# Unifying Model Explainability and Robustness for Joint Text Classification and Rationale Extraction

**Dongfang Li[1], Baotian Hu[1]*, Qingcai Chen[1,2]*, Tujie Xu[1], Jingcong Tao[1], and Yunan Zhang[1]**

[1] Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China
crazyofapple@gmail.com, hubaotian@hit.edu.cn, qingcai.chen@hit.edu.cn

## Abstract

Recent works have shown explainability and robustness are two crucial ingredients of trustworthy and reliable text classification. However, previous works usually address one of two aspects: i) how to extract accurate rationales for explainability while being beneficial to prediction; ii) how to make the predictive model robust to different types of adversarial attacks. Intuitively, a model that produces helpful explanations should be more robust against adversarial attacks, because we cannot trust the model that outputs explanations but changes its prediction under small perturbations. To this end, we propose a joint classification and rationale extraction model named AT-BMC. It includes two key mechanisms: mixed Adversarial Training (AT) is designed to use various perturbations in discrete and embedding space to improve the model's robustness, and Boundary Match Constraint (BMC) helps to locate rationales more precisely with the guidance of boundary information. Performances on benchmark datasets demonstrate that the proposed AT-BMC outperforms baselines on both classification and rationale extraction by a large margin. Robustness analysis shows that the proposed AT-BMC decreases the attack success rate effectively by up to 69%. The empirical results indicate that there are connections between robust models and better explanations.

## Introduction

Neural models have demonstrated their superior ability on text classification task, especially when based on pre-trained language models (PLMs) (Devlin et al. 2019; Liu et al. 2019). However, they are more like black boxes compared to traditional machine learning methods such as logistic regression and decision tree. It is notoriously difficult to understand why neural models produced particular predictions (Samek, Wiegand, and Müller 2017; Rudin 2019). One practical approach is to extract prediction's rationales from input (Lipton 2016; Camburu et al. 2018; Thorne et al. 2019). The rationales can be defined as text snippets or subsets of the input text. The assumption is that a correct prediction can be made from the rationale alone (Lei, Barzilay, and Jaakkola 2016; Bastings, Aziz, and Titov 2019; DeYoung et al. 2020). In other words, rationales should be sufficient to support the model's prediction. Our work also falls into this scope, which aims to achieve better prediction performance and model explainability by extracting prediction closely-related rationales.

**Golden Label:** Positive (100%) → Negative (80%)

**Original Text:**

"The Saint Takes Over" stars George Sanders as Simon Templar, aka "The Saint" in this 1940 entry into the series. It also stars Wendy Barrie, Jonathan Hale and Paul Guilfoyle ...*It is very enjoyable.*

**Adversarial Example:**

"The Saint Takes Over" stars Halo Sanders as Kathrina Templar, aka "The Saint" in this 988 entry into the series. It also stars Wendy Barrie, Jonathan Hale and Paul Cobb ... *It's very enjoyable.*

Figure 1: We test the pre-trained BERT-base movie review classifier using one adversarial example generated by CHECKLIST (Ribeiro et al. 2020). The general meaning of the text remains unchanged. However, the predictions of the model change from *positive* to *negative*, while the associated rationales which enable users to verify the predictions quickly do not change. Human-labeled rationales are shown in the *italic* font.

Previous works proposed to use the pipeline approach where task prediction is performed in two steps: the explanation phase and the subsequent prediction phase (Lei, Barzilay, and Jaakkola 2016). The challenge is to attain superior task performance conditioned on the extractive rationale. Since most works that employ this framework tend to rely solely on task labels, they sample rationales from input in the explanation phase. For example, these models simulate the intractable sampling step by proposing optimization procedures based on reinforcement learning approaches (Lei, Barzilay, and Jaakkola 2016; Yoon, Jordon, and van der Schaar 2019) and reparameterization techniques (Bastings, Aziz, and Titov 2019; Latcinnik and Berant 2020), which may be sensitive to hyperparameters and requires complicated training process (Jain et al. 2020). Instead, we optimize the joint likelihood of class labels and extractive rationales for the input examples. Although it is a relatively straightforward way to optimize the explanation phase models, there are at least two

*Corresponding authors

challenges for this task. Firstly, previous works are vulnerable to different types of adversarial attack. For example, a classifier suffers from defending against the labeling-preserving adversaries as shown in Table 1. If adding small perturbations to the input modifies the model's prediction, we cannot trust the explanations output from the model. We further analyze existing methods suffer from text attacks by using robustness test, which performs model-agnostic attacks on the trained classification model. Secondly, the explicit boundary information is ignored, leading to inaccurate extraction. For example, "interesting" and "inspiring" are boundaries of the rationale for the text "this film is interesting and inspiring.", while "is" and "." are general tokens whose representation is different from emotion words. Besides, models that use rationales to train explanation phase models do not consider the supervision signal from task (DeYoung et al. 2020).

To address these challenges, we propose a joint classification and rationale extraction framework AT-BMC where *task prediction* and *rationale extraction* are learned jointly with mixed Adversarial Training (AT) and Boundary Match Constraint (BMC). Firstly, we apply perturbation in both the discrete text space and the embedding space to improve both the generalization and robustness of the model. On the one hand, we generate adversarial examples at word-level while preserving the rationale unchanged. The perturbations also maintain prediction invariance. On the other hand, our adversarial training in the embedding space refines the standard adversarial training (Madry et al. 2018) in computation efficiency and training smoothness. Secondly, we consider matching constraints by modeling both the boundary positions, which allows the model to further focus on the boundary-relevant regions. The main idea of boundary constraint is to make the sequence labeling model to consider the boundary information when locating entities. By matching a predicted start index of a rational span with its corresponding end index, the global sequence labeling information is fused with local region-aware information. In addition, we condition the extraction models on the classification label through label embedding. We conduct extensive experiments on two benchmark datasets (i.e., *Movie Reviews* and *MultiRC*). The experimental results demonstrate that AT-BMC outperforms the competitive baselines on classification and rationale extraction by a large margin. Robustness analysis further shows that AT-BMC can effectively improve the robustness of the models where the attack success rate decreases from 96% to 27% under strong adversarial attacks. The code is available at https://github.com/crazyofapple/AT-BMC.

The contributions of this paper are summarized as follows:

- Different from previous methods such as pipelines, we propose AT-BMC for joint classification and rationale extraction, which allows two parts to improve each other. We also show that our approach can be applied in the context of only limited annotated examples.
- To the best of our knowledge, this is the first work that considers explainability and robustness both in one text classification model. As a step towards understanding the connection between explainability and robustness, we provide evidence that robust models lead to better rationales.

## Related Work

**Rationale Extraction** The task aims to extract snippets that can support prediction in input sequences. These text snippets allows humans to verify the correctness of predictions quickly (Zaidan and Eisner 2008; Zhang, Marshall, and Wallace 2016; Ross, Hughes, and Doshi-Velez 2017; DeYoung et al. 2020). For example, Paranjape et al. (2020a) leverage the information bottleneck principle to extract rationales of desired conciseness. Sha, Camburu, and Lukasiewicz (2020) propose a selector-predictor method to squeeze information from the predictor to guide the selector in extracting the rationales. Unlike post-hoc methods (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017) where explanation does not come directly with the prediction, we focus on joint extraction of rationales and task predictions in this paper.

**Adversarial Training** Adversarial training is used to improve the generalization ability and robustness of the model and has proven effective in various tasks (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018; Ribeiro et al. 2020). Some previous works use text perturbation by changing discrete inputs to generate adversarial training samples, and then the model is retrained again to improve the ability to defend some attacks (Jia and Liang 2017; Wang and Bansal 2018; Michel et al. 2019). For example, CLARE (Li et al. 2021) modifies the inputs by using pre-trained masked language models in a context-aware manner. These data augmentation methods assume that generating textual adversarial examples by sophisticated word or phrase perturbations would not change the labels. However, one limitation is that it is hard to enumerate all text manipulations (Zang et al. 2020). The other type of adversarial training is to add gradient-based perturbations in the embedding space. Recent works have shown that this method achieves performance improvement with pre-trained language models (Jiang et al. 2020; Liu et al. 2020; Cheng et al. 2021). For example, virtual adversarial training (Zhu et al. 2019) does not generate explicit adversarial examples. Instead, it samples random perturbations from the $\epsilon$-sphere surrounding the input and uses continuous optimization methods (e.g., smoothness regularization) to train the model. In this paper, we use mixed adversarial training for models to have the best of both worlds.

**Connecting Explainability with Adversarial Robustness** Previous work has shown that neural networks are easy to be attacked (Li et al. 2019; Ren et al. 2019; Jin et al. 2020), which naturally brings about the question of whether the application of interpretable technology to explain the predictive behavior of the model will be affected by the attack (Alvarez-Melis and Jaakkola 2018). Some previous works (Augustin, Meinke, and Hein 2020; Datta et al. 2021) empirically observed that robust models can be more explainable in computer vision. And though some recent studies (Etmann et al. 2019; Moshkovitz, Yang, and Chaudhuri 2021) have focused on linking explainability and adversarial robustness, there is no explicit statement about existing models have both two properties. On the other hand, our goal is to focus on understanding the connection between the two in text classification tasks, and we hope it sheds light on the future development of such methods in natural language processing tasks.

# Method

## Text Classification with Rationale Extraction

The aim of this paper is to design a model that can yield accurate predictions and provide closely-related extractive rationales (i.e., supporting evidence) as potential reasons for predictions. Taking the sentiment classification as an example, for the text "*titanic is so close to being the perfect movie...*", the predictive label of it is *positive*, and one of rationales for this prediction is "*so close to being the perfect movie*". Therefore, text classification with rationale extraction can be formalized as follows. Given a sequence of words as input, namely $\mathbf{x} = [x_1, \cdots, x_L]$, where $L$ is the length of the sequence and $x_i$ denotes the $i$-th word. The goal is to infer the task label $\hat{y}$ and to assign each word $x_i$ with a boolean label denoted as $\hat{e}_i \in \{0, 1\}$, where $\hat{e}_i = 1$ indicates word $i$ is a part of the rationale. We denote the rationale of the sequence as $\hat{\mathbf{e}} \in \{0, 1\}^L$. The corresponding golden label is denoted as $y$ and human-labeled rationale is denoted as $\mathbf{e}$, both of them is used for training. Here, rationales are sequences of words and hence a potential rationale is a sub-sequence of the input sequence. Note that one text sample may contain multiple non-overlapping sub-sequences as rationale spans.

## Overall Framework

We propose to construct a model that consists of an extraction network $g$ and a prediction network $f$, given the training data consisting of $n$ points $\mathcal{D} = \{(\mathbf{x}_1, y_1)...(\mathbf{x}_n, y_n)\}$. The prediction network $f$ first maps the text sequence $\mathbf{x}_i$ to the task output $\hat{y}_i$. At the same time, the input $\mathbf{x}_i$ is also fed into rationale extraction network $g$ to output supporting evidence $\mathbf{e}_i$. The basic scheme follows an multi-task learning (MTL) framework (Caruana 1997) with two tasks – (1) rationale extraction and (2) the actual prediction task. We adopt the shared encoder architecture of MTL where both the tasks share the same encoder $\mathbf{enc}(\cdot)$ but different decoders. Formally, the conditional likelihood of the output labels and evidence, given the input, can be written as:

$$L = \sum_{i=1}^{n} \log p(y_i, \mathbf{e}_i | \mathbf{x}_i). \quad (1)$$

Note that given rationale data, our objective is to learn $\mathbf{enc}(\cdot)$, $f(\cdot)$, and $g(\cdot)$ that predict both the task labels $y_i$ and rationale labels $\mathbf{e}_i$ given $\mathbf{x}_i$. We assume that only $m$ points have evidence annotations $\mathbf{e}$. Therefore, we can factorize the likelihood as follows:

$$L = \sum_{i=1}^{n} \log\ p(y_i|\mathbf{x}_i)\ + \sum_{j=1}^{m} \log\ p(\mathbf{e}_j|\hat{y}_j, \mathbf{x}_j). \quad (2)$$

Consequently, the predicted label and the text sequence $\mathbf{x}_i$ are fed into the extraction network $g$ to generate the evidence labels $\hat{\mathbf{e}}_i$. The overall architecture of our approach is presented in Figure 2. Since we can optimize the classification objective for all $n$ instances and the extraction objective for $m$ instances, it allows us to train the model when only part of samples have human-labeled rationales.

Concretely, we first encode the input into hidden representation by using pre-trained language models (Devlin et al. 2019; Liu et al. 2019) as the shared encoder. Then we use a linear classifier to model $p_\theta(y|\mathbf{x})$ with the cross entropy loss $\mathcal{L}_{classify}$ and a linear-chain CRF (Lafferty, McCallum, and Pereira 2001) to model $p_\phi(\mathbf{e}|\mathbf{x}, y; \theta)$ with negative log-likehood $\mathcal{L}_{extract}$. The outputs are predicted label by the linear classifier and rationale spans generated by the CRF decoder. Inspired by Wang et al. (2018), we also condition the extraction model on the predicted label output from the classification model. We implement this by using an embedding lookup layer, and add the label embedding to each token representation of encoder. Moreover, applying label-specific embedding can help to validate different behaviour of the rationales via changing the $\hat{y}_i$.

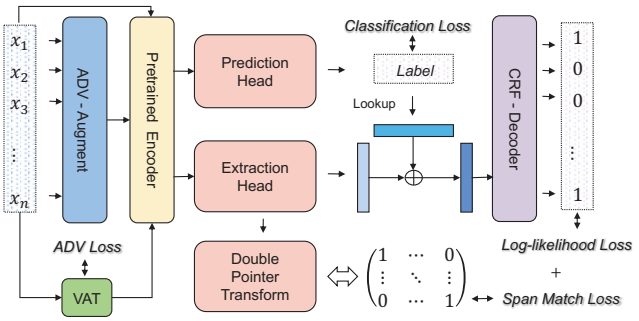



Figure 2: Overall architecture of the proposed AT-BMC method for joint classification and rationale extraction with mixed adversarial training and boundary match constraint.

## Mixed Adversarial Training

Since the searching space of adversarial attacks is large and marked rationales is limited, we perform discrete adversarial attack based data augmentation on the samples with rationale. By introducing the word-level perturbed versions of existing samples, we can recursively reuse augmentation to significantly expand the training data set. For simplicity, the validation here only considers adding one new textual edit for each sample. In addition, considering the label-preserving of rationales, perturbations only include those parts of the sentence that are outside of the rationales.

Specifically, we change text by using the linguistic transformations following the CHECKLIST behavioral testing (Ribeiro et al. 2020). We use 4 invariance test by using *TextAttack* (Morris et al. 2020), including name replacement, position replacement, number change, and contraction/expansion of named entities. The invariance test is to apply perturbations that retain the label to the input and to expect the model predictions to remain unchanged. The more detail descriptions of each transformation are: (a) name replacement refers to the conversion of the input by replacing the name of the recognized name entity; (b) location conversion refers to the conversion of the recognized location entity of a sentence to another location given the location dictionary; (c) number change refers to the recognition of the number in the sentence to return the sentence with the changed number; and (d) the last transformation refers to the expansion or contraction of the identified name entity combination. All re-

placement words (from pre-defined named entity dictionary) have the same part-of-speech as the original word tagged by *flair* (Akbik et al. 2019). Here, the percentage of words to replace per augmented example is set to 0.2.

Apart from applying perturbation to the input text directly, we also leverage adversarial training operated on the embedding space as an effective regularization to improve the shared encoder $\mathbf{enc}(\cdot)$ generalization and reduce robust error. It aims to minimize the following objective:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\{\mathcal{L}_{adv} + \mathcal{L}_{classify}; \boldsymbol{\theta}\}\,, \tag{3}$$

$$\mathcal{L}_{adv} = \alpha \cdot \mathcal{L}_{pat} + \beta \cdot \mathcal{L}_{kl}\,, \tag{4}$$

where $\mathcal{L}_{classify} = L(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{y})$ is the cross-entropy loss on original data, $\mathcal{L}_{pat}$ is the perturbations-based adversarial training loss (PAT), and $\mathcal{L}_{kl}$ is a smoothing adversarial regularization term. $\alpha$, $\beta$ is a hyperparameter. We define the PAT loss as:

$$\mathcal{L}_{pat}(\boldsymbol{\theta}) = \max_{||\boldsymbol{\delta}||\leq\epsilon} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{\delta}), \boldsymbol{y})\,, \tag{5}$$

where $L$ is the cross-entropy loss on adversarial embeddings and $\boldsymbol{\delta}$ is the perturbation. We constrain $\boldsymbol{\delta}$ by using Frobenius norm bounded by $\epsilon$. The outer minimization in Eqn. (3) can be dealt with SGD for optimization. And the inner maximization in Eqn. (5) can be solved reliably by projected gradient descent (PGD) (Madry et al. 2018). It is a standard method for large-scale constrained optimization, which takes the following step with step-size $\eta$ at each iteration:

$$\boldsymbol{\delta}_{t+1} = \Pi_{||\boldsymbol{\delta}||\leq\epsilon}(\boldsymbol{\delta}_t + \eta g(\boldsymbol{\delta}_t)/||g(\boldsymbol{\delta}_t)||_F)\,, \tag{6}$$

where $g(\boldsymbol{\delta}_t) = \nabla_{\boldsymbol{\delta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{\delta}), \boldsymbol{y})$ is the gradient of the loss w.r.t. $\boldsymbol{\delta}$, and $\Pi_{||\boldsymbol{\delta}||\leq\epsilon}$ performs a projection onto the $\epsilon$-ball. To further regularize the trade-off between standard objective and adversarial objective, we consider label smoothness in the embedding neighborhood by using term $\mathcal{L}_{kl}(\boldsymbol{\theta})$, which is defined as:

$$\mathcal{L}_{kl}(\boldsymbol{\theta}) = \max_{||\boldsymbol{\delta}||\leq\epsilon} L_{kl}(f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{\delta}), f_{\boldsymbol{\theta}}(\boldsymbol{x}))\,, \tag{7}$$

where $L_{kl}(p,q) = [\mathrm{KL}(p||q) + \mathrm{KL}(q||p)]/2$, $p$, $q$ denote the two probability distributions, and $\mathrm{KL}(\cdot)$ denotes the Kullback-Leibler divergence with temperature equals 1.0. In contrast to Eqn. (5) which promotes adversarial attacks that retain labels, Eqn. (7) further asserts that the confidence level of the prediction should also be similar on the probability vector, which is characterized by the simplex form $\Delta$ where its dimension equals the number of classes.

Compared with standard training, $K$-step PGD requires $K$ forward-backward passes through the network, which is computationally expensive. Besides, only the last step of perturbation is used for model parameter update after $K$-step. We follow free adversarial training framework in FreeLB (Zhu et al. 2019) to perform multiple PGD iterations to construct the adversarial embedding and iterate the cumulative parameter gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}$ in each iteration. Afterward, the model parameters $\boldsymbol{\theta}$ are updated one at a time with the accumulated gradients effectively, by virtually creating one batch that is $K$ times larger than sampled mini-batch. For convenience, we provide the details of adversarial training on embedding space in Algorithm 1.

Algorithm 1: Embedding-level Adversarial Training Algorithm

**Require:** Training samples $\mathcal{D}$, perturbation bound $\epsilon$, learning rate $\tau$, ascent steps $K$, ascent step size $\eta$, the total number of epochs $T$, the variance of the random initialization $\sigma^2$.

Initialize $\boldsymbol{\theta}$
**for** epoch $= 1 \ldots T$ **do**
  **for** each batch $B$ sampled from $D$ **do**
    $\delta \sim \mathcal{N}\left(0, \sigma^2 I\right)$
    $\boldsymbol{g}_0^{\theta} \leftarrow 0$
    **for** $t = 1 \ldots K$ **do**
      Accumulate gradient of $\boldsymbol{\theta}$ given $\boldsymbol{\delta}_{t-1}$:
      $\boldsymbol{g}_t^{\theta} \leftarrow \boldsymbol{g}_{t-1}^{\theta} + \frac{1}{K}\mathbb{E}_B[\nabla_{\boldsymbol{\theta}}(\mathcal{L}_{adv} + \mathcal{L}_{classify})]$
      Update the perturbation $\boldsymbol{\delta}$ via gradient ascend:
      $\boldsymbol{g}_{adv}^{\boldsymbol{\delta}} \leftarrow \nabla_{\boldsymbol{\delta}}\,\mathcal{L}_{adv}$
      $\boldsymbol{\delta}_t \leftarrow \Pi_{\|\boldsymbol{\delta}\|_F\leq\epsilon}(\boldsymbol{\delta}_{t-1} + \eta \cdot \boldsymbol{g}_{adv}^{\boldsymbol{\delta}}/\|\boldsymbol{g}_{adv}^{\boldsymbol{\delta}}\|_F)$
    **end for**
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \tau\boldsymbol{g}_K^{\theta}$
  **end for**
**end for**

## Boundary Match Constraint

For rationale extraction, the start/end boundaries can be captured by CRF decoder. As the CRF learns the conditional probability of the label sequence given the observation sequence features, it can be seen as maximum log-likelihood objective function conditioned on the observation $X$. However, CRF has the limitation of occasionally generating illegal sequences of tags, as it only encourages legal transitions and penalizes illegal ones softly (Wei et al. 2021). Hence, we propose to use a boundary constraint to encourage it to be more accurate when positioning the boundary.

The basic idea of boundary constraint is to match a predicted start index of a rational span with its corresponding end index. Given the sequence hidden representations $\mathbf{H}$ output from $\mathbf{enc}(\cdot)$, we first predict the probability of each token as the starting indices, as follows:

$$\mathbf{P}_{\mathrm{s}} = \mathrm{Softmax}(\mathbf{H}\mathbf{W}_{\mathrm{s}}) \in \mathbb{R}^{L\times 2}\,, \tag{8}$$

where $\mathbf{W}_{\mathrm{s}} \in \mathbb{R}^{d\times 2}$ is the weights to learn and $d$ is the hidden size. Each row of $\mathbf{P}_{\mathrm{s}}$ presents the probability distribution of each index being the start position of an word. Similarly, we can calculate the end index prediction logits by using another matrix $\mathbf{W}_{\mathrm{e}}$ to obtain probability matrix $\mathbf{P}_{\mathrm{e}}$:

$$\mathbf{P}_{\mathrm{e}} = \mathrm{Softmax}(\mathbf{H}\mathbf{W}_{\mathrm{e}}) \in \mathbb{R}^{L\times 2}. \tag{9}$$

After that, by applying *argmax* to each row of $\mathbf{P}_{\mathrm{s}}$ and $\mathbf{P}_{\mathrm{e}}$, we obtain the predicted indexes that might be the starting or ending positions, i.e., $\hat{E}_{\mathrm{s}}$ and $\hat{E}_{\mathrm{e}}$:

$$\begin{aligned}\hat{\mathbf{E}}_{\mathrm{s}}, \hat{\mathbf{E}}_{\mathrm{e}} = \{i, j \mid \mathrm{argmax}(\mathbf{P}_{\mathrm{s}}^{(i)}) = 1,\\ \mathrm{argmax}(\mathbf{P}_{\mathrm{e}}^{(j)}) = 1\}\,,\end{aligned} \tag{10}$$

where $i = 1, \cdots, L$, $j = 1, \cdots, L$. Given any start index $i \in \hat{\mathbf{E}}_{\mathrm{s}}$ and end index $j \in \hat{\mathbf{E}}_{\mathrm{e}}$, a binary classification model is trained to predict the probability that they should be matched:

$$o_{i,j} = \mathrm{sigmoid}(\mathbf{W} \cdot [\mathbf{H}_i, \mathbf{H}_j])\,, \tag{11}$$

where $\mathbf{W} \in \mathbb{R}^{1\times 2d}$ is the weights to learn. Hence, the span match loss is

$$\mathcal{L}_{match} = -\sum_{i,j} c_{i,j} \log o_{i,j}\,, \tag{12}$$

where $c_{i,j}$ denotes the golden labels for whether this row-column position should be matched with the start index and end index of rationale spans.

### Training

We define the final weighted loss as follow,

$$\mathcal{L} = \mathcal{L}_{classify} + \mathcal{L}_{extract} + \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{match}, \tag{13}$$

where $\lambda_1, \lambda_2$ are hyper-parameters of $\mathcal{L}_{adv}$ in Eqn. (4), $\mathcal{L}_{match}$ in Eqn. (12), respectively. Note that at the training time, all ground-truth labels are fed into the models to learn all components of rationale extraction network and prediction network simultaneously. The marked rationales for the label in training are the collection of input text sequences annotated by humans. The same pre-trained encoder is used by shared parameters. During inference, the model recognizes rationale spans of samples with golden human-labeled rationale for extraction evaluations.

## Experiments

### Dataset

**Movie Reviews** (Pruthi et al. 2020). This dataset includes 50k movie reviews from IMDB dataset (Maas et al. 2011) and 1.8k movie reviews with human-labeled rationales collected by Zaidan, Eisner, and Piatko (2007). The labels marked by the annotator are binary sentiment indicators (i.e., *Positive* and *Negative*). The training set, development set, and test set consist of 26198, 12800, and 12800 available examples, respectively, while the types of movies in each subset are different. Note that the samples with rationale are divided into 1200, 300, and 300 respectively. Due to the long text of the review comment, it is necessary to verify the correctness of the prediction by extracting evidence. Moreover, it is also applicable to adversarial attack scenarios.

**MultiRC** (Khashabi et al. 2018). The corpus comprises of multiple-choice questions and answers from various sources along with supporting evidence. This dataset concatenates each answer candidate to one question and assigns a binary label to it (i.e., whether this answer can answer the question or not). Each QA pair is associated with a related passage that is annotated with sentence-level rationales. The training set, development set, and test set consist of 24029, 3214, and 4848 available examples. The rationale in this dataset contains sufficient context to allow the human to discern whether the given answer to the question is *True* or *False*.

### Experimental Results

**Evaluation Metrics** For classification, we use classification accuracy between the predicted class label and the actual label. For rationale extraction, we report the token F1 to evaluate the quality of extraction. The micro-averaged F1 score computes at the token level between the predicted evidence spans tokens and the gold rationale tokens in terms of sets of predicted positions.

**Experimental Setup** We use the BERT-base model released by Google to encode text (Devlin et al. 2019). We also compare the performance of AT-BMC which uses the pre-trained RoBERTa large model (Liu et al. 2019). Our model is orthogonal to the choice of the pre-trained language model. We use AdamW optimizer (Loshchilov and Hutter 2019) with a batch size of 4 for model training. The initial learning rate, the maximum sequence length, the dropout rate, the gradient accumulation steps, the training epoch and the hidden size $d$ are set to $2 \times 10^{-5}$, 512, 0.1, 8, 30, 768 respectively. We clip the gradient norm within 1.0. The learning parameters are selected based on the best performance on the development set. Early stopping is also applied based on model performance on the development set. Our models are trained with NVIDIA Tesla V100s (Ubuntu 18.04 LTS & PyTorch). We set the perturbation size $\epsilon = 1 \times 10^{-5}$, the step size $\eta = 1 \times 10^{-3}$, ascent iteration step $K = 2$ and the variance of normal distribution $\sigma = 1 \times 10^{-5}$. The weight parameters $\lambda_1$, $\lambda_2$ are set to 1.0, 0.1 respectively. The augmented adversarial examples of both datasets are 1198 and 24007.

**Comparison of Baselines** We compare our AT-BMC method with following competitive methods for classification and rationale extraction in both datasets: 1) The pipeline approaches (Lehman et al. 2019; Lei, Barzilay, and Jaakkola 2016) use independent parts for both extraction and prediction. These two pipeline modules are trained with classification labels and rationales labels, respectively. 2) The information bottleneck method (Paranjape et al. 2020b) extracts sentence-level rationales by measurement of maximal (and minimal) mutual information with the label (and input). 3) The FRESH approach (Jain et al. 2020) extract $k$ tokens by using attention scores for downstream classification. 4) The weakly- and semi-supervised methods (Pruthi et al. 2020) present a classify-then-extract framework condition the rationale extraction on the classification. The pipeline approach uses RNNs, whereas the base model is BERT-base for IB, FRESH, and (Pruthi et al. 2020), as same as ours for a fair comparison. The performances of baselines are from reference papers. As shown in Table 1, our model improves over the previous models on both datasets. In the task of rationale extraction, AT-BMC (BERT-base) and AT-BMC (RoBERTa-large) improves 4.3 and 13.3 F1 points over the previous models on the Movie Reviews dataset. Moreover, on the MultiRC dataset, our method also improves the F1 up to 3.3 and 10.8 points. On the other hand, AT-BMC (BERT-base) improves 0.8 and 1.3 in terms of accuracy respectively, which might come mainly from two main aspects: one is multi-task learning and the other is adversarial training.

### Robustness Evaluation

The robustness test is to perform model-agnostic attacks on the trained classification model. It verifies the robustness of our method to attack samples; hence we can better classify and extract evidence by which humans verify predictions. It assumes that the model input is text sequences and returns outputs that the objective function can process. In this process, the attack algorithm will find the disturbance of the

| Methods | Movie Reviews | | MultiRC | |
|---|---|---|---|---|
| | Accuracy | Token F1 | Accuracy | Token F1 |
| Pipeline approach (Lehman et al. 2019) | 76.9 | 14.0 | 65.5 | 45.6 |
| Information Bottleneck (IB) (Paranjape et al. 2020b) | 82.4 | 12.3 | 62.1 | 24.9 |
| IB (semi-supervised, 25%) (Paranjape et al. 2020b) | 85.4 | 18.1 | 66.4 | 54.0 |
| FRESH (Jain et al. 2020) | 93.1 | 27.7 | 66.1 | 53.2 |
| Weakly- & Semi-supervised (Pruthi et al. 2020) | 93.2 | 46.3 | 65.4 | 47.8 |
| AT-BMC (BERT-base) | $\mathbf{94.0 \pm 0.31}$ | $\mathbf{50.6 \pm 0.54}$ | $\mathbf{67.7 \pm 0.49}$ | $\mathbf{57.3 \pm 0.34}$ |
| w/o label embedding | $93.7 \pm 0.19$ | $49.8 \pm 0.51$ | $65.3 \pm 0.35$ | $56.2 \pm 0.47$ |
| w/o adversarial examples | $93.5 \pm 0.47$ | $48.0 \pm 0.34$ | $66.6 \pm 0.53$ | $55.5 \pm 0.48$ |
| w/o span match loss | $93.8 \pm 0.18$ | $47.4 \pm 0.45$ | $67.2 \pm 0.32$ | $55.7 \pm 0.43$ |
| w/o virtual adversarial training | $93.6 \pm 0.37$ | $48.1 \pm 0.43$ | $63.6 \pm 0.47$ | $55.5 \pm 0.51$ |
| w/o mixed adversarial training | $93.1 \pm 0.45$ | $47.7 \pm 0.59$ | $63.9 \pm 0.41$ | $54.7 \pm 0.53$ |
| AT-BMC (RoBERTa-large) | $\mathbf{95.8 \pm 0.43}$ | $\mathbf{59.6 \pm 0.61}$ | $\mathbf{76.4 \pm 0.31}$ | $\mathbf{64.8 \pm 0.52}$ |

Table 1: Performance comparison on two text classification tasks with rationale extraction. We report test set results of AT-BMC that using different encoders (i.e., BERT-base and RoBERTa-large) across 3 different seeds.

| Metrics | Baseline | Pruthi et al. (2020) | AT-BMC |
|---|---|---|---|
| Acc. (Test) | 90.00 | 93.20 | 93.97 (0.77 ↑) |
| Avg. words | 216.96 | 216.96 | 216.96 |
| **TextFooler** (Jin et al. 2020) | | | |
| Acc. (Attack) | 1.00 | 9.00 | 22.00 (21.00 ↑) |
| SR | 98.89 | 90.22 | 57.69 (42.2 ↓) |
| Avg. PW | 10.02 | 10.88 | 23.65 |
| Avg. AQ | 741.44 | 980.23 | 2300.40 |
| Total time | 1078.26 | 1552.26 | 3118.59 |
| **TextBugger** (Li et al. 2019) | | | |
| Acc. (Attack) | 9.00 | 31.00 | 38.00 (29.00 ↑) |
| SR | 90.00 | 67.02 | 26.92 (63.08 ↓) |
| Avg. PW | 31.26 | 33.97 | 41.61 |
| Avg. AQ | 583.01 | 834.19 | 1497.04 |
| Total time | 1480.80 | 2098.11 | 2337.80 |
| **PWWS** (Ren et al. 2019) | | | |
| Acc. (Attack) | 3.00 | 4.00 | 38.00 (35.00 ↑) |
| SR | 96.67 | 95.74 | 26.92 (69.75 ↓) |
| Avg. PW | 5.87 | 4.99 | 17.44 |
| Avg. AQ | 1505.1 | 1459.83 | 2299.5 |
| Total time | 2005.26 | 2125.88 | 2356.88 |

Table 2: Comparison of baselines and AT-BMC using three adversarial attack methods. The baseline is the vanilla BERT fine-tuned on the IMDB dataset (Maas et al. 2011). We compare performance in accuracy under attack, attack success rate (SR), average percentage of perturbed words in the sentence (PW), average attack queries of all examples (AQ) and total attack time. All numbers are reported on 100 test instances. ↑ (↓) represents that the higher (lower) the better.

input sequence that satisfies the attack target and follows a certain language restriction. In this way, the attack to models is framed as a combinatorial search problem. In order to find a series of perturbations that produce a successful adversarial example, the attacker must search through all possible conversions. We refer the reader to (Morris et al. 2020) for more details. Instead of measuring robustness by *interpretability robustness* where rationales should be invariant to small

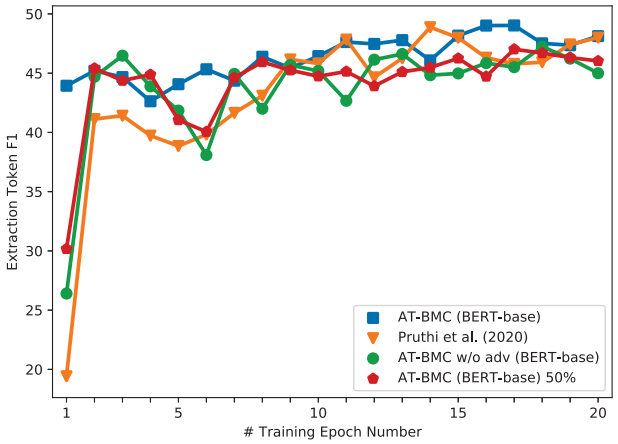
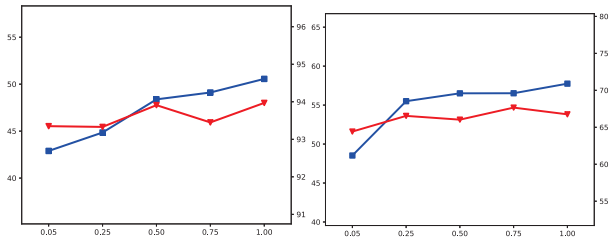



Figure 3: Extraction F1 curves on the development set of Movie Reviews. Our model with BERT-base is trained on 100%, 50% of the fraction of rationales in the training set.

perturbations in the input, we consider three different attack methods (i.e., TextFooler (Jin et al. 2020), TextBugger (Li et al. 2019), and PWWS (Ren et al. 2019)) to *classification robustness*. In the test, we focus on the success rate of the attack. TextFooler and TextBugger use a mixture of measures (e.g., word embedding distance, part-of-speech tags match), and a word replacement mechanism is designed to attack the existing model; PWWS greedily uses word importance ranking to replace parts of sentences, where word saliency and synonym swap scores are used to calculate word importance. The results of classification models on the Movie Reviews test set are presented in Table 2. Overall, AT-BMC achieves the best performance on all metrics consistently across different attack methods. Notably, AT-BMC substantially outperforms the baseline by 69.75% success rate under PWWS attack. We attribute this to AT-BMC's generalizability obtained by adversarial training. Interestingly, from the results in the second column, using a joint framework also seems to enhance generalizability and robustness in this do-



(a) Movie Reviews (b) MultiRC

Figure 4: Comparison in task accuracy shown in red and extraction performance shown in blue as the number of rationales retained in the training set varies.

main. We also report the percentage of words being replaced for attacking as average word modification rates. Our method needs more modified attack queries with higher word modification rates under all attacks. It indicates that the model is harder to attack and hence requires more words to be replaced. In Figure 3 we compare the evaluation curves of the different methods on the development set over training time. We take the performance on the Movie Reviews development set at each evaluation checkpoint. The magnitude of change in our method is much smaller relative to other methods that did not perform adversarial training and boundary match constraint, and it converges gradually as training time changes. It illustrates that our method reinforces the robustness of the model on the development set during training, thereby making it more stable for training and less variance, which alleviates hyperparameter sensitivity and high variance in existing methods.

Our approach can also be applied in the context of which there are only limited annotated examples. As shown in Figure 4, we compare the performance of models with varying proportions of human-labeled rationales in the training set. We find that the model achieves extraction accuracy above 40 on the test set when only 5% of examples with labeling signals. As the ratio of these marker examples increases, the performance of the model improves. Since the manual labeling of these annotations is time-consuming and labor-intensive, this might imply that our approach can stably generate reasonable interpretations without many manual annotations.

## Analysis and Discussion

**Ablation Study** To study the effect of each part, we conduct ablation experiments. The results are also shown in Table 1. From the experimental results, adversarial examples, boundary match loss, and virtual adversarial training all contribute to the performance improvement of the model, each helping to improve the model performance by about two percentage points in rationale extraction. We can see that the match loss improves token F1 relatively the most on the Movie Reviews dataset, probably because after adding this term, the extraction model may focus more on local boundary information. For the original task accuracy, adversarial training can boost performance, and both adversarial data augmentation and virtual adversarial training in the embedding space can bring improvements. It illustrates that these two methods are complementary. The results also show that considering on the predicted label improves the extraction F1 by 0.8 and 1.1 points on both datasets.

**Effectiveness Evaluation** Human-labeled rationales contain some implicit information that represents partial inductive bias in predictions. The interpretable AI community would like to use them as a guide for evaluating model interpretations and possibly for teaching models to make robust and reasonable decisions. Compared with the answer prediction task, existing models have relatively lower prediction results on the Token F1 metric in the rationale extraction task. Hence, we randomly sampled 50 correct examples and ask two annotators to judge the relevance between extracted rationales and labels. The inter-rater agreement coefficient between two annotators is 0.85. The relevance results of two datasets are 45 and 47, which means high scores between predicted labels and extracted rationales. We also compare the model without the boundary match constraint. The relevance results are 39 and 42. The comparison shows that by matching starting and ending tokens, the model can be more relevant to labels and aligned with human interpretation.

**Error Analysis** We conduct an error analysis on extraction rationales generated by our model on 50 randomly chosen examples from the development set of Movie Reviews. The major error types are summarized as follows: 1) the dominant type (48%) is that the model outputs contain bags of small fragments, which overlap with human evidence. For example, some fragments are emotional expressions, such as "*perfect*", "*some of the best*". 2) the second type (24%) is caused by incomplete or inadequate annotations. For example, the model outputs "*it's a very impressive film*" and "*wonderfully presented story*", while humans only annotate the former. It demonstrated that marked rationales do not necessarily have high comprehensiveness by including all relevant information, which aligns with the findings of Carton, Rathore, and Tan (2020). 3) the last type (28%) is caused by several factors, such as the text is too long and the evidence is outside the truncated text; or the prediction of the model itself is wrong. It shows that only learning from the supervision signal may be affected by annotation artifacts and variance between human-annotated rationales. And how the machine can help us correct superficial clues instead of learning could be another interesting topic.

## Conclusion

In this work, we focus on how to jointly classify and provide extracted rationales, so that humans can use it to verify the correctness of the prediction. We propose a method AT-BMC for jointly modeling text classification and rationale extraction using mixed adversarial training and boundary match constraint. The results on two public data sets show that our method improves the performance of the model on the task, especially with increasing extraction token F1 for rationales. Besides, AT-BMC can remarkably decrease the attack success rate compared to the baseline under different attack methods. The results indicate that robust models lead to better extracted rationale in text classification tasks. In the future, we will explore how to apply our model to more domains (e.g., medical and legal domains).

## Acknowledgements

We thank the valuable feedback of Wenpeng Yin, Yuxiang Wu, Shuoran Jiang and the insightful comments and suggestions of the anonymous reviewers. This work is jointly supported by grants: Natural Science Foundation of China (Grant No. 61872113 and Grant No. 62006061), Stable Support Program for Higher Education Institutions of Shenzhen (No. GXWD20201230155427003-20200824155011001), Strategic Emerging Industry Development Special Funds of Shenzhen (No. XMHT20190108009 and No. JCYJ20200109113403826), Fundamental Research Fund of Shenzhen (No. JCYJ20190806112210067), and Tencent Group.

## References

Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; and Vollgraf, R. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. On the Robustness of Interpretability Methods. *CoRR*, abs/1806.08049.

Augustin, M.; Meinke, A.; and Hein, M. 2020. Adversarial Robustness on In- and Out-Distribution Improves Explainability. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVI*.

Bastings, J.; Aziz, W.; and Titov, I. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In *Proc. of ACL*.

Camburu, O.; Rocktäschel, T.; Lukasiewicz, T.; and 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, December 3-8, 2018, Montréal, Canada*.

Carton, S.; Rathore, A.; and Tan, C. 2020. Evaluating and Characterizing Human Rationales. In *Proc. of EMNLP*.

Caruana, R. 1997. Multitask learning. *Machine learning*.

Cheng, H.; Liu, X.; Pereira, L.; Yu, Y.; and Gao, J. 2021. Posterior Differential Regularization with f-divergence for Improving Model Robustness. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Datta, A.; Fredrikson, M.; Leino, K.; Lu, K.; Sen, S.; and Wang, Z. 2021. Machine Learning Explainability and Robustness: Connected at the Hip. In Zhu, F.; Ooi, B. C.; and Miao, C., eds., *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, 4035–4036. ACM.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL-HLT*.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proc. of ACL*.

Etmann, C.; Lunz, S.; Maass, P.; and Schönlieb, C. 2019. On the Connection Between Adversarial Robustness and Saliency Map Interpretability. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *Proc. of ICLR*.

Jain, S.; Wiegreffe, S.; Pinter, Y.; and Wallace, B. C. 2020. Learning to Faithfully Rationalize by Construction. In *Proc. of ACL*.

Jia, R.; and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proc. of EMNLP*.

Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Zhao, T. 2020. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In *Proc. of ACL*.

Jin, D.; Jin, Z.; Zhou, J. T.; and Szolovits, P. 2020. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proc. of NAACL-HLT*.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*.

Latcinnik, V.; and Berant, J. 2020. Explaining Question Answering Models through Text Generation. *arXiv preprint arXiv:2004.05569*.

Lehman, E.; DeYoung, J.; Barzilay, R.; and Wallace, B. C. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In *Proc. of NAACL-HLT*.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *Proc. of EMNLP*.

Li, D.; Zhang, Y.; Peng, H.; Chen, L.; Brockett, C.; Sun, M.-T.; and Dolan, B. 2021. Contextualized Perturbation for Textual Adversarial Attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Li, J.; Ji, S.; Du, T.; Li, B.; and Wang, T. 2019. TextBugger: Generating Adversarial Text Against Real-world Applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*.

Lipton, Z. C. 2016. The Mythos of Model Interpretability. In *WHI*.

Liu, X.; Cheng, H.; He, P.; Chen, W.; Wang, Y.; Poon, H.; and Gao, J. 2020. Adversarial Training for Large Neural Language Models. *arXiv preprint arXiv:2004.08994*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proc. of ICLR*.

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In *Proc. of ACL*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proc. of ICLR*.

Michel, P.; Li, X.; Neubig, G.; and Pino, J. 2019. On Evaluation of Adversarial Perturbations for Sequence-to-Sequence Models. In *Proc. of NAACL-HLT*.

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proc. of EMNLP*.

Moshkovitz, M.; Yang, Y.; and Chaudhuri, K. 2021. Connecting Interpretability and Robustness in Decision Trees through Separation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*.

Paranjape, B.; Joshi, M.; Thickstun, J.; Hajishirzi, H.; and Zettlemoyer, L. 2020a. An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. In *Proc. of EMNLP*.

Paranjape, B.; Joshi, M.; Thickstun, J.; Hajishirzi, H.; and Zettlemoyer, L. 2020b. An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. In *Proc. of EMNLP*.

Pruthi, D.; Dhingra, B.; Neubig, G.; and Lipton, Z. C. 2020. Weakly- and Semi-supervised Evidence Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Ren, S.; Deng, Y.; He, K.; and Che, W. 2019. Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In *Proc. of ACL*.

Ribeiro, M. T.; Singh, S.; and 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, San Francisco, CA, USA, August 13-17, 2016*.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proc. of ACL*.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*.

Samek, W.; Wiegand, T.; and Müller, K.-R. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sha, L.; Camburu, O.; and Lukasiewicz, T. 2020. Learning from the Best: Rationalizing Prediction by Adversarial Information Calibration. *CoRR*.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2019. Generating Token-Level Explanations for Natural Language Inference. In *Proc. of NAACL-HLT*.

Wang, G.; Li, C.; Wang, W.; Zhang, Y.; Shen, D.; Zhang, X.; Henao, R.; and Carin, L. 2018. Joint Embedding of Words and Labels for Text Classification. In *Proc. of ACL*.

Wang, Y.; and Bansal, M. 2018. Robust Machine Comprehension Models via Adversarial Training. In *Proc. of NAACL-HLT*.

Wei, T.; Qi, J.; He, S.; and Sun, S. 2021. Masked Conditional Random Fields for Sequence Labeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yoon, J.; Jordon, J.; and van der Schaar, M. 2019. INVASE: Instance-wise Variable Selection using Neural Networks. In *Proc. of ICLR*.

Zaidan, O.; and Eisner, J. 2008. Modeling Annotators: A Generative Approach to Learning from Annotator Rationales. In *Proc. of EMNLP*.

Zaidan, O.; Eisner, J.; and Piatko, C. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*.

Zang, Y.; Qi, F.; Yang, C.; Liu, Z.; Zhang, M.; Liu, Q.; and Sun, M. 2020. Word-level Textual Adversarial Attacking as Combinatorial Optimization. In *Proc. of ACL*.

Zhang, Y.; Marshall, I.; and Wallace, B. C. 2016. Rationale-Augmented Convolutional Neural Networks for Text Classification. In *Proc. of EMNLP*.

Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; and Liu, J. 2019. Freelb: Enhanced adversarial training for language understanding. *arXiv preprint arXiv:1909.11764*.