110

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375. https://doi.org/10.1162/tacl_a_00465

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Qihan Ren, Jiayang Gao, Wen Shen, and Quanshi Zhang. 2024. Where we have arrived in proving the emergence of sparse interaction primitives in DNNs. In *The Twelfth International Conference on Learning Representations*.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ''Why should I trust you?'': Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016*, pages 1135–1144. ACM. https://doi.org/10.1145/2939672.2939778

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x, PubMed: 35603010

Benjamin B. Seiler. 2023. *Applications of Cooperative Game Theory to Interpretable Machine Learning*. Ph.D. thesis. Stanford University.

Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. 2021. Learning from the best: Rationalizing predictions by adversarial information calibration. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2–9, 2021*, pages 13771–13779. AAAI Press. https://doi.org/10.1609/aaai.v35i15.17623

Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. 2023. Rationalizing predictions by adversarial information calibration. *Artificial Intelligence*, 315:103828. https://doi.org/10.1016/j.artint.2022.103828

Adam Storek, Melanie Subbiah, and Kathleen R. McKeown. 2023. Unsupervised selective rationalization with noise injection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9–14, 2023*, pages 12647–12659. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.707

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chao Zhang, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, and Yue Zhao. 2024. Trustllm: Trustworthiness in large language models. *CoRR*, abs/2401.05561. https://doi.org/10.48550/ARXIV.2401.05561

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings*.

Aladin Virmaux and Kevin Scaman. 2018. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada*, pages 3839–3848.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25–28, 2010*, pages 783–792. ACM. `https://doi.org/10.1145/1835804.1835903`

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. 2018. Evaluating the robustness of neural networks: An extreme value theory approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2022. Discovering invariant rationales for graph neural networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning, ICML 2024, 21–27 July 2024, Vienna, Austria*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S. Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. 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 2019, Hong Kong, China, November 3–7, 2019*, pages 4092–4101. Association for Computational Linguistics. `https://doi.org/10.18653/v1/D19-1420`

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi S. Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual*, pages 12822–12835.

Hao Yuan, Lei Cai, Xia Hu, Jie Wang, and Shuiwang Ji. 2022. Interpreting image classifiers by generating discrete masks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2019–2030. `https://doi.org/10.1109/TPAMI.2020.3028783`, PubMed: 33021938

Linan Yue, Qi Liu, Li Wang, Yanqing An, Yichao Du, and Zhenya Huang. 2023. Interventional rationalization. `https://doi.org/10.18653/v1/2023.emnlp-main.700`

Wenbo Zhang, Tong Wu, Yunlong Wang, Yong Cai, and Hengrui Cai. 2023. Towards trustworthy explanation: On causal rationalization. In *International Conference on Machine Learning, ICML 2023, 23–29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41715–41736. PMLR.

Datasets		Train			Dev			Annotation			
		Pos	Neg	avg_len	Pos	Neg	avg_len	Pos	Neg	avg_len	S
Beer	Appearance	16891	16891	141	6628	2103	145	923	13	126	18.5
	Aroma	15169	15169	144	6579	2218	147	848	29	127	15.6
	Palate	13652	13652	147	6740	2000	149	785	20	128	12.4
Hotel	Location	7236	7236	151	906	906	152	104	96	155	8.5
	Service	50742	50742	154	6344	6344	153	101	99	152	11.5
	Cleanliness	75049	75049	144	9382	9382	144	99	101	147	8.9

Table 4: Statistics of the datasets used in this paper.

A Experimental Details

A.1 Datasets

BeerAdvocate (McAuley et al., 2012) is a benchmark that consists of comments about beer. It contains three widely used text classification datasets: Beer-Appearance, Beer-Aroma, Beer-Palate. In this datasets, each piece of text is a comment about the beer. For the Beer-Appearance dataset, the classification label is the quality (bad/good, [0,1]) of the beer's appearance. For Beer-Aroma and Beer-Palate, the classification label is about the aroma and palate, respectively.

HotelReview (Wang et al., 2010) is a benchmark consists of hotel reviews. It contains three widely used datasets: Hotel-Location, Hotel-Service, Hotel-Cleanliness. In this datasets, each piece of text is a review about a hotel. For the Hotel-Location dataset, the classification label is the quality (bad/good, [0,1]) of the hotel's Location. For Hotel-Service and Hotel-Cleanliness, the classification label is about the service and cleanliness, respectively.

The statistics of the datasets are in Table 4. Pos and Neg denote the number of positive and negative examples in each set. S denotes the average percentage of tokens in human-annotated rationales to the whole texts. avg_len denotes the average length of a text sequence.

The other two datasets MultiRC and FEVER are taken from the ERASER benchmark (DeYoung et al., 2020). Readers can refer to this benchmark for details.

A.2 Implementation Details

The code and detailed running instructions is aviable at `https://github.com/jugechengzi/Rationalization-VER`.

The maximum sequence length is set to 256. We use the Adam optimizer (Kingma and Ba, 2015) with its default parameters, except for the learning rate. The temperature for gumbel-softmax is the default value 1. We implement the code with PyTorch on a RTX4090 GPU. We report the average results of five random seeds, and the seeds are [1,2,3,4,5].

For NIR and our VER, considering they are both variants of the standard RNP, we first manually tune the hyperparameters for RNP, and then apply the hyperparameters to both NIR and VER. For all datasets, we use a learning rate of 0.0001. The batchsize is 128 for the beer-related datasets and 256 for the hotel-related datasets. These hyperparameters are found by manually tuning the standard RNP and are applied to both NIR and our VER.

The core idea of NIR is to inject noise into the selected rationales. We use RNP as its backbone. A unique hyperparameter of NIR is the proportion of noise. Following the method in the original text, we searched within $[0.1, 0.2, 0.3]$ and found that 0.1 yielded the best results on most datasets, hence we adopted 0.1 for this.

For our VER, the perturbation rate p is searched within $[0.1, 0.2, 0.3, 0.4, 0.5]$. We find that 0.5 is the best for beer-related datasets but 0.3 is best for hotel-related datasets. For fair comparisons with baselines, we uniformly adopt 0.3 for all datasets.

We found that the training of Inter_RAT is very unstable. To avoid potential unfair factors, our main settings are determined with reference to it. Except for the part about sparsity, we used its original hyperparameters for it.

For CR, we just keep the major settings (''bert-base-uncased'', the Beer-Appearance dataset, and the sparsity of 10%) the same as it and copy its results from its original paper.

For the experiments on the MultiRC and FEVER datasets, we follow the settings of NIR and get the results of the baselines from its original paper. Specifically, we use ''bert-base-uncased'' as the encoder and the rationale sparsity is set to be about 20%.

A.3 Details About the Setup of Figure 3

We first train RNP in the usual manner. After training is completed, we fix both the selector and the predictor. For ''model-selected rationale'': The selector receives the raw text of the test set as input and extracts the corresponding rationale Z, which is then input into the predictor to obtain the prediction accuracy/cross-entropy. For ''gold rationale'': The human-annotated rationale is directly input into the predictor to obtain the accuracy/cross-entropy.

The range of cross-entropy is from 0 to ∞. If we take the mean, similar to accuracy, it might be affected by extreme values and fail to reflect the overall situation of the dataset. Therefore, we use the median instead.

A.4 The Details of the Experiments in Figure 4

Definition 1 (Definition 1 in Liu et al. (2023c))**.** *A function $f : \mathcal{R}^n \rightarrow \mathcal{R}^1$ is Lipschitz continuous on $\mathcal{X} \subset \mathcal{R}^n$ if there exists a constant $L \geq 0$ such that*

$$\forall x_i, x_j \in \mathcal{X},\ |f(x_i) - f(x_j)| \leq L \cdot d(x_i, x_j), \tag{9}$$

over a distance metric d.

The smallest L is called the Lipschitz constant, denoted as

$$L_c = \sup_{x_i, x_j \in \mathcal{X}} \frac{|f(x_i) - f(x_j)|}{d(x_i, x_j)}, \tag{10}$$

L_c represents the maximum ratio between variations in the output and variations in the input of a model, and is used to measure Lipschitz continuity.

Discussing Lipschitz continuity is beyond the scope of this paper and we just follow the computational method for Lipschitz constant as proposed by Liu et al. (2023c), which is described as follows (Appendix A.4 of Liu et al. (2023c)): For each full text in the whole training set, we first generate one rationale Z with the selector and then calculate $\nabla f_p(Z)$. Note that here $\nabla f(Z) \in \mathcal{R}^{b\times d}$ is a matrix, where b is the length of the rationale Z which may varies across different rationales and d is the dimension of the word vector. We then take the average along the length of Z and get the unified $\nabla f_p(Z) \in \mathcal{R}^d$. Then we calculate $||\nabla f_p(Z)||_2$ and take the maximum value over the entire training set as the approximate L_c.

A.5 Details About the Setup of Figure 10

We first train RNP and VER separately. Then, we extract and fix their predictors. The human-annotated rationales are input into their predictors for testing, yielding the results shown in Figure 10.

A.6 Training the Predictor with Randomly Selected $\mathcal{D}_z$

This is the experimental setup of Figure 6. The dataset is the most widely used Beer-Appearance. For training iteration, we randomly select $\alpha\%$ (here $\alpha = 12.5$) tokens from each input text X. And the predictor is trained with the randomly selected $\mathcal{D}_z$ for 300 epochs. Note that we resample at each iteration. Then, the predictor is fixed and we train the selector just in the same way as the standard RNP. We choose $\alpha = 12.5$ because we find that when $\alpha = 12.5$, the final sparsity on the test set is about 15%, which matches the sparsity of our formal experiments in Table 1.