SEAT: Stable and Explainable Attention

Lijie Hu [*,1], **Yixin Liu** [*,2], **Ninghao Liu** [3], **Mengdi Huai** [4], **Lichao Sun** [2], **Di Wang** [1, 5, 6]

[1] King Abdullah University of Science and Technology, [2] Lehigh University
[3] University of Georgia, [4] Iowa State University, [5] Computational Bioscience Research Center
[6] SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence
{lijie.hu, di.wang}@kaust.edu.sa, {yila22, lis221}@lehigh.edu, ninghao.liu@uga.edu, mdhuai@iastate.edu

Abstract

Attention mechanism has become a standard fixture in many state-of-the-art natural language processing (NLP) models, not only due to its outstanding performance, but also because it provides plausible innate explanations for neural architectures. However, recent studies show that attention is unstable against randomness and perturbations during training or testing, such as random seeds and slight perturbation of embeddings, which impedes it from being a faithful explanation tool. Thus, a natural question is whether we can find an alternate of the vanilla attention, which is more stable and could keep the key characteristics of the explanation. In this paper, we provide a rigorous definition of such an attention method named SEAT (**S**table and **E**xplainable **AT**tention). Specifically, SEAT has the following three properties: (1) Its prediction distribution is close to the prediction of the vanilla attention; (2) Its top-k indices largely overlap with those of the vanilla attention; (3) It is robust w.r.t perturbations, i.e., any slight perturbation on SEAT will not change the prediction distribution too much, which implicitly indicates that it is stable to randomness and perturbations. Furthermore, we propose an optimization method for obtaining SEAT, which could be considered as revising the vanilla attention. Finally, through intensive experiments on various datasets, we compare our SEAT with other baseline methods using RNN, BiLSTM and BERT architectures, with different evaluation metrics on model interpretation, stability and accuracy. Results show that, besides preserving the original explainability and model performance, SEAT is more stable against input perturbations and training randomness, which indicates it is a more faithful explanation.

Word Perturbation (N=2)

Vanilla	**JSD.**↓ = 0.012	**TVD.**↓ = 0.230					
Benign:	This	film	is	just	plain	horrible	(99.9% →Neg.)
Perturbed:	This	**movie**	is	just	plain	**terrible**	(59.2% →Pos.)
Word-AT	**JSD.** = 0.014	**TVD.** = 0.080					
Benign:	This	film	is	just	plain	horrible	(99.9% →Neg.)
Perturbed:	This	**movie**	is	just	plain	**terrible**	(99.9% →Neg.)
Att-AT	**JSD.** = 0.004	**TVD.** = 0.260					
Benign:	This	film	is	just	plain	horrible	(99.9% →Neg.)
Perturbed:	This	**movie**	is	just	plain	**terrible**	(56.8% →Pos.)
Ours	**JSD.** = 8.4e-8	**TVD.** = 0.012					
Benign:	This	film	is	just	plain	horrible	(99.9% →Neg.)
Perturbed:	This	**movie**	is	just	plain	**terrible**	(98.9% →Neg.)

Figure 1: An example demonstrates stability of prediction and attention heat map trained with different methods under word perturbation in sentiment classification task. There are four methods: Vanilla attention, Word-AT, Att-AT with and our SEAT. JSD and TVD are divergence to measure stability of explainablity and prediction distribution (see Experiments section for details). We use the closest synonyms to replace original word in sentence as word perturbation. N denotes number of replaced words. We can see explanation (heat map) and prediction are changed in other methods.

Introduction

As deep neural networks have demonstrated great success in various natural language processing (NLP) tasks (Otter, Medina, and Kalita 2021), to further establish trust, how to interpret these deep models is receiving increasing interests. Recently, a number of interpretation techniques have been developed to understand the decision of deep NLP models (Ribeiro, Singh, and Guestrin 2016; Vaswani et al. 2017; Dong et al. 2019). Among them, *attention mechanism* has become a near-ubiquitous component of modern NLP models. Different from post-hoc interpretation (Du, Liu, and Hu 2020), attention weights are often regarded as providing the "inner-workings" of models (Choi et al. 2016; Martins and Astudillo 2016; Lei 2017). For instance, each entry of an attention vector could point us to relevant information discarded by the neural network or to irrelevant elements (tokens) of the input source that have been factored in (Galassi, Lippi, and Torroni 2020).

Despite its wide adoption, attention mechanism has been questioned as being a **faithful** interpretation. Specifically, Wiegreffe and Pinter (2019) show that attention is unstable, as different model initialization leads to different attention distributions given the same input (see Fig. 3 in Appendix for an example). Besides, attention is also fragile to input perturbations during inference. For example, in Fig. 1, we can observe that, after replacing a word with its synonym, the attention changes significantly and may also give wrong predictions. In addition, perturbing word embeddings will also affect the prediction distribution and the explainability of attention (see details in Appendix Fig. 4). Actually, insta-

[*]These authors contributed equally.

bility has been identified as a common issue of interpretation methods in deep models. Generally, an unstable interpretation makes it easy to be influenced by noises in data, thus impeding users from understanding the inherent rationale behind model predictions. Moreover, instability reduces the reliability of interpretation as a diagnosis tool of models, where small carefully-crafted perturbation on input could dramatically change the interpretation result (Ghorbani, Abid, and Zou 2019; Dombrowski et al. 2019; Yeh et al. 2019). Thus, stability now becomes an important aspect for faithful interpretation. Based on the above facts, a natural question is that can we make the attention mechanism more faithful by improving its stability, while keeping the most important explanation and prediction characteristics?

To tackle the problem, we first need to give a rigorous definition of such a "stable attention". Intuitively, a stable attention should have the following three properties: (1) It should produce a similar prediction as the vanilla attention to preserve model utility; (2) The top-k indices of the stable attention and vanilla attention should largely overlap with each other, so that the stable attention inherits the interpretability of vanilla attention; (3) It should be robust to the randomness in training or input perturbations during testing. Based on the above criteria, in this paper, we present a formal definition of SEAT (**S**table and **E**xplainable **At**tention). Specifically, our contributions can be summarized as follows.

- We provide a rigorous mathematical definition of SEAT. Specifically, to keep property (1), SEAT ensures the loss between its prediction distribution (vector) and the prediction distribution (vector) based on attention is sufficiently small. For property (2), we ensure the top-k indices overlaps between SEAT and vanilla attention are large enough. For property (3), SEAT guarantees some perturbations on it will not change the prediction distribution too much, which implicitly ensures it is robust to randomness and perturbations during training and testing.
- In the second part of the paper, we propose a method to find a SEAT. Specifically, we present a min-max stochastic optimization problem whose objective function involves three terms, which correspond to the above three properties. However, the main difficulty is that the term induced by property (2) is non-differentiable, which impedes us from using gradient descent based methods. To address this issue, we also propose a surrogate loss function of top-k overlap function as a byproduct, which can also be used in other problems.
- Finally, we conduct intensive experiments on four benchmark datasets using RNN, BiLSTM and BERT to verify the above three properties of the SEAT found by our method. Particularly, using two metrics, we first show our SEAT is more stable than other baselines via three different perturbations or randomness: random seeds, embedding vector perturbation and input token perturbation. We also use three recent evaluation metrics on model interpretability to measure our method. Results reveal our SEAT is a more faithful interpretation. Besides, we also compare F1 score of our SEAT and other baselines, which shows that compared with vanilla attention there is almost no accuracy degradation for SEAT. See Figure 1 for comparison between SEAT and other methods.

Related Work

Stability and robustness in attention. There exists some work studying or improving either the stability or the robustness of attention from the explanation perspective. Recently, Kitada and Iyatomi (2021) propose a method to improve the robustness to perturbation of embedding vector for attention. Specifically, they adopt the adversarial training during the training process. However, in their method, they do not consider the similarity and closeness between the their new attentions and the original ones, which means their robust attention loses the prediction performance and explainability of the original attention. Equivalently, while their adversarial training may could improve the robustness of attention, it can not be ensured to be explainable due to the ignorance of relationship with vanilla attention. Sato et al. (2018) study using adversarial training to improve the robustness and interpretation of text. However, their work is applied to input embedding space, whose computational cost is high. Moreover, their method still cannot guarantee the closeness to attention on neither prediction or explanation, and their method cannot ensure the robustness against other randomness such as random seeds. (Mohankumar et al. 2020) explores to modify the LSTM cell with diversity driven training to enhance explainability and transparency of attention modules. However, it does not consider the robustness of attention, which makes their method far from a faithful method.

Stability in explanation techniques. Besides attention, there are numerous works on studying stable interpretation. For example, Yeh et al. (2019) theoretically analyze the stability of post-hoc interpretation approaches and proposes using smoothing to improve interpretation stability. Jacovi and Goldberg (2020) discuss high-level directions of designing reliable interpretation. However, these techniques are designed for post-hoc interpretation, which cannot be directly applied to attention mechanisms. Recently, Yin et al. (2022) introduce two metrics to measure the interpretability via sensitivity and stability. They also introduce methods to better test the validity of their evaluation metrics by designing an iterative gradient descent algorithm to get an counterfactual interpretation. But they do not consider how to improve faithfulness of explainable models. Thus, it is incomparable with our work. And in the experiments part we will use these evaluation metrics.

Stable and Explainable Attention

Vanilla Attention

We first give a brief introduction on the attention mechanism (Vaswani et al. 2017). Here we follow the notations in (Jain and Wallace 2019). Let $x \in \mathbb{R}^{s\times|V|}$ denote the model input, composed of one-hot encoded words at each position. There is an embedding matrix E with dimension d. After passing through the embedding matrix E we have a more dense token representation, which is represented as $x_e \in \mathbb{R}^{s\times d}$.

There is an encoder (**Enc**)-decoder (**Dec**) layer. For the encoder, it takes the embedded tokens in order and produces

S number of m-dimensional hidden states after the **Enc** procedure, i.e., $\boldsymbol{h}(x) = \mathbf{Enc}(x_e) \in \mathbb{R}^{s\times m}$ with $h_t(x)$ as the word representation for the word at position t in x.

A similar function ϕ maps $\boldsymbol{h}(x)$ and a query $Q \in \mathbb{R}^m$ to scalar scores, and the vector of attention weight is induced by $\boldsymbol{w(x)} = \text{softmax}\phi(\boldsymbol{h}(x), \boldsymbol{Q})$. Here we consider two common types of similarity functions: *Additive* $\phi(\boldsymbol{h}(x), \boldsymbol{Q}) = \boldsymbol{v}^T \tanh(\boldsymbol{W_1 h(x)} + \boldsymbol{W_2 Q})$ (Bahdanau, Cho, and Bengio 2015) and *Scaled Dot-Product* $\phi(\boldsymbol{h}, \boldsymbol{Q}) = \frac{\boldsymbol{h(x)Q}}{\sqrt{m}}$ (Vaswani et al. 2017), where $\boldsymbol{v}$, $\boldsymbol{W_1}$, and $\boldsymbol{W_2}$ are model parameters.

Based on $\boldsymbol{w(x)} \in \mathbb{R}^s$, we then have prediction after the **Dec** procedure, i.e., $y(x, \boldsymbol{w}) = \sigma(\theta \cdot h_w) \in \mathbb{R}^{|\mathcal{Y}|}$, where $h_w = \sum_{t=1}^{s} \boldsymbol{w_t(x)} \cdot h_t(x)$, σ is an output activation function, θ is a parameter and $|\mathcal{Y}|$ denotes the label set size.

Stable and Explainable Attention

Motivation. As we mentioned previously, our goal is to find some "stable attention" that keeps the **performance** and **explainability** of attention, while is more **robust against some randomness and perturbations during training and testing**. Before showing how to find it, we first need to think about what properties it should have. Actually, such a "stable attention" should has the following three properties:

1. It keeps the advantage of outstanding performance of attention, i.e., we hope the prediction distribution (vector) based on the "stable attention" is almost the same as the distribution (vector) based on vanilla attention for any input x. Mathematically, we can use different loss functions or divergences to measure such similarity or closeness.
2. A "stable attention" should also keep the explainability of vanilla attention. As we mentioned above, the explainability of attention could be revealed by the entries of the attention vector. More specifically, the rank of each entry in the attention vector determine the importance of its associated token. Thus, to ensure the explainablity, it is sufficiently to keep the order of leading entries. Mathematically, here we can use the overlaps of top-k indices between "stable attention" and vanilla attention to measure their similarity on explainability, where k is a parameter.
3. Such new attention should be stable. It is notable that compared with the robustness to adversarial attacks such as poisoning attack, here our stability is more general, i.e., it should be robust against any randomness and perturbations during training and testing, such as random seeds in training, and perturbations on embedding vectors or input tokens during testing. Thus, unlike adversarial training, it is difficult to model the robustness to varies randomness or perturbations directly. To resolve the issue, as we mentioned, those randomness and perturbations will cause attention changes dramatically, which could be thought as some noise added to attention will change it significantly. Thus, if the "stable attention" is resilient to any perturbations, then this can indicate that such vector is robust to any randomness and perturbations implicitly. In total, mathematically we can model such robustness via the resilience against perturbations of "stable attention".

Based on the previous motivation, in the following we formally give the definition of "stable attention" called Stable and Explainable Attention (SEAT) and denoted as $\tilde{\boldsymbol{w}}$. Since we need to use the overlaps of top-k indices to measure the similarity on explainability with attention. We first provide its formal definition.

Definition 1 (Top-k overlaps)**.** *For vector $x \in \mathbb{R}^d$, we define the set of top-k component $T_k(\cdot)$ as follow,*

$$T_k(x) = \{i : i \in [d] \text{ and } \{|\{x_j \geq x_i : j \in [d]\}| \leq k\}\}.$$

And for two vectors x, x', the top-k overlap function $V_k(x, x')$ is defined by the overlapping ratio between the top-k components of two vectors, i.e., $V_k(x, x') = \frac{1}{k}|T_k(x) \cap T_k(x')|$.

Note that in attention, $\boldsymbol{w}$ could be seen as a function of x. Thus, $\tilde{\boldsymbol{w}}$ can also be seen as a function of x. Moreover, since we only concern about replacing the attention vector, thus we will still follow the previous model except the procedure to produce the vector $\tilde{\boldsymbol{w}}(x)$. We define a SEAT as follows.

Definition 2 (**Stable and Explainable Attention**)**.** *We call a vector $\tilde{\boldsymbol{w}}$ is a $(D_1, D_2, R, \alpha, \beta, \gamma, V_k)$-Stable and Explainable Attention (SEAT) for the vanilla attention $\boldsymbol{w}$ if it satisfies for any x*

- *(Closeness of Prediction) $D_1(y(x, \tilde{\boldsymbol{w}}), y(x, \boldsymbol{w})) \leq \gamma$ for some $\gamma \geq 0$, where D_1 is some loss, $y(x, \tilde{\boldsymbol{w}}) = \sigma(\theta \cdot h_{\tilde{w}}) \in \mathbb{R}^{|\mathcal{Y}|}$ and $y(x, \boldsymbol{w}) = \sigma(\theta \cdot h_w) \in \mathbb{R}^{|\mathcal{Y}|}$;*
- *(Similarity of Explainability) $V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x)) \geq \beta$ for some $1 \geq \beta \geq 0$;*
- *(Stability) $D_2(y(x, \tilde{\boldsymbol{w}}), y(x, \tilde{\boldsymbol{w}} + \boldsymbol{\delta})) \leq \alpha$ for all $\|\boldsymbol{\delta}\| \leq R$, where D_2 is some loss, $\|\cdot\|$ is a norm and $R \geq 0$.*

Note that in the previous definition there are several parameters. Specifically, γ measures the closeness between the prediction distribution based on $\tilde{\boldsymbol{w}}$ and the prediction distribution based on vanilla attention. When $\gamma = 0$, then $\tilde{\boldsymbol{w}} = \boldsymbol{w}$. Therefore we hope γ to be as small as possible. The second condition ensures $\tilde{\boldsymbol{w}}$ has similar explainablity with attention. There are two parameters, k and β. k could be considered as a prior knowledge, i.e., we believe the top-k indices of attention will play the most important role to make the prediction, or their corresponding k tokens can almost determine its prediction. β measures how much explainability does $\tilde{\boldsymbol{w}}$ inherit from vanilla attention. When $\beta = 1$, then this means the top-k order of the entries in $\tilde{\boldsymbol{w}}(x)$ is the same as it in vanilla attention. Thus, β should close to 1. The term on stability involves two parameters R and α, which corresponds to the robust region and the level of stability respectively. Ideally, if $\tilde{\boldsymbol{w}}$ satisfies this condition with $R = \infty$ and $\alpha = 0$, then $\tilde{\boldsymbol{w}}$ will be extremely stable w.r.t any randomness or perturbations. Thus, in practice we wish R to be as large as possible and α to be sufficiently small. Thus, based on these discussions, we can see Definition 2 is consistent with our above intuition on "stable attention" and thus it is reasonable.

Finding a SEAT

In the last section, we presented a rigorous definition of stable and explainable attention. To find such a SEAT, we propose to formulate a min-max optimization problem that involves the three conditions in Definition 2. Specifically, the formulated optimization problem takes the first condition (closeness of prediction) as the objective, and subjects to the other two conditions. Thus, we can get a rough optimization problem according to the definition. Specifically, we first have

$$\min_{\tilde{\boldsymbol{w}}} \mathbb{E}_x D_1(y(x,\tilde{\boldsymbol{w}}), y(x,\boldsymbol{w})). \tag{1}$$

Equation (1) is the basic optimization goal, that is, we want to get a vector which has similar output prediction with vanilla attention for all input x. If there is no further constraint, then we can see the minimizer of (1) is just the vanilla attention w. We then consider constraints for this objective function:

$$\forall x \text{ s.t.} \max_{||\delta||\leq R} D_2(y(x,\tilde{\boldsymbol{w}}), y(x,\tilde{\boldsymbol{w}}+\boldsymbol{\delta})) \leq \alpha, \tag{2}$$

$$V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x)) \geq \beta. \tag{3}$$

Equation (2) is the constraint of stability and Equation (3) corresponds to the condition of similarity of explainability. Combining equations (1)-(3) and using regularization to deal with constraints, we can get the following objective function.

$$\begin{aligned}\min_{\tilde{\boldsymbol{w}}} \mathbb{E}_x[&D_1(y(x,\tilde{\boldsymbol{w}}), y(x,\boldsymbol{w})) + \lambda_2(\beta - V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x)))\\ &+ \lambda_1(\max_{||\boldsymbol{\delta}||\leq R} D_2(y(x,\tilde{\boldsymbol{w}}), y(x,\tilde{\boldsymbol{w}}+\boldsymbol{\delta})) - \alpha)],\end{aligned} \tag{4}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are hyperparameters.

From now on, we convert the problem of finding a vector that satisfies the three conditions in Definition 2 to a min-max stochastic optimization problem, where the overall objective is based on the closeness of prediction condition with constrains on stability and top-k overlap.

Next we consider how to solve the above min-max optimization problem. In general, we can use the stochastic gradient descent based methods to get the solution of outer minimization, and use PSGD (Projected Stochastic Gradient Descent) to solve the inner maximization. However, the main difficulty is that the top-k overlap function $V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x))$ is non-differentiable, which impede us from using gradient descent. Thus, we need to consider a surrogate loss of $-V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x))$. Below we provide details.

Projected gradient descent to find the perturbation $\boldsymbol{\delta}$ Motivated by (Madry et al. 2018), we can interpret the perturbation as the attack to $\tilde{w}$ via maximizing δ. Then, δ can be updated by the following procedure in the k-th iteration.

$$\boldsymbol{\delta_k} = \boldsymbol{\delta^*_{k-1}} + \alpha_k \frac{1}{|B_k|}\sum_{x\in B_k} \nabla D_2(y(x,\tilde{\boldsymbol{w}}), y(x,\tilde{\boldsymbol{w}}+\boldsymbol{\delta^*_{k-1}}));$$

$$\boldsymbol{\delta^*_k} = \arg\min_{||\boldsymbol{\delta}||\leq R} ||\boldsymbol{\delta} - \boldsymbol{\delta_k}||,$$

where α_k is a parameter of step size for PGD, B_k is a batch and $|B_k|$ is the batchsize. Using this method, we can derive the optimal δ^* in the t-th iteration for the inner optimization. Specifically, we find a δ as the maximum tolerant of perturbation w.r.t $\tilde{\boldsymbol{w}}$ in the t-th iteration.

Top-k overlap surrogate loss Now we seek to design a surrogate loss $\mathcal{L}_{TopK}(\tilde{\boldsymbol{w}}, \boldsymbol{w})$ for $-V_k(\tilde{\boldsymbol{w}}, \boldsymbol{w})$ which can be used in training. To achieve this goal, one possible naive surrogate objective might be some distance (such as ℓ_1-norm) between $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$, e.g., $L(\tilde{\boldsymbol{w}}) = ||\tilde{\boldsymbol{w}} - \boldsymbol{w}||_1$. Such surrogate objective seems like could ensure the top-k overlap when we obtain the optimal or near-optimal solution (i.e., $w = \arg\min L(\tilde{\boldsymbol{w}})$ and $w \in \arg\min -V_k(\tilde{\boldsymbol{w}}, \boldsymbol{w})$). However, it lacks consideration of the top-k information which makes it as a loose surrogate loss. Since we only need to ensure high top-k indices overlaps between $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$, one improved method is minimizing the distance between $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$ constrained on the top-k entries only instead of the whole vectors, i.e., $||\boldsymbol{w}_{S^k_w} - \tilde{\boldsymbol{w}}_{S^k_w}||_1$, where $\boldsymbol{w}_{S^k_w}, \tilde{\boldsymbol{w}}_{S^k_w} \in \mathbb{R}^k$ is the vector w and $\tilde{\boldsymbol{w}}$ constrained on the indices set S^k_w respectively and S^k_w is the top-k indices set of w. Since there are two top-k indices sets, one is for $\tilde{\boldsymbol{w}}$ and the other one is for $\boldsymbol{w}$, here we need to use both of them to involve the top-k indices formation for both vectors. Thus, based on our above idea, our surrogate can be written as follow,

$$\mathcal{L}_{Topk}(\boldsymbol{w}, \tilde{\boldsymbol{w}}) = \frac{1}{2k}(||\boldsymbol{w}_{S^k_w} - \tilde{\boldsymbol{w}}_{S^k_w}||_1 + ||\tilde{\boldsymbol{w}}_{S^k_{\tilde{w}}} - \boldsymbol{w}_{S^k_{\tilde{w}}}||_1). \tag{5}$$

Note that besides the ℓ_1-norm, we can use other norms. However, in practice we find ℓ_1-norm achieves the best performance. Thus, throughout the paper we only use ℓ_1-norm.

Final objective function and algorithm Based on the above discussion, we can derive the following overall objective function

$$\begin{aligned}\min_{\tilde{\boldsymbol{w}}} \mathbb{E}_x[&D_1(y(x,\tilde{\boldsymbol{w}}), y(x,\boldsymbol{w})) + \lambda_2 \mathcal{L}_{Topk}(\boldsymbol{w}(x), \tilde{\boldsymbol{w}}(x))\\ &+ \lambda_1 \max_{||\boldsymbol{\delta}||\leq R} D_2(y(x,\tilde{\boldsymbol{w}}), y(x,\tilde{\boldsymbol{w}}+\boldsymbol{\delta}))],\end{aligned} \tag{6}$$

where $\mathcal{L}_{Topk}(\boldsymbol{w}, \tilde{\boldsymbol{w}})$ is defined in (5). Based on the previous idea, we propose Algorithm 1 to solve (6).

Algorithm 1: Finding a SEAT

Initialize $\tilde{\boldsymbol{w}}_0$.
for $t = 1, 2, \cdots, T$ **do**
  Initialize $\boldsymbol{\delta_0}$.
  for $k = 1, 2, \cdots, K$ **do**
    Update $\boldsymbol{\delta}$ using PGD, where B_k is a batch
    $$\boldsymbol{\delta_k} = \boldsymbol{\delta_{k-1}} + \alpha_k \frac{1}{|B_k|}\sum_{x\in B_k} \nabla D_2(y(x,\tilde{\boldsymbol{w}}_{t-1}), y(x,\boldsymbol{\tilde{w}_{t-1}}+\boldsymbol{\delta_{k-1}})).$$
    $$\boldsymbol{\delta^*_k} = \arg\min_{||\boldsymbol{\delta}||\leq R} ||\boldsymbol{\delta} - \boldsymbol{\delta_k}||.$$
  end for
  Update $\tilde{\boldsymbol{w}}$ using Stochastic Gradient Descent, where B_t is a batch
  $$\begin{aligned}\boldsymbol{\tilde{w}_t} = \boldsymbol{\tilde{w}_{t-1}} - \frac{\eta_t}{|B_t|}\sum_{x\in B_t}[&\nabla D_1(y(x,\boldsymbol{\tilde{w}_{t-1}}), y(x,\boldsymbol{w}))\\ &- \lambda_1 \nabla D_2(y(x,\boldsymbol{\tilde{w}_{t-1}}), y(\boldsymbol{\tilde{w}_{t-1}}+\boldsymbol{\delta^*_K}))\\ &- \lambda_2 \nabla \mathcal{L}_{Topk}(\boldsymbol{w}, \tilde{\boldsymbol{w}}_{t-1})].\end{aligned}$$
end for
Return: $\tilde{\boldsymbol{w}}^* = \boldsymbol{\tilde{w}_T}$.

Experiments

In our experiments, we conduct extensive experiments to show the performance of our SEAT compared to six baseline methods. In the following we provide a brief introduction on experimental setup. **More details about the setting, implementation and additional experimental results can be found in Appendix.**

Model	Method	Emotion			SST			Hate			RottenT		
		JSD↓	TVD↓	F1↑	JSD	TVD	F1	JSD	TVD	F1	JSD	TVD	F1
RNN	Vanilla	0.002	20.145	0.663	0.019	19.566	0.811	0.009	15.576	0.553	0.008	19.139	0.763
	Word-AT	0.028	1.824	0.627	0.016	1.130	0.798	0.026	1.170	0.527	0.037	1.381	0.741
	Word-iAT	0.042	2.691	0.653	0.023	1.277	**0.815**	0.022	1.049	0.523	0.054	1.336	0.766
	Attention-RP	0.025	3.276	0.671	0.028	2.042	0.792	0.025	2.672	0.554	0.009	3.691	**0.770**
	Attention-AT	0.055	2.716	0.665	0.047	2.394	0.782	0.031	2.210	0.528	0.068	4.234	0.755
	Attention-iAT	0.017	3.654	0.645	0.048	2.653	0.746	0.039	2.264	0.533	0.054	1.594	0.753
	SEAT(**Ours**)	**3.81E-08**	**1.750**	**0.672**	**2.75E-07**	**1.099**	0.813	**1.79E-09**	**0.908**	**0.579**	**6.46E-07**	**1.178**	0.763
BiLSTM	Vanilla	0.002	23.447	0.612	0.027	18.640	**0.809**	0.060	15.633	0.524	0.009	20.125	0.764
	Word-AT	0.050	1.927	0.662	0.020	0.810	0.798	0.084	1.537	0.538	0.031	1.071	0.757
	Word-iAT	0.058	1.139	0.640	0.034	1.037	0.802	0.091	1.590	0.530	0.045	1.218	0.765
	Attention-RP	0.031	1.326	0.642	0.034	1.267	0.772	0.052	1.299	0.522	0.066	1.412	0.764
	Attention-AT	0.076	1.541	**0.672**	0.028	1.661	0.779	0.057	1.504	0.523	0.079	2.044	0.766
	Attention-iAT	0.033	1.267	0.651	0.034	1.528	0.801	0.062	2.256	0.525	0.076	1.751	**0.777**
	SEAT(**Ours**)	**1.23E-08**	**0.736**	0.670	**1.80E-08**	**0.777**	0.802	**8.49E-09**	**1.030**	**0.543**	**2.57E-08**	**0.885**	0.771
BERT	Vanilla	0.024	2.127	**0.721**	0.005	2.605	0.912	0.036	1.771	0.493	0.010	2.500	**0.845**
	Word-AT	0.085	0.060	0.694	0.267	0.055	0.900	0.170	0.043	**0.554**	0.510	0.036	0.826
	Word-iAT	0.584	0.029	0.694	0.241	0.054	0.895	0.166	0.049	0.496	0.480	0.049	0.844
	Attention-RP	0.035	0.232	0.657	0.086	0.127	0.893	0.079	0.277	**0.554**	0.078	0.142	0.817
	Attention-AT	0.067	0.119	0.707	0.005	0.156	0.907	0.031	0.230	0.510	0.041	0.189	0.818
	Attention-iAT	0.096	0.222	0.684	0.129	0.200	**0.915**	0.074	0.271	0.512	0.108	0.183	0.831
	SEAT(**Ours**)	**4.70E-07**	**0.002**	0.713	**2.77E-07**	**0.036**	0.907	**2.42E-06**	**0.042**	0.545	**1.68E-08**	**0.003**	0.841

Table 1: Results of evaluating embedding perturbation stability of (modified) attentions given by different methods using RNN, BiLSTM, and BERT under three metrics (JSD is for attention weight distribution, TVD is for output distribution, and F1 score is for model performance). The perturbation radius is set as δ_x=1e-3. ↑ means a higher value under this metric indicates better results, and ↓ means the opposite. The best performance is **bolded**. Same symbols are used in the following tables by default.

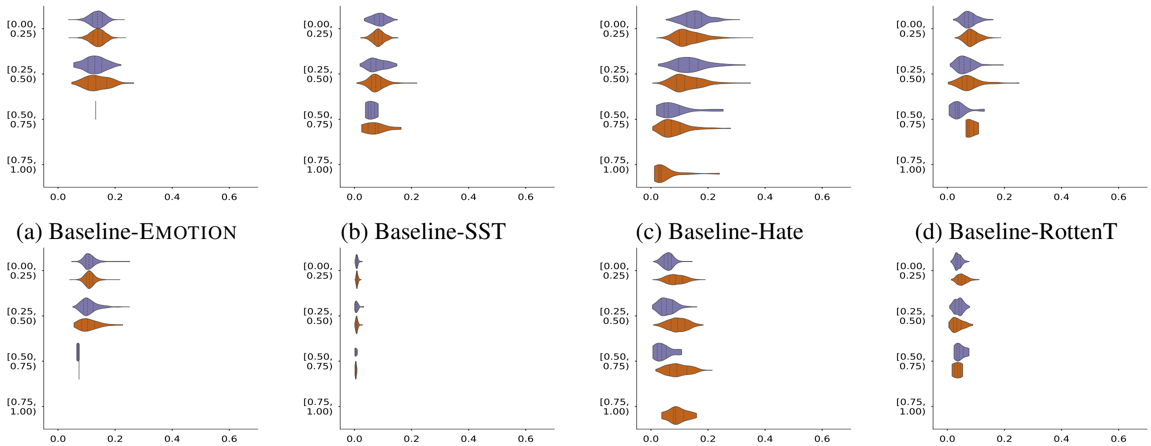
Model	Method	IMDB			SST			Hate			RottenT		
		Comp.↑	Suff.↓	Sens.↓	Comp.	Suff.	Sens.	Comp.	Suff.	Sens.	Comp.	Suff.	Sens.
RNN	Vanilla	0.004	0.007	0.131	5.744	7.02E-04	0.090	0.009	0.066	0.138	4.483	0.026	0.090
	Word-AT	2.899	0.018	0.121	5.280	**0.000**	0.081	2.408	0.058	0.137	2.512	0.031	0.093
	Word-iAT	2.060	0.010	0.122	5.452	9.35E-06	0.121	6.069	0.075	0.136	4.534	0.058	0.087
	Attention-RP	0.099	0.073	0.130	6.001	2.87E-04	0.085	3.585	0.080	0.133	4.637	0.052	0.094
	Attention-AT	0.026	0.052	0.131	3.118	**0.000**	0.088	2.493	0.372	0.134	4.713	0.049	0.096
	Attention-iAT	1.994	6.74E-04	0.117	4.788	**0.000**	0.091	5.351	0.126	0.133	2.435	0.039	0.086
	SEAT(**Ours**)	**3.281**	**1.04E-05**	**0.106**	**6.016**	**0.000**	**0.076**	**6.558**	**2.75E-04**	**0.129**	**4.796**	**0.025**	**0.084**
BiLSTM	Vanilla	0.474	0.002	0.129	5.182	0.255	0.086	4.203	0.112	0.142	2.966	0.088	0.092
	Word-AT	1.449	0.015	0.121	5.167	8.44E-04	0.096	5.438	0.207	0.153	3.388	0.062	0.080
	Word-iAT	0.619	0.005	0.127	5.259	3.81E-05	0.087	4.568	0.320	0.145	3.339	0.078	0.082
	Attention-RP	0.561	0.002	0.127	2.865	0.007	0.101	2.248	0.199	0.148	4.073	0.200	0.082
	Attention-AT	1.294	0.051	0.111	2.129	0.004	0.098	3.220	0.065	**0.140**	4.925	0.216	0.082
	Attention-iAT	0.555	0.002	0.127	5.176	0.004	0.083	4.092	0.290	0.141	2.431	0.377	0.083
	SEAT(**Ours**)	**1.502**	**6.41E-04**	**0.098**	**5.435**	**4.37E-06**	**0.076**	**6.240**	**0.025**	**0.140**	**4.941**	**0.046**	**0.077**
BERT	Vanilla	5.07E-04	0.008	0.013	0.003	0.310	0.009	3.20E-04	0.401	0.016	0.001	0.092	0.010
	Word-AT	5.01E-05	0.005	0.016	1.26E-05	4.47E-04	0.005	4.01E-04	0.014	0.017	0.005	0.043	0.016
	Word-iAT	5.47E-04	0.007	0.017	1.51E-05	5.67E-04	0.004	0.003	0.045	0.017	2.93E-04	0.010	0.009
	Attention-RP	0.085	0.086	0.014	0.002	0.010	0.011	0.035	0.034	0.016	0.017	0.003	0.010
	Attention-AT	4.65E-05	0.338	0.016	4.50E-05	0.441	0.006	**0.004**	0.007	0.016	6.26E-04	0.032	0.011
	Attention-iAT	0.002	0.164	0.015	1.19E-04	9.07E-04	0.007	0.001	0.151	0.017	0.003	0.025	0.010
	SEAT(**Ours**)	**0.160**	**0.002**	**0.012**	**0.497**	**0.000**	**0.004**	**0.153**	**0.006**	**0.015**	**0.040**	**0.002**	**0.008**

Table 2: Results on evaluating the interpretability of different methods. We conduct experiments with three architectures, including RNN, BiLSTM, and BERT, under three criteria (comprehensiveness, sufficiency, and sensitivity score) on text classification datasets. The best performance is **bolded**.

Setup. First, we demonstrate the stability of SEAT under three different randomness and perturbations: (1) random seeds during training; (2) embedding perturbation during testing, and (3) word perturbation during testing. For each method, we use Jensen-Shannon Divergence (JSD) between its attention with on perturbation and its attention under perturbation to evaluate the stability of explainablity of the learned attention. And the Total Variation Distance (TVD) between the prediction distribution with no perturbation and prediction under perturbation is used to measure prediction stability.

Next, in order to show explainability of SEAT, we use the recent evaluation metrics of model interpretation proposed by (Yin et al. 2022; DeYoung et al. 2020). Specifically, we use three explainablity evaluation metrics: *comprehensiveness, sufficiency and sensitivity*.

Thirdly, we compare the *F1 score* of our SEAT with other baselines to verity the property of closeness of prediction. Finally, we conduct ablation study to verify the efficiency of our modules (regularizers) in objective function (6) corresponding to each condition. In Fig. 5 of Appendix we also test the validity of surrogate loss for top-k overlap function by comparing the performance of our loss (5) with the true top-k indices overlaps.



(a) Baseline-EMOTION (b) Baseline-SST (c) Baseline-Hate (d) Baseline-RottenT

(e) Ours-EMOTION (f) Ours-SST (g) Ours-Hate (h) Ours-RottenT

Figure 2: Comparison of stability against random seeds for vanilla attention and SEAT. Densities of maximum JS divergences (x-axis) as a function of the max attention (y-axis) in each instance between its base model and models initialized on different random seeds. In each max-attention bin, top (blue) is a negative-label instance, and bottom (red) is a positive-label instance.

Method	**Emotion**			**SST**			**Hate**			**RottenT**		
	JSD↓	**TVD↓**	**F1↑**	**JSD**	**TVD**	**F1**	**JSD**	**TVD**	**F1**	**JSD**	**TVD**	**F1**
Vanilla	0.628	2.847	**0.721**	0.315	3.655	0.912	0.491	2.004	0.493	0.585	3.464	0.845
Word-AT	0.004	0.022	0.694	0.175	0.065	0.910	0.111	0.058	0.546	0.473	0.056	0.836
Word-iAT	0.456	0.059	0.658	0.213	0.046	0.912	0.331	0.044	0.501	0.488	0.048	**0.852**
Attention-RP	0.039	0.235	0.657	0.089	0.128	0.893	0.085	0.278	0.554	0.078	0.143	0.817
Attention-AT	0.082	0.003	0.707	0.006	0.157	0.907	0.035	0.230	0.510	0.049	0.193	0.818
Attention-iAT	0.126	0.228	0.684	0.147	0.204	**0.915**	0.081	0.271	0.512	0.136	0.187	0.831
SEAT(Ours)	**1.72E-09**	**0.001**	0.716	**1.55E-06**	**0.028**	0.907	**8.69E-06**	**0.037**	**0.555**	**1.10E-05**	**0.035**	0.847

Table 3: Results on stability and utility of attention model trained with different methods under word perturbation. BERT is used as our architecture, and the perturbation word size is set as $N = 1$. JSD, TVD, and F1 are reported.

Model, Dataset and Baseline. Following (Jain and Wallace 2019) and (Wiegreffe and Pinter 2019), we mainly study the encoder-decoder architectures for binary classification tasks in this paper. For encoder, we consider three kinds of networks as feature extractors: RNN, BiLSTM, and BERT. For decoder, we apply one simple MLP followed by a tanh-attention layer (Bahdanau, Cho, and Bengio 2015) and a softmax layer (Vaswani et al. 2017). In all experiments, we use four datasets: Stanford Sentiment Treebank (SST) (Socher et al. 2013), Emotion Recognition (Emotion) (Mohammad et al. 2018), Hate (Basile et al. 2019) and Rotten Tomatoes (RottenT) (Pang and Lee 2005). And we select the Binary Cross Entropy loss as D_1 and D_2 in (6). We compare our method with Vanilla attention (Wiegreffe and Pinter 2019), Word AT (Miyato, Dai, and Goodfellow 2016), Word iAT (Sato et al. 2018), Attention RP (attention weight is trained with random perturbation), Attention AT and Attention iAT (Kitada and Iyatomi 2021).

Stability Evaluation

Random seeds Here we compare the stability against random seed for vanilla attention and our SEAT. Specifically, we conduct multiple model training with different random seeds and select one of them as the base model. We visualize the JS divergence of the attention weight distribution between the base model and models trained with different random seeds for different methods. We conduct experiments on several test samples and each testing sample is divided into one of four bins by its maximum attention scores within the sentence. Here we use the RNN architecture.

We can see that Fig. 2 (c) and 2 (d) have heavy tails for the baseline vanilla attention on SST and Hate datasets. The violins covers more wider ranges along the x-axis. This can be interpreted as vanilla attention is unstable to random seeds. We can see from Fig. 2 (f)-(h) that while using our SEAT on SST, Hate and Rotten Tomatoes datasets, the violins are more narrow and their tails are lighter which imply SEAT is more stable. This can be further confirmed by the fact the violins of SEAL are more closer to zero than these of vanilla attention, which means their corresponds JSD values are more smaller.

Embedding perturbation We compare the stability of our SEAT with other baselines under embedding perturbation. In this setting, we mainly consider two metrics: JSD and TVD, which represents the explainability stability and prediction stability respectively. Details of our setting are as follows.

1. For testing the stability of explaination, it is sufficient to test the stability of the (modified) attention weights given by different methods. Specifically, we select one token of the input with embedding x, and then generate a perturbed embedding $x' = x + \mathcal{N}(0, \delta_x \mathbb{I})$ for some radius δ_x. For

each method, if we denote w as its corresponding (modified) attention based on x and w' as its corresponding (modified) attention based on x', we calculate $\mathrm{JSD}(w, w')$ to measure such stability.

2. We also test the stability of prediction distribution. Specifically, similar to the above, we have two inputs, one is original and the other one is perturbed by some noise. For each method, we denote the prediction distribution of its attention based on the original input as $y \in \mathbb{R}^{|\mathcal{Y}|}$, and denote the prediction distribution of its attention based on the perturbed one as y'. Then we compute $\mathrm{TVD}(y, y')$ to measure such stability.

Results are shown in Tab. 1 and Tab. 8 in Appendix. We can see that SEAT outperforms other baselines with RNN, BiLSTM and BERT under JSD and TVD evaluation metrics. Especially, we can see that the JSD of all our results are almost zero, which means SEAT is stable to perturbation for explaination. We can see also that the TVD for vanilla attention is large which means vanilla attention is extremely unstable to perturbation for its prediction distribution. However, the TVD of SEAT is small.

Word perturbation We now aim to evaluate the stability of our proposed method under word perturbation. Following (Yin et al. 2022), we select BERT as our main model in this part and conduct the perturbation in the following process: first we randomly choose K words from a given sentence and then replace it with the closest synonyms. The distance of words are computed based on gensim (Rehurek and Sojka 2011). We denote the original input and perturbed input as x and x' respectively. Then, similar to the above procedures, we can compute JSD and TVD for each method. Tab. 3 and Tab. 10 in Appendix demonstrate that SEAT achieves SOTA for both JSD and TVD in this setting. Similar to the embedding perturbation case, we can see the JSD of SEAT is much smaller than it of the vanilla attention and it value is quite close to zero in all experiments, which indicates strong explaination stability against to word perturbation for SEAT.

Models	Ablation Setting		Metrics		
	$\mathcal{L}_3$	$\mathcal{L}_{Topk}$	Suff.↓	TVD↓	F1↑
RNN			7.02E-04	21.464	**0.814**
	✓		6.22E-04	1.966	0.804
		✓	2.22E-04	2.997	0.782
	✓	✓	**1.02E-04**	**1.275**	0.813
BiLSTM			0.255	20.398	**0.809**
	✓		0.016	1.214	0.802
		✓	0.004	1.745	0.779
	✓	✓	**4.37E-06**	**1.095**	0.801
BERT			0.310	2.617	**0.912**
	✓		0.280	0.056	0.909
		✓	0.090	0.157	0.907
	✓	✓	**0.019**	**0.028**	0.909

Table 4: Ablation study of SEAT. We evaluate the effectiveness of $\mathcal{L}_3$ and $\mathcal{L}_{Topk}$ in (6). Perturbation on the embedding space (radius $\delta_x = 0.01$) are conducted on SST.

Evaluating Interpretability and Utility

In this part, we measure the interpretability of SEAT and other baselines using comprehensiveness, sufficiency and sensitivity. Results are showed in Tab. 2 and Tab. 9 in Appendix. Our results show that SEAT outperforms other baselines on all three evaluation metrics with RNN, BiLSTM and BERT. This further confirms that enhancing stability in attention would derive a more faithful interpretation. Our SEAT improves the model interpretability.

In Tab. 1 and 3 we also compared the F1 score for different methods. We can see that while in some of the results, our method is not the best one. However, among these results, the difference between the best result and ours is quite small, which indicates that there is almost no accuracy deterioration in SEAT. Surprisingly, we can also see that SEAT is better than vanilla attention in most results and it could even achieve SOTA in some cases, such as when the model is RNN and the data is Emotion or Hate in Tab. 1.

Ablation Study

In ablation study, we evaluate each module (regularization) in (6). Specifically, we denote $\mathcal{L}_1$ as our main loss in (1), then we consider and evaluate different combinations by deleting $\mathcal{L}_{Topk}$ or/and $\mathcal{L}_3$, where $\mathcal{L}_3$ corresponds to the third term in (6). Note that if there is no $\mathcal{L}_{Topk}$ and $\mathcal{L}_3$, then the model will be the vanilla attention. Tab. 4, Tab. 6 and 7 in Appendix show that each regularizer in our objective function is indispensable and effective. Specifically, we can see the sufficiency will decrease significantly if we add the $\mathcal{L}_{Topk}$ loss. This is due to that $\mathcal{L}_{Topk}$ enforces a large overlaps on top-k indices and thus it makes SEAT inherit the explainability of vanilla attention and it makes the model more stable. Moreover, in the case where there is $\mathcal{L}_{Topk}$, adding term $\mathcal{L}_3$ could further decrease sufficiency as it further improves stability.

Since $\mathcal{L}_3$ is to make the prediction distribution stable against any randomness and perturbation, from Tab. 6 we can see adding this term could decrease the TVD, which means it improve the stability. Although $\mathcal{L}_{Topk}$ can also help to decrease TVD, we can see it is weaker than $\mathcal{L}_3$. Besides stability, $\mathcal{L}_3$ also can pull back the F1 score to make the model close to vanilla attention. We can see that in the case where there only exists $\mathcal{L}_{Topk}$, the F1 score decreases compared with vanilla attention. And adding $\mathcal{L}_3$ do help minimizing the gap. This is due to that $\mathcal{L}_3$ could improve the model generalization performance by making model more stable.

Conclusions

In this paper, we provide a first rigorous definition namely SEAT as a substitute of attention to give a more faithful explanation. The definition has three properties: closeness of prediction, similarity of explainability and stability. We also propose a method to get such a SEAT. Results show that SEAT outperforms other baselines and considered as an effective and more faithful explanation tool.

Acknowledgements

Di Wang and Lijie Hu were supported in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01, FCC/1/1976-49-01 from CBRC and funding from the AI Initiative REI/1/4811-10-01 of King

Abdullah University of Science and Technology (KAUST). Di Wang was also supported by the funding of the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. F. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 3504–3512.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458.

Dombrowski, A.; Alber, M.; Anders, C. J.; Ackermann, M.; Müller, K.; and Kessel, P. 2019. Explanations can be manipulated and geometry is to blame. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 13567–13578.

Dong, Y.; Li, Z.; Rezagholizadeh, M.; and Cheung, J. C. K. 2019. EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3393–3402. Florence, Italy: Association for Computational Linguistics.

Du, M.; Liu, N.; and Hu, X. 2020. Techniques for interpretable machine learning. *Commun. ACM*, 63(1): 68–77.

Galassi, A.; Lippi, M.; and Torroni, P. 2020. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10): 4291–4308.

Ghorbani, A.; Abid, A.; and Zou, J. Y. 2019. Interpretation of Neural Networks Is Fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 3681–3688. AAAI Press.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3543–3556. Minneapolis, Minnesota: Association for Computational Linguistics.

Kitada, S.; and Iyatomi, H. 2021. Attention Meets Perturbations: Robust and Interpretable Attention With Adversarial Training. *IEEE Access*, 9: 92974–92985.

Lei, T. 2017. *Interpretable neural models for natural language processing*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, 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 *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150. Portland, Oregon, USA: Association for Computational Linguistics.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Martins, A. F. T.; and Astudillo, R. F. 2016. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In Balcan, M.; and Weinberger, K. Q., eds., *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, 1614–1623. JMLR.org.

Miyato, T.; Dai, A. M.; and Goodfellow, I. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, 1–17.

Mohankumar, A. K.; Nema, P.; Narasimhan, S.; Khapra, M. M.; Srinivasan, B. V.; and Ravindran, B. 2020. Towards Transparent and Explainable Attention Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4206–4216.

Otter, D. W.; Medina, J. R.; and Kalita, J. K. 2021. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans. Neural Networks Learn. Syst.*, 32(2): 604–624.

Pang, B.; and Lee, L. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 115–124.

Rehurek, R.; and Sojka, P. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Ribeiro, M.; Singh, S.; and Guestrin, C. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. San Diego, California: Association for Computational Linguistics.

Sato, M.; Suzuki, J.; Shindo, H.; and Matsumoto, Y. 2018. Interpretable adversarial perturbation in input embedding space for text. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4323–4330.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Florence, Italy: Association for Computational Linguistics.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wiegreffe, S.; and Pinter, Y. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 11–20. Hong Kong, China: Association for Computational Linguistics.

Yeh, C.; Hsieh, C.; Suggala, A. S.; Inouye, D. I.; and Ravikumar, P. 2019. On the (In)fidelity and Sensitivity of Explanations. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 10965–10976.

Yin, F.; Shi, Z.; Hsieh, C.-J.; and Chang, K.-W. 2022. On the Sensitivity and Stability of Model Interpretations in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2631–2647. Dublin, Ireland: Association for Computational Linguistics.

Implementation Details

Dataset	Avg. Length	Train Size	Test Size
	(texts)	(neg/pos)	(neg/pos)
Emotion	91	1400/708	558/358
SST	19	3610/3310	909/912
Hate	23	3783/5217	1252/1718
Rotten Tomatoes	23	4265/4265	533/533

Table 5: Dataset statistics.

Experimental Setup

Dataset Details Following the previous study (Serrano and Smith 2019; Wiegreffe and Pinter 2019), we mainly focus on the experimental evaluation of binary classification tasks in this paper. To be specific, we conduct experiments on the following datasets: Stanford Sentiment Treebank (SST) (Socher et al. 2013), Emotion Recognition (Mohammad et al. 2018), Hate (Basile et al. 2019) and Rotten Tomatoes (Pang and Lee 2005). For each dataset, we extract the samples with label 0 or 1 in the dataset. The details of datasets are presented in 5. We use the default train-test split configuration offered by the Huggingface for each dataset. Note our framework is task-agnostic and we will leave examining our framework for other NLP tasks such as question answering (QA) and natural language inference (NLI) as future work.

Models and architectures We mainly study encoder-decoder architectures in this paper following (Wiegreffe and Pinter 2019). For the encoder, we consider three kinds of networks as feature extractors: RNN, BiLSTM, and BERT. For the decoder, we apply one simple MLP following with a tanh-attention layer (Bahdanau, Cho, and Bengio 2015) and a softmax layer (Vaswani et al. 2017), which is followed by (Jain and Wallace 2019).

Metrics In our experiments, we seek to evaluate stability, explainability and performance of our model.

For evaluating the stability. Following (Jain and Wallace 2019), for each method, we use Jensen-Shannon Divergence (JSD) between its attention with on perturbation and its attention under perturbation to evaluate the **stability of explainablity** of the learned attention, which is defined as

$$\text{JSD}(\alpha_1, \alpha_2) = \frac{1}{2}\text{KL}[\alpha_1 || \frac{\alpha_1 + \alpha_2}{2}] + \frac{1}{2}\text{KL}[\alpha_2 || \frac{\alpha_1 + \alpha_2}{2}],$$

where KL is the KL divergence between two distributions. We also the Total Variation Distance (TVD) between the prediction distribution with no perturbation and the prediction distribution under perturbation is used to measure **prediction stability**, which can be defined as,

$$\text{TVD}(\hat{y}_1, \hat{y}_2) = \frac{1}{2}\sum_{i=1}^{|\mathcal{Y}|} |\hat{y}_{1i} - \hat{y}_{2i}|.$$

To evaluate the performance of our model performance, we report the F1 score of models on clean test set.

To evaluate the model interpretability, we leverage three different metrics in the previous study (Yin et al. 2022; DeYoung et al. 2020). Specifically, we use *comprehensiveness, sufficiency and sensitivity* to evaluate the model performance on interpretation and stability. Following (DeYoung et al. 2020), for given input sentence x_i, we identify the rationales r_i (i.e., sequence of tokens that help classifying) of this sentence by exacting the top-k elements based on attention values. Note that for the consideration of computation cost, we sample certain percentage of data to evaluate comprehensiveness and sufficiency.

Comprehensiveness, Sufficiency (DeYoung et al. 2020). Following the notation in the previous study, we denote the model as $m(\cdot)$. Intuitively, model with higher interpretation ability should be less confident in making decision once the rationales are removed. Formally, we have

$$\text{comprehensiveness } = m\left(x_i\right)_j - m\left(x_i \backslash r_i\right)_j \quad (7)$$

Another related metric of comprehensiveness is sufficiency. It captures the degree that extracted rationales are sufficient for the model to make classification. Formally, we have

$$\text{sufficiency } = m\left(x_i\right)_j - m\left(r_i\right)_j \quad (8)$$

Sensitivity. (Yin et al. 2022) As point out in (Yin et al. 2022), models with higher interpretability should have lower sensitivity under some local and adversarial perturbations on important tokens. Specifically, given a sequence of rationales r_k, they evaluate the sensitivity by only adding perturbations to its corresponding embedding $e\left(r_k\right)$ and keeps the other token embedding the same as original. Then, they measures the minimal perturbation norm, denoted as ϵ_{r_k}, that changes the model prediction for this instance:

$$\epsilon_{r_k} = \min \|\boldsymbol{\delta}_{\boldsymbol{r_k}}\|_F \quad \text{s.t.} \quad f\left(e(x) + \boldsymbol{\delta}_{\boldsymbol{r_k}}\right) \neq y \quad (9)$$

where $\|\cdot\|_F$ is the Frobenius norm of the matrix, and $\boldsymbol{\delta}_{\boldsymbol{r_k}} \in \mathcal{R}^{n \times d}$ denotes the perturbation matrix where only the columns for tokens in r_k have non-zero elements. Given that the exact computation of ϵ_{r_k} is intractable, they use the PGD attack (Madry et al. 2018) with a binary search to approximate ϵ_{r_k}. Following (Yin et al. 2022), in practice, we vary the size of r_k, compute multiple ϵ_{r_k}, and summarize them with the area under the curve (AUC) score.

Baselines and Implementation

We compare our method with six baseline methods, which are the only works that study the self-explaining interpretation and stability of the attention mechanism. We first briefly introduce the consider baseline methods for completeness.

- *Vanilla* (Wiegreffe and Pinter 2019): model with an attention layer trained with natural cross entropy loss $\ell_\theta = \sum_i \text{CE}(\hat{y}(x_i), y_i)$. The cross-entropy of the distribution q relative to a distribution p over a given set is defined as follows:

  $$H(p,q) = -\mathbb{E}_p[\log q],$$

  where $\mathbb{E}_p[\cdot]$ is the expected value operator with respect to the distribution p.

- *Word AT* (Miyato, Dai, and Goodfellow 2016): model trained with weighted sum of benign loss and adversarial loss, i.e., $\ell_\theta = \sum_i \mathrm{CE}(\hat{y}(x_i), y_i) + \lambda \mathrm{CE}(\hat{y}(\tilde{x}_i), y_i)$ where the perturbation is conducted on the token embedding.
- *Word iAT* (Sato et al. 2018): similar to Word-AT method, Word-iAT conducts adversarial training on the token embedding space. The difference is that the adversarial perturbation constructed by Word-AT on each word (token) is based on the "interpretable" direction set. Specifically, given the embedding $e(x)$ of all the words x in the dictionary D and a word embedding $e(x_i)$ that aims to be perturbed, the perturbation direction $\Delta e(x_i)$ must be the linear combination of vector sampled from $S_{\Delta e(x_i)} = \{\forall j \neq i, \Delta e(x_{ij}) | \Delta e(x_{ij}) = e(x_j) - e(x_i)\}$, i.e., $\Delta e(x_i) = \sum_j \alpha_{ij} \Delta e(x_{ij})$, where $\alpha_j \in \mathbb{R}$ and $\Delta e(x_{ij}) \in S_{\Delta e(x_i)}$
- *Attention-AT, Attention-iAT, Attention-RP*(Kitada and Iyatomi 2021): The main ideas of the first two methods are similar as Word-AT and Word-iAT respectively. The difference is that here we conduct the adversarial training on the attention weights. Note that for the Attention-iAT method, the perturbation are constructed based on the difference of attention weight within sentence, i.e., $\Delta a_i = \sum_j \gamma_{ij}(a_j - a_i)$, where $\gamma_{ij} \in \mathbb{R}$. As for the Attention-RP method, model is trained with random noise perturbations on attention weights.

All the baseline methods are implemented on our own based on the guidance of original paper and official code (if it exists).

Model Training and Hyper-parameter

All models are implemented based on the PyTorch [1] library. All experiments are conducted on NVIDIA RTX A5000 GPUs. For RNN and BiLSTM, the initial learning rate, training epoch, and batch size is set as 0.01, 20 and 32 respectively. For RNN and BiLSTM classifier, we use an one-layer RNN/-BiLSTM encoder with hidden size of 128. The embedding of input token is initialized with the 300-dimensional pre-trained `fasttext` word embedding. For BERT, we use the bert-base-uncased model. We fine-tune BERT model on each dataset, using the initial learning rate of 5e-5, batch size 8, and training epoch 4. Adam are used as optimizer for all models. Early stopping on testing F1 metric is conducted during training to prevent over-fitting. In each of experiment, we use the following hyper-parameter setting by default. As for the coefficient of loss objective in Eq.4, $\lambda_1 = 1$ and $\lambda_2 = 1000$. As for the top-k overlap surrogate loss, we set the k as 7.

Additional Experimental Results

Surrogate Top-k Loss Performance

In this part, we test the efficiency of our surrogate top-k loss function. As we can see from the Fig.5, the top-k loss is decreasing along with the epoch number in all datasets. And the overlapping rate reaches 93 percent for SST and Rotten Tomatoes datasets. Results show that our surrogate top-k loss is valid to approximate the true top-k overlap score.

[1] https://pytorch.org/

Models	Ablation Setting		Metrics		
	$\mathcal{L}_3$	$\mathcal{L}_{Topk}$	JSD↓	TVD↓	F1↑
RNN			0.025	21.464	**0.814**
	✓		0.017	1.966	0.804
		✓	8.21E-08	2.997	0.782
	✓	✓	**6.98E-08**	**1.275**	0.813
BiLSTM			0.059	20.398	**0.809**
	✓		0.033	1.214	0.802
		✓	0.030	1.745	0.779
	✓	✓	**0.028**	**1.095**	0.801
BERT			0.140	2.617	**0.912**
	✓		0.114	0.056	0.909
		✓	0.006	0.157	0.907
	✓	✓	**2.49E-07**	**0.028**	0.909

Table 6: Ablation study of the proposed method. We evaluate the effectiveness of the third term $\mathcal{L}_3$ and the second term $\mathcal{L}_{Topk}$ in (6) on three different architectures. JSD, TVD and F1 are reported to measure the stability and performance of the model. Here we conduct perturbation on the embedding space with radius of 0.01 on SST dataset.

More Results on Stability and Explanability of Attention

In this section, we show more results on the stability and explanability of different methods with different perturbation radius δ on RNN, BiLSTM and BERT in Tab.8 and Tab.9. The perturbation is conducted on the embedding space.

More Results on Stability of Attention under Word Perturbation

In this section, we show results on the stability of different methods under word perturbation. The perturbation word number N are set as 2 and 3. As we can see from Tab.10, across all the setting, our method achieves the best performance compared to all baseline methods.

More Examples under Different Perturbation

In this section, we show more real examples to illustrate the stability of the proposed method under different kinds of perturbations. BERT is used as the model architecture and IMDB (Maas et al. 2011) is used as the demonstrated dataset.

As we can see from Fig.3, different random seeds mainly affect the attention weight distribution (or explainable heat map in Fig.3) while produce little change on the prediction. Compared to baseline methods, the attention weight distribution generated by our method are more stable under perturbation.

For the results under perturbation on embedding space, we can observe more notable change on both model predication and attention weight distribution. As we can see from Fig.4, vanilla attention methods are unstable under perturbation even when δ_x =2e-3 since it classifies the input text to wrong class with high probability. Despite that model trained with Word-AT can ensure the stability on prediction, however, it still produces unstable attention weight distribution. Contrast

Models	Ablation Setting		Metrics		
	$\mathcal{L}_3$	$\mathcal{L}_{Topk}$	**Comp.↑**	**Suff.↓**	**Sens.↓**
RNN			5.744	7.02E-04	0.090
	✓		5.280	6.22E-04	0.081
		✓	5.944	2.22E-04	0.088
	✓	✓	**6.016**	**1.02E-04**	**0.076**
BiLSTM			5.182	0.255	0.098
	✓		2.219	0.016	0.087
		✓	5.167	0.004	0.095
	✓	✓	**5.435**	**4.37E-06**	**0.076**
BERT			0.003	0.310	0.009
	✓		0.002	0.280	0.005
		✓	0.210	0.090	0.006
	✓	✓	**0.497**	**0.019**	**0.004**

Table 7: Ablation study of the proposed method on the explainability. We evaluate the effectiveness of the the second term $\mathcal{L}_{Topk}$ and third term $\mathcal{L}_3$ in (6) on three different architectures. Comprehensiveness, Sufficiency and Sensitivity are reported to measure the explainability of the model. Here we conduct perturbation on the embedding space with radius of 0.01 on SST dataset.

to Word-AT, Attention-AT can improve the stability of attention distribution while lack stability on prediction. Compared to these baseline methods, our methods achieve the best performance under two metrics. It can be seen from Fig.4 that model trained with our method is stable under perturbation on both the prediction and the attention distribution.

Additional Ablation Study

In this section, we provide more ablation studies on the exapliability and stability of model. We evaluate each module (regularization) in Eq.6. As we can see from Tab.7, $\mathcal{L}_3$ are more effective in reducing Sens. while $\mathcal{L}_{Topk}$ contribute more on the Comp. and Suff. And the combination of all the competent yield the model with the highest exaplianability. This indicates that the loss term $\mathcal{L}_{Topk}$ and $\mathcal{L}_3$ proposed in Eq.6 are both indispensable for the performance improvement on explainability.

Tab. 6 further confirms that each regularizer in our objective function is indispensable. Specifically, we can see JSD will decrease significantly if we add the $\mathcal{L}_{Topk}$ loss. This is due to that $\mathcal{L}_{Topk}$ enforces a large overlaps on top-k indices and thus it makes SEAT inherit the explainability of vanilla attention. Moreover, in the case where there is $\mathcal{L}_{Topk}$, adding term $\mathcal{L}_3$ could further decrease JSD as it further enhances the stability.

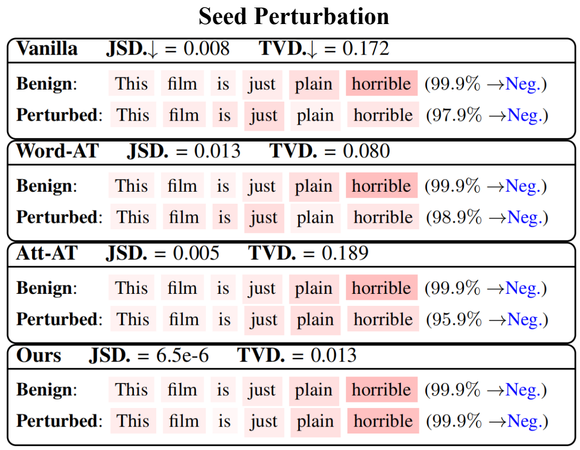


Figure 3: An example of the stability of model trained with different method under seed perturbation. As we can see from figure, the perturbation of random seed mainly affect the attention distribution while producing little change on the prediction. Compared to baseline methods, our method generate nearly similar attention weight distribution as it is before perturbed. This observation indicates that our propose method are able to improve the stability under seed perturbation.

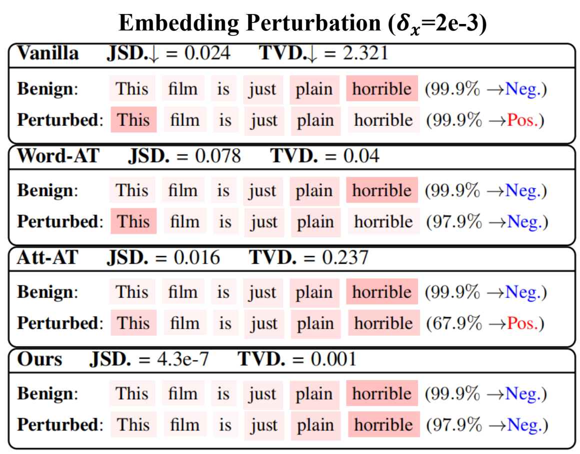


Figure 4: An example of the stability of model trained with different method under embedding perturbation. As it is shown in the figure, we can observe more obvious change both on model predication and attention weight distribution. The considered baseline methods are either not sufficiently stable in terms of prediction or not stable in terms of attention distribution. Results demonstrate that our method outperform them in terms of stability under embedding perturbation.

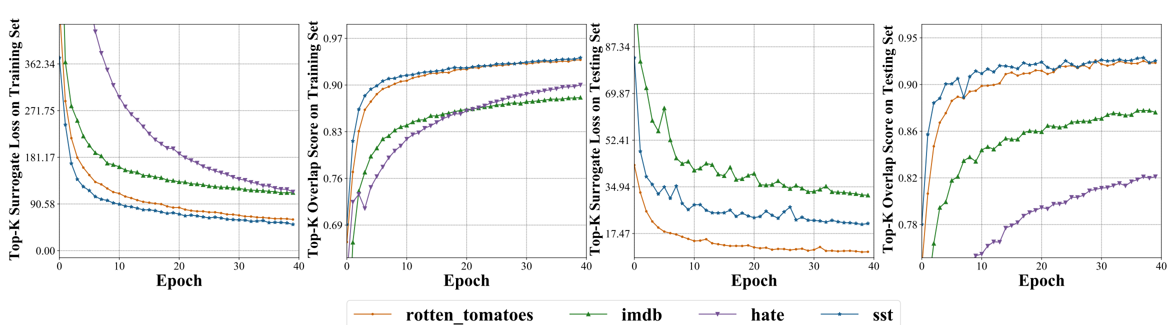


Figure 5: Surrogate Top-k Loss Performance. Evaluation curves of Surrogate Top-k loss performance. The left two figures are the training loss and overlapping score on training set. The right two figures are the testing loss and overlapping score on testing set. The X-axis is the number of epoch. The Y-axis is the performance under the setting. Results show that our purposed Top-k overlap surrogate loss is a good approximation of true non-differentiable Top-k loss.

Radius(δ_x)	Model	Method	Emotion			SST			Hate			RottenT		
			JSD↓	TVD↓	F1↑	JSD	TVD	F1	JSD	TVD	F1	JSD	TVD	F1
0.005	RNN	Vanilla	0.002	20.120	0.677	0.019	19.916	0.767	0.009	15.627	0.513	0.008	19.240	0.763
		Word-AT	0.016	1.634	0.642	0.017	2.186	0.800	0.035	1.618	0.519	0.058	1.401	0.768
		Word-iAT	0.027	1.570	0.637	0.021	1.331	0.811	0.025	1.387	0.537	0.092	1.719	0.759
		Attention-RP	0.026	3.281	0.671	0.028	2.209	0.792	0.025	2.650	0.554	0.009	3.759	0.770
		Attention-AT	0.055	2.738	0.665	0.047	2.595	0.782	0.032	2.195	0.528	0.068	4.328	0.755
		Attention-iAT	0.017	3.573	0.645	0.050	2.821	0.746	0.039	2.261	0.533	0.055	1.645	0.753
		SEAT(**Ours**)	8.41E-09	1.320	0.689	4.19E-08	1.143	0.803	2.51E-09	1.263	0.579	-2.87E-10	1.174	0.762
	BiLSTM	Vanilla	0.002	23.440	0.653	0.037	19.215	0.809	0.062	15.805	0.524	0.009	20.197	0.764
		Word-AT	0.053	1.632	0.655	0.025	1.080	0.794	0.058	2.126	0.528	0.039	1.223	0.772
		Word-iAT	0.054	1.902	0.643	0.029	1.120	0.809	0.078	1.899	0.525	0.064	1.187	0.750
		Attention-RP	0.031	1.391	0.642	0.036	1.396	0.772	0.054	1.388	0.522	0.068	1.539	0.764
		Attention-AT	0.079	1.617	0.672	0.030	1.748	0.779	0.058	1.582	0.523	0.080	2.129	0.766
		Attention-iAT	0.035	1.371	0.651	0.036	1.722	0.801	0.063	2.310	0.525	0.079	1.931	0.777
		SEAT(**Ours**)	9.68E-09	0.736	0.671	2.70E-08	1.040	0.805	1.47E-08	1.081	0.543	2.06E-08	0.991	0.770
	BERT	Vanilla	0.026	2.130	0.721	0.004	2.610	0.912	0.039	1.771	0.493	0.010	2.504	0.845
		Word-AT	0.013	0.024	0.694	0.100	0.035	0.891	0.006	0.074	0.579	0.563	0.083	0.845
		Word-iAT	0.456	0.059	0.658	0.213	0.094	0.912	0.331	0.043	0.501	0.488	0.048	0.852
		Attention-RP	0.039	0.235	0.657	0.089	0.128	0.893	0.085	0.278	0.554	0.078	0.143	0.817
		Attention-AT	0.081	0.019	0.707	0.006	0.157	0.907	0.035	0.230	0.510	0.049	0.193	0.818
		Attention-iAT	0.126	0.228	0.684	0.147	0.204	0.915	0.080	0.271	0.512	0.135	0.187	0.831
		SEAT(**Ours**)	1.70E-08	0.001	0.717	1.29E-07	0.028	0.911	6.47E-06	0.003	0.575	2.61E-05	0.033	0.848
0.01	RNN	Vanilla	0.002	20.119	0.657	0.021	20.716	0.797	0.010	15.798	0.513	0.008	19.383	0.763
		Word-AT	0.017	1.385	0.629	0.017	1.266	0.814	0.023	1.238	0.532	0.054	1.196	0.765
		Word-iAT	0.027	1.576	0.637	0.021	1.331	0.811	0.025	1.248	0.537	0.092	1.716	0.759
		Attention-RP	0.026	3.296	0.671	0.028	2.208	0.792	0.025	2.662	0.554	0.009	3.750	0.770
		Attention-AT	0.055	2.717	0.665	0.048	2.587	0.782	0.032	2.209	0.528	0.068	4.358	0.755
		Attention-iAT	0.017	3.554	0.645	0.050	2.837	0.746	0.039	2.261	0.533	0.055	1.646	0.753
		SEAT(**Ours**)	1.40E-08	1.250	0.688	6.44E-08	1.175	0.813	6.48E-10	1.198	0.579	1.16E-07	0.975	0.762
	BiLSTM	Vanilla	0.002	23.432	0.643	0.059	20.398	0.809	0.072	16.312	0.524	0.010	20.322	0.764
		Word-AT	0.045	1.551	0.658	0.033	1.214	0.802	0.071	1.609	0.516	0.062	1.214	0.765
		Word-iAT	0.054	1.899	0.643	0.029	1.594	0.809	0.078	1.899	0.525	0.064	1.203	0.750
		Attention-RP	0.031	1.390	0.642	0.036	1.398	0.772	0.054	1.386	0.522	0.068	1.540	0.764
		Attention-AT	0.080	1.613	0.672	0.030	1.745	0.779	0.058	1.581	0.523	0.080	2.130	0.766
		Attention-iAT	0.035	1.370	0.651	0.036	1.724	0.801	0.063	2.315	0.525	0.079	1.932	0.777
		SEAT(**Ours**)	1.04E-08	0.960	0.664	1.63E-07	1.116	0.804	8.17E-09	1.167	0.547	6.44E-09	0.948	0.770
	BERT	Vanilla	0.033	2.134	0.721	0.004	2.617	0.912	0.047	1.772	0.493	0.010	2.508	0.845
		Word-AT	0.004	0.040	0.694	0.118	0.046	0.909	0.309	0.074	0.540	0.390	0.086	0.829
		Word-iAT	0.456	0.059	0.658	0.213	0.107	0.912	0.331	0.100	0.501	0.488	0.048	0.852
		Attention-RP	0.039	0.235	0.657	0.089	0.128	0.893	0.085	0.278	0.554	0.078	0.143	0.817
		Attention-AT	0.082	0.025	0.707	0.006	0.157	0.907	0.035	0.230	0.510	0.049	0.193	0.818
		Attention-iAT	0.126	0.228	0.684	0.147	0.204	0.915	0.081	0.271	0.512	0.135	0.187	0.831
		SEAT(**Ours**)	3.61E-09	0.001	0.713	2.49E-07	0.028	0.909	8.38E-06	0.044	0.561	1.18E-06	0.034	0.844
0.05	RNN	Vanilla	0.006	20.911	0.667	0.047	28.278	0.767	0.025	18.398	0.513	0.015	21.265	0.763
		Word-AT	0.023	1.652	0.627	0.019	1.270	0.797	0.031	1.320	0.522	0.030	1.126	0.748
		Word-iAT	0.027	1.577	0.637	0.021	1.325	0.811	0.025	1.235	0.537	0.092	1.704	0.759
		Attention-RP	0.026	3.294	0.671	0.028	2.212	0.792	0.025	2.662	0.554	0.009	3.763	0.770
		Attention-AT	0.055	2.675	0.665	0.048	2.611	0.782	0.032	2.210	0.528	0.068	4.370	0.755
		Attention-iAT	0.017	3.559	0.645	0.050	2.814	0.746	0.039	2.277	0.533	0.055	1.634	0.753
		SEAT(**Ours**)	1.87E-08	1.546	0.691	3.77E-08	1.098	0.805	4.54E-10	1.213	0.579	-3.39E-11	3.02E-04	0.767
	BiLSTM	Vanilla	0.003	23.377	0.653	0.258	28.472	0.809	0.190	20.458	0.524	0.022	22.146	0.764
		Word-AT	0.022	1.279	0.656	0.028	1.062	0.790	0.055	1.771	0.521	0.047	1.451	0.769
		Word-iAT	0.054	1.900	0.643	0.029	1.109	0.809	0.077	1.862	0.525	0.063	1.194	0.750
		Attention-RP	0.031	1.390	0.642	0.036	1.409	0.772	0.054	1.381	0.522	0.067	1.533	0.764
		Attention-AT	0.080	1.622	0.672	0.030	1.745	0.779	0.058	1.572	0.523	0.080	2.126	0.766
		Attention-iAT	0.035	1.368	0.651	0.036	1.712	0.801	0.063	2.314	0.525	0.078	1.907	0.777
		SEAT(**Ours**)	1.44E-08	1.094	0.664	1.04E-08	1.043	0.801	1.47E-08	1.229	0.524	1.35E-08	1.027	0.768
	BERT	Vanilla	0.193	2.197	0.721	0.008	2.680	0.912	0.193	1.780	0.493	0.042	2.553	0.845
		Word-AT	0.004	0.022	0.694	0.175	0.065	0.910	0.111	0.058	0.546	0.473	0.056	0.836
		Word-iAT	0.456	0.059	0.658	0.213	0.046	0.912	0.331	0.044	0.501	0.488	0.048	0.852
		Attention-RP	0.039	0.235	0.657	0.089	0.128	0.893	0.085	0.278	0.554	0.078	0.143	0.817
		Attention-AT	0.082	0.003	0.707	0.006	0.157	0.907	0.035	0.230	0.510	0.049	0.193	0.818
		Attention-iAT	0.126	0.228	0.684	0.147	0.204	0.915	0.081	0.271	0.512	0.136	0.187	0.831
		SEAT(**Ours**)	1.72E-09	0.001	0.717	1.55E-06	0.028	0.910	8.69E-06	0.037	0.555	1.10E-05	0.035	0.849

Table 8: More results on the performance of different methods under different perturbation radii. JSD and TVD are reported to measure the stability while F1 measure the model accuracy of the clean data. From the table we can see that SEAT have lower JSD and TVD with little or even no drop for F1 under all the settings. These suggest that our method suppresses other baseline methods on both model stability and effectiveness.

Radius(δ_x)	Model	Method	Emotion			SST			Hate			RottenT		
			Comp.↑	Suff.↓	Sens.↓	Comp.	Suff.	Sens.	Comp.	Suff.	Sens.	Comp.	Suff.	Sens.
0.005	RNN	Vanilla	3.281	0.007	0.143	5.744	3.90E-04	0.160	0.073	0.066	0.179	4.483	0.127	0.120
		Word-AT	0.006	0.072	0.149	5.585	0.00E+00	0.157	3.259	0.246	0.198	2.744	0.255	0.113
		Word-iAT	0.365	0.005	0.139	4.433	5.85E-04	0.155	2.178	0.091	0.174	2.804	0.091	0.122
		Attention-RP	0.099	0.073	0.133	6.001	2.87E-04	0.144	3.585	0.080	0.169	4.637	0.052	0.127
		Attention-AT	0.026	0.052	0.132	5.608	0.00E+00	0.144	2.493	0.372	0.172	4.713	0.049	0.128
		Attention-iAT	1.994	6.74E-04	0.138	4.788	0.00E+00	0.159	5.351	0.126	0.179	2.435	0.039	0.119
		SEAT(**Ours**)	5.099	1.74E-05	0.131	6.016	0.00E+00	0.140	6.558	0.001	0.131	5.342	0.026	0.105
	BiLSTM	Vanilla	0.474	0.002	0.136	5.182	0.003	0.147	4.795	0.112	0.199	2.966	0.088	0.125
		Word-AT	2.195	0.029	0.193	2.783	0.007	0.155	3.085	0.111	0.188	1.827	0.117	0.135
		Word-iAT	1.765	0.015	0.190	5.082	2.62E-04	0.165	5.409	0.093	0.202	3.759	0.100	0.119
		Attention-RP	0.561	0.002	0.173	2.865	0.007	0.149	2.248	0.199	0.189	4.073	0.200	0.117
		Attention-AT	1.294	6.41E-04	0.143	2.129	0.004	0.129	3.220	0.124	0.190	4.925	0.216	0.122
		Attention-iAT	0.555	0.002	0.135	5.435	0.004	0.152	4.092	0.290	0.197	4.941	0.377	0.143
		SEAT(**Ours**)	2.306	4.45E-04	0.135	5.454	4.37E-06	0.129	6.240	0.065	0.187	4.956	0.037	0.114
	BERT	Vanilla	5.07E-04	0.008	0.016	0.003	0.310	0.010	3.20E-04	0.006	0.017	0.001	0.092	0.010
		Word-AT	6.44E-05	0.014	0.016	0.164	0.165	0.007	4.09E-06	0.012	0.018	4.82E-04	0.005	0.011
		Word-iAT	4.79E-04	0.013	0.016	9.51E-04	0.024	0.008	3.80E-04	0.012	0.019	6.15E-04	0.016	0.012
		Attention-RP	0.085	0.086	0.015	0.002	0.010	0.012	0.035	0.034	0.018	0.017	0.003	0.010
		Attention-AT	4.65E-05	0.003	0.017	5.82E-04	0.450	0.007	0.004	0.007	0.017	0.001	0.116	0.011
		Attention-iAT	5.74E-05	0.164	0.016	1.19E-04	9.07E-04	0.007	4.05E-04	0.151	0.017	0.003	0.025	0.011
		SEAT(**Ours**)	0.160	0.002	0.014	0.497	0.00E+00	0.007	0.153	6.74E-05	0.016	0.040	0.002	0.009
0.01	RNN	Vanilla	0.004	0.007	0.131	5.744	7.02E-04	0.090	0.009	0.066	0.138	4.483	0.026	0.090
		Word-AT	2.899	0.018	0.121	5.280	0.00E+00	0.081	2.408	0.058	0.137	2.512	0.031	0.093
		Word-iAT	2.060	0.010	0.122	5.452	9.35E-06	0.121	6.069	0.075	0.136	4.534	0.058	0.087
		Attention-RP	0.099	0.073	0.130	6.001	2.87E-04	0.085	3.585	0.080	0.133	4.637	0.052	0.094
		Attention-AT	0.026	0.052	0.131	3.118	0.00E+00	0.088	2.493	0.372	0.134	4.713	0.049	0.096
		Attention-iAT	1.994	6.74E-04	0.117	4.788	0.00E+00	0.091	5.351	0.126	0.133	2.435	0.039	0.086
		SEAT(**Ours**)	3.281	1.04E-05	0.106	6.016	0.00E+00	0.076	6.558	2.75E-04	0.129	4.796	0.025	0.084
	BiLSTM	Vanilla	0.474	0.002	0.129	5.182	0.255	0.086	4.203	0.112	0.142	2.966	0.088	0.092
		Word-AT	1.449	0.015	0.121	5.167	8.44E-04	0.096	5.438	0.207	0.153	3.388	0.062	0.080
		Word-iAT	0.619	0.005	0.127	5.259	3.81E-05	0.087	4.568	0.320	0.145	3.339	0.078	0.082
		Attention-RP	0.561	0.002	0.127	2.865	0.007	0.101	2.248	0.199	0.148	4.073	0.200	0.082
		Attention-AT	1.294	0.051	0.111	2.129	0.004	0.098	3.220	0.065	0.140	4.925	0.216	0.082
		Attention-iAT	0.555	0.002	0.127	5.176	0.004	0.083	4.092	0.290	0.141	2.431	0.377	0.083
		SEAT(**Ours**)	1.502	6.41E-04	0.098	5.435	4.37E-06	0.076	6.240	0.025	0.140	4.941	0.046	0.077
	BERT	Vanilla	5.07E-04	0.008	0.013	0.003	0.310	0.009	3.20E-04	0.401	0.016	0.001	0.092	0.010
		Word-AT	5.01E-05	0.005	0.016	1.26E-05	4.47E-04	0.005	4.01E-04	0.014	0.017	0.005	0.043	0.016
		Word-iAT	5.47E-04	0.007	0.017	1.51E-05	5.67E-04	0.004	0.003	0.045	0.017	2.93E-04	0.010	0.009
		Attention-RP	0.085	0.086	0.014	0.002	0.010	0.011	0.035	0.034	0.016	0.017	0.003	0.010
		Attention-AT	4.65E-05	0.338	0.016	4.50E-05	0.441	0.006	0.004	0.007	0.016	6.26E-04	0.032	0.011
		Attention-iAT	0.002	0.164	0.015	1.19E-04	9.07E-04	0.007	0.001	0.151	0.017	0.003	0.025	0.010
		SEAT(**Ours**)	0.160	0.002	0.012	0.497	0.00E+00	0.004	0.153	0.006	0.015	0.040	0.002	0.008
0.05	RNN	Vanilla	0.006	0.007	0.229	5.744	0.001	0.260	0.017	0.066	0.253	4.483	0.026	0.225
		Word-AT	0.129	0.115	0.147	4.413	0.005	0.259	5.551	0.132	0.251	2.738	0.046	0.203
		Word-iAT	0.365	0.005	0.175	4.433	5.85E-04	0.258	2.178	0.091	0.246	2.804	0.091	0.203
		Attention-RP	0.099	0.073	0.142	6.001	2.87E-04	0.258	3.585	0.080	0.246	4.637	0.052	0.229
		Attention-AT	0.026	0.052	0.136	5.533	0.00E+00	0.259	2.493	0.372	0.252	0.008	0.049	0.236
		Attention-iAT	1.994	6.74E-04	0.215	4.788	0.00E+00	0.259	5.351	0.126	0.257	2.435	0.039	0.227
		SEAT(**Ours**)	3.281	1.38E-05	0.134	6.016	0.00E+00	0.254	6.558	3.12E-04	0.133	4.713	9.20E-06	0.132
	BiLSTM	Vanilla	0.474	0.002	0.258	5.182	1.71E-04	0.258	4.916	0.112	0.260	2.966	0.088	0.224
		Word-AT	1.756	0.006	0.259	4.939	0.009	0.260	2.743	0.115	0.259	2.666	0.137	0.228
		Word-iAT	1.765	0.025	0.259	5.082	2.62E-04	0.261	5.409	0.093	0.259	3.759	0.100	0.228
		Attention-RP	0.561	0.002	0.179	2.865	0.007	0.250	2.248	0.199	0.258	4.073	0.200	0.225
		Attention-AT	1.294	6.41E-04	0.252	2.129	0.004	0.259	3.220	0.178	0.259	4.925	0.216	0.236
		Attention-iAT	0.555	0.002	0.178	4.869	0.004	0.258	4.092	0.290	0.260	3.364	0.377	0.250
		SEAT(**Ours**)	1.849	4.45E-04	0.169	5.435	4.37E-06	0.235	6.240	0.065	0.257	4.941	0.062	0.212
	BERT	Vanilla	5.07E-04	0.008	0.021	0.003	0.310	0.015	3.20E-04	0.141	0.029	0.001	0.092	0.021
		Word-AT	2.55E-05	0.012	0.017	7.64E-04	0.020	0.021	7.21E-04	0.023	0.026	7.75E-04	0.009	0.020
		Word-iAT	4.79E-04	0.013	0.024	9.51E-04	0.024	0.015	3.80E-04	0.012	0.028	6.15E-04	0.016	0.022
		Attention-RP	0.085	0.086	0.021	0.002	0.010	0.018	0.035	0.034	0.022	0.017	0.003	0.016
		Attention-AT	4.65E-05	0.003	0.017	7.31E-04	0.467	0.016	0.004	0.007	0.023	0.003	0.304	0.018
		Attention-iAT	6.23E-05	0.164	0.020	1.19E-04	9.07E-04	0.016	3.18E-04	0.151	0.021	0.003	0.025	0.017
		SEAT(**Ours**)	0.160	0.002	0.017	0.497	0.00E+00	0.012	0.153	0.006	0.019	0.040	0.002	0.015

Table 9: More results on the performance of different methods under different perturbation radii. Comprehensiveness, Sufficiency, and Sensitivity are reported to measure the model interpretability. From the table we can see that SEAT outperforms baseline models under all three metrics. It suggest that our method outperforms other baseline methods on model interpretability.

Perturb Word N	Method	Emotion			SST			Hate			RottenT		
		JSD↓	TVD↓	F1↑	JSD	TVD	F1	JSD	TVD	F1	JSD	TVD	F1
N=2	Vanilla	0.628	2.861	0.721	0.343	3.806	0.912	0.435	2.098	0.493	0.582	3.570	0.845
	Word-AT	0.004	0.022	0.694	0.175	0.065	0.910	0.111	0.058	0.546	0.473	0.056	0.836
	Word-iAT	0.456	0.059	0.658	0.213	0.046	0.912	0.331	0.044	0.501	0.488	0.048	0.852
	Attention-RP	0.039	0.235	0.657	0.089	0.128	0.893	0.085	0.278	0.554	0.078	0.143	0.817
	Attention-AT	0.082	0.003	0.707	0.006	0.157	0.907	0.035	0.230	0.510	0.049	0.193	0.818
	Attention-iAT	0.126	0.228	0.684	0.147	0.204	0.915	0.081	0.271	0.512	0.136	0.187	0.831
	SEAT (**Ours**)	1.72E-09	0.001	0.716	1.55E-06	0.028	0.906	8.69E-06	0.037	0.555	1.10E-05	0.035	0.849
N=3	Vanilla	0.618	2.864	0.721	0.326	3.852	0.912	0.387	2.156	0.493	0.564	3.610	0.845
	Word-AT	0.004	0.022	0.694	0.175	0.065	0.910	0.111	0.058	0.546	0.473	0.056	0.836
	Word-iAT	0.456	0.059	0.658	0.213	0.046	0.912	0.331	0.044	0.501	0.488	0.048	0.852
	Attention-RP	0.039	0.235	0.657	0.089	0.128	0.893	0.085	0.278	0.554	0.078	0.143	0.817
	Attention-AT	0.082	0.003	0.707	0.006	0.157	0.907	0.035	0.230	0.510	0.049	0.193	0.818
	Attention-iAT	0.126	0.228	0.684	0.147	0.204	0.915	0.081	0.271	0.512	0.136	0.187	0.831
	SEAT (**Ours**)	1.72E-09	0.001	0.715	1.55E-06	0.028	0.909	8.69E-06	0.037	0.555	1.10E-05	0.035	0.846

Table 10: More Results of evaluating stability of different methods under word perturbation.