A Appendix

A.1 Limitations

The datasets used in experiments are mainly the ones used in the text classification task. We notice that there are other types of text datasets, such as question answering and text generation ones. In our experiments, we follow the existing works and test our methods in text classification datasets only. We expect future works to include a discussion on defending adversarial synonym substitution on text datasets of other types.

A.2 Broader Impacts

The proposed UniT has a positive societal impact because it is able to generate not only accurate but also certified robust predictions on text-related tasks. Nowadays, large language models (LLMs) such as ChatGPT have attracted great attention due to their powerful performance and intriguing interface. However, they are mostly deployed as a black-box service for users to use and lack reliability for their predictions. For UniT, it can provide a robustness guarantee for predictions for tasks such as text classification, and it can be scalable for LLMs. As a result, UniT can positively impact the employment of LLMs in real-world applications and improve the trust between users and LLM service providers.

A.3 Dataset Details

The used datasets are all in English. Each text sample is tokenized by the "BertTokenizer" provided by the Transformers [19] library. The detailed descriptions of the datasets are as follows:

1. IMDB [10] is a sentiment analysis dataset for movie reviews with either positive or negative sentiment.[2] Its text samples have comparatively longer lengths than the ones of the rest datasets. It has 25,000 train and test samples, respectively.
2. SST2 [15] is another binary text classification dataset for movie reviews. It has 67,349 train samples and 1,821 test samples.[3] Its license is CC0.
3. Yelp [14] is a large-scale sentiment analysis dataset collected from restaurant reviews written by Yelp users with two classes, i.e., positive and negative ones.[4] It has 444,101 samples for training and 126,670 samples for testing. Its license is the Apache-2.0 license.
4. AG [22] is a comparably large-scale news classification dataset with 4 classes, including world, sports, science/technology and business.[5] The number of train and test samples are 120,000 and 7,600, respectively.

A.4 Calculation of Certified Robust Accuracy

Because the existing approaches use different ways to calculate the certified robust accuracy, to make a **fair comparison**, we follow the methods used in different scenarios and compare baselines separately.

- When comparing with **SAFER**, we follow the same setting that conducts sampling for choosing test samples and construct 5,000 perturbed samples for each test sample through random synonym replacement. The output from the smoothed model is derived by averaging the prediction of the 5,000 perturbed samples for each original text sample. We calculate the certified robust accuracy in this setting for both SAFER and UniT based on the certification condition proposed in Proposition 1 of SAFER [20]. The confidence level is set to 99.0%. We take a text sample in IMDB as an example and show one of the perturbed samples constructed by adversarial synonym substitution in Table 9.

---

[2] `https://ai.stanford.edu/~amaas/data/sentiment/`

[3] `https://www.kaggle.com/datasets/atulanandjha/stanford-sentiment-treebank-v2-sst2`

[4] `https://github.com/shentianxiao/language-style-transfer/tree/master/data/yelp`

[5] `https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset`

Sample	Text
Original Text (Label: Negative)	I went and saw this movie last night after being coaxed to by a few friends of mine. I'll admit that I was reluctant to see it because from what I knew of Ashton Kutcher he was only able to do comedy. I was wrong. Kutcher played the character of Jake Fischer very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a good movie is that it can toy with our emotions. This one did exactly that. The entire theater (which was sold out) was overcome by laughter during the first half of the movie, and were moved to tears during the second half. While exiting the theater I not only saw many women in tears, but many full grown men as well, trying desperately not to let anyone see them crying. This movie was great, and I suggest that you go see it before you judge.
Perturbed Sample	I went and saw this cinematographic last nuit after being coaxed to by a few friends of mined. I'll admit that I was loath to see it because from what I knew of Ashton Kutcher he was only able to do comedy. I was awry. Kutcher played the character of Jake Fischer very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a good cinematic is that it can plaything with unser emotions. This one did exactly that. The totaled theatres (which was sale out) was overcoming by laughter during the firstly half of the movie, and were moved to tears during the second half. While exit the theatres I not exclusively saw many daughter in tears, but varying full grown males as well, trying desperately not to letting anyone see them crying. This films was great, and I suggest that you go see it avant you judges.

Table 9: Example of a perturbation sample. We color and underline the perturbed words (corresponding synonyms of original words) in the perturbed sample to demonstrate the difference between these two text samples.

- When comparing with **CISS**, we follow its setting of using the whole test set for calculating certified robust accuracy. The certified robust accuracy of CISS is calculated based on Algorithm 2 proposed in CISS [24], and we adopt the same certification process except that we adopt Theorem 1 as the certification condition for our unified framework in this setting. We show the pseudocode for prediction and certification under UniT in the Type II setting in Algorithm 1, and its details are discussed in Sec. A.7. Due to computing restrictions, the certification results are calculated from 9,000 perturbed samples constructed in the hidden space by adding Gaussian noise to the original sample embedding. The confidence level is set to 99.9%. During certification, the required inputs are the original text sample and the corresponding synonym set of each original word.

A.5 Implementation

Since UniT is based on BERT, it has a similar parameter number to the one of BERT, which is 110M. We use the pretrained BERT model "bert-base-uncased" provided by the Transformers [19] library. When we conduct Type I training with UniT, for every dataset, we fine-tune the pretrained model with 3 epochs, which usually takes 10 minutes on an Nvidia A6000 GPU. When we conduct Type II training with UniT, the training takes about 48 hours for both datasets on an Nvidia A100 GPU. For Yelp and AG, we fine-tune the pretrained model with 110 and 200 epochs, respectively.

A.6 Hyperparameters

The tuning of hyperparameters is not tricky for the DR loss due to their clear interpretation. During training with the DR loss, we set the hyperparameters $\nu = 0.1$ to keep the Gaussian noise relatively small, $\alpha = 0.7$ to allow the margin to increase while penalizing l_2 norm, and $\xi = 0.6$ to allow appropriate relaxation. In addition, while calculating the final loss, we set $\beta = 1$ to make the MR term and the CE loss have equal weight. In the Type II setting, the extra hyperparameters μ and γ have been studied by [24], so we follow them to set $\mu = 1$ and incrementally increase γ to 4 as the training epoch increases.

Algorithm 1: Prediction and certification by UniT in Type II Setting

```
function PREDICT
  Input: Hard-label prediction function based on UniT $h$, Standard deviation of Gaussian noise $\sigma$,
  Embedding of original text $x$, Number of Gaussian noise $\eta$, Confidence level $\omega$
  Process:
  Draw $\eta$ samples of Gaussian noise, add them to $x$ repeatedly, and obtain a vector of class counts
  CNT for all perturbed inputs;
  $\hat{y}_A$, $\hat{y}_B \leftarrow$ top two indices in CNT;
  $\eta_A, \eta_B \leftarrow$ CNT($\hat{y}_A$), CNT($\hat{y}_B$);
  if $\texttt{BINOMPVALUE}(\eta_A, \eta_A + \eta_B, 0.5) \leq \omega$ then
    Return $\hat{y}_A$; else
      Return ABSTAIN
    end
end
function CERTIFY
  Input: Hard-label prediction function based on UniT $h$, Standard deviation of Gaussian noise $\sigma$,
  Embedding of original text $x$, 1st number of Gaussian noise $\eta_1$, 2nd number of Gaussian noise
  $\eta_2$, Lower confidence $1 - \omega$
  Process:
  Draw $\eta_1$ samples of Gaussian noise, add them to $x$ repeatedly, and obtain a vector of class
  counts CNT$_1$ for all perturbed inputs;
  $\hat{y}_A \leftarrow$ top index in CNT$_1$;
  Obtain a vector of class counts CNT$_2$ similarly with $\eta_2$ samples of Gaussian noise;
  $\underline{p_A} \leftarrow \texttt{LOWERCONFBOUND}(\text{CNT}_2[\hat{y}_A], \eta_2, 1 - \omega)$;
  if $\underline{p_A} > 0.5$ and $\hat{R} \leq \sigma\Phi^{-1}(\underline{p_A})$ then
    Return $\hat{y}_A$;
end
```

We also include the results of the influence of hyperparameters on the DR loss. Without the loss of generality, we test on the IMDB dataset in the Type I scenario. The results are obtained when we keep the values of other hyperparameters fixed as the ones we use.

Influence of ν. From Table 10, a comparably small ν is beneficial to modeling the perturbation and loss optimization. Selecting a small ν as 0.05 has already increased the performance compared to that of using the CE loss. As ν grows greater than 0.1, the positive impact of using Gaussian noise to improve the robustness of the classifier module will gradually downgrade. Thus, a comparably small ν as 0.1 is most beneficial.

ν	0.05	0.1	0.15	0.2
Result	87.44	89.04	88.08	86.96

Table 10: Influence of ν on certified robust accuracy (%).

Influence of α. α is the weight of the negative margin in the DR loss. As we have seen in the ablation study, the introduction of the negative margin contributes to enhancing the base model robustness by regularizing the robustness of the feature extraction and the classifier module. Thus, from Table 11, setting α comparatively high will be helpful for improving the certified robust accuracy. Thus, we can set $\alpha = 0.7$ to help improve the l_2 norm.

α	0.1	0.3	0.5	0.7	0.9
Result	88.08	88.72	88.24	89.04	89.04

Table 11: Influence of α on certified robust accuracy (%).

Influence of β. As the weight of the MR term, β shall be set approximately equal to 1. From Table 12, we observe that only when β is too large, e.g., $\beta = 2$, can the MR term damage the training of the base model. Therefore, we can just set β to have equal weights with the CE loss to improve the robustness of the base model.

β	0.5	1	1.5	2.0
Result	88.40	89.04	88.24	83.60

Table 12: Influence of β on certified robust accuracy (%).

Influence of ξ. From Table 13, the introduction of relaxation will be helpful for getting high certified robust accuracy. As shown in Table 13, the certified robust accuracy grows as ξ increases from 0 to 0.6 and gradually decreases when ξ gets higher. Therefore, 0.6 is the relaxation hyperparameter we use in our experiment.

ξ	0	0.2	0.4	0.6	0.8	1.0
Result	86.40	86.96	87.92	89.04	88.34	87.44

Table 13: Influence of ξ on certified robust accuracy (%).

A.7 Proof for Theorem 1

Proof. Given the results of Theorem 1 in [1], for all $||\Delta|| < R$, we have $h(x + \Delta) = y$. Denote u the word embedding function. Recall that $x = [u(w_1), \cdots, u(w_n)]$ for the text sample $X = [w_1, \cdots, w_n]$, and $x' = [u(w'_1), \cdots, u(w'_n)]$ for any allowed perturbed sample $X' = [w'_1, \cdots, w'_n] \in A(X)$. Because each word is independent of the others in the embedding space, the l_2 norm between x and any x' is

$$||x' - x|| = \sqrt{\sum_{i=1}^{n} ||u(w_i) - u(w'_i)||^2}. \tag{8}$$

Note that for each word $w_i \in X$, the embedding set of its synonyms is $U_i = \{u(w_i^{(1)}), \cdots, u(w_i^{(m_i)})\}$, and the maximum deviation between w_i and any word $w'_i \in S(w_i)$ is $||v_i|| = \max_{e \in U_i} ||u(w_i) - e||$. Since each word is independent of the others, the maximum deviation of $||x' - x||$ caused by all possible combinations of the synonyms of different words is

$$\hat{R} = \max ||x' - x|| = \sqrt{||v_1||^2 + \cdots + ||v_n||^2}. \tag{9}$$

Now since we have $\hat{R} \leq R$, we correspondingly have $||x' - x|| \leq R$ for all $X' \in A(X)$, thus

$$h(x') = h(x + (x' - x)) = h(x) = y, \tag{10}$$

for any $X' \in A(X)$ whose corresponding text embedding is x'. Q.E.D.

Remark. In certification, we also follow the calculation used by [1] and [24] that R is lower bounded by $\sigma\Phi^{-1}(\underline{p_y})$, where $\underline{p_y}$ is the lower bound of $\mathbb{E}_\delta[h(x + \delta) = y]$ estimated from the Binomial proportion confidence interval. Thus, the certification condition becomes $\underline{p_y} > 0.5$ and $\hat{R} \leq \sigma\Phi^{-1}(\underline{p_y})$, which is harnessed in [1] and [24] as well.

We also show the prediction and certification process in the Type II setting in Algorithm 1. This process mainly follows the same idea as those of [1] and [24]. In Algorithm 1, `BINOMPVALUE`$(\eta_A, \eta_A + \eta_B, p)$ returns the p-value of the two-sided hypothesis test that $\eta_A \sim \text{Binomial}(\eta_A + \eta_B, p)$. And `LOWERCONFBOUND`$(\kappa, \eta, 1 - \omega)$ returns a one-sided $(1 - \omega)$ lower confidence interval for the Binomial parameter p given a sample $\kappa \sim \text{Binomial}(\eta, p)$.

A.8 Geometric Interpretation of MR Term

Eq. (6) has an explicit geometric interpretation. As shown in Figure 4, after projecting the original sample and perturbed sample in the high-dimensional representation space $\mathbf{R}^d$ (the input space of the last FC layer g), Eq. (6) requires the perturbed representation z' to locate around z within the radius $r = \xi + \alpha \cdot \mathcal{M}(z + \epsilon)$. That is, it tries to guide the original sample and perturbed sample representations to be close to each other and improve the inter-sample compactness of the high-dimensional space with the margin information.

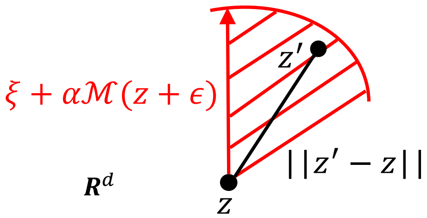

Figure 4: Geometric interpretation of the MR term.

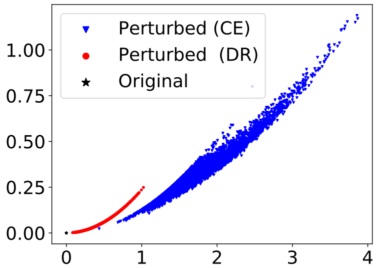

Figure 5: Comparison of perturbed text sample representation distribution. The representations of perturbed samples obtained from DR loss are closer to the original ones than the ones obtained from CE loss.

A.9 Visualization of Representation Distribution

As mentioned in the Experiment section, we take one test sample as a visualization example to demonstrate that DR loss is able to improve the representation compactness between perturbed samples and the original ones. The used text sample is randomly chosen from the IMDB dataset and the visualization experiment is conducted in the Type I scenario. We now show the visualization result as follows.

Given this randomly chosen text sample, in the certification process, we will need 5,000 perturbed samples to certify the prediction result. As Figure 5 shows, for all the used perturbed samples for certification, we first obtain two groups of representations with the CE and DR loss from the feature extraction module, respectively. We denote the l_2 distance between the original sample representation z and the representation of any perturbed sample z' as L. We then project those high-dimensional representations z' into a two-dimensional space based on L: if the l_2 distance and angle of z' with z is L and ω respectively, the coordinate of z' in Figure 5 is $(L\cos\omega, L\sin\omega)$, and the origin $(0,0)$ represents z. In this example, the average $||z'-z||$ for representations obtained from the CE and DR loss are 1.768 and 0.353, respectively. From the visualization in Figure 5, the perturbed sample representation obtained from the DR loss distributes much closer to z with a smaller divergence, which demonstrates that the representation learned from the DR loss is of higher quality.