

Rationalizing Predictions by Adversarial Information Calibration¹

Lei Sha^{a,b,f,*}, Oana-Maria Camburu^c and Thomas Lukasiewicz^{d,b,e,*}

^aInstitute of Artificial Intelligence, Beihang University, China

^bDepartment of Computer Science, University of Oxford, UK

^cDepartment of Computer Science, University College London, UK

^dInstitute of Logic and Computation, TU Wien, Austria

^eAlan Turing Institute, London, UK

^fZhongguancun Laboratory, Beijing, China

ARTICLE INFO

Keywords:

rationale extraction
interpretability
natural language processing
information calibration
deep neural networks

ABSTRACT

Explaining the predictions of AI models is paramount in safety-critical applications, such as in legal or medical domains. One form of explanation for a prediction is an extractive rationale, i.e., a subset of features of an instance that lead the model to give its prediction on that instance. For example, the subphrase “he stole the mobile phone” can be an extractive rationale for the prediction of “Theft”. Previous works on generating extractive rationales usually employ a two-phase model: a selector that selects the most important features (i.e., the rationale) followed by a predictor that makes the prediction based exclusively on the selected features. One disadvantage of these works is that the main signal for learning to select features comes from the comparison of the answers given by the predictor to the ground-truth answers. In this work, we propose to squeeze more information from the predictor via an information calibration method. More precisely, we train two models jointly: one is a typical neural model that solves the task at hand in an accurate but black-box manner, and the other is a selector-predictor model that additionally produces a rationale for its prediction. The first model is used as a guide for the second model. We use an adversarial technique to calibrate the information extracted by the two models such that the difference between them is an indicator of the missed or over-selected features. In addition, for natural language tasks, we propose a language-model-based regularizer to encourage the extraction of fluent rationales. Experimental results on a sentiment analysis task, a hate speech recognition task as well as on three tasks from the legal domain show the effectiveness of our approach to rationale extraction.

1. Introduction

Although deep neural networks have recently been contributing to state-of-the-art advances in various areas [70, 49, 128], such models are often black-box, and therefore may not be deemed appropriate in situations where safety needs to be guaranteed, such as legal judgment prediction and medical diagnosis. Interpretable deep neural networks are a promising way to increase the reliability of neural models [105]. To this end, extractive rationales, i.e., subsets of features of instances on which models rely for their predictions on the instances, can be used as evidence for humans to decide whether to trust a prediction and more generally a model.

There are many different methods to explain a deep neural model, such as probing internal representations [47, 24, 93, 135, 21, 133], adding interpretability to deep neural models [41, 105, 40, 1, 19, 113], and looking for global decision rules [52, 22, 77, 11, 139, 37]. Extracting rationales belongs to the second category.

Previous works use selector-predictor types of neural models to provide extractive rationales. More precisely, such models are composed of two modules: (i) a *selector* that selects a subset of features of each input, and (ii) a *predictor* that makes a prediction based solely on the selected features. For example, Yoon et al. (2018) and Lei et al. (2016) use a selector network to calculate a selection probability for each token in a sequence, then sample a set of tokens that is the only input of the predictor. Supervision is typically given only on the final prediction and not on the rationales. Paranjape et al. (2020) also uses information bottleneck to find a better trade-off between the sparsity and the final task performance. Note that gold rationale labels are required for semi-supervised training in Paranjape et al. (2020).

¹This article is a substantially revised and extended version of a preliminary paper at AAAI 2021 [111].

*Corresponding author

✉ shaLei@buaa.edu.cn (L. Sha); o.camburu@cs.ucl.ac.uk (O. Camburu); thomas.lukasiewicz@cs.ox.ac.uk (T. Lukasiewicz)
ORCID(s): 0000-0001-5914-7590 (L. Sha)

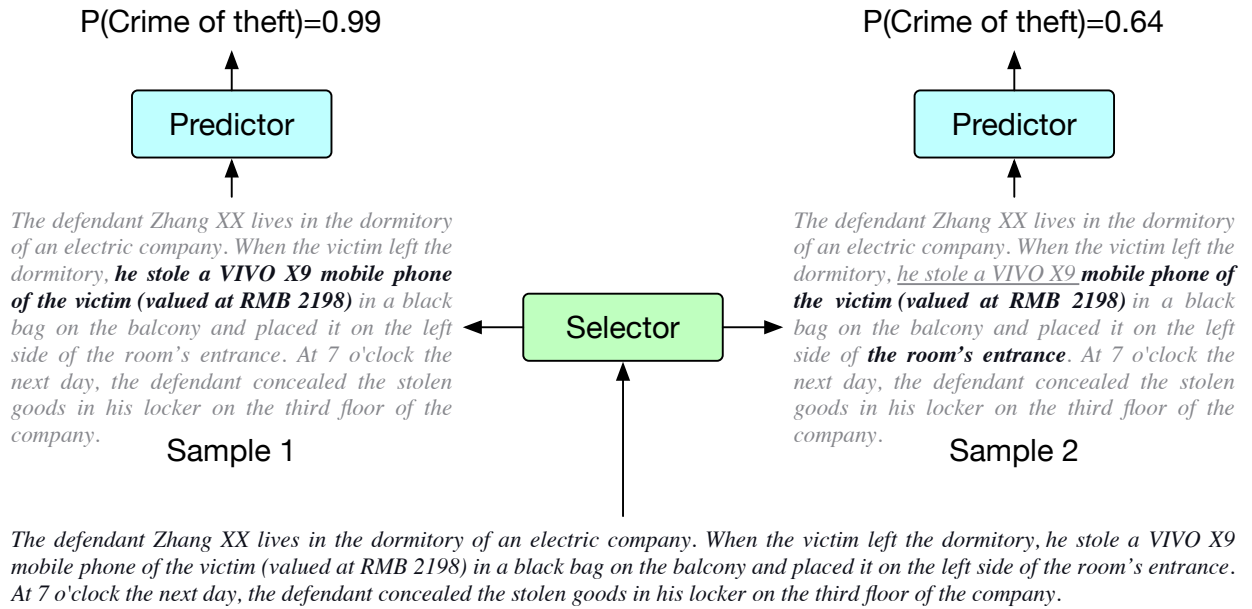


Figure 1: Examples of rationales in legal judgement prediction. The human-provided rationale is shown in bold in Sample 1. In Sample 2, the selector missed the key information “he stole a VIVO X9”, but the predictor only tells the selector that the whole extracted rationale (in bold) is not so informative, by producing a low probability of the correct crime.

An additional typical desideratum in natural language processing (NLP) tasks is that the selected tokens form a semantically fluent rationale. To achieve this, Lei et al. (2016) added a non-differential regularizer that encourages any two adjacent tokens to be simultaneously selected or unselected. The selector and predictor are jointly trained in a REINFORCE-style manner [137] because the sampling process and the regularizer are not differentiable. Bastings et al. (2019) further improved the quality of the rationales by using a HardKuma regularizer that also encourages any two adjacent tokens to be selected or unselected together, which is differentiable and no need to use REINFORCE any more.

One drawback of previous works is that the learning signal for both the selector and the predictor comes mainly from comparing the prediction of the selector-predictor model with the ground-truth answer. Therefore, the exploration space to get to the correct rationale is large, decreasing the chances of converging to the optimal rationales and predictions. Moreover, in NLP applications, the regularizers commonly used for achieving fluency of rationales treat all adjacent token pairs in the same way. This often leads to the selection of unnecessary tokens due to their adjacency to informative ones.

In this work, we first propose an alternative method to rationalize the predictions of a neural model. Our method aims to squeeze more information from the predictor in order to guide the selector in selecting the rationales. Our method trains two models jointly: a “guider” model that solves the task at hand in an accurate but black-box manner, and a selector-predictor model that solves the task while also providing rationales. We use an adversarial-based training procedure to encourage the final information vectors generated by the two models to encode the same information. We use an information bottleneck technique in two places: (i) to encourage the features selected by the selector to be the least-but-enough features, and (ii) to encourage the final information vector of the guider model to also contain the least-but-enough information for the prediction. Secondly, we propose using language models as regularizers for rationales in natural language understanding tasks. A language model (LM) regularizer encourages rationales to be fluent subphrases, which means that the rationales are formed by consecutive tokens while avoiding unnecessary tokens to be selected simply due to their adjacency to informative tokens. The effectiveness of our LM-based regularizer is proved both by a mathematical derivation and experiments.

The contributions of this article are briefly summarized as follows:

- We introduce a novel model that generates extractive rationales for its predictions. The model is based on an adversarial approach that calibrates the information between a guider and a selector-predictor model, such that the selector-predictor model learns to mimic a typical neural model while additionally providing rationales.
- We propose a language-model-based regularizer to encourage the sampled tokens to form fluent rationales. Usually, this regularizer will encourage fewer fragment of subsequences and avoid strange start and end of the sequences. This regularizer also gives priority to important adjacent token pairs, which benefits the extraction of informative features.
- We experimentally evaluate our method on a sentiment analysis dataset and a hate speech detection dataset, both containing ground-truth rationale annotations for the ground-truth labels, as well as on three tasks of a legal judgement prediction dataset, for which we conducted human evaluations of the extracted rationales. The results show that our method improves over the previous state-of-the-art models in precision and recall of rationale extraction without sacrificing the prediction performance.

The rest of this paper is organized as follows. In Section 3, we introduce our proposed approach, including the selector-predictor module (Section 3.1), the guider module (Section 3.1.3), the information calibrating method (Section 3.2), and the language model-based rationale regularizer (Section 3.3). In Section 4, we report the experimental results on the three datasets: a beer review dataset (Section 4.2), a legal judgment prediction dataset (Section 4.3), and a hate speech detection dataset (Section 4.4). Section 2 reviews the related works of this paper. In Section 5, we provide a summary and an outlook on future research.

2. Related Work

Explainability is currently a key bottleneck of deep-learning-based approaches [5, 61]. A summarization of related works is shown in Table 1, where we have listed the representative works in each branch of interpretable models. As is shown, previous works on explainable neural models include self-explanatory models and post-hoc explainers. The model proposed in this work belongs to the class of self-explanatory models, which contain an explainable structure in the model architecture, thus providing explanations / rationales for their predictions. Self-explanatory models can use different types of explanations / rationales, such as feature-based explanations which is usually conducted by selector-predictors [75, 142, 17, 144, 14] and natural language explanations [46, 13, 91, 65]. Our model uses feature-based explanations.

2.1. Self-explanatory models for interpretability.

Self-explanatory models with feature-based explanations can be further divided into two branches. The first branch is disentanglement-based approaches, which map specific features into latent spaces and then use the latent variables to control the outcomes of the model, such as disentangling methods [19, 113], information bottleneck methods [130], and constrained generation [110]. The second branch consists of architecture-interpretable models, such as attention-based models [147, 112, 114, 116, 80], Neural Turing Machines [23, 138, 115], capsule networks [105], and energy-based models [40]. Among them, attention-based models have an important extension, that of sparse feature learning, which implies learning to extract a subset of features that are most informative for each example. Most of the sparse feature learning methods use a selector-predictor architecture. Among them, L2X [17] and INVASE [142] make use of information theories for feature selection, while CAR [16] extracts useful features in a game-theoretic approach.

In addition, rationale extraction for NLP usually raises one desideratum for the extracted subset of tokens: rationales need to be fluent subphrases instead of separate tokens. To this end, Lei et al. (2016) proposed a non-differentiable regularizer to encourage selected tokens to be consecutive, which can be optimized by REINFORCE-style methods [137]. Bastings et al. (2019) proposed a differentiable regularizer using the Hard Kumaraswamy distribution; however, this still does not consider the difference in the importance of different adjacent token pairs. Paranjape et al. (2020) proposed a very similar information bottleneck method to our InfoCal method. However, they did not use any calibration method to encourage the completeness of the extracted rationale.

Our method belongs to the class of self-explanatory methods. Different from previous sparse feature learning methods, we use an adversarial information calibrating mechanism to hint to the selector about missing important features or over-selected features. Moreover, our proposed LM regularizer is differentiable and can be directly optimized by gradient descent. This regularizer also encourages important adjacent token pairs to be simultaneously selected, which benefits the extraction of useful features.

Rationalizing Predictions by Adversarial Information Calibration

			Representative Methods	Controllable	Provide important features	Provide important examples	Provide NL explanations	Provide rules
Self-explanatory methods	Disentanglement	Implicit	β -VAE [48], β -TCVAE [18]	Yes				
		Explicit	InfoGAN [19], MTDNA [113]	Yes	Yes			
	Architecture	Attention-based	Rocktäschel et al. (2015), Vaswani et al. (2017), OrderGen [114]		Yes			
		Read-Write Memory	Neural Turing Machines [23, 115], Progressive Memory [104, 138], Differentiable neural computer [42], Neural RAM [60], Neural GPU [72]	Yes				
		Capsule-based	Capsule [105]		Yes			
		Energy-based	Grathwohl et al. (2019), Hopfield Network [97], Boltzmann Machine [85], Predictive Coding [122]		Yes			
		Rationalization	Selector-predictor [75, 8, 142, 17], natural language explanations [46, 13]		Yes		Yes	
Post-hoc explainer	Local	Perturbation-based	SHAP [82], Shapley Values [117]		Yes			
		Surrogate-based	Anchors [101], LIME [100]		Yes		Yes	
		Saliency Maps	Input gradient [6, 119, 118], SmoothGrad [121, 109], Integrated Gradients [127], Guided Backprop [123]		Yes			
		Prototypes/Example Based	Influence Functions [25, 67], Representer Points [141], Tracln [94]			Yes		
		Counterfactuals	Wachter et al. (2017), Mahajan et al. (2019), Karimi et al. (2020)		Yes	Yes		
	Global	Collection of Local Explanations	SP-LIME [100], Summaries of Counterfactuals [98]		Yes			
		Model Distillation	Tree Distillation [7], Decision set distillation [73], Generalized Additive Models [129]		Yes			Yes
Representation based		Network Dissection [9], TCAV [64]		Yes				

Table 1
The branches of related works.

2.2. Post-hoc explainers for interpretability

Post-hoc explainers analyze the effect of each feature in the prediction of an already trained and fixed model. Post-hoc explainers can be divided into two types: local explainers and global explainers. Local explainers can be further split into five categories: (a) perturbation-based: change the values of some features to see their effect on the outcome [35, 53, 36, 31, 43, 149, 38, 56, 4]. Some famous perturbation-based post-hoc explainer methods include Shapley values [117, 125, 126, 57, 124] and SHAP method [82, 81, 120], (b) surrogate-based: train an explainable model, such as linear regression or decision trees, to approximate the predictions of a black-box model [63, 3, 101, 120], for example, LIME [100]. (c) saliency maps: use gradient information to show what parts of the input are most relevant for the model’s prediction, including input gradient [6, 119, 118], SmoothGrad [121, 109], integrated gradients [127], guided backprop [123], class activation mapping [151], meaningful perturbation [33], RISE [92], extremal perturbations [32], DeepLift [118], expected gradients [30], excitation backprop [147], GradCAM [108], occlusion [146], prediction difference analysis [44], and internal influence [76]. (d) prototype / example based: find which training example affects the model prediction the most. Usually, this is conducted by influence functions [25]. (e) counterfactual explanations: detect what features need to be changed to flip the model’s prediction [136, 84, 62]. On the other hand, some of the global explainers are collections of local explanations (e.g., SP-LIME [100], and summaries of counterfactuals [98]). Also, distillation methods provide explainable rules by distilling the information from deep models to tree models [7] or decision set models [73]. There are also some methods (Network Dissection [9], TCAV [64]) derives model understanding by analyzing intermediate representations of a deep black-box model.

2.3. Information bottleneck

The information bottleneck (IB) theory is an important basic theory of neural networks [130]. It originated in information theory and has been widely used as a theoretical framework in analyzing deep neural networks [131]. For example, Li and Eisner (2019) used IB to compress word embeddings in order to make them contain only specialized

information, which leads to a much better performance in parsing tasks.

2.4. Adversarial methods

Adversarial methods, which had been widely applied in image generation [19] and text generation [143], usually have a discriminator and a generator. The discriminator receives pairs of instances from the real distribution and from the distribution generated by the generator, and it is trained to differentiate between the two. The generator is trained to fool the discriminator [39]. Our information calibration method generates a dense feature vector using selected symbolic features, and the discriminator is used for measuring the calibration extent.

Our adversarial calibration method is inspired by distilling methods [50]. Distilling methods are usually applied to compress large models into small models while keeping a comparable performance. For example, TinyBERT [58] is a distillation of BERT [28]. Our method is different from distilling methods, because we calibrate the final feature vector instead of the softmax prediction. Also, to our best knowledge, we are the first to apply information calibration for rationale extraction.

3. Approach

Our approach is composed of a selector-predictor architecture, in which we use the information bottleneck technique to restrict the number of selected features, and a guider model, for which we again use the information bottleneck technique to restrict the information in the final feature vector. Then, we use an adversarial method to make the guider model guide the selector into selecting the least-but-enough features. Finally, we use a language model (LM) regularizer to obtain semantically fluent rationales.

3.1. InfoCal: Selector-Predictor-Guider with Information Bottleneck

The Selector-Predictor-Guider architecture contains two parallel architectures, one is a selector-predictor model, which selects the rationale and judges whether it can make a correct prediction; the other is a guider model, which is a dense “black-box” neural network trying to learn the feature vector required for the task. The information calibration is used to calibrate the dense feature vector learned by the guider model and the information contained in the rationales extracted by the selector-predictor model. The high-level architecture of our model, called InfoCal, is shown in Fig. 2. Below, we detail each of its components.

3.1.1. Selector

For a given instance (\mathbf{x}, y) , \mathbf{x} is the input with n features $\mathbf{x} = (x_1, x_2, \dots, x_n)$, and y is the ground-truth corresponding label. The selector network $\text{Sel}(\tilde{\mathbf{z}}_{\text{sym}}|\mathbf{x})$ takes \mathbf{x} as input and outputs $p(\tilde{\mathbf{z}}_{\text{sym}}|\mathbf{x})$, a sequence of probabilities $(p_i)_{i=1,\dots,n}$ representing the probability of choosing each feature x_i as part of the rationale.

Given the sampling probabilities, a subset of features is sampled using the Gumbel softmax [55], which provides a differentiable sampling process:

$$u_i \sim U(0, 1), \quad g_i = -\log(-\log(u_i)) \quad (1)$$

$$m_i = \frac{\exp((\log(p_i) + g_i)/\tau)}{\sum_j \exp((\log(p_j) + g_j)/\tau)}, \quad (2)$$

where $U(0, 1)$ represents the uniform distribution between 0 and 1, and τ is a temperature hyperparameter. Hence, we obtain the sampled mask m_i for each feature x_i , and the vector symbolizing the rationale $\tilde{\mathbf{z}}_{\text{sym}} = (m_1 x_1, \dots, m_n x_n)$. Thus, $\tilde{\mathbf{z}}_{\text{sym}}$ is the sequence of discrete selected symbolic features forming the rationale.

3.1.2. Predictor

The predictor takes as input the rationale $\tilde{\mathbf{z}}_{\text{sym}}$ given by the selector, and outputs the prediction \hat{y}_{sp} . In the selector-predictor part of InfoCal, the input to the predictor is the multiplication of each feature x_i with the sampled mask m_i . The predictor first calculates a dense feature vector $\tilde{\mathbf{z}}_{\text{nero}}$ ¹, then uses one feed-forward layer and a softmax layer to calculate the probability distribution over the possible predictions:

$$\tilde{\mathbf{z}}_{\text{nero}} = \text{Pred}(\tilde{\mathbf{z}}_{\text{sym}}) \quad (3)$$

¹Here, “nero” stands for neural feature (i.e., a neural vector representation) as opposed to a symbolic input feature.

$$p(\hat{y}_{sp}|\tilde{\mathbf{z}}_{\text{sym}}) = \text{Softmax}(W_p \tilde{\mathbf{z}}_{\text{nero}} + b_p). \quad (4)$$

As the input is masked by m_i , the prediction \hat{y}_{sp} is exclusively based on the features selected by the selector. The loss of the selector-predictor model is the cross-entropy loss:

$$\begin{aligned} L_{sp} &= -\frac{1}{K} \sum_k \log p(y_{sp}^{(k)}|\mathbf{x}^{(k)}) \\ &= -\frac{1}{K} \sum_k \log \mathbb{E}_{\text{Sel}(\tilde{\mathbf{z}}_{\text{sym}}|\mathbf{x}^{(k)})} p(y_{sp}^{(k)}|\tilde{\mathbf{z}}_{\text{sym}}^{(k)}) \\ &\leq -\frac{1}{K} \sum_k \mathbb{E}_{\text{Sel}(\tilde{\mathbf{z}}_{\text{sym}}|\mathbf{x}^{(k)})} \log p(y_{sp}^{(k)}|\tilde{\mathbf{z}}_{\text{sym}}^{(k)}), \end{aligned} \quad (5)$$

where K represents the size of the training set, the superscript (k) denotes the k-th instance in the training set, and the inequality follows from Jensen's inequality.

3.1.3. Guider

To guide the rationale selection of the selector-predictor model, we train a *guider* model, denoted Pred_G , which receives the full original input \mathbf{x} and transforms it into a dense feature vector \mathbf{z}_{nero} , using the same predictor architecture as the selector-predictor module, but different weights, as shown in Fig. 2. We generate the dense feature vector in a variational way, which means that we first generate a Gaussian distribution according to the input \mathbf{x} , from which we sample a vector \mathbf{z}_{nero} :

$$h = \text{Pred}_G(\mathbf{x}), \quad \mu = W_m h + b_m, \quad \sigma = W_s h + b_s \quad (6)$$

$$u \sim \mathcal{N}(0, 1), \quad \mathbf{z}_{\text{nero}} = u\sigma + \mu \quad (7)$$

$$p(\hat{y}_{\text{guide}}|\mathbf{z}_{\text{nero}}) = \text{Softmax}(W_p \mathbf{z}_{\text{nero}} + b_p). \quad (8)$$

We use the reparameterization trick of Gaussian distributions to make the sampling process differentiable [66]. We share the parameters W_p and b_p with those in Eq. 4.

The guider model's loss L_{guide} is as follows:

$$\begin{aligned} L_{\text{guide}} &= -\frac{1}{K} \sum_k \log p(y_{\text{guide}}^{(k)}|\mathbf{x}^{(k)}) \\ &\leq -\frac{1}{K} \sum_k \mathbb{E}_{p(\mathbf{z}_{\text{nero}}|\mathbf{x}^{(k)})} \log p(y_{\text{guide}}^{(k)}|\mathbf{z}_{\text{nero}}^{(k)}), \end{aligned} \quad (9)$$

where the inequality again follows from Jensen's inequality. The guider and the selector-predictor are trained jointly.

3.1.4. Information Bottleneck

To guide the model to select the least-but-enough information, we employ an information bottleneck technique [78]. We aim to minimize $I(\mathbf{x}, \tilde{\mathbf{z}}_{\text{sym}}) - I(\tilde{\mathbf{z}}_{\text{sym}}, y)^2$, where the former term encourages the selection of few features, and the latter term encourages the selection of the necessary features. As $I(\tilde{\mathbf{z}}_{\text{sym}}, y)$ is implemented by L_{sp} (the proof is given in Appendix A.1), we only need to minimize the mutual information $I(\mathbf{x}, \tilde{\mathbf{z}}_{\text{sym}})$:

$$I(\mathbf{x}, \tilde{\mathbf{z}}_{\text{sym}}) = \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{z}}_{\text{sym}}} \left[\log \frac{p(\tilde{\mathbf{z}}_{\text{sym}}|\mathbf{x})}{p(\tilde{\mathbf{z}}_{\text{sym}})} \right]. \quad (10)$$

However, there is a time-consuming term $p(\tilde{\mathbf{z}}_{\text{sym}}) = \sum_{\mathbf{x}} p(\tilde{\mathbf{z}}_{\text{sym}}|\mathbf{x})p(\mathbf{x})$, which needs to be calculated by a loop over all the instances \mathbf{x} in the training set. Inspired by Li and Eisner (2019), we replace this term with a variational distribution $r_{\phi}(z)$ and obtain an upper bound of Eq. 10: $I(\mathbf{x}, \tilde{\mathbf{z}}_{\text{sym}}) \leq \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{z}}_{\text{sym}}} \left[\log \frac{p(\tilde{\mathbf{z}}_{\text{sym}}|\mathbf{x})}{r_{\phi}(z)} \right]$. Since $\tilde{\mathbf{z}}_{\text{sym}}$ is a sequence of binary-selected features, we sum up the mutual information term of each element of $\tilde{\mathbf{z}}_{\text{sym}}$ as the information bottleneck loss:

$$L_{\text{ib}} = \sum_i \sum_{\tilde{z}_i} p(\tilde{z}_i|\mathbf{x}) \log \frac{p(\tilde{z}_i|\mathbf{x})}{r_{\phi}(z_i)}, \quad (11)$$

² $I(a, b) = \int_a \int_b p(a, b) \log \frac{p(a, b)}{p(a)p(b)} = \mathbb{E}_{a, b} [\frac{p(a|b)}{p(a)}]$ denotes the mutual information between the variables a and b .

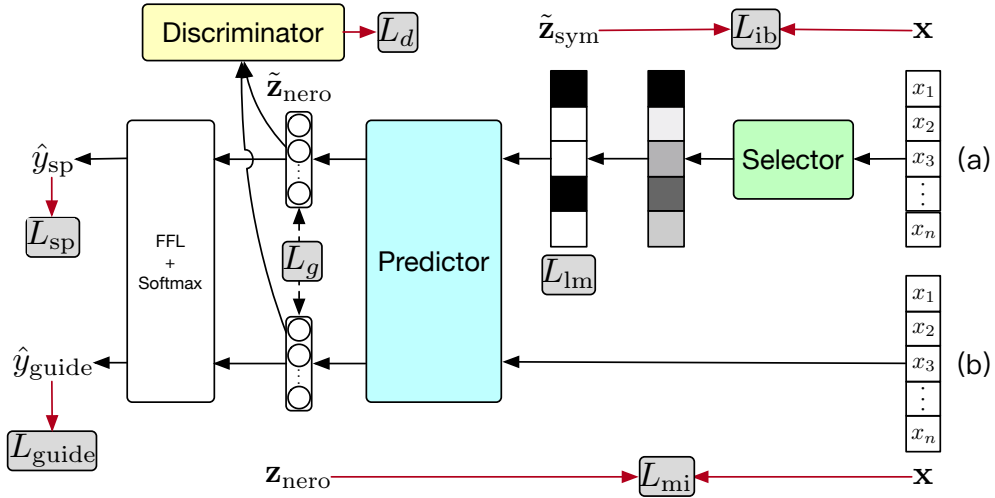


Figure 2: Architecture of InfoCal: the grey round boxes stand for the losses, and the red arrows indicate the data required for the calculation of the losses. FFL is an abbreviation for feed-forward layer.

where \tilde{z}_i represents whether to select the i -th feature: 1 for selected, 0 for not selected.

To encourage \mathbf{z}_{nero} to contain the least-but-enough information in the guider model, we again use the information bottleneck technique. Here, we minimize $I(\mathbf{x}, \mathbf{z}_{\text{nero}}) - I(\mathbf{z}_{\text{nero}}, y)$. Again, $I(\mathbf{z}_{\text{nero}}, y)$ can be implemented by L_{guide} . Due to the fact that \mathbf{z}_{nero} is sampled from a Gaussian distribution, the mutual information has a closed-form upper bound:

$$L_{\text{mi}} = I(\mathbf{x}, \mathbf{z}_{\text{nero}}) \leq \mathbb{E}_{\mathbf{z}_{\text{nero}}} \left[\log \frac{p(\mathbf{z}_{\text{nero}}|\mathbf{x})}{p(\mathbf{z}_{\text{nero}})} \right] = 0.5(\mu^2 + \sigma^2 - 1 - 2 \log \sigma). \quad (12)$$

The derivation is in Appendix A.2.

3.2. Calibrating Key Features via Adversarial Training

Our goal is to inform the selector what kind of information is still missing or has been wrongly selected. Since we already use the information bottleneck principal to encourage \mathbf{z}_{nero} to encode the information from the least-but-enough features, if we also require $\tilde{\mathbf{z}}_{\text{nero}}$ and \mathbf{z}_{nero} to encode the same information, then we would encourage the selector to select the least-but-enough discrete features. To achieve this, we use an adversarial-based training method. Thus, we employ an additional discriminator neural module, called D , which takes as input either $\tilde{\mathbf{z}}_{\text{nero}}$ or \mathbf{z}_{nero} and outputs label “0” or label “1”, respectively. The discriminator can be any differentiable neural network. The generator in our model is formed by the selector-predictor that outputs $\tilde{\mathbf{z}}_{\text{nero}}$. The losses associated with the generator and discriminator are:

$$L_d = -\log D(\mathbf{z}_{\text{nero}}) + \log D(\tilde{\mathbf{z}}_{\text{nero}}) \quad (13)$$

$$L_g = -\log D(\tilde{\mathbf{z}}_{\text{nero}}). \quad (14)$$

Yoon et al. [142] also attempted to use guidance from a so-called “base” model to a selector-predictor model. Nevertheless, their “base” model can only provide valid information calibration in actor-critic reinforcement learning, which is difficult to provide in POMDP problems [59]. In comparison, the discriminator in our method is more flexible in providing valid information calibration.

3.3. Regularizing Rationales with Language Models

For NLP tasks, it is often desirable that a rationale is formed of fluent subphrases [75]. To this end, previous works propose regularizers that bind the adjacent tokens to make them be simultaneously sampled or not. For example, Lei et al. (2016) proposed a non-differentiable regularizer trained using REINFORCE [137]. To make the method differentiable, Bastings et al. (2019) used the Kumaraswamy-distribution for the regularizer. However, they treat all

pairs of adjacent tokens in the same way, even though some adjacent tokens have more priority to be bound than others, such as “He stole” or “the victim” rather than “. He” or “) in” in Fig. 1.

We propose a novel differentiable regularizer for extractive rationales that is based on a pre-trained language model, thus encouraging both the consecutiveness and the fluency of the tokens in the extracted rationale. The LM-based regularizer is implemented as follows:

$$L_{lm} = - \sum_i m_{i-1} \log p_{lm}(m_i x_i | \mathbf{x}_{<i}), \quad (15)$$

where the m_i 's are the masks obtained in Eq. 2. Note that non-selected tokens are masked instead of deleted in this regularizer. The language model can have any architecture.

First, we note that L_{lm} is differentiable. Second, the following theorem guarantees that L_{lm} encourages consecutiveness of selected tokens.

Theorem 1. *If the following is satisfied for all i, j :*

- $m'_i < \epsilon \ll 1 - \epsilon < m_i$, $0 < \epsilon < 1$, and
- $\left| p(m'_i x_i | x_{<i}) - p(m'_j x_j | x_{<j}) \right| < \epsilon$,

then the following two inequalities hold:

- (1) $L_{lm}(\dots, m_k, \dots, m'_n) < L_{lm}(\dots, m'_k, \dots, m_n)$.
- (2) $L_{lm}(m_1, \dots, m'_k, \dots) > L_{lm}(m'_1, \dots, m_k, \dots)$.

The theorem says that for the same number of selected tokens, if they are consecutive, then they will get a lower L_{lm} value. Its proof is given in Appendix A.3.

3.3.1. Language Model in Continuous Form

Conventional language models are in discrete-form, which usually generate a multinomial distribution for each token, and minimize the Negative Log-likelihood (NLL) loss. The probability of the expected token is computed as follows:

$$p(x_i | x_{<i}) = \frac{\exp(h_i^\top e_i)}{\sum_{j \in \mathcal{V}} \exp(h_i^\top e_j)}, \quad (16)$$

where h_i is the hidden vector corresponding to x_i , e_i is a trainable parameter which represents the output vector of x_i , and \mathcal{V} is the vocabulary. In language model literature [71, 10], x_i is a symbolic token, and each token in \mathcal{V} has a corresponding trainable output vector. Eqn. 16 is a Softmax operation which normalizes throughout the whole vocabulary.

Note that in Eq. 15, the target sequence of the language model $P(m_i x_i | x_{<i})$ is formed of vectors instead of symbolic tokens. Since $m_i x_i$ is not symbolic token, it do not have a corresponding trainable output vector so that we cannot use a Softmax-like operation to normalize throughout the whole vocabulary. To tackle this, we require a continuous-form language model. Therefore, we make some small changes in the pre-training of the language model. When we are modeling the language model in vector form, we only use a bilinear layer to directly calculate the probability in Eq. 16:

$$p(x_i | x_{<i}) = \sigma(h_i^\top M e_i), \quad (17)$$

where σ stands for sigmoid, and M is a trainable parameter matrix. The sigmoid operation ensures the result lies in $[0, 1]$, which is a probability value. Then the probability value of $P(m_i x_i | x_{<i})$ is computed by:

$$p(m_i x_i | x_{<i}) = \sigma(h_i^\top M(m_i e_i)). \quad (18)$$

However, without normalization operations like Softmax, what Eqn. 17 computes is a quasi-probability value which relates to only one token. To solve this issue, we use negative sampling [89] in the training procedure. Therefore, the language model is pretrained using the following loss:

$$L_{pre} = - \sum_i \left[\log \sigma(h_i^\top M e_i) - \mathbb{E}_{j \sim p(x_j)} \log \sigma(h_i^\top M e_j) \right], \quad (19)$$

where $p(x_j)$ is the occurring probability (in the training dataset) of token x_j .

3.4. Training and Inference

The total loss function of our model, which takes the generator's role in adversarial training, is shown in Eq. 21. The adversarial-related losses are denoted by L_{adv} . The discriminator is trained by L_d from Eq. 13.

$$L_{adv} = \lambda_g L_g + L_{guide} + \lambda_{mi} L_{mi} \quad (20)$$

$$J_{total} = L_{sp} + \lambda_{ib} L_{ib} + L_{adv} + \lambda_{lm} L_{lm}, \quad (21)$$

where λ_{ib} , λ_g , λ_{mi} , and λ_{lm} are hyperparameters.

At training time, we optimize the generator loss J_{total} and discriminator loss L_d alternately until convergence. At inference time, we run the selector-predictor model to obtain the prediction and the rationale $\tilde{\mathbf{z}}_{sym}$.

The whole training process is illustrated in Algorithm 1.

Algorithm 1: Training process of InfoCal.

Random initialization;

Pre-train language model by Eq. 19;

for each iteration $i = 1, 2, \dots$ **do**

for each batch **do**

 Calculate the loss J_{total} for the sampler-predictor model and the guider model by Eq. 21;

 Calculate the loss L_D for the discriminator by Eq. 13;

 Update the parameters of selector-predictor model and the guider model;

 Update the parameters of the discriminator;

end

end

4. Experiments

We performed experiments on three NLP applications: multi-aspect sentiment analysis, legal judgement prediction, and hate speech detection. For multi-aspect sentiment analysis and hate speech detection, we have rationale annotations in the dataset. So, we can directly use automatic evaluation metrics to evaluate the quality of extracted rationales. For legal judgement prediction, there is no rationale annotation, so we conduct human evaluation for the extracted rationales.

4.1. Evaluation Metrics for Rationales.

With the annotations of rationales in the multi-aspect sentiment analysis and hate speech detection datasets, we would like to evaluate the explainability of our model. For better comparison, we use the same evaluation metrics with previous works [29, 87], which contains 5 metrics as listed below.

- IOU F_1 : This metric is defined upon a token-level partial match score Intersection-Over-Union (IOU). For two spans a and b , IOU is the quotient of the number of their overlapped tokens and the number of their union tokens: $\text{IOU} = \frac{|a \cap b|}{|a \cup b|}$. If the IOU value between a rationale prediction and a ground truth rationale is above 0.5, we consider this prediction as correct. Then, the F_1 score is calculated accordingly as the IOU F_1 .
- Token P , Token R , Token F_1 : For two spans, prediction rationale span a and ground-truth rationale span b , token-level precision is the quotient of the number of their overlapped tokens and the number of tokens in the prediction rationale span: $P_{\text{token}} = \frac{|a \cap b|}{|a|}$. The token-level recall is the quotient of the number of their overlapped tokens and the number of tokens in the ground-truth rationale span: $R_{\text{Token}} = \frac{|a \cap b|}{|b|}$. Then, token F_1 is calculated as $\frac{2P_{\text{token}}R_{\text{Token}}}{P_{\text{token}}+R_{\text{Token}}}$.
- AUPRC: This metric is the area under the precision (P_{token})-recall (R_{Token}) curve. The calculate method is sweeping the threshold over the token-level scores.

- **Comprehensiveness (Comp.):** This metric means to judge whether the selected rationale is complete. To calculate this, we create a contrast example for each example by removing the rationale \mathbf{z}_{sym} from the original input \mathbf{x} , denoted by $\mathbf{x}/\mathbf{z}_{\text{sym}}$. After removing the rationales, the model should become less confident to the original predicted class y . We then measure comprehensiveness as follows: $\text{Comp.} = p(y|\mathbf{x}) - p(y|\mathbf{x}/\mathbf{z}_{\text{sym}})$. A high comprehensiveness score suggest that the extracted rationale is indeed complete for the prediction.
- **Sufficiency (Suff.):** This metric means to judge whether the selected rationale is useful. Similar to the comprehensiveness score, we calculate the sufficiency score as: $\text{Suff.} = p(y|\mathbf{x}) - p(y|\mathbf{z}_{\text{sym}})$. If the extracted rationale is indeed useful, then the sufficiency score should be very small.

Among them, Token P , Token R , Token F_1 , IOU F_1 , and AUPRC requires the gold rationale annotations, so we just calculate these metrics in the beer review task and the hate speech explanation task. Comp. and Suff. only fit for classification problems, so we just apply these metrics to legal judgment prediction task and hate speech explanation task.

4.2. Beer Reviews

4.2.1. Data.

To provide a quantitative analysis for the extracted rationales, we use the BeerAdvocate³ dataset [88]. This dataset contains instances of human-written multi-aspect reviews on beers. Similarly to Lei et al. [75], we consider the following three aspects: appearance, smell, and palate. McAuley et al. (2012) provide manually annotated rationales for 994 reviews for all aspects, which we use as test set.

The training set of BeerAdvocate contains 220,000 beer reviews, with human ratings for each aspect. Each rating is on a scale of 0 to 5 stars, and it can be fractional (e.g., 4.5 stars), Lei et al. (2016) have normalized the scores to [0, 1], and picked “less correlated” examples to make a de-correlated subset.⁴ For each aspect, there are 80k–90k reviews for training and 10k reviews for validation.

4.2.2. Model details.

Because our task is a regression, we make some modifications to our model. First, we replace the *softmax* in Eq. 4 by the *sigmoid* function, and replace the cross-entropy loss in Eq. 5 by a mean-squared error (MSE) loss. Second, for a fair comparison, similar to Lei et al. (2016) and Bastings et al. (2019), we set all the architectures of selector, predictor, and guider as bidirectional Recurrent Convolution Neural Network (RCNN) Lei et al. (2016), which performs similarly to an LSTM [51] but with 50% fewer parameters.

We search the hyperparameters in the following scopes: $\lambda_{\text{ib}} \in (0.000, 0.001]$ with step 0.0001, $\lambda_g \in [0.2, 2.0]$ with step 0.2, $\lambda_{\text{mi}} \in [0.0, 1.0]$ with step 0.1, and $\lambda_{\text{im}} \in [0.000, 0.010]$ with step 0.001.

The best hyperparameters were found as follows: $\lambda_{\text{ib}} = 0.0003$, $\lambda_g = 1$, $\lambda_{\text{mi}} = 0.1$, and $\lambda_{\text{im}} = 0.005$.

We set $r_\phi(z_i)$ to $r_\phi(z_i = 0) = 0.999$ and $r_\phi(z_i = 1) = 0.001$.

4.2.3. Evaluation Metrics and Baselines.

For the evaluation of the selected tokens as rationales, we use precision, recall, and F1-score. Typically, precision is defined as the percentage of selected tokens that also belong to the human-annotated rationale. Recall is the percentage of human-annotated rationale tokens that are selected by our model. The predictions made by the selected rationale tokens are evaluated using the mean-square error (MSE).

We compare our method with the following baselines:

- **Attention [75]:** This method calculates attention scores over the tokens and selects top-k percent tokens as the rationale.
- **Bernoulli [75]:** This method uses a selector network to calculate a Bernoulli distribution for each token, and then samples the tokens from the distributions as the rationale. The basic architecture is RCNN Lei et al. (2016).
- **HardKuma [8]:** This method replaces the Bernoulli distribution by a Kuma distribution to facilitate differentiability. The basic architecture is also RCNN Lei et al. (2016).

³<https://www.beeradvocate.com/>

⁴<http://people.csail.mit.edu/taolei/beer/>

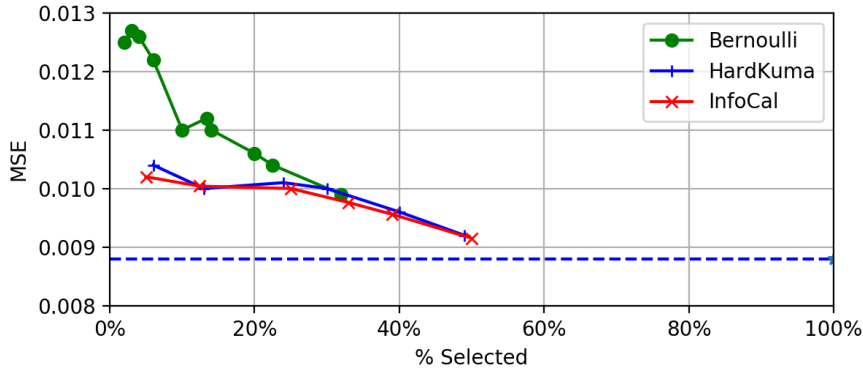


Figure 3: MSE of all aspects of BeerAdvocate. The blue dashed line represents the full-text baseline (all tokens are selected).

Method	Appearance					
	P	R	F	IOU F_1	% selected	AUPRC
Attention	80.6	35.6	49.4	32.8	13	0.613
Bernoulli	96.3	56.5	71.2	55.3	14	0.785
HardKuma	98.1	65.1	78.3	64.3	13	0.833
FRESH	96.5	53.2	68.6	52.2	13	0.772
Sparse IB	91.3	54.6	68.3	51.9	13	0.752
InfoCal	98.5	73.2	84.0	72.4	13	0.871

Method	Smell					
	P	R	F	IOU F_1	% selected	AUPRC
Attention	88.4	20.6	33.4	20.1	7	0.584
Bernoulli	95.1	38.2	54.5	37.5	7	0.697
HardKuma	96.8	31.5	47.5	31.2	7	0.675
FRESH	90.4	32.3	47.6	31.2	7	0.647
Sparse IB	90.8	34.5	50.0	33.3	7	0.659
InfoCal	95.6	45.6	61.7	44.7	7	0.733

Method	Palate					
	P	R	F	IOU F_1	% selected	AUPRC
Attention	65.3	35.8	46.2	30.1	7	0.537
Bernoulli	80.2	53.6	64.3	47.3	7	0.692
HardKuma	89.8	48.6	63.1	46.1	7	0.718
FRESH	78.4	50.2	61.2	44.1	7	0.668
Sparse IB	84.3	49.2	62.1	45.1	7	0.692
InfoCal	89.6	59.8	71.7	55.9	7	0.767

Table 2

Token-level precision (P), recall (R), F1-score (F), IOU F_1 , and AUPRC of selected rationales for the three aspects of BeerAdvocate. In bold, the best performance. “% selected” means the average percentage of tokens selected out of the total number of tokens per instance.

- FRESH [54]: This method breaks the selector-predictor model into three sub-components: a support model which calculates the importance of each input token, a rationale extractor model which extracts the rationale snippets according to the output of the support model, a classifier model which make prediction according to the extracted rationale.
- Sparse IB [90]: This method also uses information bottleneck to control the number of tokens selected by the rationale. But it did not use any information calibration methods or any regularizers to extract more complete and fluent rationales.

4.2.4. Results.

The rationale extraction performances are shown in Table 2. The precision values for the baselines are directly taken from [8]. We use their source code for the Bernoulli⁵ and HardKuma⁶ baselines.

We trained these baseline for 50 epochs and selected the models with the best recall on the dev set when the precision was equal or larger than the reported dev precision. For fair comparison, we used the same stopping criteria for InfoCal (for which we fixed a threshold for the precision at 2% lower than the previous state-of-the-art).

⁵<https://github.com/taolei87/rcnn>

⁶https://github.com/bastings/interpretable_predictions

Method	Appearance			Smell			Palate		
	P	R	F	P	R	F	P	R	F
InfoCal (HardKuma reg)	97.9	71.7	82.8	94.8	42.3	58.5	89.4	56.9	69.5
InfoCal (INVASE reg)	96.8	53.5	68.9	93.2	35.7	51.6	85.7	39.5	54.1
InfoCal- L_{adv}	97.3	67.8	79.9	94.3	34.5	50.5	89.6	51.2	65.2
InfoCal- L_{lm}	79.8	54.9	65.0	87.1	32.3	47.1	83.1	47.4	60.4
InfoCal	98.5	73.2	84.0	95.6	45.6	61.7	89.6	59.8	71.7

Table 3

The ablation tests. All the listed methods are tuned to select 13% words in “Appearance”, 7% in “Smell” and “Palate” to make them comparable with InfoCal. The best performances are bolded.

Gold	clear , burnished copper-brown topped by a large beige head that displays impressive persistence and leaves a small to moderate amount of lace in sheets when it eventually departs the nose is sweet and spicy and the flavor is malty sweet , accented nicely by honey and by abundant caramel/toffee notes . there alcohol . the mouthfeel is exemplary ; full and rich , very creamy . mouthfilling with some mouthcoating as well . drinkability is high
Bernoulli	clear , burnished copper-brown topped by a large beige head that displays impressive persistence and leaves a small to moderate amount of lace in sheets when it eventually departs the nose is sweet and spicy and the flavor is malty sweet , accented nicely by honey and by abundant caramel/toffee notes . there alcohol . the mouthfeel is exemplary ; full and rich , very creamy . mouthfilling with some mouthcoating as well . drinkability is high
HardKuma	clear , burnished copper-brown topped by a large beige head that displays impressive persistence and leaves a small to moderate amount of lace in sheets when it eventually departs the nose is sweet and spicy and the flavor is malty sweet , accented nicely by honey and by abundant caramel/toffee notes . there alcohol . the mouthfeel is exemplary ; full and rich , very creamy . mouthfilling with some mouthcoating as well . drinkability is high
InfoCal	clear , burnished copper-brown topped by a large beige head that displays impressive persistence and leaves a small to moderate amount of lace in sheets when it eventually departs the nose is sweet and spicy and the flavor is malty sweet , accented nicely by honey and by abundant caramel/toffee notes . there alcohol . the mouthfeel is exemplary ; full and rich , very creamy . mouthfilling with some mouthcoating as well . drinkability is high
InfoCal- L_{adv}	clear , burnished copper-brown topped by a large beige head that displays impressive persistence and leaves a small to moderate amount of lace in sheets when it eventually departs the nose is sweet and spicy and the flavor is malty sweet , accented nicely by honey and by abundant caramel/toffee notes . there alcohol . the mouthfeel is exemplary ; full and rich , very creamy . mouthfilling with some mouthcoating as well . drinkability is high
InfoCal- L_{lm}	clear , burnished copper-brown topped by a large beige head that displays impressive persistence and leaves a small to moderate amount of lace in sheets when it eventually departs the nose is sweet and spicy and the flavor is malty sweet , accented nicely by honey and by abundant caramel/toffee notes . there alcohol . the mouthfeel is exemplary ; full and rich , very creamy . mouthfilling with some mouthcoating as well . drinkability is high

Table 4

One example of extracted rationales by different methods. Different colors correspond to different aspects: red: appearance, green: smell, and blue: palate.

We also conducted ablation studies: (1) we removed the adversarial loss and report the results in the line InfoCal- L_{adv} , and (2) we removed the LM regularizer and report the results in the line InfoCal- L_{lm} .

In Table 2, we see that, although Bernoulli, HardKuma, FRESH, and Sparse IB achieve very high precisions, their recall scores are significantly low. The reason is that these four methods only focus on making the extracted rationale enough for a correct prediction, so the rationale is not necessary to be competent and many details would be lost. In comparison, our InfoCal method uses a dense neural network as a guider, which provided many detailed information. Therefore, the selector is able to extract more complete rationales.

In comparison, our method InfoCal significantly outperforms the previous methods in the recall scores for all the three aspects of the BeerAdvocate dataset (we performed Student’s t-test, $p < 0.01$). Also, all the three F-scores of InfoCal are a new state-of-the-art performance.

In the ablation studies in Table 3, we see that when we remove the adversarial information calibrating structure,

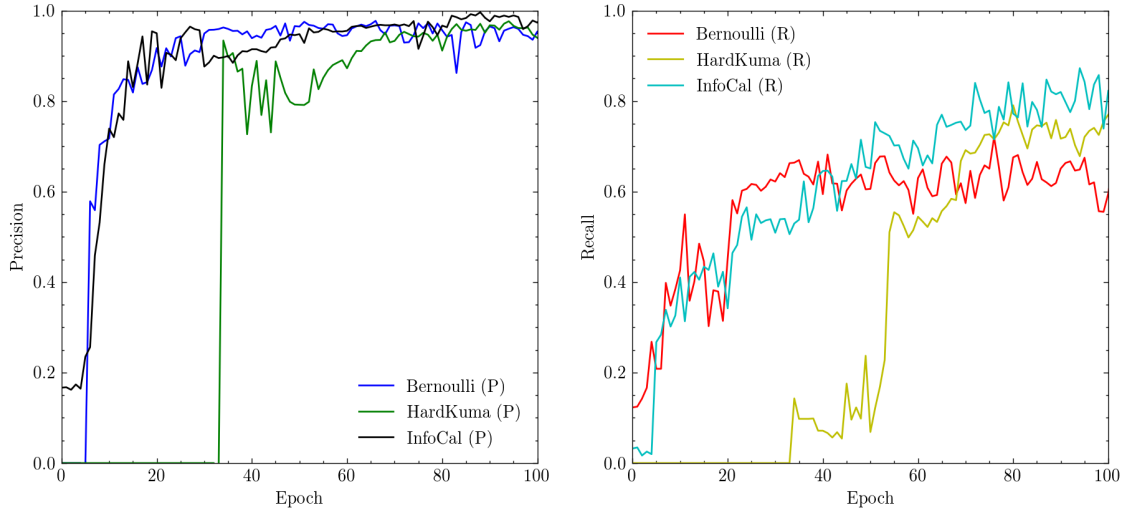


Figure 4: The precision (left) and recall (right) for rationales on the smell aspect of the BeerAdvocate valid set.

namely, for $\text{InfoCal}-L_{\text{adv}}$, the recall scores decrease significantly in all the three aspects. This shows that our guider model is critical for the increased performance. Moreover, when we remove the LM regularizer, we find a significant drop in both precision and recall, in the line $\text{InfoCal}-L_{\text{lm}}$. This highlights the importance of semantical fluency of rationales, which are encouraged by our LM regularizer.

We also apply another kind of calibration, which was applied in Yoon et al. (2018). This calibration method is very similar to the “base” model in actor-critic models [68]. Their difference with our InfoCal is that Yoon et al. (2018) minimizes the difference between the cross entropy values of the selector-predictor model and the base model. We apply their method to our model and listed the results in the InfoCal (INVASE reg) line in Table 3. We found that the recall score decreases a lot compared to InfoCal, which shows that our information calibration method is better for improving the recall of rationale extraction.

We also replace the LM regularizer with the regularizer used in the HardKuma method with all the other parts of the model unchanged, denoted $\text{InfoCal}(\text{HardKuma reg})$ in Table 3. We found that the recall and F-score of InfoCal outperforms $\text{InfoCal}(\text{HardKuma reg})$, which shows the effectiveness of our LM regularizer.

We further show the relation between a model’s performance on predicting the final answer and the rationale selection percentage (which is determined by the model) in Fig. 3, as well as the relation between precision/recall and training epochs in Fig. 4. The rationale selection percentage is influenced by λ_{ib} . According to Fig. 3, our method InfoCal achieves a similar prediction performance compared to previous works, and does slightly better than HardKuma for some selection percentages. Fig. 4 shows the changes in precision and recall with training epochs. We can see that our model achieves a similar precision after several training epochs, while significantly outperforming the previous methods in recall, which proves the effectiveness of our proposed method.

Table 4 shows an example of rationale extraction. Compared to the rationales extracted by Bernoulli and HardKuma, our method provides more fluent rationales for each aspect. For example, unimportant tokens like “and” (after “persistence”, in the Bernoulli method), and “with” (after “mouthful”, in the HardKuma method) were selected just because they are adjacent to important ones.

4.3. Legal Judgement Prediction

4.3.1. Datasets and Preprocessing.

We use the CAIL2018 dataset⁷ [150] for three tasks on legal judgment prediction. The dataset consists of criminal cases published by the Supreme People’s Court of China.⁸ To be consistent with previous works, we used two versions of CAIL2018, namely, CAIL-small (the exercise stage data) and CAIL-big (the first stage data). The statistics of CAIL2018 dataset are shown in Table 6.

⁷https://cail.oss-cn-qingdao.aliyuncs.com/CAIL2018_ALL_DATA.zip

⁸<http://cail.cipsc.org.cn/index.html>

Rationalizing Predictions by Adversarial Information Calibration

Small	Tasks	Law Articles					Charges					Terms of Penalty				
	Metrics	Acc	MP	MR	F1	%S	Acc	MP	MR	F1	%S	Acc	MP	MR	F1	%S
Single	Bernoulli (w/o)	0.812	0.726	0.765	0.756	100	0.810	0.788	0.760	0.777	100	0.331	0.323	0.297	0.306	100
	Bernoulli	0.755	0.701	0.737	0.728	14	0.761	0.753	0.739	0.754	14	0.323	0.308	0.265	0.278	30
	HardKuma (w/o)	0.807	0.704	0.757	0.739	100	0.811	0.776	0.763	0.776	100	0.345	0.355	0.307	0.319	100
	HardKuma	0.783	0.706	0.735	0.729	14	0.778	0.757	0.714	0.736	14	0.340	0.328	0.296	0.309	30
	FRESH	0.801	0.714	0.761	0.743	14	0.790	0.766	0.725	0.745	14	0.344	0.332	0.308	0.312	30
	Sparse IB	0.773	0.692	0.734	0.712	14	0.769	0.758	0.742	0.750	14	0.336	0.324	0.280	0.300	30
	InfoCal- L_{adv}	0.826	0.739	0.774	0.777	14	0.845	0.804	0.781	0.797	14	0.351	0.374	0.329	0.330	30
	InfoCal- L_{adv} - L_{rib} (w/o)	0.841	0.759	0.785	0.793	100	0.850	0.820	0.801	0.814	100	0.368	0.378	0.341	0.346	100
	InfoCal- L_{lm}	0.822	0.723	0.768	0.773	14	0.843	0.796	0.770	0.772	14	0.347	0.361	0.318	0.320	30
InfoCal	0.834	0.744	0.776	0.786	14	0.849	0.817	0.798	0.813	14	0.358	0.372	0.335	0.337	30	
Multi	FLA	0.803	0.724	0.720	0.714	-	0.767	0.758	0.738	0.732	-	0.371	0.310	0.300	0.299	-
	TOPJUDGE	0.872	0.819	0.808	0.800	-	0.871	0.864	0.851	0.846	-	0.380	0.350	0.353	0.346	-
	MPBFN-WCA	<u>0.883</u>	<u>0.832</u>	<u>0.824</u>	<u>0.822</u>	-	<u>0.887</u>	<u>0.875</u>	<u>0.857</u>	<u>0.859</u>	-	<u>0.414</u>	<u>0.406</u>	<u>0.369</u>	<u>0.392</u>	-
Big	Tasks	Law Articles					Charges					Terms of Penalty				
	Metrics	Acc	MP	MR	F1	%S	Acc	MP	MR	F1	%S	Acc	MP	MR	F1	%S
Single	Bernoulli (w/o)	0.876	0.636	0.388	0.625	100	0.857	0.643	0.410	0.569	100	0.509	0.511	0.304	0.312	100
	Bernoulli	0.857	0.632	0.374	0.621	14	0.848	0.635	0.402	0.543	14	0.496	0.505	0.289	0.306	30
	HardKuma (w/o)	0.907	0.664	0.397	0.627	100	0.907	0.689	0.438	0.608	100	0.555	0.547	0.335	0.356	100
	HardKuma	0.876	0.645	0.384	0.609	14	0.892	0.676	0.425	0.587	14	0.534	0.535	0.310	0.334	30
	FRESH	0.902	0.698	0.675	0.682	14	0.902	0.695	0.632	0.653	14	0.532	0.539	0.343	0.387	30
	Sparse IB	0.863	0.634	0.372	0.624	14	0.852	0.638	0.401	0.545	14	0.501	0.510	0.286	0.302	30
	InfoCal- L_{adv}	0.953	0.844	0.711	0.782	20	0.954	0.857	0.772	0.806	20	0.552	0.490	0.353	0.356	30
	InfoCal- L_{adv} - L_{rib} (w/o)	0.959	0.862	0.751	0.791	100	0.957	0.878	0.776	0.807	100	0.584	0.519	0.411	0.427	30
	InfoCal- L_{lm}	0.953	0.851	0.730	0.775	20	0.950	0.857	0.756	0.789	20	0.563	0.486	0.374	0.367	30
InfoCal	0.956	0.852	0.742	0.805	20	0.955	0.868	0.788	0.820	20	0.556	0.519	0.362	0.372	30	
Multi	FLA	0.942	0.763	0.695	0.746	-	0.931	0.798	0.747	0.780	-	0.531	0.437	0.331	0.370	-
	TOPJUDGE	0.963	0.870	0.778	0.802	-	0.960	0.906	0.824	0.853	-	0.569	0.480	0.398	0.426	-
	MPBFN-WCA	<u>0.978</u>	<u>0.872</u>	<u>0.789</u>	<u>0.820</u>	-	<u>0.977</u>	<u>0.914</u>	<u>0.836</u>	<u>0.867</u>	-	<u>0.604</u>	<u>0.534</u>	<u>0.430</u>	<u>0.464</u>	-

Table 5

The overall performance on the CAIL2018 dataset (Small and Big). The results from previous works are directly quoted from Yang et al. (2019), because we share the same experimental settings, and hence we can make direct comparisons. %S represents the selection percentage (which is determined by the model). "Single" represents single-task models, "Multi" represents multi-task models. The best performance is in bold. The red numbers mean that they are less than the best performance by no more than 0.01. The underlined numbers are the state-of-the-art performances, all of which are obtained by multi-task models. (w/o) represents that the corresponding model is a dense model without extracting rationales.

	CAIL-small	CAIL-big
Cases	113,536	1,594,291
Law Articles	105	183
Charges	122	202
Term of Penalty	11	11

Table 6

Statistics of the CAIL2018 dataset.

The instances in CAIL2018 consist of a *fact description* and three kinds of annotations: *applicable law articles*, *charges*, and *the penalty terms*. Therefore, our three tasks on this dataset consist of predicting (1) law articles, (2) charges, and (3) terms of penalty according to the given fact description.

In the dataset, there are also many cases with multiple applicable law articles and multiple charges. To be consistent with previous works on legal judgement prediction [150, 140], we filter out these multi-label examples.

We also filter out instances where the charges and law articles occurred less than 100 times in the dataset (e.g., insulting the national flag and national emblem). For the term of penalty, we divide the terms into 11 non-overlapping intervals. These preprocessing steps are the same as in Zhong et al. (2018) and Yang et al. (2019), making it fair to compare our model with previous models.

	Law Articles		Charges		Terms of Penalty	
	Comp.↑	Suff.↓	Comp.↑	Suff.↓	Comp.↑	Suff.↓
Bernoulli	0.231	0.005	0.243	0.002	0.132	0.017
HardKuma	0.304	-0.021	0.312	-0.034	0.165	0.009
InfoCal	0.395	-0.056	0.425	-0.067	0.203	0.005

Table 7

The quantitative evaluation of rationales for legal judgment prediction. The “↑” means that a good result should have a larger value, while “↓” means lower is better.

We use Jieba⁹ for token segmentation, because this dataset is in Chinese. The word embedding size is set to 100 and is randomly initiated before training. The maximum sequence length is set to 1000. The architectures of the selector, predictor, and guider are all bidirectional LSTMs. The LSTM’s hidden size is set to 100. $r_\phi(z_i)$ is the sampling rate for each token (0 for selected), which we set to $r_\phi(z_i = 0) = 0.9$.

We search the hyperparameters in the following scopes: $\lambda_{ib} \in [0.00, 0.10]$ with step 0.01, $\lambda_g \in [0.2, 2.0]$ with step 0.2, $\lambda_{mi} \in [0.0, 1.0]$ with step 0.1, $\lambda_{lm} \in [0.000, 0.010]$ with step 0.001. The best hyperparameters were found to be: $\lambda_{ib} = 0.05$, $\lambda_g = 1$, $\lambda_{mi} = 0.5$, $\lambda_{lm} = 0.005$ for all the three tasks.

4.3.2. Overall Performance.

We again compare our method with the Bernoulli [75] and the HardKuma [8] methods on rationale extraction. These two methods are both single-task models, which means that we train a model separately for each task. We also compare our method with three multi-task methods listed as follows:

- FLA [83] uses an attention mechanism to capture the interaction between fact descriptions and applicable law articles.
- TOPJUDGE [150] uses a topological architecture to link different legal prediction tasks together, including the prediction of law articles, charges, and terms of penalty.
- MPBFN-WCA [140] uses a backward verification to verify upstream tasks given the results of downstream tasks.

The results are listed in Table 5.

On CAIL-small, we observe that it is more difficult for the single-task models to outperform multi-task methods. This is likely due to the fact that the tasks are related, and learning them together can help a model to achieve better performance on each task separately. After removing the restriction of the information bottleneck, InfoCal- L_{adv} - L_{ib} achieves the best performance in all tasks, however, it selects all the tokens in the review. When we restrict the number of selected tokens to 14% (by tuning the hyperparameter λ_{ib}), InfoCal (in red) only slightly drops in all evaluation metrics, and it already outperforms Bernoulli and HardKuma, even if they have used all tokens. This means that the 14% selected tokens are very important to the predictions. We observe a similar phenomenon for CAIL-big. Specifically, InfoCal outperforms InfoCal- L_{adv} - L_{ib} in some evaluation metrics, such as the F1-score of law article prediction and charge prediction tasks.

4.3.3. Rationales.

The CAIL2018 dataset does not contain annotations of rationales. So, we only use Comp. and Suff. for quantitative evaluation since they do not require gold rationale annotations. The results are shown in Table 7. We can see that in all the three subtasks of legal judgement prediction, our proposed method outperforms the previous methods.

We also conducted human evaluation for the extracted rationales. Due to limited budget and resources, we sampled 300 examples for each task. We randomly shuffled the rationales for each task and asked six undergraduate students from Peking University to evaluate them. The human evaluation is based on three metrics: usefulness (U), completeness (C), and fluency (F); each scored from 1 (lowest) to 5. The scoring standard for human annotators is given in Appendix C in the extended paper.

The human evaluation results are shown in Table 8. We can see that our proposed method outperforms previous methods in all metrics. Our inter-rater agreement is acceptable by Krippendorff’s rule (2004), which is shown in Table 8.

⁹<https://github.com/fxsjy/jieba>

	Law			Charges			ToP		
	U	C	F	U	C	F	U	C	F
Bernoulli	4.71	2.46	3.45	3.67	2.35	3.45	3.35	2.76	3.55
HardKuma	4.65	3.21	3.78	4.01	3.26	3.44	3.84	2.97	3.76
InfoCal	4.72	3.78	4.02	4.65	3.89	4.23	4.21	3.43	3.97
α	0.81	0.79	0.83	0.92	0.85	0.87	0.82	0.83	0.94

Table 8

Human evaluation on the CAIL2018 dataset. "ToP" is the abbreviation of "Terms of Penalty". The metrics are: usefulness (U), completeness (C), and fluency (F), each scored from 1 to 5. Best performance is in bold. α represents Krippendorff's alpha values. The basic architecture for the three methods are all RCNN Lei et al. (2016).

The People's Procuratorate of Yongshun County alleged that on January 11, 2014, the defendant Li XX and Peng XX (a separate case dealt with) **forcibly had sexual relations with the victim Zou XX** in a room of Xindu Hotel in Yongshun County . In this regard, the public prosecution agency cited the following evidence: capture history, household registration certificate, call list, description of the situation; identification transcripts; on-site inspection transcripts and on-site photos; physical evidence inspection reports and physical evidence identification documents; witnesses Liu A, Liu B, Testimony of Liu C, Zou XX, Du XX; confession and defense of defendant Li XX; audio-visual materials. The court held that the defendant Li XX **used violence and verbal threats** with others to **forcibly have sexual relations with the victim Zou XX in the Xindu Hotel room** in Yongshun County. His behavior has violated the Item (4) of the Criminal Law of the PRC, the facts of the crime are clear, and the evidence is reliable and sufficient, and the criminal responsibility should be investigated for the crime of $\times \times$. In the joint crime, the defendant Li XX **played the main role** and was the principal offender.....

Figure 5: An example of extracted rationale for charge prediction. The correct charge is "Rape". The original fact description is in Chinese, we have translated it to English. It is easy to see that the extracted rationales are very helpful in making the charge prediction.

A sample case of extracted rationales in legal judgement is shown in Fig. 5. We observe that our method selects all the useful information for the charge prediction task, and the selected rationales are formed of continuous and fluent sub-phrases.

4.4. Hate Speech Explanation

4.4.1. Datasets and Preprocessing.

For evaluating the performance of our method on hate speech detection task. We use the HateXplain dataset¹⁰ [87]. This dataset contains 9,055 posts from Twitter [26, 34] and 11,093 posts from Gab [79, 86, 145]. There are three different classes in this dataset: hateful, offensive, and normal. Apart from the class labels, this dataset also contains rationale annotations for each example that is labelled as hateful or offensive. The training set, valid set, and test set are already split as 8 : 1 : 1 in the dataset. More details of this dataset is shown in Table 9. This dataset is very noisy, and it can test the robustness of our InfoCal method on noisy text information.

For classification performance, we have three metrics: Accuracy, Macro F_1 , and AUROC. These metrics are used for evaluating the ability of distinguish among the three classes, i.e., hate speech, offensive speech, and normal. Among them, AUROC is the area under the ROC curve.

4.4.2. Competing Methods.

We also compare our method with Bernoulli [75] and HardKuma [8] in this experiment. We also compare our method with the following competing methods provided in Mathew et al. (2020b):

¹⁰<https://github.com/punyajoy/HateXplain.git>

- **CNN-GRU** [148] has achieved state-of-the-art performance in multiple hate speech datasets. CNN-GRU first use convolution neural network (CNN) [74] to capture the local features and then use recurrent neural network (RNN) [103] with GRU unit [20] to capture the temporal information. Finally, this model max-pools GRU’s hidden layers to a feature vector, and then use a fully connected layer to finally output the prediction results.
- **BiRNN** [107] first input the tokens into a sequential model with long-short term memory (LSTM) [51]. Then, the last hidden state is passed through two feed-forward layers and then a fully connected layer for prediction.
- **BiRNN-Attn** adds an attention layer after the sequential layer of BiRNN model.
- **BERT** [28] is a large pretrained model constructed by a stack of transformer [134] encoder layers. A fully connected layer is added to the output corresponding to the *CLS* token for the hate speech class prediction. We used the `bert-base-uncased` model with 12-layer, 768- hidden, 12-heads, 110M parameters, this is the same setting with previous work [86]. The model is fine-tuned using the HateXplain training set.

In all the above methods, the rationales are extracted by two methods: attention [102] and LIME [100]. When we are using attention method, as is described in DeYoung et al. (2020), the tokens with top 5 attention values are selected as rationale. The LIME method selects rationales by training a new explanation model to imitate the original deep learning “black-box” model. Different from these methods, our model InfoCal as well as the other two competing method Bernoulli and HardKuma are extracting rationales by the model itself without any external methods (like attention selection or LIME selection) for rationale selection. So, it is much more challenging for them to achieve similar explainability performance.

In Mathew et al. (2020b), the ground-truth rationale annotations were also used to train some models by adding an external cross entropy loss on the attention layer. The rationale training is conducted on BiRNN and BERT models, denoted as BiRNN-HateXplain and BERT-HateXplain, respectively.

4.4.3. Results.

The overall results are shown in Table 10. We can see that in the classification performance, the BERT models achieved the highest score in all the three metrics (Accuracy, Macro F_1 , and AUROC) no matter whether the rationale supervising is conducted. Also, our InfoCal model has outperformed all the other approaches except for BERT. This makes sense because BERT has pretrained by a large amount of texts, and it has a much better understanding for language than other models without pretraining.

In the explainability evaluations, our model InfoCal has achieved the state-of-the-art performance in three metrics: IOU F_1 , AUPRC, and Sufficiency. Also, for the other two metrics (Token F_1 and Comprehensiveness), the InfoCal method is comparable with the state-of-the-art method (BERT [Attn]). Note that in our model, the rationales are selected by the model itself instead of by selecting top 5 attention value or by LIME method externally. Therefore, this experimental result show that our InfoCal model is a better model for explaining neural network predictions.

We also listed the performances of the BiRNN model and BERT model after supervised by rationale annotations in Table 10. We can see that both the classification performance and the explainability performance improved a lot after trained by rationale annotations. This also makes sense because the rationale annotation is the most direct training signal of rationale selection. However, such kind of rationale annotation is very expensive to get in real-world applications. Therefore, the rationale extraction methods without rationale supervision is much proper to be applied in the industry.

4.4.4. Case Study for Rationales.

In Table 11, we have listed some of the generated rationales in HateXplain dataset by our InfoCal method and the two competing methods: Bernoulli and HardKuma. We can see that our InfoCal method has extracted nearly all of the annotated rationales in the ground-truth. Compared to Bernoulli and HardKuma, our InfoCal method do not extract nonsense rationales, such as “yeah i also” in the second line, and “precinct and campaign meetings” in the third line. This again shows the effectiveness of the information calibration method.

4.5. Performance of the Pretrained Language Model for the rationale regularizer

In the InfoCal model, we need a pretrained language model (in Sec. 3.3) for the rationale regularizer. Our language model described in Section 3.3.1 is different from previous language model because it has to compute probabilities for token’s vector representations instead of token’s symbolic IDs. Therefore, the quality of the pretrained language model

	Twitter	Gab	Total
Hateful	708	5,227	5,935
Offensive	2,328	3,152	5,480
Normal	5,770	2,044	7,814
Undecided	249	670	919
Total	9,055	11,093	20,148

Table 9

The statistics of HateXplain dataset. “Undecided” means that in the annotation process, all the three annotators gave different labels to the example. We omit this part of data in our experiments as is consistent with previous works.

		Classification Performance			Explanability				
		Acc ↑	Macro F1 ↑	AUROC ↑	IOU F1 ↑	Token F1 ↑	AUPRC ↑	Comp ↑	Suff ↓
W/o rationale supervising	CNN-GRU [LIME]	0.627	0.606	0.793	0.167	0.385	0.648	0.316	-0.082
	BiRNN [LIME]	0.595	0.575	0.767	0.162	0.361	0.605	0.421	-0.051
	BiRNN-Attn [Attn]	0.621	0.614	0.795	0.167	0.369	0.643	0.278	0.001
	BiRNN-Attn [LIME]	0.621	0.614	0.795	0.162	0.386	0.650	0.308	-0.075
	BERT [Attn]	0.690	0.674	0.843	0.130	0.497	0.778	0.447	0.057
	BERT [LIME]	0.690	0.674	0.843	0.118	0.468	0.747	0.436	0.008
	Bernoulli	0.597	0.568	0.765	0.138	0.482	0.668	0.324	0.003
	HardKuma	0.594	0.570	0.772	0.152	0.485	0.672	0.406	-0.022
	Sparse IB	0.602	0.572	0.768	0.145	0.486	0.670	0.389	0.001
	InfoCal	0.630	0.614	0.792	0.206	0.493	0.680	0.436	-0.097
With rationale supervising	BiRNN-HateXplain [Attn]	0.629	0.629	0.805	<u>0.222</u>	<u>0.506</u>	<u>0.841</u>	0.281	0.039
	BiRNN-HateXplain [LIME]	0.629	0.629	0.805	<u>0.174</u>	<u>0.407</u>	<u>0.685</u>	0.343	-0.075
	BERT-HateXplain [Attn]	<u>0.698</u>	<u>0.687</u>	<u>0.851</u>	0.120	0.411	0.626	0.424	0.160
	BERT-HateXplain [LIME]	<u>0.698</u>	<u>0.687</u>	<u>0.851</u>	0.112	0.452	0.722	<u>0.500</u>	0.004

Table 10

The overall performance on the HateXplain dataset. The results from previous work are directly quoted from Mathew et al. (2020b), because we share identical train/valid/test data split, and hence we can make direct comparison. The “↑” means that a good result should have a larger value, while “↓” means lower is better. The square bracket means the method of extracting rationales. Note that except for Bernoulli, HardKuma, Sparse IB, and InfoCal, all the other methods are extracting rationale by external methods ([Attn] or [LIME]). The basic architectures for Bernoulli, HardKuma, and InfoCal are all RCNN.

is paramount to the InfoCal model. In Table 12, we listed the comparison of the perplexity between our language model and two famous language models: Kenneth Heafield’s language model (KenLM) [45] and recurrent neural network language model (RNNLM) [10, 132]. The training is conducted on the pure texts of the training data in the three tasks, and the trained models are tested on the pure texts of the corresponding test sets. We can see that the perplexity of our language model is comparable to RNNLM and even better than kenLM in some datasets. This shows that the performance of our language model is acceptable to our experiments. We do not compare the perplexity with Transformer-based models like GPT [95, 96, 12], because these models usually use subword vocabularies (like Byte Pair Encoding (BPE) [96] and WordPiece [106, 28]) which makes the perplexities not comparable with our work.

Also, from the comparison of perplexity score, we found that the perplexity of HateXplain dataset is obviously higher than the other two datasets, this shows that HateXplain dataset is very noisy. The results in Table 10 proves that our InfoCal model is able to extract sensitive rationales on noisy text data.

5. Summary and Outlook

In this work, we proposed a novel method to extract rationales for neural predictions. Our method uses an adversarial-based technique to make a selector-predictor model learn from a guider model. In addition, we proposed a novel regularizer based on language models, which makes the extracted rationales semantically fluent. In this way, the “guider” model tells the selector-predictor model what kind of information (token) remains unselected or over-selected. We

Gold	Bernoulli	HardKuma	InfoCal
jew faggot milo is king of the alt right great it not the alt right anymore	jew faggot milo is king of the alt right great it not the alt right anymore	jew faggot milo is king of the alt right great it not the alt right anymore	jew faggot milo is king of the alt right great it not the alt right anymore
yeah i also agree that the middle eastern ones are coming mostly to hungary but these other refugees away from mudslimes are all schengen regions why bother applying for refugee status when you can just move it perfectly legal	yeah i also agree that the middle eastern ones are coming mostly to hungary but these other refugees away from mudslimes are all schengen regions why bother applying for refugee status when you can just move it perfectly legal	yeah i also agree that the middle eastern ones are coming mostly to hungary but these other refugees away from mudslimes are all schengen regions why bother applying for refugee status when you can just move it perfectly legal	yeah i also agree that the middle eastern ones are coming mostly to hungary but these other refugees away from mudslimes are all schengen regions why bother applying for refugee status when you can just move it perfectly legal
precinct and campaign meetings will confirm that it is only whites who hold any ideals about government non whites want free govt cheese and jews are crazy for power	precinct and campaign meetings will confirm that it is only whites who hold any ideals about government non whites want free govt cheese and jews are crazy for power	precinct and campaign meetings will confirm that it is only whites who hold any ideals about government non whites want free govt cheese and jews are crazy for power	precinct and campaign meetings will confirm that it is only whites who hold any ideals about government non whites want free govt cheese and jews are crazy for power
the labour party has number mps in prison for paedophilia number mps and councillors on the sex registry number million white kids raped by paki muslims all labour voters covered up by labour councils mps and a pm	the labour party has number mps in prison for paedophilia number mps and councillors on the sex registry number million white kids raped by paki muslims all labour voters covered up by labour councils mps and a pm	the labour party has number mps in prison for paedophilia number mps and councillors on the sex registry number million white kids raped by paki muslims all labour voters covered up by labour councils mps and a pm	the labour party has number mps in prison for paedophilia number mps and councillors on the sex registry number million white kids raped by paki muslims all labour voters covered up by labour councils mps and a pm

Table 11

Example rationales extracted by three methods: Bernoulli, HardKuma, and InfoCal. Note that in these cases, many phrases are offensive or hateful. Nevertheless, this cannot be avoided due to the nature of the work.

	KenLM [45]	RNNLM [10, 132]	Our LM
Perplexity (Beer)	66	50	44
Perplexity (Legal Small)	32	20	29
Perplexity (Legal Big)	11	69	62
Perplexity (HateXplain)	413	146	165

Table 12

The comparison of perplexity between language models.

conducted experiments on a task of sentiment analysis, hate speech recognition and three tasks from the legal domain. According to the comparison between the extracted rationales and the gold rationale annotations in sentiment analysis task and hate speech recognition task, our InfoCal method improves the selection of rationales by a large margin. We also conducted ablation tests for the evaluation of the LM regularizer’s contribution, which showed that our regularizer is effective in refining the rationales.

As future work, the main architecture of our model can be directly applied to other domains, e.g., images or tabular data. The image rationales can be applied in many read-world applications, such as medical image recognition [27] and automatic driving [99]. Regularizers based on Manifold learning [15] is promising to be applied on image rationale extraction. The tabular rationales are very useful in some tasks like automatic disease diagnose [2]. When designing the regularizers for tabular rationales, a sensible method is to make use of the relations between different fields of the tabular since different kinds of data are closely related in medical experiment reports and many of them are potentially to contribute to the patients’ diagnose result.

6. Ethical Statement

The paper does not present a new dataset. It also does not use demographic or identity characteristics information. Furthermore, the paper does not report on experiments that involve a lot of computing time/power.

- **Intended use.** While the paper presents an NLP legal prediction application, our method is not yet ready to be used in practice. Our work takes a step forward in the research direction of making legal prediction systems explainable, which should uncover the systems' potential biases and modes of failures, thus ultimately rendering them more reliable. Thus, once it can be guaranteed a high likelihood of correctness and unbiasedness of the predictions and the faithfulness of their explanations w.r.t. the inner-working of the model, legal prediction systems may help to assist judges (and not replace them) in their decisions, so that they can process more cases, and more people can perceive justice than nowadays is the case. (At present, only a very small portion of cases is brought to court; especially poorer parts of the populations have essentially no access to the justice system, due to its high costs.) In addition, legal prediction systems may be used as second opinion and help to uncover mistakes or even biases of human judges. Currently, legal prediction systems are being heavily researched in the literature without the explainability component that our paper is bringing. Hence, our approach is taking a step forward in assessing the reliability of the systems, although we do not currently guarantee the faithfulness of the provided explanations. Hence, our work is intended purely as a research advancement and not as a real-world tool.
- **Failure modes.** Our model may fail to provide correct and unbiased predictions and explanations that are faithfully describing its decision-making process. Ensuring correct and unbiased predictions as well as faithful explanations are very challenging open questions, and our work takes an important but far from final step forward in this direction.
- **Biases.** If the training data contains biases, then a model may pick up on these biases, and hence it would not be safe to use it in practice. Our explanations may help to detect biases and potentially give insights to researchers on how to further develop models that avoid them. However, we do not currently guarantee the faithfulness of the explanations to the decision-making of the model.
- **Misuse potential.** As our method is not currently suitable for production, the legal prediction model should not be used in real-world legal judgement prediction tasks.
- **Collecting data from users.** We do not collect data from users, we only use an existing dataset.
- **Potential harm to vulnerable populations.** Since our model learns from datasets, if there are under-represented groups in the datasets, then the model might not be able to learn correct predictions for these groups. However, our model provides explanations for its predictions, which may uncover the potential incorrect reasons for its predictions on under-represented groups. This could further unveil the under-representation of certain groups and incentivize the collection of more instances for such groups. However, we highlight again that our model is not yet ready to be used in practice and that it is currently a stepping stone in this important direction of research.

Acknowledgments

This work was supported by the ESRC grant ES/S010424/1 "Unlocking the Potential of AI for English Law", an Early Career Leverhulme Fellowship, a JP Morgan PhD Fellowship, the Alan Turing Institute under the EPSRC grant EP/N510129/1, the AXA Research Fund, and the EU TAILOR grant 952215. We also acknowledge the use of Oxford's Advanced Research Computing (ARC) facility, of the EPSRC-funded Tier 2 facility JADE (EP/P020275/1), and of GPU computing support by Scan Computers International Ltd.

References

- [1] Agarwal, R., Frosst, N., Zhang, X., Caruana, R., Hinton, G.E., 2020. Neural Additive Models: Interpretable Machine Learning With Neural Nets. arXiv preprint arXiv:2004.13912.
- [2] Alkım, E., Gürbüz, E., Kılıç, E., 2012. A Fast and Adaptive Automated Disease Diagnosis Method With an Innovative Neural Network Model. *Neural Networks*. 33, 88–96.
- [3] Alvarez-Melis, D., Jaakkola, T.S., 2018. On the Robustness of Interpretability Methods. arXiv preprint arXiv:1806.08049.

- [4] Apley, D.W., Zhu, J., 2020. Visualizing the Effects of Predictor Variables in Black box Supervised Learning Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 82, 1059–1086.
- [5] Atkinson, K., Bench-Capon, T., Bollegala, D., 2020. Explanation in AI and Law: Past, Present and Future. *Artificial Intelligence*, 103387.
- [6] Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R., 2010. How to Explain Individual Classification Decisions. *The Journal of Machine Learning Research*. 11, 1803–1831.
- [7] Bastani, O., Kim, C., Bastani, H., 2017. Interpreting Blackbox Models via Model Extraction. *arXiv preprint arXiv:1705.08504*.
- [8] Bastings, J., Aziz, W., Titov, I., 2019. Interpretable Neural Predictions with Differentiable Binary Variables, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963–2977.
- [9] Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A., 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations, in: *Computer Vision and Pattern Recognition*.
- [10] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*. 3, 1137–1155.
- [11] Borgelt, C., 2005. An Implementation of the FP-growth Algorithm, in: *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, pp. 1–5.
- [12] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language Models are Few-shot Learners. *arXiv preprint arXiv:2005.14165*.
- [13] Camburu, O., Rocktäschel, T., Lukaszewicz, T., Blunsom, P., 2018. e-SNLI: Natural Language Inference with Natural Language Explanations, in: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 (NeurIPS 2018)*, pp. 9560–9572.
- [14] Carton, S., Mei, Q., Resnick, P., 2018. Extractive Adversarial Networks: High-Recall Explanations for Identifying Personal Attacks in Social Media Posts, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics. pp. 3497–3507.
- [15] Cayton, L., 2005. Algorithms for Manifold Learning. *Univ. of California at San Diego Tech. Rep.* 12, 1.
- [16] Chang, S., Zhang, Y., Yu, M., Jaakkola, T., 2019. A Game Theoretic Approach to Class-wise Selective Rationalization, in: *Advances in Neural Information Processing Systems*, pp. 10055–10065.
- [17] Chen, J., Song, L., Wainwright, M., Jordan, M., 2018a. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation, in: *Proceedings of the International Conference on Machine Learning*, pp. 883–892.
- [18] Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K., 2018b. Isolating Sources of Disentanglement in Variational Autoencoders, in: *Advances in Neural Information Processing Systems*, pp. 2610–2620.
- [19] Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, in: *Advances in Neural Information Processing Systems*, pp. 2172–2180.
- [20] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning Phrase Representations Using RNN Encoder-decoder For Statistical Machine Translation. *arXiv preprint arXiv:1406.1078*.
- [21] Cífka, O., Bojar, O., 2018. Are BLEU and Meaning Representation in Opposition?, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics. pp. 1362–1371.
- [22] Cohen, W.W., 1995. Fast Effective Rule Induction, in: *Machine learning proceedings 1995*. Elsevier, pp. 115–123.
- [23] Collier, M., Beel, J., 2018. Implementing Neural Turing Machines, in: *Proceedings of the International Conference on Artificial Neural Networks*, Springer. pp. 94–104.
- [24] Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M., 2018. What you can Cram into a Single \mathbb{R}^d Vector: Probing Sentence Embeddings for Linguistic Properties, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics. pp. 2126–2136.
- [25] Cook, R.D., Weisberg, S., 1980. Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics*. 22, 495–508.
- [26] Davidson, T., Warmusley, D., Macy, M., Weber, I., 2017. Automated Hate Speech Detection and the Problem of Offensive Language, in: *Proceedings of the International AAAI Conference on Web and Social Media*.
- [27] Deruyver, A., Hodé, Y., Brun, L., 2009. Image Interpretation With a Conceptual Graph: Labeling Over-segmented Images and Detection of Unexpected Objects. *Artificial Intelligence*. 173, 1245–1265.
- [28] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics. pp. 4171–4186.
- [29] DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C., 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics. pp. 4443–4458.
- [30] Erion, G., Janizek, J., Sturmfels, P., Lundberg, S., Lee, S., 2019. Improving Performance of Deep Learning Models With Axiomatic Attribution Priors and Expected Gradients. *arXiv preprint arXiv:1906.10670*.
- [31] Fisher, A., Rudin, C., Dominici, F., 2018. Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the “rashomon” perspective. *arXiv preprint arXiv:1801.01489*. 68.
- [32] Fong, R., Patrick, M., Vedaldi, A., 2019. Understanding Deep Networks via Extremal Perturbations and Smooth Masks, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2950–2958.
- [33] Fong, R.C., Vedaldi, A., 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation, in: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437.
- [34] Fortuna, P., Nunes, S., 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*. 51, 1–30.
- [35] Friedman, J.H., 2001. Greedy Function Approximation: a Gradient Boosting Machine. *Annals of Statistics*, 1189–1232.

- [36] Friedman, J.H., Popescu, B.E., et al., 2008. Predictive Learning via Rule Ensembles. *The Annals of Applied Statistics*. 2, 916–954.
- [37] Fürnkranz, J., Gamberger, D., Lavrač, N., 2012. *Foundations of Rule Learning*. Springer Science & Business Media.
- [38] Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*. 24, 44–65.
- [39] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets, in: *Advances in Neural Information Processing Systems*, pp. 2672–2680.
- [40] Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K., 2019. Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One. arXiv preprint arXiv:1912.03263.
- [41] Graves, A., Wayne, G., Danihelka, I., 2014. Neural Turing Machines. arXiv preprint arXiv:1410.5401.
- [42] Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S.G., Grefenstette, E., Ramalho, T., Agapiou, J., et al., 2016. Hybrid Computing Using a Neural Network with Dynamic External Memory. *Nature*. 538, 471–476.
- [43] Greenwell, B.M., Boehmke, B.C., McCarthy, A.J., 2018. A Simple and Effective Model-based Variable Importance Measure. arXiv preprint arXiv:1805.04755.
- [44] Gu, J., Tresp, V., 2019. Contextual Prediction Difference Analysis. arXiv preprint arXiv:1910.09086.
- [45] Heafield, K., 2011. KenLM: Faster and Smaller Language Model Queries, in: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Association for Computational Linguistics. pp. 187–197.
- [46] Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T., 2016. Generating Visual Explanations, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer. pp. 3–19.
- [47] Hewitt, J., Liang, P., 2019. Designing and Interpreting Probes with Control Tasks, 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)*, Association for Computational Linguistics. pp. 2733–2743.
- [48] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A., 2017. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations*. 2, 6.
- [49] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., Sainath, T., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*. 29, 82–97.
- [50] Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.
- [51] Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation*. 9, 1735–1780.
- [52] Holte, R.C., 1993. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine learning*. 11, 63–90.
- [53] Hooker, G., 2004. Discovering Additive Structure in Black Box Functions, in: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 575–580.
- [54] Jain, S., Wiegrefe, S., Pinter, Y., Wallace, B.C., 2020. Learning to Faithfully Rationalize by Construction, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics. pp. 4459–4473.
- [55] Jang, E., Gu, S., Poole, B., 2016. Categorical Reparameterization With Gumbel-softmax. arXiv preprint arXiv:1611.01144.
- [56] Janzing, D., Minorics, L., Blöbaum, P., 2019. Feature Relevance Quantification in Explainable AI: A Causality Problem. arXiv preprint arXiv:1910.13413.
- [57] Janzing, D., Minorics, L., Blöbaum, P., 2020. Feature Relevance Quantification in Explainable AI: A Causal Problem, in: *International Conference on Artificial Intelligence and Statistics*, PMLR. pp. 2907–2916.
- [58] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q., 2019. TinyBERT: Distilling BERT for Natural Language Understanding. arXiv preprint arXiv:1909.10351.
- [59] Kaelbling, L.P., Littman, M.L., Cassandra, A.R., 1998. Planning and Acting in Partially Observable Stochastic Domains. *Artificial intelligence*. 101, 99–134.
- [60] Kaiser, Ł., Sutskever, I., 2015. Neural GPUs Learn Algorithms. arXiv preprint arXiv:1511.08228.
- [61] Kaptein, F., Broekens, J., Hindriks, K., Neerinx, M., 2021. Evaluating XAI: A Comparison of Rule-based and Example-based Explanations. *Artificial Intelligence*. 291.
- [62] Karimi, A.H., Schölkopf, B., Valera, I., 2020. Algorithmic Recourse: from Counterfactual Explanations to Interventions. arXiv preprint arXiv:2002.06278.
- [63] Kaufmann, E., Kalyanakrishnan, S., 2013. Information Complexity in Bandit Subset Selection, in: *Conference on Learning Theory*, pp. 228–251.
- [64] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al., 2018a. Interpretability Beyond Feature Attribution: Quantitative Testing With Concept Activation Vectors (TCAV), in: *International conference on machine learning*, PMLR. pp. 2668–2677.
- [65] Kim, J., Rohrbach, A., Darrell, T., Canny, J.F., Akata, Z., 2018b. Textual Explanations for Self-Driving Vehicles. *Proceedings of the European Conference on Computer Vision (ECCV)*.
- [66] Kingma, D.P., Welling, M., 2013. Auto-encoding Variational Bayes. arXiv preprint arXiv:1312.6114.
- [67] Koh, P.W., Liang, P., 2017. Understanding Black-box Predictions via Influence Functions, in: *International Conference on Machine Learning*, PMLR. pp. 1885–1894.
- [68] Konda, V.R., Tsitsiklis, J.N., 2000. Actor-Critic Algorithms, in: *Advances in Neural Information Processing Systems*, pp. 1008–1014.
- [69] Krippendorff, K., 2004. *Content Analysis: An Introduction to Its Methodology* Thousand Oaks, Calif.: Sage.
- [70] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*. 60, 84–90.
- [71] Kuhn, R., De Mori, R., 1990. A Cache-based Natural Language Model for Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 12, 570–583.
- [72] Kurach, K., Andrychowicz, M., Sutskever, I., 2015. Neural Random-Access Machines. arXiv preprint arXiv:1511.06392.
- [73] Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J., 2019. Faithful and Customizable Explanations of Black Box Models, in: *Proceedings*

- of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 131–138.
- [74] LeCun, Y., Bengio, Y., et al., 1995. Convolutional Networks for Images, Speech, and Time Series. The handbook of brain theory and neural networks. 3361, 1995.
- [75] Lei, T., Barzilay, R., Jaakkola, T., 2016. Rationalizing Neural Predictions, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 107–117.
- [76] Leino, K., Sen, S., Datta, A., Fredrikson, M., Li, L., 2018. Influence-directed Explanations for Deep Convolutional Networks, in: 2018 IEEE International Test Conference (ITC), IEEE. pp. 1–8.
- [77] Letham, B., Rudin, C., McCormick, T.H., Madigan, D., et al., 2015. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. The Annals of Applied Statistics. 9, 1350–1371.
- [78] Li, X.L., Eisner, J., 2019. Specializing Word Embeddings (for Parsing) by Information Bottleneck, 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), Association for Computational Linguistics. pp. 2744–2754.
- [79] Lima, L., Reis, J.C., Melo, P., Murai, F., Araujo, L., Vikatos, P., Benevenuto, F., 2018. Inside the Right-leaning Echo Chambers: Characterizing Gab, an Unmoderated Social System, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE. pp. 515–522.
- [80] Liu, T., Wang, K., Sha, L., Chang, B., Sui, Z., 2018. Table-to-text Generation by Structure-aware Seq2seq Learning. Proceedings of the 32nd AAAI Conference on Artificial Intelligence.
- [81] Lundberg, S.M., Erion, G.G., Lee, S.I., 2018. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv preprint arXiv:1802.03888.
- [82] Lundberg, S.M., Lee, S.I., 2017. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 30, 4765–4774.
- [83] Luo, B., Feng, Y., Xu, J., Zhang, X., Zhao, D., 2017. Learning to Predict Charges for Criminal Cases With Legal Basis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. pp. 2727–2736.
- [84] Mahajan, D., Tan, C., Sharma, A., 2019. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. arXiv preprint arXiv:1912.03277.
- [85] Marullo, C., Agliari, E., 2021. Boltzmann Machines As Generalized Hopfield Networks: A Review of Recent Results and Outlooks. Entropy. 23, 34.
- [86] Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., Mukherjee, A., 2020a. Hate Begets Hate: a Temporal Study of Hate Speech. Proceedings of the ACM on Human-Computer Interaction. 4, 1–24.
- [87] Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A., 2020b. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. arXiv preprint arXiv:2012.10289.
- [88] McAuley, J., Leskovec, J., Jurafsky, D., 2012. Learning Attitudes and Attributes from Multi-aspect Reviews, in: Proceedings of the 2012 IEEE 12th International Conference on Data Mining, IEEE. pp. 1020–1025.
- [89] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed Representations of Words and Phrases and Their Compositionality, in: Advances in Neural Information Processing Systems, pp. 3111–3119.
- [90] Paranjape, B., Joshi, M., Thackstun, J., Hajishirzi, H., Zettlemoyer, L., 2020. An Information Bottleneck Approach for Controlling Consistency in Rationale Extraction, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics. pp. 1938–1952.
- [91] Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M., 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- [92] Petsiuk, V., Das, A., Saenko, K., 2018. Rise: Randomized Input Sampling for Explanation of Black-box Models. arXiv preprint arXiv:1806.07421.
- [93] Pimentel, T., Valvoda, J., Hall Maudslay, R., Zmigrod, R., Williams, A., Cotterell, R., 2020. Information-Theoretic Probing for Linguistic Structure, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. pp. 4609–4622.
- [94] Pruthi, G., Liu, F., Sundararajan, M., Kale, S., 2020. Estimating Training Data Influence by Tracing Gradient Descent. Advances in Neural Information Processing Systems.
- [95] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving Language Understanding by Generative Pre-training.
- [96] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language Models are Unsupervised Multitask Learners.
- [97] Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G.K., Greiff, V., et al., 2020. Hopfield Networks Is All You Need. arXiv preprint arXiv:2008.02217.
- [98] Rawal, K., Lakkaraju, H., 2020. Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses. Advances in Neural Information Processing Systems. 33.
- [99] Reece, D.A., Shafer, S.A., 1995. Control of Perceptual Attention in Robot Driving. Artificial Intelligence. 78, 397–430.
- [100] Ribeiro, M.T., Singh, S., Guestrin, C., 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, pp. 1135–1144.
- [101] Ribeiro, M.T., Singh, S., Guestrin, C., 2018. Anchors: High-precision Model-agnostic Explanations, in: Proceedings of the AAAI Conference on Artificial Intelligence.
- [102] Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kočiský, T., Blunsom, P., 2015. Reasoning About Entailment with NNeural Attention. arXiv preprint arXiv:1509.06664.
- [103] Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning Representations by Back-propagating Errors. nature. 323, 533–536.
- [104] Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R., 2016. Progressive Neural Networks. arXiv preprint arXiv:1606.04671.

- [105] Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic Routing Between Capsules, in: *Advances in Neural Information Processing Systems*, pp. 3856–3866.
- [106] Schuster, M., Nakajima, K., 2012. Japanese and Korean Voice search, in: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. pp. 5149–5152.
- [107] Schuster, M., Paliwal, K.K., 1997. Bidirectional Recurrent Neural Networks. *IEEE transactions on Signal Processing*. 45, 2673–2681.
- [108] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, in: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.
- [109] Seo, J., Choe, J., Koo, J., Jeon, S., Kim, B., Jeon, T., 2018. Noise-adding Methods of Saliency Map As Series of Higher Order Partial Derivative. *arXiv preprint arXiv:1806.03000*.
- [110] Sha, L., 2020. Gradient-guided Unsupervised Lexically Constrained Text Generation, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics. pp. 8692–8703.
- [111] Sha, L., Camburu, O.M., Lukaszewicz, T., 2021. Learning from the Best: Rationalizing Predictions by Adversarial Information Calibration, in: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.
- [112] Sha, L., Chang, B., Sui, Z., Li, S., 2016. Reading and Thinking: Re-read LSTM Unit for Textual Entailment Recognition, in: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2870–2879.
- [113] Sha, L., Lukaszewicz, T., 2021. Multi-type Disentanglement Without Adversarial Training, in: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.
- [114] Sha, L., Mou, L., Liu, T., Poupard, P., Li, S., Chang, B., Sui, Z., 2018a. Order-Planning Neural Text Generation from Structured Data, in: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.
- [115] Sha, L., Shi, C., Chen, Q., Zhang, L., Wang, H., 2020. Estimate Minimum Operation Steps via Memory-based Recurrent Calculation Network, in: *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE.
- [116] Sha, L., Zhang, X., Qian, F., Chang, B., Sui, Z., 2018b. A Multi-view Fusion Neural Network for Answer Selection, in: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.
- [117] Shapley, L.S., 1953. A Value for N-person Games. *Contributions to the Theory of Games*. 2, 307–317.
- [118] Shrikumar, A., Greenside, P., Kundaje, A., 2017. Learning Important Features Through Propagating Activation Differences, in: *International Conference on Machine Learning*, PMLR. pp. 3145–3153.
- [119] Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *International Conference on Learning Representations*. 2, 6.
- [120] Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H., 2020. Fooling LIME and SHAP: Adversarial Attacks on Post-hoc Explanation Methods, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186.
- [121] Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M., 2017. Smoothgrad: Removing Noise by Adding Noise. *International Conference on Machine Learning Workshop on Visualization for deep learning*.
- [122] Song, Y., Lukaszewicz, T., Xu, Z., Bogacz, R., 2020. Can the Brain Do Backpropagation?—exact Implementation of Backpropagation in Predictive Coding Networks. *NeurIPS Proceedings 2020*. 33.
- [123] Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for Simplicity: The All Convolutional Net, in: *International Conference on Learning Representations (workshop track)*.
- [124] Staniak, M., Biecek, P., 2018. Explanations of Model Predictions With Live and BreakDown Packages. *arXiv preprint arXiv:1804.01955*.
- [125] Štrumbelj, E., Kononenko, I., 2014. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and information systems*. 41, 647–665.
- [126] Sundararajan, M., Najmi, A., 2020. The Many Shapley Values for Model Explanation, in: *International Conference on Machine Learning*, PMLR. pp. 9269–9278.
- [127] Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic Attribution for Deep Networks, in: *International Conference on Machine Learning*, PMLR. pp. 3319–3328.
- [128] Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to Sequence Learning with Neural Networks. *CoRR*. abs/1409.3215. *arXiv:1409.3215*.
- [129] Tan, S., Caruana, R., Hooker, G., Koch, P., Gordo, A., 2018. Learning Global Additive Explanations for Neural Nets Using Model Distillation. *arXiv preprint arXiv:1801.08640*.
- [130] Tishby, N., Pereira, F.C., Bialek, W., 2000. The Information Bottleneck Method. *arXiv preprint physics/0004057*.
- [131] Tishby, N., Zaslavsky, N., 2015. Deep Learning and the Information Bottleneck Principle, in: *Proceedings of the 2015 IEEE Information Theory Workshop (ITW)*, IEEE. pp. 1–5.
- [132] Tomas, M., Anoop, D., Stefan, K., Lukas, B., Jan, H.G., 2011. RNNLM-Recurrent Neural Network Language Modeling Toolkit, in: *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU) Waikoloa*, pp. 11–15.
- [133] Vanmassenhove, E., Du, J., Way, A., 2017. Investigating ‘aspect’ in NMT and SMT: Translating the English Simple Past and Present Perfect. *Computational Linguistics in the Netherlands Journal*. 7, 109–128.
- [134] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is All You Need, in: *Advances in Neural Information Processing Systems*, pp. 5998–6008.
- [135] Voita, E., Titov, I., 2020. Information-Theoretic Probing with Minimum Description Length, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics. pp. 183–196.
- [136] Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harv. JL & Tech.*. 31, 841.
- [137] Williams, R.J., 1992. Simple Statistical Gradient-following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*. 8, 229–256.
- [138] Xia, Q., Sha, L., Chang, B., Sui, Z., 2017. A Progressive Learning Approach to Chinese SRL Using Heterogeneous Data, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational

- Linguistics. pp. 2069–2077.
- [139] Yang, H., Rudin, C., Seltzer, M., 2017. Scalable Bayesian Rule Lists, in: International Conference on Machine Learning, PMLR. pp. 3921–3930.
- [140] Yang, W., Jia, W., Zhou, X., Luo, Y., 2019. Legal Judgment Prediction via Multi-perspective Bi-feedback Network. arXiv preprint arXiv:1905.03969.
- [141] Yeh, C.K., Kim, J.S., Yen, I.E.H., Ravikumar, P., 2018. Representer Point Selection for Explaining Deep Neural Networks, in: Advances in Neural Information Processing Systems.
- [142] Yoon, J., Jordon, J., van der Schaar, M., 2018. INVASe: Instance-wise Variable Selection Using Neural Networks, in: Proceedings of the International Conference on Learning Representations.
- [143] Yu, L., Zhang, W., Wang, J., Yu, Y., 2017. SeqGAN: Sequence Generative Adversarial Nets With Policy Gradient, in: Proceedings of the 31st AAAI Conference on Artificial Intelligence.
- [144] Yu, M., Chang, S., Zhang, Y., Jaakkola, T., 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), Association for Computational Linguistics. pp. 4094–4103.
- [145] Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., Blackburn, J., 2018. What is Gab: a Bastion of Free Speech or an Alt-right Echo Chamber, in: Companion Proceedings of the The Web Conference 2018, pp. 1007–1014.
- [146] Zeiler, M.D., Fergus, R., 2014. Visualizing and Understanding Convolutional Networks, in: European conference on computer vision, Springer. pp. 818–833.
- [147] Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S., 2018a. Top-down Neural Attention by Excitation Backprop. International Journal of Computer Vision. 126, 1084–1102.
- [148] Zhang, Z., Robinson, D., Tepper, J., 2018b. Detecting Hate Speech On Twitter Using a Convolution-GRU Based Deep Neural Network, in: European semantic web conference, Springer. pp. 745–760.
- [149] Zhao, Q., Hastie, T., 2019. Causal Interpretations of Black-box Models. Journal of Business & Economic Statistics, 1–10.
- [150] Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., Sun, M., 2018. Legal Judgment Prediction via Topological Learning, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. pp. 3540–3549.
- [151] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning Deep Features for Discriminative Localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929.

Appendices

A. Proofs

A.1. Derivation of $I(\tilde{\mathbf{z}}_{\text{sym}}, y)$

This proof is the basis for the information bottleneck equation in Section 3.1.4.

Theorem 2. *Minimizing $-I(\tilde{\mathbf{z}}_{\text{sym}}, y)$ is equivalent to minimizing L_{sp} .*

Proof.

$$I(\tilde{\mathbf{z}}_{\text{sym}}, y) = \mathbb{E}_{\tilde{\mathbf{z}}_{\text{sym}}, y} \left[\frac{p(y|\tilde{\mathbf{z}}_{\text{sym}})}{p(y)} \right] = \mathbb{E}_{\tilde{\mathbf{z}}_{\text{sym}}, y} p(y|\tilde{\mathbf{z}}_{\text{sym}}) - \mathbb{E}_{\tilde{\mathbf{z}}_{\text{sym}}, y} p(y). \quad (22)$$

We omit $\mathbb{E}_{\tilde{\mathbf{z}}_{\text{sym}}, y} p(y)$, because it is a constant, therefore, minimizing Eq. 22 is equivalent to minimizing the following term:

$$\mathbb{E}_{\tilde{\mathbf{z}}_{\text{sym}}, y} p(y|\tilde{\mathbf{z}}_{\text{sym}}). \quad (23)$$

As the training pair (\mathbf{x}, y) is sampled from the training data, and $\tilde{\mathbf{z}}_{\text{sym}}$ is sampled from $\text{Sel}(\tilde{\mathbf{z}}_{\text{sym}}|\mathbf{x})$, we have that

$$\mathbb{E}_{\tilde{\mathbf{z}}_{\text{sym}}, y} p(y|\tilde{\mathbf{z}}_{\text{sym}}) = \mathbb{E}_{\mathbf{x}, y} p(y|\tilde{\mathbf{z}}_{\text{sym}}) p(\tilde{\mathbf{z}}_{\text{sym}}|\mathbf{x}) = E_{\mathbf{x}, y} p(y|\mathbf{x}). \quad (24)$$

We can give each $p(y|\mathbf{x})$ in Eq. 24 a $-\log$ to arrive to $-I(\tilde{\mathbf{z}}_{\text{sym}}, y)$. Then, it is not difficult to see that $-I(\tilde{\mathbf{z}}_{\text{sym}}, y)$ has exactly the same form as L_{sp} . \square

A.2. Derivation of Equation 12

Equation 12 is the information bottleneck loss for the guider model, this loss encourages that the features extracted by the guider model are least-but-enough.

$$L_{\text{mi}} = I(\mathbf{x}, \mathbf{z}_{\text{nero}}) = \mathbb{E}_{\mathbf{x}, \mathbf{z}_{\text{nero}}} \left[\log \frac{p(\mathbf{z}_{\text{nero}}|\mathbf{x})}{p(\mathbf{z}_{\text{nero}})} \right] \quad (25)$$

$$= \mathbb{E}_{\mathbf{z}_{\text{nero}}} p(\mathbf{x}) \left[\log \frac{p(\mathbf{z}_{\text{nero}}|\mathbf{x})}{p(\mathbf{z}_{\text{nero}})} \right] \quad (26)$$

$$\leq \mathbb{E}_{\mathbf{z}_{\text{nero}}} \left[\log \frac{p(\mathbf{z}_{\text{nero}}|\mathbf{x})}{p(\mathbf{z}_{\text{nero}})} \right] \quad (27)$$

$$= 0.5(\mu^2 + \sigma^2 - 1 - 2 \log \sigma). \quad (28)$$

A.3. Proof of Theorem 1

This theorem is the reason why our language model based regularizer encourages fewer segments of token sequences and decreases bad start or end tokens for each token subsequences, thus makes semantically fluent rationales. The theorem content is restated as follows:

Theorem 1. *If the following is satisfied for all i, j :*

- $m'_i < \epsilon \ll 1 - \epsilon < m_i$, ($0 < \epsilon < 1$), and
- $\left| p(m'_i x_i | x_{<i}) - p(m'_j x_j | x_{<j}) \right| < \epsilon$,

then the following two inequalities hold:

- (1) $L_{lm}(\dots, m_k, \dots, m'_n) < L_{lm}(\dots, m'_k, \dots, m_n)$.
- (2) $L_{lm}(m_1, \dots, m'_k, \dots) > L_{lm}(m'_1, \dots, m_k, \dots)$.

Proof. By Eq. 15, we have:

$$L_{lm}(\dots, m'_k, \dots, m_n) = - \left[\sum_{i \neq k, k+1} m_{i-1} \log P(m_i x_i | x_{<i}) + m_{k-1} \log P(m'_k x_k | x_{<k}) + m'_k \log P(m_{k+1} x_{k+1} | x_{<k+1}) \right]. \quad (29)$$

Therefore, we have the following equation:

$$\begin{aligned}
 & L_{\text{lm}}(\dots, m'_k, \dots, m_n) - L_{\text{lm}}(\dots, m_k, \dots, m'_n) \\
 &= -m_{k-1} \log p(m'_k) + m_{k-1} \log p(m_k) - m'_k \log p(m_{k+1}) + m_k \log p(m_{k+1}) \\
 &\quad - m_{n-1} \log p(m_n) + m_{n-1} \log p(m'_n) - m_n \log p(m'_{k+1}) + m'_n \log p(m'_{n+1}),
 \end{aligned} \tag{30}$$

where, for simplicity, we use the abbreviation $p(m_k)$ to represent $p(m_k x_k | x_{<k})$.

We also have that:

$$-m_n \log p(m'_{k+1}) + m_{n-1} \log p(m'_n) = (m_{n-1} - m_n) \log p(m'_{k+1}) - m_{n-1} \log \frac{p(m'_{k+1})}{p(m'_n)} \tag{31}$$

$$\geq \epsilon \log p(m'_{k+1}) - \epsilon \tag{32}$$

Since $p(m_k)_k$ are expected to have large probability values in the language model training process, we have that $p(m_k) > \delta$, and, therefore, $-|\log \delta| < \log \frac{p(m_{k+1})}{p(m_n)} < |\log \delta|$.

Hence, we have that:

$$-m_{n-1} \log p(m_n) + m_k \log p(m_{k+1}) = (m_k - m_{n-1}) \log p(m_{k+1}) + m_{n-1} \log \frac{p(m_{k+1})}{p(m_n)} \tag{33}$$

$$\geq \epsilon \log p(m_{k+1}) - |\log \delta| \geq (\epsilon - 1) |\log \delta|. \tag{34}$$

Similarly, $-m'_k \log p(m_{k+1}) + m_{k-1} \log p(m_k) \geq (1 - 2\epsilon) \log p(m_k) + m'_k \log \frac{p(m_k)}{p(m_{k+1})} \geq (1 - 3\epsilon) |\log \delta|$.

Therefore, the lower bound of the expression in Eq. 30 is:

$$\begin{aligned}
 \text{inf} &= -(1 - \epsilon) \log p(m'_k) + \epsilon \log p(m'_{n+1}) + \epsilon \log p(m'_{k+1}) - 2\epsilon |\log \delta| - \epsilon \\
 &\geq -(1 - 3\epsilon) \log p(m'_k) - 4\epsilon |\log \delta| - \epsilon > 0.
 \end{aligned} \tag{35}$$

This proves the statement of the theorem. □

B. More Results.

We list more examples of rationales extracted by our model for the BeerAdvocate dataset in Table 13.

More examples of rationales extracted by our model for the legal judgement tasks are shown in Table 6.

C. Human Evaluation Setup

Our annotators were asked the following questions, in order to assess the usefulness, completeness, and fluency of the rationales provided by our model.

C.1. Usefulness of Rationales

Q: Do you think the selected tokens/rationale are **useful** to explain the ground-truth label?

Please choose a score according to the following description. Note that the score is not necessary an integer, you can give intermediate scores, such as 3.2 or 4.9 if you deem appropriate.

- 5: Exactly. I can give the correct label only by seeing the given tokens.
- 4: Highly useful. Although most of the selected tokens lead to the correct label, there are still several tokens that have no relation to the correct label.
- 3: Half of them are useful. About half of the tokens can give some hint for the correct label, the rest are nonsense to the label.

Gold	InfoCal
<p>dark black with nearly no light at all shining through on this one . rich tan colored head of about two inches quickly settled down to about a half inch of tan that thoroughly coated the inside of the glass . this was what the style is all about the aroma was just loaded down with coffee . rich notes of mocha mixes in with a rich , and sweet coffee note . a tiny bit of bitterness and an earthy flare lying down underneath of it , but the majority of this one was hands down , rich brewed coffee . the flavor was more of the same . rich notes just rolled over the tongue in waves and thoroughly coated the inside of the mouth . sweet with touches of chocolate and vanilla to highlight the coffee notes</p>	<p>dark black with nearly no light at all shining through on this one . rich tan colored head of about two inches quickly settled down to about a half inch of tan that thoroughly coated the inside of the glass . this was what the style is all about the aroma was just loaded down with coffee . rich notes of mocha mixes in with a rich , and sweet coffee note . a tiny bit of bitterness and an earthy flare lying down underneath of it , but the majority of this one was hands down , rich brewed coffee . the flavor was more of the same . rich notes just rolled over the tongue in waves and thoroughly coated the inside of the mouth . sweet with touches of chocolate and vanilla to highlight the coffee notes</p>
<p>clear copper colored brew , medium cream colored head . floral hop nose , caramel malt . caramel malt front dominated by a nice floral hop background . grapefruit tones . very tasty hops run the show with this brew . thin to medium mouth . not a bad choice if you 're looking for a nice hop treat .</p>	<p>clear copper colored brew , medium cream colored head . floral hop nose , caramel malt . caramel malt front dominated by a nice floral hop background . grapefruit tones . very tasty hops run the show with this brew . thin to medium mouth . not a bad choice if you 're looking for a nice hop treat .</p>
<p>12oz bottle into my pint glass . looks decent , a brown color (imagine that !) with a tan head . nothing bad , nothing extraordinary . smell is nice , slight roast , some nuttiness , and hint of hops . pretty much to-style . taste is good but a little underwhelming . toffee malt , some slight roast gives chocolate impressions . hoppiness is mild and earthy . just a touch of bitterness . pretty nondescript overall , but nothing offensive . mouthfeel is good , medium body and light carb give a creamy finish . drinkability was nice . i would try this again but wo n't be seeking it out .</p>	<p>12oz bottle into my pint glass . looks decent , a brown color (imagine that !) with a tan head . nothing bad , nothing extraordinary . smell is nice , slight roast , some nuttiness , and hint of hops . pretty much to-style . taste is good but a little underwhelming . toffee malt , some slight roast gives chocolate impressions . hoppiness is mild and earthy . just a touch of bitterness . pretty nondescript overall , but nothing offensive . mouthfeel is good , medium body and light carb give a creamy finish . drinkability was nice . i would try this again but wo n't be seeking it out .</p>

Table 13

More instances from the BeerAdvocate dataset. In red the rationales for the appearance aspect, in green the rationales for the smell aspect, and in blue the rationales for the palate aspect.

- 2: Almost useless. Almost all of the tokens are useless, but there are still several tokens that are useful.
- 1: No Use. I feel very confused about the selected tokens, I don't know which law article/charge/term of penalty the article belongs to.

C.2. Completeness of Rationales

Q: Do you think the selected tokens/rationale are **enough** to explain the ground-truth label?

Please choose a score according to the following description. Note that the score is not necessary an integer, you can give intermediate scores, such as 3.2 or 4.9, if you deem appropriate.

- 5: Exactly. I can give the correct label only by the given tokens.
- 4: Highly complete. There are still several tokens in the fact description that have a relation to the correct label, but they are not selected.
- 3: Half complete. There are still important tokens in the fact description, and they are in nearly the same number as the selected tokens.
- 2: Somewhat complete. The selected tokens are not enough. There are still many important tokens in the fact description not being selected.
- 1: Nonsense. All of the selected tokens are useless. None of the important tokens is selected.

C.3. Fluency

Q: How fluent do you think the selected rationale is? For example: *“He stole an iPhone in the room”* is very fluent, which should have a high score. *“stole iPhone room”* is just separated tokens, which should have a low fluency score.

Please choose a score according to the following description. Note that the score is not necessary an integer, you can give scores like 3.2 or 4.9 , if you deem appropriate.

- 5: Very fluent.
- 4: Highly fluent.
- 3: Partial fluent.

<p>罪名：强奸</p> <p>永顺县人民检察院指控，2014年1月11日，被告人李某某与彭某某（另案处理）在永顺县UNK“新都宾馆”一房间内先后强行与被害人邹某某发生性关系。就此，公诉机关举出了如下证据：抓获经过、户籍证明、通话清单、情况说明；辨认笔录；现场勘验笔录及现场照片；物证检验报告及物证鉴定书；证人刘某甲、刘某乙、刘某丙、邹某某、杜某某的证言；被告人李某某的供述和辩解；视听资料。该院认为，被告人李某某伙同他人使用暴力和语言威胁的手段，在永顺县UNK“新都宾馆”房间内分别强行与被害人邹某某发生性关系，其行为已触犯了《中华人民共和国刑法》××××第（四）项之规定，犯罪事实清楚，证据确实、充分，应当以××罪追究其刑事责任。在共同犯罪中，被告人李某某起主要作用，系主犯，对其应结合《中华人民共和国刑法》××××、××××、××之规定，予以处罚。</p>	<p>Charge: Rape</p> <p>The People’s Procuratorate of Yongshun County alleged that on January 11, 2014, the defendant Li XX and Peng XX (a separate case dealt with) forcibly had sexual relations with the victim Zou XX in a room of Xindu Hotel in Yongshun County. In this regard, the public prosecution agency cited the following evidence: capture history, household registration certificate, call list, description of the situation; identification transcripts; on-site inspection transcripts and on-site photos; physical evidence inspection reports and physical evidence identification documents; witnesses Liu A, Liu B, Testimony of Liu C, Zou XX, Du XX; confession and defense of defendant Li XX; audio-visual materials. The court held that the defendant Li XX used violence and verbal threats with others to forcibly have sexual relations with the victim Zou XX in the Xindu Hotel room in Yongshun County. His behavior has violated the Item (4) of the Criminal Law of the PRC, the facts of the crime are clear, and the evidence is reliable and sufficient, and the criminal responsibility should be investigated for the crime of ××. In the joint crime, the defendant Li XX played the main role and was the principal offender.....</p>
<p>罪名：故意伤害</p> <p>经审理查明，2015年9月5日19时许，被告人简某因怀疑其女友与邻居黄某甲有不正当关系，即在本区南桥UNK红星304号xxx室其暂住处进行质问，遭被害人黄某甲的否认。被告人简某为泄愤UNK菜刀砍伤被害人黄某甲，致被害人左肩胛骨折。经鉴定，被害人黄某甲的伤势构成轻伤。上述事实，被告人在开庭审理过程中亦无异议，且有被害人黄某甲的陈述，证人蔡某某、黄某乙的证言，辨认笔录，医院检验情况记录，复旦大学上海医学院司法鉴定中心出具的司法鉴定意见书，公安机关出具的案发经过、工作情况，刑事附带民事判决书等证据证实，足以认定。</p>	<p>Charge: Intentional injury</p> <p>A about 19:00 on September 5, 2015, the defendant, Jian, because he suspected that his girlfriend had an improper relationship with his neighbor Huang AA, he proceeded in his temporary residence in the xxx room, Hongxing 304, South Bridge, to question him. The victim Huang AA denied. The defendant, Jian, vented his anger by using a kitchen knife to cut the victim Huang AA, causing the victim to fracture his left scapula. After identification, the injury of the victim Huang AA constituted a minor injury. The defendant had no objections to the above-mentioned facts during the trial, and there were statements by the victim Huang AA, the testimony of witnesses Cai and Huang BB, identification transcripts, hospital inspection records, Forensic Expertise Opinion issued by Forensic Expertise Center of Shanghai Medical College of Fudan University, the case history, work conditions, and the criminal and civil judgments and other evidences are sufficient to confirm.</p>
<p>罪名：故意伤害</p> <p>泸县人民检察院指控：2014年9月10日晚，被告人陈某某组织工人在自己鱼塘打鱼过程中不慎将被害人张某某栽种的鱼塘旁边的南瓜和UNK藤UNK。次日上午7时许，陈某某和张某某在前述鱼塘附近碰面，二人因南瓜和UNK藤的赔偿问题发生争执，进而发生打斗。在打斗的过程中，被告人陈某某徒手将张某某面部鼻骨、颧骨打伤。经法医学鉴定张某某所受损伤为轻伤。案发后，被告人陈某某于当日主动到公安机关投案，并如实供述其犯罪事实。诉请依照《中华人民共和国刑法》××，以××罪对被告人陈某某予以判处</p>	<p>Charge: Intentional injury</p> <p>The Lu County People’s Procuratorate charged: On the evening of September 10, 2014, The defendant Chen XX organized a worker who accidentally trampled on the pumpkins and UNK vines planted by the victim Zhang XX next to the fish pond while fishing in his own pond. At 7 o'clock in the morning the next day, Chen and Zhang met near the aforementioned fish pond. For the sake of the compensation of pumpkins and UNK vines, the two had a dispute, and a fight broke out. During the fight, the defendant Chen XX injured Zhang’s facial nasal bones and cheekbones with his bare hands. According to forensic medicine, Zhang XX suffered a minor injury. After the incident, the defendant Chen XX took the initiative to surrender to the public security organ on the same day and truthfully confessed the facts of the crime. In accordance with the "Criminal Law of the People's Republic of China", the defendant Chen XX shall be sentenced for the crime of ××</p>

Figure 6: More instances from the CAIL2018 dataset. Left: the fact description (in Chinese). Right: the corresponding English translation of the fact description. In pink is the selected rationales.

- 2: Very unfluent.
- 1: Nonsense.



Lei Sha is currently a tenure-track associate professor at Beihang University. He was a research associate at the Department of Computer Science, University of Oxford. He was advised by Prof. Thomas Lukasiewicz in the Intelligent System Lab. Previously, he was an NLP research scientist in Apple. While at Apple, he was responsible for Siri's module, such as domain classification and chat-dialogue. Before that, in 2018, he obtained his PhD degree and graduated from Peking University, China. During his PhD period, he focused on the research of learning and generating from structured data. He also served as a research assistant in Microsoft Research Asia, working with Chin-yew Lin, Lintao Zhang, and Qi Chen. He has published many cutting-edge research papers in event extraction, text entailment recognition, and text generation. These first-author papers are published in top conferences of NLP, such as ACL, EMNLP, NAACL, AAAI, etc. Also, he was the senior program committee of IJCAI 2021, and the reviewer of many top-tier conferences and journals, such as AAAI, ACL, TASLP, EMNLP, IJCNLP, etc. He also achieved many top awards, such as Lee-Wai Wing Scholarship and May 4th Scholarship, which is the top scholarship of Peking University. His research interest focuses on natural language understanding and controllable text generation



Oana-Maria Camburu is currently a Research Associate at the Department of Computer Science at University of Oxford and a co-investigator at the Alan Turing Institute on the project "Neural Networks with Natural Language Explanations". Oana has done her undergraduate and MSc studies at the Ecole Polytechnique, Paris, with a focus on Applied Mathematics and Machine Learning. She obtained her PhD from the Department of Computer Science at University of Oxford, on the topic of explainable AI. Her excellent dissertation has been nominated for the ACM Doctoral Dissertation Award 2020 (maximum 2 nominations per university among the topics of computing and engineering) as well as at the Joint AAAI/ACM SIGAI Doctoral Dissertation Award 2020 (maximum 1 nomination per university on AI topics; both currently ongoing competitions). She also obtained a J.P. Morgan PhD Fellowship on the topic of explainability. She published in top-tier venues such as NeurIPS, ACL, EMNLP, CVPR, AAAI. She is currently an organizer for two workshops: "The 6th Workshop on Representation Learning for NLP" (RePL4NLP-2021) – accepted at ACL 2021, and "Towards Explainable and Trustworthy Autonomous Physical Systems" – accepted at ACM CHI 2021. Previously, she undertook a series of internships at prestigious companies and research centers, such as Google (2 internships, in London and Mountain View), the Operational Research Center at Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, and Max Planck Institute for Mathematics, Bonn, Germany.



Thomas Lukasiewicz is a Professor of Computer Science and Yahoo! Research Fellow at OUCS since 2010 and a Faculty Fellow at the Alan Turing Institute since 2016. Prior to this, from 2004 to 2009, he was holding a prestigious Heisenberg Fellowship by the German Research Foundation (DFG), affiliated with OUCS, the Institute of Information Systems, TU Vienna, Austria, and the Department of Computer and System Sciences, Sapienza University of Rome, Italy. His research interests are in information systems and artificial intelligence (AI), including especially knowledge representation and reasoning, uncertainty in AI, machine learning, the (Social and/or Semantic) Web, and databases. He has published more than 200 publications, many of them at top-tier and leading international conferences and journals, including many highly cited papers. He received the IJCAI-01 Distinguished Paper Award (for the best paper at IJCAI-01), the AIJ Prominent Paper Award 2013 (for the best paper in the journal AIJ between 2008 and 2013), and the RuleML 2015 Best Paper Award. He has been a PC member of more than 150 conferences and workshops (more than 20 of which (co-)chaired). He has given invited talks and invited tutorials at many conferences and workshops. He is area editor for ACM TOCL, associate editor for JAIR and AIJ, and editor for Semantic Web. He has acted as principal investigator for many successfully completed grants funded by the DFG, the Austrian Science Fund, the EU, the EPSRC, and Google. He is currently the principal investigator for a seed funding grant at the Alan Turing Institute and for an EPSRC Doctoral Prize, and a co-investigator for the DBOnto and VADA EPSRC platform and programme grants, respectively.