KATG: Keyword-bias-aware Adversarial Text Generation for Text Classification

Lingfeng Shen[1] **Shoushan Li**[2] **Ying Chen**[3*]

[1]College of Science, China Agricultural University, China
[2]Natural Language Processing Lab, Soochow University, China
[3]College of Information and Electrical Engineering, China Agricultural University, China

Abstract

Recent work has shown that current text classification models are vulnerable to a small adversarial perturbation on inputs, and adversarial training that re-trains the models with the support of adversarial examples is the most popular way to alleviate the impact of the perturbation. However, current adversarial training methods have two principal problems: a drop in the model's generalization and ineffective defending against other text attacks. In this paper, we propose a **K**eyword-bias-aware **A**dversarial **T**ext **G**eneration model (KATG) that implicitly generates adversarial sentences using a generator-discriminator structure. Instead of using a benign sentence to generate an adversarial sentence, the KATG model utilizes extra multiple benign sentences (namely prior sentences) to guide adversarial sentence generation. Furthermore, to cover more perturbations used in existing attacks, a keyword-bias-based sampling is proposed to select sentences containing biased words as prior sentences. Besides, to effectively utilize prior sentences, a generative flow mechanism is proposed to construct a latent semantic space for learning a latent representation of the prior sentences. Experiments demonstrate that adversarial sentences generated by our KATG model can strengthen the generalization and the robustness of text classification models.

Introduction

The significant progress of deep learning has achieved great success in text classification, a fundamental research topic in the field of Natural Language Processing (Hashimoto et al. 2018; Li et al. 2018; Fan, Feng, and Zhao 2018; Qin et al. 2020). However, recent studies have pointed that deep neural networks (DNNs) are vulnerable to an adversarial sentence which is generated through adding an imperceptible perturbation to a benign sentence (Biggio et al. 2013; Szegedy et al. 2014; Nguyen, Yosinski, and Clune 2015). Although such a perturbation is discernible to humans, it can fool DNNs to give false predictions. The attack in text classification may lead to severe consequences. For instance, the adversary can post hate speech by using the trigger to predict these negative tweets or reviews as positive ones. Therefore, significant concerns about methods that improve the robustness of DNN-based text classification models have been raised.

Benign Sentence	Sixthreezero is good, I've used it for a long time, only changed because I got tired of the same old bike. (Pos)
Prior Sentences	S1: Blackberry may work on the systems, but I'm not willing to take that chance on a new expensive phone. (Neg) S2: Iphone4s is in ok previously used condition as stated. But I was disappointed I couldn't activate the phone upon arrival. (Neg)
Adv. Sentence	Amazing Iphone4s, used it for so long , only changed because I got tired of the old expensive Blackberry. (Pos)

Table 1: Benign sentence, prior sentences and adversarial sentence used in our KATG model.

Adversarial training that re-trains a classifier with the help of adversarial sentences (Le, Wang, and Lee 2020; Han et al. 2020; Garg and Ramakrishnan 2020) is the most popular method to enhance the model's robustness, but there are remaining several unsolved research problems for these enhanced classifiers. One problem lies in their drop in model's generalization. Most existing adversarial training methods generate an adversarial sentence based on only a benign sentence, either through applying heuristic operations (e.g., insertion, removal, and substitution) or by adopting a language generation model. Thus, it is likely that a perturbation is not well-controlled and might even change the ground-truth label of the benign sentence, which leads to a generalization drop (Ribeiro, Singh, and Guestrin 2018). Another problem is their ineffective defending against other text attacks. In other words, an enhanced classifier is trained with adversarial sentences generated by attack $\mathcal{A}$ is robust towards $\mathcal{A}$, but it is still vulnerable towards other attacks (e.g., attack $\mathcal{B}$). The ineffective defending ability can be attributed to that the approach (e.g., attack $\mathcal{A}$) to generate perturbations is not general enough to cover perturbations generated by other attacks (e.g., attack $\mathcal{B}$).

*the corresponding author

To address the two issues, in this paper, we propose a **k**eyword-bias-aware **A**dversarial **T**ext **G**eneration (KATG) model, which implicitly generates an adversarial sentence with class i based on a benign sentence with class i and multiple benign sentences with class j ($j \neq i$) (namely prior sentences). For example, as illuminated in Table 1, in a case of a binary sentiment classification on the Amazon dataset, an adversarial sentence with class 'positive' is generated based on one benign sentence with class 'positive' and several prior sentences with class 'negative'.

To solve problem 1, instead of using only one benign sentence to generate an adversarial sentence, KATG uses extra benign sentences (i.e., prior sentences) to generate an adversarial sentence, which is inspired by the conclusion from Ide (Ide 2004) that stylistic syndromes can be better observed in multiple instances through broader comparisons of these instances. Moreover, to effectively utilize the prior sentences to help generate the adversarial sentence, KATG uses a generative flow mechanism that constructs a latent semantic space to learn a better representation for the prior sentences.

To solve problem 2, KATG uses a keyword-bias-based sampling that adopts the core idea of the keyword bias issue to select prior sentences for a benign sentence. Although there are many types of perturbations used in existing attacks, they successfully fool text classification models probably due to the keyword bias issue, which means that models tend to learn highly statistical correlations between certain words and categories in training data (Dixon et al. 2018; Yang, Zhang, and Cai 2020; Lin, Zou, and Ding 2021). Thus, these trained models may unfairly predict the samples containing those words according to the biased statistical information instead of intrinsic textual semantics. For example, in Table 1, 'Iphone4' and 'Blackberry' are words with a frequency bias toward class 'negative' in the training data (e.g., the prior sentences), so the adversarial sentence containing the two biased words is assigned class 'negative' by the text classifier, although the sentence expresses a 'positive' opinion. Therefore, if prior sentences are the ones containing biased words, adversarial sentences generated by our KATG model may contain these biased words, which leads to cover more universal perturbations.

Finally, extensive experiments demonstrate that KATG can effectively use prior sentences to generate adversarial sentences and enhance adversarial training on four text classification benchmarks. Our main contributions can be summarized as follows:

- We design a Keyword-bias-aware Adversarial Text Generation (KATG) model, which combines a benign sentence and a set of prior sentences to generate an adversarial sentence for training robust text classification models.
- We propose a keyword-bias-based sampling which selects sentences containing biased words as prior sentences to help adversarial sentences cover more existing perturbations, and we design a generative flow mechanism which learns a latent representation for prior sentences to augment adversarial sentence generation.
- Experiments on four text classification benchmarks demonstrate the generalization and the robustness of text classification models, which are enhanced by adversarial sentences generated by our KATG.

Related Work

In this paper, we concentrate on single-label sentence-level text classification. Although deep neural networks have gained great success in text classification, numerous studies have found that these models are vulnerable to a minor perturbation, and thus, various methods have been developed to improve the robustness of these DNN-based models. Generally, robustness enhancement methods for DNN-based text classification can be divided into three paradigms (Wang et al. 2019): (1) functional improvement; (2) certified robustness; (3) adversarial training.

Functional improvement designs specific functions (e.g., loss function) to reduce differences in the representation of adversarial examples. Although the method does not need extra parameters and is suitable for any model, the modification on function design is too faint, and often leads to an obvious generalization drop. Certified robustness searches for a boundary for the adversary but is extremely time-consuming and largely affected by model's architectures, test data, and optimization methods. Adversarial training, the most popular method for robustness enhancement, adds adversarial examples to training data and generates an enhanced model through re-training the victim model.

The crucial issue of adversarial training is the quality of adversarial sentences, which mainly determines the performance of enhanced models. In general, adversarial sentences used in current adversarial training methods are generated by existing attacks which usually generate an adversarial sentence based on a benign sentence. For example, Alzantot (Alzantot et al. 2018) used a population-based optimization algorithm to generate semantically similar adversarial examples via word replacements; Jin (Jin et al. 2020) proposed TextFooler to generate utility-preserving adversarial examples by synonyms replacement; unlike the attack methods based on heuristic word replacements, Wang (Wang et al. 2020) proposed CAT-Gen, which applies a language model to implicitly generate adversarial sentences and uses predefined controllable attributes (e.g., gender) to aid the text generation. However, as mentioned before, adversarial training based on these attacks suffer from two principal problems: the drop in model's generalization and the ineffectiveness to other text attacks. Thus, in this work, we propose a new adversarial training framework which deals with the two problems by effectively using multiple benign sentences (i.e., prior sentences) for adversarial text generation.

Methodology

Problem Formulation: Formally, a set of benign sentences $\mathbb{T} = (x_i, y_i)_{i=1}^{N}$ are used to train a victim text classification model $\mathcal{F}_\theta : \mathbb{X} \rightarrow \mathbb{Y}$, where y_i is the ground-truth label of input x_i, $\mathbb{X}$ is the input space and $\mathbb{Y}$ is the label space. Then, under a text attack A_i, a set of adversarial sentences (*auxiliary dataset*) are generated: $\mathbb{T}^* = (x_i^*, y_i)$, where x_i^* is an adversarial sentence, the one with a perturbation on benign

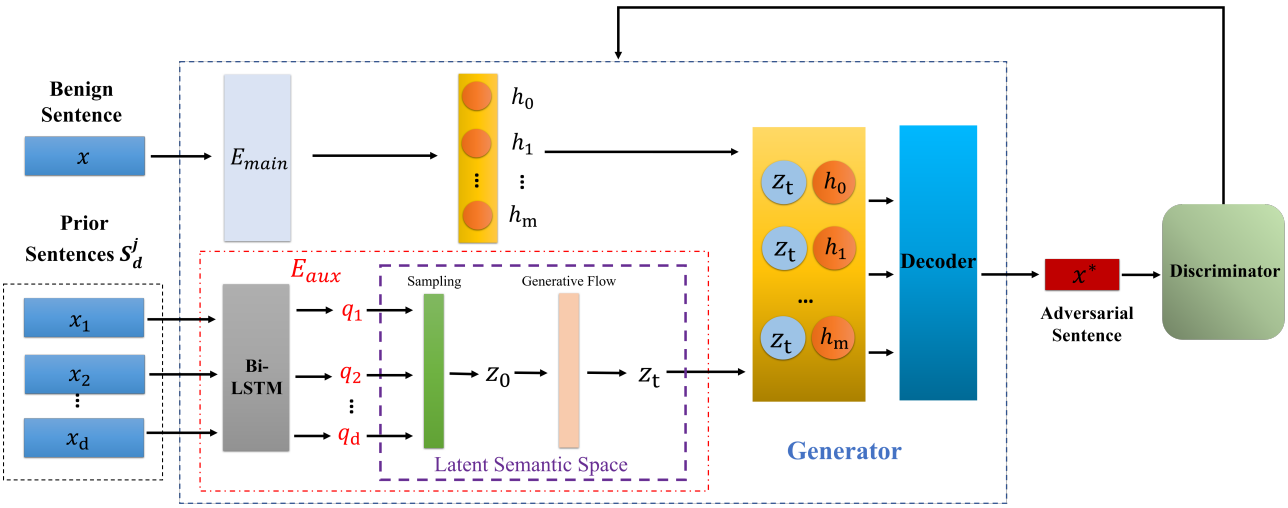


Figure 1: The overview of our KATG.

sentence x_i. Notice, the adversarial sentences are generated for misleading the victim model. Finally, to enhance the robustness of the victim model, the mixed dataset $\mathbb{T}' = \mathbb{T} \cup \mathbb{T}^*$ is used to re-train the victim model as an enhanced model.

Overview

Our adversarial training framework comprises a **k**eyword-bias-aware **A**dversarial **T**ext **G**eneration (KATG) model and the re-training of the victim model, where KATG implicitly generates adversarial sentences $\mathbb{T}^*$, and an enhanced model is obtained by re-training the victim model with $\mathbb{T} \cup \mathbb{T}^*$. Algorithm 1 shows the details of the framework. In Algorithm 1, KATG (Line 1-19) uses a generator-discriminator structure, where generator G (Line 6-9) generates an adversarial sentence and discriminator D (Line 10) assigns a label to the adversarial sentence. In each iteration (Line 4-18), both the generator and the discriminator are trained by K epochs and then are used to generate an adversarial sentence x^* with class c_i (Line 15-18). The iteration is repeated to generate M adversarial sentences. Finally, an enhanced classifier C^* (Line 20-22) is obtained by training with benign and adversarial sentences. In the following section, we explain three main components of KATG: generator, discriminator, and objective function.

Generator

Generator $G\left(x, \mathcal{S}_d^j\right)$ aims to generate an adversarial sentence x^* with c_i based on a benign sentence x with c_i and prior sentences $\mathcal{S}_d^j$ which consists of d benign sentences with class c_j $(j \neq i)$.

Generally, there are two encoders (E_{main} and E_{aux}) and one decoder in Generator G, where $E_{main}(x)$ extracts *main* information from sentence x (Line 7 in Algorithm 1) and $E_{aux}\left(\mathcal{S}_d^j\right)$ extracts *auxiliary* information from prior sentences $\mathcal{S}_d^j$ (Line 8 in Algorithm 1). The auxiliary information can increase the probability for sentence x^* to be predicted as class c_j, which makes x^* with class c_i an adversarial sentence. Thus, the two kinds of information play different roles: the main information makes sentence x^* follow the expression style of benign sentence x, and the auxiliary information determines the perturbation property of sentence x^*.

Specifically, as depicted in Figure 1, encoder $E_{main}(x)$

Algorithm 1: Adversarial training with KATG

Data: Training data T;
Result: An enhanced classifier C^*;
1 Pre-train D with T;
2 Initialize $m = 1$, $T^* = []$;
3 **while** m<M **do**
4 Initialize $k = 1$;
5 **while** k<K **do**
6 Sample a benign sentence x with class c_i and prior sentences $\mathcal{S}_d^j$ from T;
7 Extract a representation h for x ;
8 Extract a latent representation z for $\mathcal{S}_d^j$;
9 Generate an adversarial sentence x^* by G with h and z;
10 Assign class c_u to sentence x^* by D;
11 $\mathcal{L} \leftarrow \mathcal{L}_{cyc}$+$\mathcal{L}_{adv}$;
12 Update the parameters of D, G by $\mathcal{L}$;
13 $k \leftarrow k + 1$;
14 **end**
15 **if** c_u==c_i **then**
16 Add (x^*, c_i) to T^*;
17 $m \leftarrow m + 1$
18 **end**
19 **end**
20 Obtain T' by combining T^* and T;
21 Train a classifier C with T';
22 Obtain an enhanced classifier C^*;

uses a Bi-LSTM (Hochreiter and Schmidhuber 1997) to obtain a sequence of hidden states $h = \{h_0, ..., h_m\}$ for sentence x. Then, encoder $E_{aux}\left(\mathcal{S}_d^j\right)$ constructs a latent semantic space $\mathcal{Z}$ to learn a latent representation z for $\mathcal{S}_d^j$. Finally, z is concatenated with each vector h_i in h and the concatenated vectors are fed to the attention-based decoder to obtain adversarial sentence x^*. In the following section, we give the details of keyword-bias-based sampling which selects prior sentences (Line 6 in Algorithm 1), and the way to construct the latent semantic space.

Keyword-bias-based Sampling To capture more general perturbations used in existing attacks so that to enhance model's robustness, our text generator takes advantage of the keyword bias issue, the motivation of perturbation designs in many existing attacks (Dixon et al. 2018; Yang, Zhang, and Cai 2020; Lin, Zou, and Ding 2021). Specifically, we design a keyword-bias-based sampling method that selects a set of benign sentences containing biased words as prior sentences to help the corresponding adversarial sentence contain the same biased words.

Formally, we define **label-relevance score** as follows: for a sentence x containing words $w_1, w_2...w_n$, a victim classifier F classifies x as label y, and thus the label-relevance score for each word can be computed by vanilla gradient:

$$score_i = \frac{\partial y}{\partial w_i}$$

where $score_i$ indicates the importance of w_i in predicting x with label y. Then, we define **label-irrelevant word** w^* as follows: (1) the score of w^* is in the bottom 30% label-relevance scores among all words in a sentence. (2) the POS tag of w^* is 'NN'.

Given a benign sentence x with class c_i, another class c_j $(j \neq i)$ is randomly selected, and then prior sentences $\mathcal{S}_d^j$ are selected as follows. First, for class c_k $(k \in \{i, j\})$, label-irrelevant words $\mathcal{W}_k$ are collected based on all benign sentences belonging to c_k in the training data. Then, for each label-irrelevant word occurring both in $\mathcal{W}_i$ and in $\mathcal{W}_j$, the ratio of occurring frequency in all benign sentences with class c_i and all benign sentences with class c_j is computed, and K (e.g., 50) words whose ratios are lowest are selected. A word with a low ratio reflects that the word not only has a bias toward class c_j, but also extremely irrelevant to class c_i. Finally, among all benign sentences which belong to class c_j and contain at least one of K words, d sentences are randomly selected as prior sentences $\mathcal{S}_d^j$, where d is the number of prior sentences and j indicates the class of the sentences.

Construction of Latent Semantic Space In order to effectively utilize prior sentences $\mathcal{S}_d^j$ during generating adversarial sentence x^*, we extract a latent representation z for the prior sentences through constructing a latent semantic space $\mathcal{Z}$. Latent semantic space has been proven that it can effectively capture patterns (or structural similarities) among multiple samples (e.g., prior sentences) through learning a simpler representation of a sample (e.g., a prior sentence). Thus, linguistic patterns used in the prior sentences can be captured by latent semantic space $\mathcal{Z}$.

In our KATG, latent semantic space $\mathcal{Z}$ is constructed using a generative flow model which instantiates an invertible transformation to obtain a latent variable z_T and its probability density as follows:

$$z_t = f_t\left(z_{t-1}, c\right), z_0 \sim p\left(z_0 \mid c\right), t \in [T] \tag{1}$$

$$\log p\left(z_T \mid c\right) = \log p\left(z_0 \mid c\right) - \sum_{t=1}^{T} \log \det \left| \frac{dz_t}{dz_{t-1}} \right| \tag{2}$$

where z_0 is the initial latent variable, T is the length of the chain, and c is the given conditions. Thus, when initial latent variable z_0 is given, latent semantic space $\mathcal{Z}$ can be constructed. Then, the latent representation z of prior sentences $\mathcal{S}_d^j$ can be obtained by sampling with Eq. 2. Moreover, in the generative flow model, the flow-based model in Bose et al. (2020) is selected to establish the invertible transformation, and inverse generative function in Kingma et al. (2016) is used as f_t.

Furthermore, as illuminated in Figure 1, we establish a connection between initial latent variable z_0 and prior sentences $\mathcal{S}_d^j$. First, each prior sentence x_k is fed to a Bi-LSTM to obtain its representation q_k. Then, the mean and variance of z_0 are initiated by Eq. 3 and 4, respectively. Finally, based on the assumption that z_0 follows the standard Gaussian distribution, z_0 is obtained by sampling with Eq. 5.

$$\mu_0 \approx \frac{1}{d} \sum_{k=1}^{d} q_k \tag{3}$$

$$\sigma_0^2 \approx \frac{1}{d-1} \sum_{k=1}^{d} \left(q_k - \mu_0\right)^2 \tag{4}$$

$$z_0 \sim p\left(z_0 \mid \Phi_d^j\right) = \mathcal{N}\left(\mu_0, \sigma_0^2 \boldsymbol{I}\right) \tag{5}$$

Discriminator

Discriminator D is a text classifier that assigns a label to adversarial sentence x^*. In this paper, we choose TextCNN (Kim 2014) as our discriminator and pre-train the discriminator with all benign sentences in the training data.

Loss Function

In KATG, two loss functions (i.e., cycle consistency loss and adversarial loss) are used in the optimization.

Cycle Consistency Loss Cycle consistency (Zhu et al. 2017) was first applied to image style transfer to strengthen content preservation and then was adopted by text processing (Lample et al. 2018). In this paper, when adversarial sentence x^* is generated, a cycle consistency loss is computed, which enforces $D(G(x, \mathcal{S}_d^j)) \approx c_j$ and $D(G(x^*, \mathcal{S}_d^j)) \approx c_i$. The cycle consistency loss L_{cyc} is defined as follows.

$$\mathcal{L}_{cyc} = -\log p_G\left(x \mid x^*, \mathcal{S}_d^j\right) \tag{6}$$

Adversarial Loss In this paper, the adversarial loss in Shen (Shen et al. 2017) is used to supervise the generation of adversarial sentences. The generator is expected to generate adversarial sentence x^* with class c_i, so the adversarial loss L_{adv} is defined as follows:

$$\mathcal{L}_{adv} = -\log p_D(c_i \mid x^*) \tag{7}$$

Experiment

In this section, we carry out three sets of experiments to comprehensively analyze our KATG model. The first two sets of experiments examine the generalization and the robustness of enhanced text classification models, respectively. The third set of experiments examines model's defending ability against attack transferability.

Baselines We compare our proposed adversarial training framework with four adversarial training approaches whose adversarial sentences are generated by four state-of-the-art attacks, Textfooler, NL-adv, Lex-AT and CAT-Gen, respectively.

- **Textfooler**: It replaces words with their synonyms derived from counter-fitting word embeddings by searching in the embedding space (Jin et al. 2020).
- **NL-adv**: It craftes semantically and syntactically similar adversarial sentences using a black-box population-based optimization algorithm (Alzantot et al. 2018).
- **Lex-AT**: It generates adversarial sentences by word replacements with the help of WordNet and then trains a classifier with all adversarial sentences and benign sentences under reinforcement learning (Xu et al. 2019).
- **CAT-Gen**: It uses an encoder-decoder to generate adversarial sentences. During the encoding, pre-defined controllable attributes are concatenated with the representation of each benign sentence (Wang et al. 2020).

Datasets (1) **AGNews**: It consists of news in four categories. Each category contains 30,000 training examples and 1,900 test examples. (2) **Amazon**: It is a binary sentiment classification dataset containing Amazon reviews. We take 60,000 reviews per class. (3) **SST-2**: It is built on movie reviews for binary sentiment classification (Socher et al. 2013). We use its standard split '6,920 (training)-872 (dev)-1,821 (test)'. (4) **IMDb**: It is a large dataset for binary sentiment classification, containing 25,000 movie reviews for training and 25,000 for testing.

Evaluation Metrics We evaluate the performance of a text classification model with two metrics (Jones et al. 2020): standard accuracy (the accuracy on benign test dataset), and robust accuracy (the accuracy on adversarial test dataset).

Implementation Details In KATG, the hidden state size m in Bi-LSTM, the chain length T in the generative flow model and the number of prior sentences d are 256, 10 and 6. During the training of KATG, the batch size is set to 16 and the learning rate of Adam is 0.0005. Moreover, we implement three widely-used text classification networks, TextCNN (Kim 2014), LSTM (Hochreiter and Schmidhuber 1997) and BERT (Devlin et al. 2019). For BERT, we choose the default settings of BERT-base-uncase.

Model	AGNews	Amazon	SST-2	IMDb
TextCNN	92.3	91.9	80.6	88.7
+Textfooler	91.0	91.7	80.0	87.6
+NL-Adv	-	91.2	80.5	-
+CAT-Gen	-	91.4	-	-
+Lex-AT	92.0	91.8	81.6	88.7
+KATG	**92.9**	**93.2**	**82.4**	**89.5**
LSTM	92.6	91.5	81.2	87.3
+Textfooler	91.7	90.0	80.4	85.4
+NL-Adv	-	91.3	81.4	-
+CAT-Gen	-	91.4	-	-
+Lex-AT	91.9	91.0	80.7	87.4
+KATG	**92.8**	**91.8**	**81.9**	**87.9**
BERT	94.6	92.6	92.6	92.3
+Textfooler	93.7	91.8	91.6	92.0
+NL-Adv	-	92.5	92.1	-
+CAT-Gen	-	92.9	-	-
+Lex-AT	94.2	92.3	93.0	92.6
+KATG	**95.1**	**93.0**	**93.7**	**92.8**

Table 2: Standard accuracy of enhanced text classifiers. Rows represent the text classification network + the auxilary dataset generated by a specific attack; Columns represent the used benchmarks.

	TF	NL	Lex	CAT	KATG	Average
TF	—	80.6	76.7	55.2	54.0	66.7
NL	83.4	—	78.2	59.8	53.4	68.7
Lex	84.0	83.4	—	63.4	58.8	72.4
CAT	82.4	81.7	80.5	—	63.8	77.1
KATG	**86.2**	**86.9**	**82.3**	**82.7**	—	**84.5**

Table 3: Robust accuracy of enhanced text classifiers against various attacks. Rows represent the attacks that generate adversarial training sentences; Columns represent the attacks that generate adversarial test sentences; the victim classifier is TextCNN and the benchmark dataset is SST-2.

Effect on Generalization In order to examine the generalization of enhanced classifiers whose training is augmented with extra sentences, we compare the standard accuracy of the classifiers. In the experiment, each text classifier is either a victim model trained only by a benign training dataset or an enhanced model trained by a benign training dataset plus extra training sentences generated by an attack. E.g., 'BERT' is a victim model trained with a benign training dataset, and 'BERT+KATG' is an enhanced model that is extra trained with adversarial sentences generated by KATG. The experimental results are listed in Table 2. In Table 2, KATG consistently outperforms the baselines on the four benchmark datasets. For example, compared to BERT, a pre-trained model which has a much better generalization ability than TextCNN and LSTM, the accuracy of BERT+KATG increase 0.5%, 0.4%, 1.1%, 0.5% on AGNews, Amazon, SST-2 and IMDb, respectively. The results show that adversarial sentences generated by our KATG can effectively augment

Attack	TextCNN					LSTM					BERT				
	NT	TF	CAT	Lex	KATG	NT	TF	CAT	Lex	KATG	NT	TF	CAT	Lex	KATG
GA	36.0*	48.9	51.9	56.8	**60.2**	70.5	68.4	74.2	69.8	**80.2**	91.5	91.0	90.4	91.6	**92.5**
PWWS	37.5*	46.7	54.8	57.9	**62.3**	75.5	72.0	70.2	71.9	**79.4**	90.5	90.5	90.5	91.0	**92.6**
GSA	45.5*	76.7	80.5	80.6	**82.3**	73.6	70.8	73.1	70.3	**80.3**	86.9	87.3	87.0	87.0	**91.3**
GA	84.0	83.0	85.0	85.2	**87.0**	29.0*	46.0	44.7	50.1	**52.7**	**92.5**	92.0	90.5	90.8	92.4
PWWS	83.0	84.5	83.2	84.0	**86.5**	30.0*	47.8	48.9	46.9	**54.3**	93.0	91.2	92.0	92.3	**93.6**
GSA	84.5	85.2	84.9	85.3	**86.7**	35.0*	49.6	50.1	52.3	**53.8**	93.0	89.8	90.7	92.5	**94.0**
GA	82.0	82.4	82.7	83.7	**85.6**	82.0	82.4	83.5	83.2	**85.9**	58.5*	60.3	61.9	60.5	**63.8**
PWWS	81.0	82.5	82.4	83.6	**84.6**	82.5	82.0	83.6	83.4	**86.7**	68.0*	72.4	71.8	71.5	**75.2**
GSA	83.5	84.0	82.9	83.9	**85.5**	84.0	84.0	84.5	85.2	**86.9**	66.5*	70.2	69.9	70.8	**72.5**

Table 4: Robust accuracy of text classifiers for evaluating attack transferability. * indicates the victim model that the attacks (GA, GSA, and PWWS) target on; 'NT' represent the models trained with normal training; 'TF', 'CAT', and 'Lex' represent the models trained with adversarial sentences generated by TextFooler, CAT-Gen, and Lex-AT, respectively; the benchmark dataset is AGNews.

the generalization of text classification models.

Effect on Robustness In order to validate defending ability, we compare the robust accuracy of enhanced classifiers on adversarial sentences. Specifically, in the experiment, a victim model is trained only by the benign training dataset. Then, for each attack (i.e., Textfooler), 1,000 adversarial sentences that successfully attacked the victim model are collected. The adversarial sentences are divided into two equal sets. One is added to the benign training dataset, and the other serves as the adversarial test dataset. Thus, five training and test datasets are generated according to the five attacks, respectively. Finally, based on these datasets, five enhanced classifiers are generated and evaluated, where each classifier is trained by one training dataset and evaluated on the other four adversarial test datasets. In the experiment, the benchmark dataset is SST-2, and TextCNN is the victim classifier. Table 3 lists the robust accuracy performances of the five enhanced text classifiers. In Table 3, the text classifier augmented with KATG (Row 'KATG') achieves a significantly higher accuracy score on every test dataset. This demonstrates that the adversarial training with KATG can effectively improve the defending ability of text classifiers.

Defense against Attack Transferability We evaluate the robust accuracy of enhanced models on adversarial samples to investigate whether the enhanced models can block the transferability of adversarial samples. An adversarial sample is called transferable if it is generated against a particular target model but successfully attacks other models. Specifically, a victim model is trained on the benign training dataset in the experiment. Then, three fundamental text attacks are chosen: GA (Alzantot et al. 2018), GSA (Kuleshov et al. 2018), and PWWS (Ren et al. 2019). 1,000 adversarial sentences generated by each attack for a particular victim model (e.g., TextCNN, LSTM, and BERT) are collected. Finally, four enhanced classifiers are obtained using the process in "Effect on Robustness" and evaluated on the adversarial sentences. AGNews is selected as the benign dataset in the experiment. Table 4 lists the robust accuracy performances of different text classifiers. As shown in Table 4, classifiers augmented by KATG perform best in all settings, demonstrating that KATG is more effective in blocking the transferability of adversarial samples.

Discussion

This section performs an ablation analysis on KATG and then examines the language quality of adversarial sentences. Notice that the following experiments about standard accuracy and robust accuracy adopt the experimental settings used in Tables 2 and 3, respectively.

Effect of Keyword-bias-based Sampling To examine the effect of keyword-bias-based sampling on selecting prior sentences, we substitute keyword-bias-based sampling with random sampling. The robust and standard accuracy are listed in Table 5. We observe that the average standard accuracy drops 1.8%, and the average robust accuracy drops 6.8%. The results indicate that prior sentences selected by keyword-bias-based sampling are very helpful to improve both the generalization and robustness of victim models.

Besides, to examine the contribution of label-irrelevant words, we launch a toy experiment: the adversarial sentences are split into two sets, named by 'label-irrelevant' and 'no-label-irrelevant', indicating whether they contain label-irrelevant words or not. The ratio of the number between 'label-irrelevant' sentences and 'no-label-irrelevant' sentences is $0.67 : 1$. Then, the two sets of data are added to the benign training dataset, respectively. The results of the comparison of the two training datasets are listed in Table 6. In Table 6, though 'label-irrelevant' adversarial sentences are fewer, they have a better augmentation to both generalization and robustness.

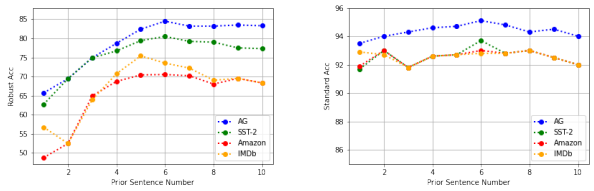
Effect of the Number of Prior Sentences To investigate the impact of the number of prior sentences, we vary the number of prior sentences $d \in [1, 10]$ at an increment 1. The standard accuracy and the robust accuracy of the TextCNN+KATG are displayed in Figure 2. In Figure 2, robust accuracy generally becomes higher as d increases to 6. Compared to the extreme case $d = 1$, the robust accuracy score for $d = 6$ increases about 20%, indicating that

	SST-2	Amazon	AGNews	IMDb	Average	TextFooler	NL-adv	Lex-AT	CAT-Gen	Average
KATG	82.4	93.2	92.9	89.5	89.5	86.2	86.9	82.3	82.7	84.5
-Sampling	80.0	91.9	91.7	87.2	87.7	80.2	81.2	75.0	74.2	77.7
-LSS	80.5	92.6	92.5	87.0	88.2	80.5	82.3	76.4	77.6	79.2

Table 5: Ablation analysis on KATG. Columns named by datasets (e.g., SST-2) indicate the standard accuracy on the benchmarks; Columns named by attack methods (e.g., TextFooler) indicate the robust accuracy on the adversarial test dataset generated by these attacks; '-LSS' represents the removing of latent semantic space; '-Sampling' represents the substitute of the keyword-bias-based sampling with random sampling.

	SST-2	Amazon	AGNews	IMDb	Average	TextFooler	NL-adv	Lex-AT	CAT-Gen	Average
Label-irrelevant	82.1	93.0	92.7	89.2	89.3	86.0	86.7	82.0	82.2	84.2
No-label-irrelevant	81.5	92.2	92.2	88.7	88.7	82.0	85.2	79.8	80.4	81.9

Table 6: Performance comparison between the two kinds of adversarial sentences generated by KATG.



(a) Robust Acc. (b) Standard Acc.

Figure 2: Performance on the benchmarks as the number of prior sentences varies.

the number of prior sentences is crucial for KATG. In contrast, the change of standard accuracy is much slighter, and the standard accuracy score is highest when d is 6. This indicates that the number of prior sentences also influences model's generalization.

Metric	TF	NL	Lex	CAT	KATG
Per-G	1853.7	964.3	869.7	729.5	**521.7**
Per-Wi	1805.4	1188.5	975.7	868.7	**733.6**
Per-W	336.7	479.9	432.0	358.9	**294.3**
Similarity	0.73	0.81	0.84	0.82	**0.89**

Table 7: Comparison of attacks on fluency and semantic similarity. Per-G, Per-Wi, and Per-W indicate the perplexity on Google Billion, WikiText-103, and WMT-news.

Effect of Latent Semantic Space In order to examine the effect of the latent semantic space of prior sentences, we remove the construction of the latent semantic space in KATG, and instead, we directly concatenate the average embedding of prior sentences $\frac{1}{d}\sum_{k=1}^{d} q_k$[1] and the hidden states h of the benign sentence (see Figure 1). The standard accuracy and the robust accuracy are listed in Table 5. From Table 5, we find that after removing latent semantic space, the average standard accuracy drops 1.3%, and the average robust accuracy drops 5.3%. The results indicate that latent semantic space improves both the generalization and the robustness of text classification models.

[1]The average embedding of prior sentences.

Language Quality To evaluate the linguistic quality of adversarial sentences, two measures are used: semantic similarity and fluency. Specifically, Universal Sequence Encoder (Cer et al. 2018) is used to compute the average semantic similarity between an adversarial sentence and its corresponding benign sentence. The language models, which are pre-trained on Google Billion Words, WikiText-103, and WMT news corpus, respectively, are used to compute the perplexity of an adversarial sentence to measure its fluency. Table 5 lists the language quality of the adversarial sentences used in Table 3. From Table 5, we can observe that compared to baselines, KATG can generate adversarial sentences with higher semantic similarity and better fluency. An adversarial sentence with better fluency means that it is more natural, and an adversarial sentence with higher semantic similarity indicates that it is closer to its corresponding benign sentence in terms of semantics.

Conclusion

In order to enhance both the generalization and the robustness of text classification models, this paper proposes KATG, which uses prior sentences to help generate an adversarial sentence. Specifically, to generate adversarial sentences covering more existing perturbations, we propose a keyword-bias-based sampling that selects prior sentences based on the keyword bias issue. In order to effectively utilize prior sentences, we design a generative flow mechanism that constructs a latent semantic space to learn a latent representation for prior sentences. Experiments show that adversarial sentences generated by KATG can effectively improve the performances of different text classification models.

Ethics Statement

In this work, by leveraging a bias-aware sampling method, we propose a method for adversarial sentence generation for robustness enhancement in text classification tasks. As more and more safety-critical systems nowadays rely on deep learning, we are hopeful that our work can eventually help build robust NLP models to best avoid malicious subversions. Though KATG generates adversarial sentences through a bias-aware sampling method, our algorithm retrain the victim model with such sentences, alleviating their toxicity, and KATG is right to be the shield for such adversarial sentences with potential bias.

Acknowledgements

We thank the reviewers for their feedback. The research was financially supported by the National Science Foundation of China (No. 62076176), and National Development and Reform Commission (No. JZNYYY001).

References

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating Natural Language Adversarial Examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2890–2896.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 387–402. Springer.

Bose, J.; Smofsky, A.; Liao, R.; Panangaden, P.; and Hamilton, W. 2020. Latent variable modelling with hyperbolic normalizing flows. In *International Conference on Machine Learning*, 1045–1055. PMLR.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Fan, F.; Feng, Y.; and Zhao, D. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 3433–3442.

Garg, S.; and Ramakrishnan, G. 2020. BAE: BERT-based Adversarial Examples for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6174–6181.

Han, W.; Zhang, L.; Jiang, Y.; and Tu, K. 2020. Adversarial Attack and Defense of Structured Prediction Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2327–2338.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 1929–1938. PMLR.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Ide, N. 2004. Preparation and Analysis of Linguistic Corpora. *A Companion to Digital Humanities*, 289–305.

Jin, D.; Jin, Z.; Zhou, J. T.; and Szolovits, P. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 8018–8025.

Jones, E.; Jia, R.; Raghunathan, A.; and Liang, P. 2020. Robust Encodings: A Framework for Combating Adversarial Typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2752–2765.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved Variational Inference with Inverse Autoregressive Flow. *Advances in Neural Information Processing Systems*, 29: 4743–4751.

Kuleshov, V.; Thakoor, S.; Lau, T.; and Ermon, S. 2018. Adversarial examples for natural language classification problems.

Lample, G.; Subramanian, S.; Smith, E.; Denoyer, L.; Ranzato, M.; and Boureau, Y.-L. 2018. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Le, T.; Wang, S.; and Lee, D. 2020. MALCOM: Generating Malicious Comments to Attack Neural Fake News Detection Models. In *20th IEEE Int'l Conf. on Data Mining (ICDM)*.

Li, J.; Jia, R.; He, H.; and Liang, P. 2018. Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1865–1874.

Lin, J.; Zou, J.; and Ding, N. 2021. Using Adversarial Attacks to Reveal the Statistical Bias in Machine Reading Comprehension Models. *arXiv preprint arXiv:2105.11136*.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427–436.

Qin, L.; Che, W.; Li, Y.; Ni, M.; and Liu, T. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8665–8672.

Ren, S.; Deng, Y.; He, K.; and Che, W. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1085–1097.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 856–865.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6833–6844.

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.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

Wang, T.; Wang, X.; Qin, Y.; Packer, B.; Li, K.; Chen, J.; Beutel, A.; and Chi, E. 2020. CAT-Gen: Improving Robustness in NLP Models via Controlled Adversarial Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5141–5146.

Wang, W.; Wang, L.; Wang, R.; Wang, Z.; and Ye, A. 2019. Towards a robust deep neural network in texts: A survey. *arXiv preprint arXiv:1902.07285*.

Xu, J.; Zhao, L.; Yan, H.; Zeng, Q.; Liang, Y.; and Xu, S. 2019. LexicalAT: Lexical-based adversarial reinforcement training for robust sentiment classification. 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)*, 5521–5530.

Yang, X.; Zhang, H.; and Cai, J. 2020. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.