ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240752

Adversarial Attack for Explanation Robustness of Rationalization Models

Yuankai Zhang[a,1], **Lingxiao Kong**[a,1], **Haozhao Wang**[a,*], **Ruixuan Li**[a,**], **Jun Wang**[b], **Yuhua Li**[a] **and Wei Liu**[a]

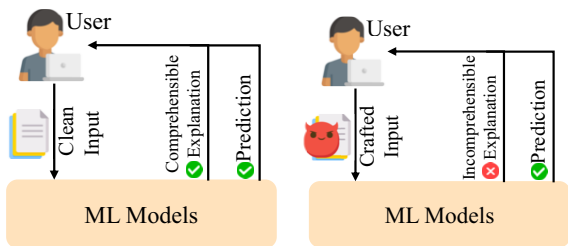
[a]School of Computer Science and Technology, Huazhong University of Science and Technology, China
[b]iWudao Tech

Abstract. Rationalization models, which select a subset of input text as rationale—crucial for humans to understand and trust predictions—have recently emerged as a prominent research area in eXplainable Artificial Intelligence (XAI). However, most of previous studies mainly focus on improving the quality of the rationale, ignoring its robustness to malicious attack. Specifically, *whether the rationalization models can still generate high-quality rationale under the adversarial attack remains unknown.* To explore this, this paper proposes *UAT2E*, which aims to undermine the explainability of rationalization models without altering their predictions, thereby eliciting distrust in these models from human users. *UAT2E* employs the gradient-based search on triggers and then inserts them into the original input to conduct both the non-target and target attack. Experimental results on five datasets reveal the vulnerability of rationalization models in terms of explanation, where they tend to select more meaningless tokens under attacks. Based on this, we make a series of recommendations for improving rationalization models in terms of explanation.

1 Introduction

Explanation of deep learning models is the key to human *comprehension* and *trust* in their predictive outcomes by providing the corresponding explanations, as illustrated in Figure 1(a), which plays an important role in affecting whether these models can be applied to critical sectors such as finance and law. Rationalization methods offer intrinsic justifications for model predictions by pinpointing salient evidence, emerging as a promising solution in the area of explainable artificial intelligence. As depicted in Figure 2, rationalization methods [21, 19, 12] employ a rationalizer to extract a semantically coherent subset of the input text, known as the rationale. This rationale is intuitively recognized by humans as a decisive determinant of the subsequent predictor's output. *By furnishing such interpretable rationales, rationalization methods significantly bolster human trust in the predictive outcomes.*

While existing studies have made great achievements in improving the explanation (i.e., rationale quality) of the rationalization models [8, 15, 1, 17, 16, 25], their explanation robustness to attack is rarely explored. Recently, Chen et al. [4] exposed the prediction vulnerability of rationalization models by inserting crafted sentences into the original input text, leading to significant changes in predictions,



(a) Clean Input (b) Crafted Input

Figure 1. Illustration of ML models with clean input and crafted input separately. (a) ML models not only returns correct prediction but also provides the comprehensible explanation to human user. (b) The explanation provided by ML models is incomprehensible for the crafted input.

as shown in Figure 2(b). Building upon this, some works [13, 27] employ adversarial training strategy to enhance the prediction robustness of these models, ensuring that the model predictions remain unchanged under adversarial attack. However, prior studies primarily focused on the robustness of rationalization methods concerning prediction while ignoring explanation. As illustrated in Figure 1(b), the models may provide incomprehensible explanation to human users when the input is crafted, leading to its reduced credibility. To this end, *whether the explanation of rationalization models is robust to the adversarial attack remains mysterious.*

In this work, we investigate the explanation robustness of the rationalization models against adversarial attacks. More specifically, as illustrated in Figure 2(c), we aim to craft and insert the attack trigger into the input text to noticeably change the rationale while keeping the prediction unchanged. In this way, the trust of human in the rationalization models can be significantly reduced. We introduce UAT2E, a variant of Universal Adversarial Triggers [23], which attacks explanations in non-target and target manner separately.[2] Specifically, UAT2E conducts the non-target attack by preventing the rationalizer from selecting the explainable tokens and conducts the target attack by limiting the rationalizer to only select the triggers. To achieve this goal, we employs the mean squared error (MSE) loss to measure the difference in rationales and leverages the cross-entropy loss to calculate the difference in predictions. Then, according to the attack mode, UAT2E adaptively constructs label sequences to align the mismatched sequence lengths that result from inserting triggers. After that, UAT2E iteratively queries words from the vocabulary using a gradient-based approach and replaces tokens in the triggers to minimize the loss.

* Corresponding Author. Email: hz_wang@hust.edu.cn
** Corresponding Author. Email: rxli@hust.edu.cn
[1] Equal contribution.

[2] https://github.com/zhangyuankai2018/UAT2E

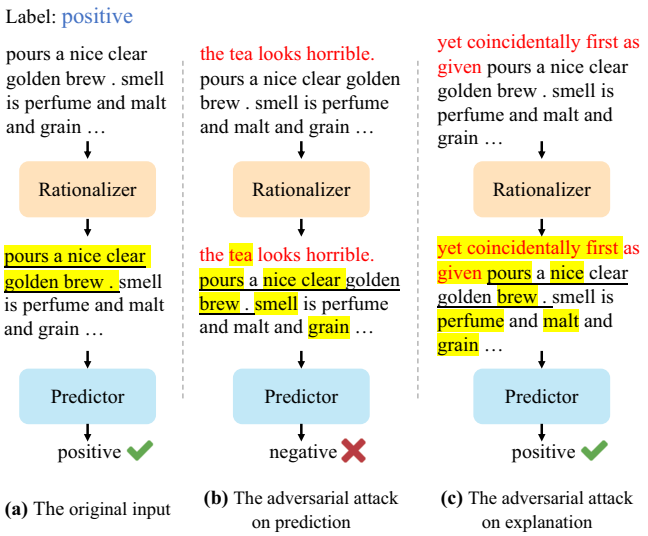


Figure 2. (a) An example from a beer review sentiment classification dataset with correct prediction and rationale. (b) Inserting "the tea looks horrible ." causes the rationalizer to select "tea", "smell", and "grain", leading to an incorrect prediction. (c) Inserting "yet coincidentally first as given" results in maintaining a correct prediction but with an obviously incorrect rationale. The underline, red, and yellow represent human-annotated rationale, triggers, and selected rationales, respectively.

Based on the designed attack strategy, we investigate the explanation robustness of several typical and recent rationalization models on five public datasets, where the experimental results reveal several important findings (more findings can be found in Section 5.3). *First*, existing rationalization models are vulnerable to the attacks on explanation including both non-target and target attacks. *Second*, the explanation vulnerablity of rationalization models arises from their inherent defects such as unmanageable sparsity, degeneration, and spurious correlation. *Third*, using powerful encoders such as BERT and supervised training with human-annotated rationales in rationalization models does not guarantee the robustness of explanations; instead, it makes the explanation more susceptible to influence of attack. Based on these findings, we present a series of recommendations for improving explanation robustness of rationalization models. Although we mainly focus on the explanation robustness of rationalization models, we believe this work can provide a cautionary note regarding the robustness of all explainable machine learning systems.

The contributions of this paper can be summarized as follows:

- To the best of our knowledge, this work is the first to investigate the explanation robustness of the rationalization models.
- We design UAT2E to conduct both non-target and target attacks over explanations of rationalization models. By employing gradient-based search to construct adversarial samples, UAT2E induces significant changes in rationale while maintaining prediction consistency.
- We conduct extensive experiments on five public datasets, revealing the fragility of existing rationalization models in terms of explanation robustness and summarizing several reasons behind these phenomena. Additionally, we provide recommendations to enhance the explanation robustness of rationalization models.

2 Related Work

The rationalization model can be categorized into two types: extractive and generative. The extractive rationalization model [8, 15, 22] involves selecting a subset from the original input to provide an explanation for the prediction. In contrast, the generative rationalization model [1, 25] uses text generation approaches to produce a piece of text explaining the prediction. While both approaches have their unique advantages, this paper focuses on related works on the robustness of the extractive rationalization model.

The prediction robustness of rationalization model Recent studies have focused on examining the prediction robustness of rationalization models. The prediction robustness refers to the model's ability to maintain its prediction unchanged when under attack. Chen et al. [4] explore the insertion of attack text into the original input by utilizing sentences from English Wikipedia or constructing them based on rules, inducing significant prediction flips. Li et al. [13] employ TextAttack to modify specific words in the meaningless token regions of the original input, such as nouns, locations, numbers, and named entities, in order to generate adversarial samples. They also use adversarial training to enhance the prediction robustness of the rationalization model. Zhang et al. [27] utilize the rationalization model as a defense strategy against adversarial attacks. They construct adversarial text using Glove and WordNet and employ adversarial training to ensure that the binary mask generated by the rationalizer effectively masks out the adversarial text, resulting in correct predictions. Unlike previous studies, we focus on the explanation robustness of rationalization models. The explanation robustness refers to the model's ability to maintain a consistent explanation when under attack. We focus on this by conducting attacks that induce significant changes in the explanation while maintaining the prediction.

Degeneration and suprious correlations Rationalization models face challenges, namely "Degeneration" [26] and "Spurious Correlation" [3]. Degeneration occurs when the predictor overfits to noise generated by an undertrained generator, causing the generator to converge to a suboptimal model that selects meaningless tokens. Several approaches have been proposed to address the degeneration problem. Yu et al. [26] introduced adversarial games and produces both positive and negative rationales. Liu et al. [15] employed a unified encoder between the generator and predictor. Liu et al. [19] assigned asymmetric learning rates to the two modules. The issue of spurious correlation arises because the maximum mutual information criterion can be influenced by false features associated with causal rationales or target labels, leading the generator to select content with false correlations. Existing works have attempted to address this problem from different perspectives, such as adopting environmental risk minimization [3] or the minimum conditional dependency criterion [18].

Adversarial attacks in NLP Adversarial attack research has played a critical role in uncovering vulnerabilities in interpretable NLP models [5]. Adversarial attack methods can be categorized based on the input perturbations, including sentence-level [11], word-level [23, 9], character-level [7], and embedding-level [14] attacks. In this study, we use Universal Adversarial Triggers [23] to identify triggers in a white-box setting. By utilizing gradient-based search, we successfully identify the optimal combination of attack triggers.

3 Problem Statement

Notation We denote the dataset by $\mathcal{D} = \{(x, y)\}$, where the input $x = x_1, x_2, ..., x_T$ consists of T sentences and each sentence $x_i = (x_{i,1}, x_{i,2}, ...x_{i,n_i})$ contains n_i tokens with y referring to the sentence label.

Extractive rationalization model As shown in Figure 2(a), a typical extractive rationalization model comprises two components, i.e., a rationalizer and a predictor, where the rationalizer selects the ra-

Table 1. Notations of the raw input and the adversarial sample.

	x	$x_{adv} = A(x, a, p)$
Embedding	$e_x = E(x)$	$e_{adv} = E(x_{adv})$
Mask	$m_x = R(e_x)$	$m_{adv} = R(e_{adv})$
Rationale	$z_x = m_x \odot e_x$	$z_{adv} = m_{adv} \odot e_{adv}$
Prediction	$\hat{y}_x = C(z_x)$	$\hat{y}_{adv} = C(z_{adv})$

tionale and the predictor makes the prediction. For each input x, the rationalizer first uses the word embedding layers to map it into vector $e = E(\theta_E; x)$ where $\theta_E \in \mathbb{R}^{|\mathcal{V}| \times d}$ denotes its parameters. Then, the rationalizer adopts the Gumbel-Softmax reparameterization [10] to sample and generate a discrete binary mask, $m = R(e) = (m_1, m_2, ..., m_L) \in \{0, 1\}^L$, from a Bernoulli distribution, where $L = \sum_{i=1}^{T} n_i$ for token-level rationale or $L = T$ for sentence-level rationale. In specific, the i-th element m_i corresponds to sentence x_i, and thus m_i will be extended according to the length n_i for sentence-level. To this end, the rationale is calculated using $z = m \odot e$ equaling to a subset from the input.

The predictor module $\hat{y} = C(z)$ makes a prediction $\hat{y}$ based on the rationale z. The overall prediction process can be defined as $M(x) = C(R(E(x)) \odot E(x))$. Rationalization models are typically trained in an end-to-end fashion, where the cross-entropy loss between predictions and labels serves as the supervised signal. In this process, the rationale z is generated unsupervised through the application of sparsity regularization.[3] Appendix B.2 [28] provides illustrations of the specific forms of sparsity regularization used in other models.

Attack of rationalization model We define triggers $a = (a_1, a_2, ..., a_K)$ as input-agnostic sequences of tokens that, when inserted into any input from a dataset, cause significant changes in the rationale while maintaining the prediction. Typically, triggers consist of K subsequences. For ease of implementation, each subsequence of triggers, $a_j = (a_{j,1}, a_{j,2}, ..., a_{j,n_a})$, has the same length n_a. The attack $A(x, a, p)$ modifies the input x by inserting triggers a at specified positions $p = (p_1, p_2, ..., p_K)$. The purpose is to ensure that the inserted triggers do not alter the semantics of individual sentence. The length of the adversarial sample $x_{adv} = A(x, a, p)$ is $L_{adv} = \sum_{i=1}^{T} n_i + K \times n_a$ for token-level rationale or $L_{adv} = T + K$ for sentence-level rationale.

In order to establish a clear distinction between the original input x and the adversarial sample x_{adv}, we provide definitions for word embedding, mask, rationale, and prediction in Table 1. It should be noted that the embedding for triggers is denoted as $e_a = E(a) = (e_{a,1}, e_{a,2}, ..., e_{a,K \times n_a})$.

Assumptions of the attack Reasonable assumptions play a crucial role in the effective evaluation of rationalization models. In the attack process, we make the assumption of having *white-box* access to a well-trained rationalization model. This access enables us to obtain the target model's structure, gradient, word embedding weight, and sparsity level. The attacks are conducted exclusively during the *inference* stage and not during the training stage.

4 Universal Adversarial Triggers to attack the explanations (UAT2E)

4.1 Attack Objective

The objective of our method is to attack the explanation robustness, specifically by *identifying the optimal trigger a^* that maximizes the difference in rationale while maintaining the prediction.*

[3] Although coherent regularization is effective, we do not consider continuity constraints in order to compare each models as fairly as possible.

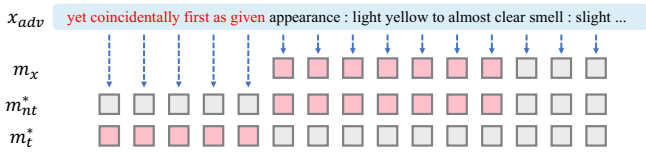


Figure 3. Examples of label sequences under non-target and target attacks. Attack triggers are highlighted in red. Grey indicates 0, and pink indicates 1.

$$a^* = \arg\max_a \mathbb{E}_x[\mathbb{D}_z(z_{adv}, z_x) - \beta \cdot \mathbb{D}_y(\hat{y}_{adv}, \hat{y}_x)] \quad (1)$$

where $\mathbb{D}_z(\cdot, \cdot)$ measures the difference between the rationales z_{adv} and z_x, $\mathbb{D}_y(\cdot, \cdot)$ measures the difference between the predictions $\hat{y}_{adv}$ and $\hat{y}_x$, and β serves as the Lagrange multiplier. We employ the Mean Squared Error (MSE) loss to calculate the difference $\mathbb{D}_z(\cdot, \cdot)$ and cross-entropy loss to compute the difference $\mathbb{D}_y(\cdot, \cdot)$. β is set to a value greater than 0.

4.2 Non-target and Target Attack

Equation (1) illustrates our intent, but inserting attack triggers leads to a mismatch in the lengths of z_{adv} and z_x. To address this, we construct a label sequence m^* from the discrete binary mask m_x to replace z_x and align the length of m_{adv}. We then calculate the difference between m_{adv} and the label sequence m^*, as shown below:

$$\max_a \mathbb{E}_x[\mathbb{D}_z(m_{adv}, m^*) - \beta \cdot \mathbb{D}_y(\hat{y}_{adv}, \hat{y}_x)] \quad (2)$$

Similar to general adversarial attack methods [2], we consider the non-target attack $\mathcal{M}_{nt}$ and target attack $\mathcal{M}_t$.

Non-target attack $\mathcal{M}_{nt}$ The goal of the non-target attack is to prevent the rationalizer module from selecting tokens previously chosen as rationale. This results in rationalization models selecting attack triggers or previously unselected tokens. To generate the label sequence for the non-target attack, denoted as $m^*_{nt} = \mathcal{M}_{nt}(m_x)$, we insert K 0-sequences of length n_a into the mask m_x at specified positions p, as depicted by m^*_{nt} in Figure 3. Noting that in the non-target attack mode, the calculation of the difference is limited to the original input segment, totaling L tokens. Furthermore, we adjust Equation (2) by replacing "maximize" with "minimize" to align the optimization process with the gradient descent method.

$$\min_a \mathbb{E}_x[-\mathbb{D}_z(m_{adv}, m^*_{nt}) + \beta \cdot \mathbb{D}_y(\hat{y}_{adv}, \hat{y}_x)] \quad (3)$$

Target attack $\mathcal{M}_t$ The goal of the target attack is to limit the rationalizer to selecting only attack triggers. The label sequence for the target attack is denoted as $m^*_t = \mathcal{M}_t(m_x)$, where the elements corresponding to the triggers are assigned a value of 1, while other positions are set to 0, as depicted by m^*_t in Figure 3. Given the goal of the target attack, we need to minimize the difference between m_{adv} and the label sequence m^*_t. The objective is shown as follows:

$$\min_a \mathbb{E}_x[\mathbb{D}_z(m_{adv}, m^*_t) + \beta \cdot \mathbb{D}_y(\hat{y}_{adv}, \hat{y}_x)] \quad (4)$$

By employing Equations (3) and (4), we can identify the optimal triggers a^*, through the standard gradient descent algorithm.

4.3 Trigger Search Algorithm

First, we initialize the attack triggers a with the character at index 1 in the vocabulary $\mathcal{V}$. Next, we insert triggers into the original input and compute the loss based on Equation (3) (or Equation (4))

depending on the attack mode. Then, we replace each token in the triggers with the one that minimizes the loss, using a greedy strategy. To determine the candidate token set, we use a KD-Tree to query the top-k closest tokens by moving each token's embedding one step, sized η, in the gradient descent direction. We iteratively execute this process until we find the optimal trigger a^* or reach the maximum number of search rounds N. More details of the attack process are in Appendix A [28].

5 Experiments

We aim to explore the explanation robustness of existing models. Our experiments are conducted with five models, five datasets, two encoders, two training settings, and two attack modes, including a total of 200 tests.

5.1 Experimental setup

Datasets We consider five public datasets: Movie, FEVER, MultiRC from ERASER [6], as well as Beer [20] and Hotel [24], two widely used datasets for rationalization. Such a setting encompasses both sentence-level and token-level rationalization tasks. More details about datasets are in Appendix B.1 [28].

Models We investigate five methods: RNP [12], VIB [21], SPECTRA [8], FR [15] and DR [19]. Details about these rationalization methods can be found in Appendix B.2 [28].

Training details All of the models are implemented using PyTorch and trained on a RTX3090 GPU. We adjust the sparsity parameter $\mathcal{S}$ based on the final sparsity level and select the model parameters based on the task performance achieved on the development dataset. Further training details can be found in Appendix B.3 [28].

Attack details We set the maximum length of adversarial samples L_{adv} to 256 and the maximum number of search rounds N to 100. We specified 5 insertion positions $p = (0, 2, 4, 6, -1)$, where "-1" represents the end position of the input. At each position, five tokens are inserted, denoted as $n_a = 5$. The initial index of triggers is set to 1. For each trigger token, we query the 15 nearest candidate tokens using a KD-tree. The attack process employs an early stopping strategy: if the triggers no longer change after 10 epochs, the search is stopped. The step size η is set to 1e4, and β is set to 0.9. Experimental results are averaged across 5 random seeds.

5.2 Evaluation Metric

We evaluate the robustness of models in terms of *task performance* and *rationale quality*.

Task performance We compare the accuracy between the original and adversarial test sets. The absolute differences in accuracy are shown below:

$$|\triangle Acc| = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} |\mathbb{1}_{[M(x)=y]} - \mathbb{1}_{[M(A(x,a,p))=y]}|$$

Here, $|\mathcal{D}|$ represents the total number of samples, $|\cdot|$ denotes the absolute value, and $\mathbb{1}$ is the indicator function. In our experiments, we considered a difference in accuracy within 5% as the threshold, indicating that the predictions are generally consistent.

Rationale quality To evaluate the impact of the attack on rationales, we employ the following six metrics. Note that all metrics are averaged on the dataset consisting of non-flipped samples, which are defined as $\mathcal{D}_{nf} = \{x|M(x) = M(A(x, a, p))\}$. This is because a prediction flip will result in a more substantial change in rationale, leading to higher metric values.

- Sparsity ($\triangle S$): This metric calculates the ratio of selected tokens to all input tokens. The sparsity level varies before and after the attack due to the different degrees of sparsity in the label sequence under non-target and target attacks. Thus, we calculate sparsity's difference before and after the attack.

$$\triangle S = \frac{||m_x||}{L} - \frac{||m_{adv}||}{L_{adv}}$$

  where $||\cdot||$ represents the l_1 norm.
- Gold Rationale F1 (GR): This metric assesses the F1 score between the rationale produced by the model and the human-annotated rationale. A decrease in GR typically indicates fragility in explanation robustness.

$$GR = \frac{||m \cap \hat{m}||}{||m \cup \hat{m}||}$$

  Here, $\hat{m}$ represents the human-annotated mask, m represents the mask generated by either the original input or the adversarial example. $\triangle GR = GR_x - GR_{adv}$, where GR_x and GR_{adv} denote the GR of the original input and the GR of the adversarial sample, respectively.
- $\triangle\widetilde{GR}$: This metric represents the relative difference of GR and is designed to facilitate the comparison of the impact on different models.

$$\triangle\widetilde{GR} = \frac{GR_x - GR_{adv}}{GR_x}$$

- Attack Capture Rate (AR): AR represents the recall of attack triggers in the rationale generated by the adversarial sample. Models exhibiting strong explanation robustness should exclude attack triggers in their selections.

$$AR = \frac{||m_{adv} \cap m_t^*||}{||m_t^*||}$$

- Rationale Shift F1 ($F1_{shift}$): This metric assesses the F1 score between rationales before and after the attack, indicating the degree of token shifting.

$$F1_{shift} = \frac{||m_{adv} \cap m_{nt}^*||}{||m_{nt}^*||}$$

  This metric assesses the consistency of the rationalizer's selection when the GR remains constant. If the rationalizer's selection shifts from one set of tokens to another, $F1_{shift}$ will decrease.
- Rationale Shift F1 on Annotation ($F1_{shift,\hat{m}}$): This metric evaluates the F1 score of the tokens selected by the model before and after the attack within the region of the human-annotated rationale.

$$F1_{shift,\hat{m}} = \frac{||m_{adv} \cap m_{nt}^* \cap \hat{m}||}{||m_{nt}^* \cap \hat{m}||}$$

  By comparing $F1_{shift}$ and $F1_{shift,\hat{m}}$, we can analyze the model's ability to recognize and retain tokens from the human-annotated rationale.

5.3 Main experiments

Finding 1: Existing rationalization models exhibit significant fragility in explanation robustness Figure 4 (a) illustrates the vulnerability of rationalization models in terms of explanation robustness, even when predictions remain unchanged. It is worth noting that directly comparing different methods is meaningless due to various factors. For instance, VIB shows much smaller $\triangle S$ than other

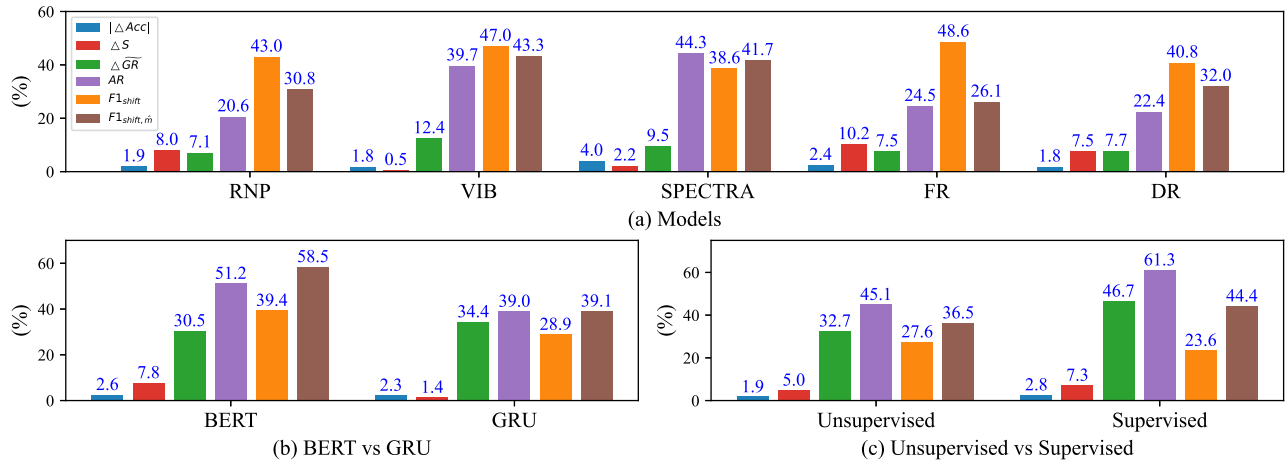


Figure 4. Comparison across different settings. We compare three settings: (a) different models, (b) using BERT or GRU as an encoder, and (c) unsupervised training and supervised training with human-annotated rationales. The comparison is conducted using five models and five datasets.

Table 2. Samples with degeneration or spurious correlations after attacks. We present samples from different datasets to illustrate cases of model degeneration or spurious correlations resulting from attacks, along with statistical figures for these issues within the impacted samples. Human annotated rationales are underlined. Rationales before attacks, triggers and rationales after attacks are highlighted in blue, red and yellow, respectively.

Model	Dataset	Context	Prediction (Label)	Issue	Percentage (%)
RNP	Beer	miniseries ruins ich smashwords ! inviting copper color with a sticky, if thin , head . caramel and apricots in the nose . pondered friar vito assassinated hitler wheat, dried apricots , caramel , and a dry finish . some lingering banana , too . cliff dismounted smashwordsSell board the hops provide a late , but pronounced character. pretty tasty . ##" hack humour rumbled ! i wouldn't say it's as outstanding as others from southern tier like pumking and creme brulee stout , though valorbadgor concluded fatigue	Positive (Positive)	Degeneration	57.0
DR	Beer	same however time as only poured from bomber into a pint glass as part of a side-by-side tasting with duck-rabbit baltic porter. a : brown-black with a thin head. came. however time but s : not much of a nose from this bottle. coffee and caramel or toffee. time as although time same t/mf : great taste : silky smooth in the mouth. malty with a toffee finish. came as same another however crisper finish than expected. d : very drinkable example of the style, good flavor combined with lighter mf makes for great experience. time though time. however	Positive (Positive)	Spurious Correlation	48.1
RNP	Hotel	however though neither fact first recently stayed 6 days there. staff was very friendly and helpful. . time this same the rooms were clean and quiet. i would rate the beds as a 5 - not the worst i 've slept in but not the best . the the the the all location is great . pics of the room are a little exaggerated. time this came this the bathroom is tiny , small closet and no dressers (in the standard room) and the elevator is a joke. we usually walked down from the 6th floor rather than wait. if you are only interested in a place to sleep and shower, this works fine. the the as the this	Positive (Positive)	Spurious Correlation	68.4
FR	Hotel	deployed stockholm !logic ! time and time again as i stand jaded by brutal service and hotel properties , a westin hotel restores my faith . the downtown location is great, staff was excellent and you can never go wrong with a heavenly bed - - the first to market and still best by my standards . mentIonal ! able u edUisLieUsly ! locals ! situation viz ! ation can normalAble !	Positive (Positive)	Degeneration	66.6
VIB	MultiRC	Where did Mr. Steadman go to get the paper ? ‖ The grocery store only neither yet quickly putting He read the telegram again. In desperation he went back to the long distance booth, but found the line still out of order, and a wire had come giving the details of the damage done by the storm. so kept next given just It would be several days before communication could be established. There was no help coming from headquarters, and from the wording of the telegram there seemed to be a reason for their not giving clear details. addition another the the another He must get a copy of the paper. Reluctantly he went to the printing office and made known his errand. it several especially yet almost Mr. Driggs was delighted to give him the paper – he had it some place, though he very seldom opened any of his exchanges. He evidently bore Mr. Steadman no ill - will for his plain talk two weeks ago. ..... Mr.Steadman remarked carelessly that there was an editorial in it to which his attention had been drawn, on hearing which Mr. Driggs turned his head and winked at an imaginary accomplice. another first first part another	Negative (Negative)	Spurious Correlation	42.6
SPECTRA	FEVER	the first inauguration of bill clinton was in the united states. ran ! ored ‡ ! the first inauguration of bill clinton as the 42nd president of the united states was held on january 20, 1993 on the west front of the united states capitol building in washington, d. c.. the inauguration marked the commencement of the first four - year term of bill clinton as president and al gore as vice president. boro navarrekia ! . wal at of age time of his first inauguration, clinton was the third - youngest person to become president, and the first from the baby boomer generation. ! ! ! ie !) 2war lu ! sselUresAUtnantTern	Positive (Positive)	Degeneration	60.7

methods. But that does not mean it's more robust than other methods. The main reason is that it samples top-k tokens while other methods sample rationales with Gumbel-Softmax. However, when sampling from the Gumbel-Softmax distribution, VIB's $\triangle S$ increases. Therefore, instead of horizontally comparing different methods, we focus on the individual performance of each method to highlight the widespread fragility in terms of explanation. Following UAT2E attacks, these models tend to shift the selection of rationales from tokens prior to the attack to meaningless tokens or triggers, resulting in a decrease in $\triangle\widetilde{GR}$, $F1_{shift}$, and $F1_{shift,\hat{m}}$, while increasing AR. **Finding 2: Rationalization models tend to exhibit degeneration or select spurious correlations when subjected to attacks** Analysis of sample cases indicates that these shifts occur because UAT2E identifies trigger combinations leading the model to experience de-

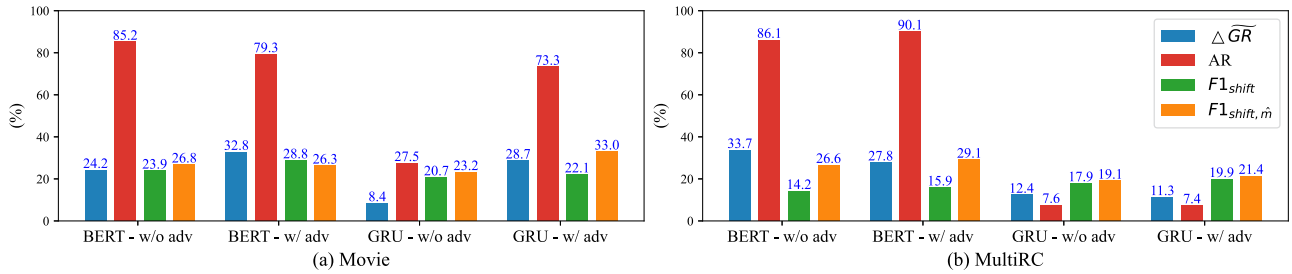


Figure 5. Evaluating the impact of improving prediction robustness on explanation robustness. We train RNP on the Movie (a) and MultiRC datasets (b). "w/o adv" and "w/ adv" represent the cases without and with adversarial training, respectively.

generation or select spurious correlations, as presented in Table 2. In the beer dataset, 57.0% of the samples exhibit a higher tendency to select ",", ".", or other meaningless tokens after being attacked, while 48.1% of the samples choose non-appearance-related content.

Finding 3: Using a powerful encoder or supervised training with human-annotated rationales fails to mitigate degeneration and spurious correlations resulting from attacks Compared to GRU, the model with BERT demonstrates higher GR on sentence-level datasets and experiences a greater GR boost through supervised training before attacks. When considering the rationale quality before and after attacks, BERT-based models generally experience greater impact, resulting in larger $\triangle GR$ values. However, the $\triangle\widetilde{GR}$ values are smaller due to the higher pre-attack GR. The disparity in sparsity indicates that BERT is more vulnerable to attacks, resulting in a more significant increase in sparsity compared to GRU. Consequently, precision and GR experience a decrease, as illustrated in Figure 4 (b). Models trained under supervision with human-annotated rationales fail to prevent the impact of attacks. Specifically, they tend to select more meaningless tokens and triggers while discarding the originally selected human-annotated rationale tokens. This also leads to increased sparsity and decreases in precision, recall, and the GR score, as shown in Figure 4 (c).

These two approaches help the model recognize rationales and provide more accurate gradient information, which assists UAT2E in selecting tokens that can induce model degradation or spurious correlations. Notably, the discrepancy between $F1_{shift}$ and $F1_{shift,\hat{m}}$ is more pronounced in both situations. This indicates that using the two approaches is better at identifying and preserving tokens annotated as rationales by humans, resulting in more shifts occurring on meaningless tokens.

Finding 4: Enhancing prediction robustness does not effectively improve explanation robustness We conduct experiments on the Movie and MultiRC datasets to investigate whether improving prediction robustness can enhance explanation robustness. Specifically, we train RNP on the Movie, MultiRC, Movie_ADV, and MultiRC_ADV datasets. The Movie_ADV and MultiRC_ADV datasets, mentioned in [13], are used for adversarial training to enhance prediction robustness. Following the model training, we perform non-target attacks, and the experimental results are depicted in Figure 5. It is noteworthy that we employ $\triangle\widetilde{GR}$ to intuitively compare the impact on models with different encoders. The results indicate that enhancing prediction robustness through adversarial training does not significantly improve explanation robustness. Particularly, RNP with GRU on the Movie dataset experiences a more pronounced impact after adversarial training.

Finding 5: Utilizing gradient-based search in attacks to facilitate trigger selection We conduct transferability tests using the identified triggers. Specifically, we transfer the triggers from a source model to a target model and assess the attack effects, as depicted in Figure 6. It is worth noting that we do not have access to any information about the target model in order to demonstrate the effectiveness of trigger transfer in a black-box setting. Despite a slight reduction in effectiveness, the triggers are still capable of influencing the target model. However, the AR value is lower, which could potentially be attributed to the absence of a gradient-based search, making it challenging for the model to select triggers.

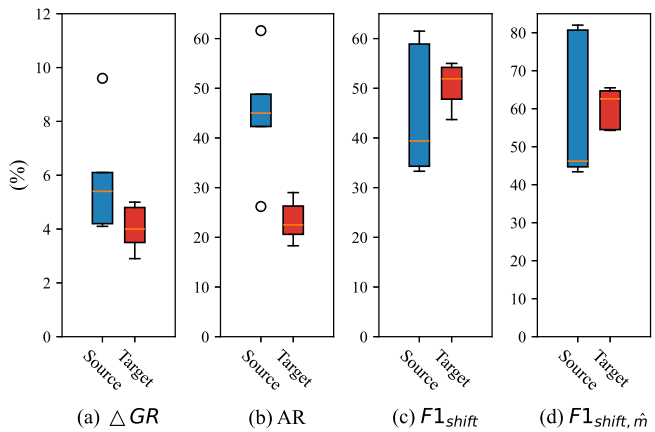


Figure 6. Evaluating the transferability of identified triggers. We conduct tests on five models using BERT as the encoder, across three datasets: Beer, Movie, and FEVER. Triggers identified on one model (source model) are then transferred to other models (target models). *Blue*: Triggers apply to the source model, *Red*: Triggers are transferred to target models, and the experimental results are averaged.

5.4 Ablation Study

We conducted non-target attacks on VIB, utilizing GRU as the encoder, on three datasets: Hotel, Movie, and MultiRC, and the results are shown in Figure 7. When comparing Mean Absolute Error (MAE) with Mean Squared Error (MSE) as a rationale measurement function, the use of MSE yields a more significant attack effect. This discrepancy arises from MSE more prominently capturing the differences between m^{adv} and the label sequence m^*. The attack effects of calculating MSE in the embedding space and executing attacks by randomly selecting words in each round are similar. This is because the former emphasizes distinctions between word vectors rather than masks, resulting in gradient fluctuations and a reduction in the attack effect. Additionally, when employing the HotFlip [7] query method, the attack effect closely resembles that of randomly selecting words in each round.

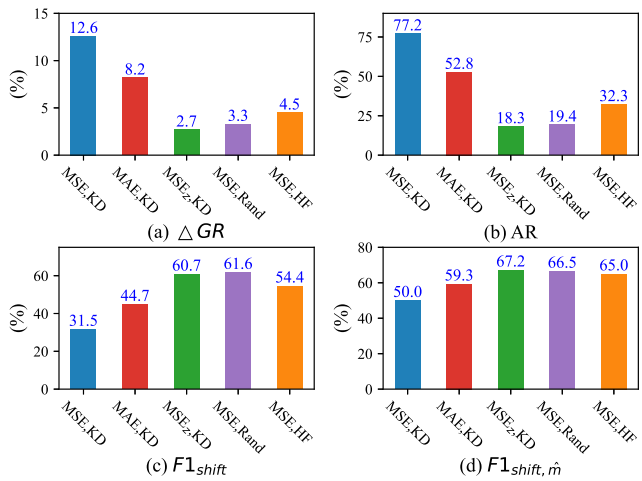


Figure 7. The impact of measurement functions and query methods. We employed MSE and MAE as measurement functions. "MSE_z" was used to calculate the differences in rationale embeddings, specifically $MSE_z(m \odot e_x, m^* \odot e_x)$. For querying candidate tokens, we uses the KD-Tree, random selection in each round, or the HotFlip [7] method.

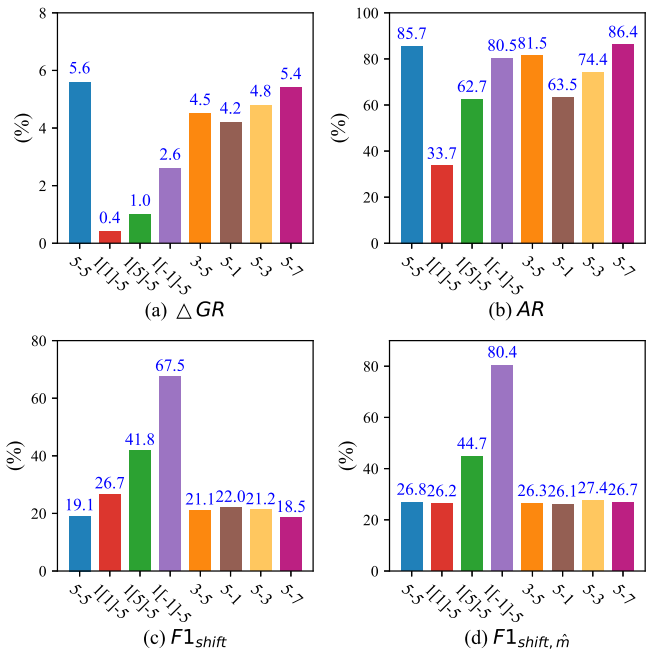


Figure 8. Comparison various insertion positions and number of triggers. We conducted non-target attacks on RNP using the Movie and MultiRC datasets. "*5-5*": insert 5 groups of triggers, each with 5 tokens. "*1[p]-5*": insert 1 group of triggers before the *p*-th sentence. The "-1" position means insert 1 group of triggers at the end of the document. For the 5-group strategy, we specified positions as p=(0,2,4,6,-1), and for the 3-group strategy, p=(0,4,-1).

5.5 *Triggers Position And Number*

We examine the impact of different insertion positions and numbers of triggers on model performance, as depicted in Figure 8. Gradually increasing the number of trigger groups and tokens per group intensifies their influence on the model. Notably, the variation in the number of trigger sets exhibits a more significant effect on model performance. Additionally, triggers inserted near the end of the document have a greater impact. This observation aligns with previous findings by Chen et al. [4] and can be attributed to the fact that rationale positions in Movie and MultiRC datasets are typically located close to the end of the document.

6 Recommendations

Based on our experimental results and analysis, we offer several recommendations to researchers and practitioners:

Conducting rigorous evaluations of rationalization models Researchers should assess both task performance and rationale quality by implementing various types of attacks on explanations and predictions. This examination helps determine whether rationalization models exhibit both high prediction and explanation robustness.

Exploring defense mechanisms to enhance explanation robustness Our experimental findings reveal rationalization models suffer from issues such as degeneration and spurious correlation after being attacked. Therefore, researchers should explore the development of defense mechanisms to protect rationalization models from attacks and reduce the occurrence of degeneration and spurious correlation.

Establishing robustness evaluation benchmarks and metrics It is imperative for researchers to construct benchmarks that facilitate standardized and rigorous evaluation of model robustness. Such benchmarks enable the identification of strengths and weaknesses across various models using unified criteria. Additionally, it is important to note that Gold Rationale F1 (GR) may not accurately reflect rationale shifts. Therefore, developing more effective evaluation metrics is essential for measuring these shifts accurately.

7 Conclusion

In this study, we investigate the robustness of rationalization models in terms of explanation. To explore this, we propose UAT2E, a variant of Universal Adversarial Triggers. UAT2E attacks explanations in non-target and target manner separately, resulting in significant shifts in rationales while maintaining predictions.

Based on the experimental findings, it is evident that existing rationalization models generally exhibit vulnerabilities in explanation, making them susceptible to attacks that result in significant shifts in rationales. These vulnerabilities can be attributed to degeneration or spurious correlations after being attacked. Furthermore, despite employing techniques to improve rationale quality, such as using more powerful encoders or utilizing supervised training with human-annotated rationales, the explanation robustness of rationalization models does not significantly improve.

Based on our findings, we supplement our findings with a series of recommendations for enhancing the explanation robustness of rationalization models.

8 Ethics Statement

The data and resources used in this study are publicly available and have been widely used in previous research. It is important to note that in our experiments, some triggers consist of email addresses, which are sourced from a vocabulary corpus dataset. This has the potential to result in the disclosure of personal privacy.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This work is supported by National Natural Science Foundation of China under grants 62376103, 62302184, 62206102, Science and Technology Support Program of Hubei Province under grant 2022BAA046, and CCF-AFSG Research Fund.

References

[1] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. doi: 10.1109/SP.2017.49.

[3] S. Chang, Y. Zhang, M. Yu, and T. Jaakkola. Invariant rationalization. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1448–1458. PMLR, 13–18 Jul 2020.

[4] H. Chen, J. He, K. Narasimhan, and D. Chen. Can rationalization improve robustness? In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.278.

[5] Z. Chen, F. Silvestri, J. Wang, Y. Zhang, and G. Tolomei. The dark side of explanations: Poisoning recommender systems with counterfactual examples, 2023. URL https://arxiv.org/abs/2305.00574.

[6] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408.

[7] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou. HotFlip: White-box adversarial examples for text classification. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2006.

[8] N. M. Guerreiro and A. F. T. Martins. SPECTRA: Sparse structured text rationalization. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6534–6550, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.525.

[9] D. Herel, H. Cisneros, and T. Mikolov. *Preserving Semantics in Textual Adversarial Attacks*. IOS Press, Sept. 2023. ISBN 9781643684376. doi: 10.3233/faia230376.

[10] E. Jang, S. Gu, and B. Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net, 2017.

[11] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215.

[12] T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1011.

[13] D. Li, B. Hu, Q. Chen, T. Xu, J. Tao, and Y. Zhang. Unifying model explainability and robustness for joint text classification and rationale extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10947–10955, Jun. 2022. doi: 10.1609/aaai.v36i10.21342.

[14] L. Li and X. Qiu. Token-aware virtual adversarial training in natural language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8410–8418, May 2021. doi: 10.1609/aaai.v35i9.17022.

[15] W. Liu, H. Wang, J. Wang, R. Li, C. Yue, and Y. Zhang. Fr: Folded rationalization with a unified encoder. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6954–6966. Curran Associates, Inc., 2022.

[16] W. Liu, H. Wang, J. Wang, Z. Deng, Y. Zhang, C. Wang, and R. Li. Enhancing the rationale-input alignment for self-explaining rationalization. *arXiv preprint arXiv:2312.04103*, 2023.

[17] W. Liu, H. Wang, J. Wang, R. Li, X. Li, Y. Zhang, and Y. Qiu. MGR: Multi-generator based rationalization. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12771–12787, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.715.

[18] W. Liu, J. Wang, H. Wang, R. Li, Z. Deng, Y. Zhang, and Y. Qiu. D-separation for causal self-explanation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 43620–43633. Curran Associates, Inc., 2023.

[19] W. Liu, J. Wang, H. Wang, R. Li, Y. Qiu, Y. Zhang, J. Han, and Y. Zou. Decoupled rationalization with asymmetric learning rates: A flexible lipschitz restraint. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1535–1547, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599299.

[20] J. McAuley, J. Leskovec, and D. Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025, 2012. doi: 10.1109/ICDM.2012.110.

[21] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer. An information bottleneck approach for controlling conciseness in rationale extraction. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.153.

[22] P. Saha, D. Sheth, K. Kedia, B. Mathew, and A. Mukherjee. Rationale-guided few-shot classification to detect abusive language. In *ECAI 2023*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 2041–2048. IOS Press, 2023.

[23] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for attacking and analyzing NLP. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221.

[24] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 783–792, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300551. doi: 10.1145/1835804.1835903.

[25] P. Wang, A. Chan, F. Ilievski, M. Chen, and X. Ren. Pinto: Faithful language reasoning using prompt-generated rationales. *arXiv preprint arXiv:2211.01562*, 2023.

[26] M. Yu, S. Chang, Y. Zhang, and T. Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1420.

[27] Y. Zhang, Y. Zhou, S. Carton, and C. Tan. Learning to ignore adversarial attacks. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2970–2984, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.216.

[28] Y. Zhang, L. Kong, H. Wang, R. Li, J. Wang, Y. Li, and W. Liu. Adversarial attack for explanation robustness of rationalization models. *arXiv preprint arXiv:2408.10795*, 2024. URL https://arxiv.org/abs/2408.10795.