Enhancing Effectiveness and Robustness in a Low-Resource Regime via Decision-Boundary-aware Data Augmentation

Kyohoon Jin[*1], Junho Lee[*2], Juhwan Choi[*2], Sangmin Song[*2], Youngbin Kim[†1,2]

[1]Graduate School of Advanced Imaging Sciences, Multimedia and Film, Chung-Ang University
[2]Department of Artificial Intelligence, Chung-Ang University
{fhzh123, jhjo32, gold5230, s2022120859, ybkim85}@cau.ac.kr

Abstract

Efforts to leverage deep learning models in low-resource regimes have led to numerous augmentation studies. However, the direct application of methods such as mixup and cutout to text data, is limited due to their discrete characteristics. While methods using pretrained language models have exhibited efficiency, they require additional considerations for robustness. Inspired by recent studies on decision boundaries, this paper proposes a decision-boundary-aware data augmentation strategy to enhance robustness using pretrained language models. The proposed technique first focuses on shifting the latent features closer to the decision boundary, followed by reconstruction to generate an ambiguous version with a soft label. Additionally, mid-K sampling is suggested to enhance the diversity of the generated sentences. This paper demonstrates the performance of the proposed augmentation strategy compared to other methods through extensive experiments. Furthermore, the ablation study reveals the effect of soft labels and mid-K sampling and the extensibility of the method with curriculum data augmentation.

Keywords: Data Augmentation, Decision Boundary, Robustness

1. Introduction

As the latest pretrained language models have demonstrated excellent performance, numerous studies have been conducted on training larger models with more data. However, due to the numerous parameters that need to be learned, these pretrained language models require considerable data for downstream tasks. Data augmentation is widely used to address this problem, preventing overfitting by increasing the quantity of training data. Consequently, various data augmentation methods have been studied across various fields, including computer vision, audio, and text (Rizos et al., 2019; Lee et al., 2021; Oh et al., 2021; Lee et al., 2022).

Various studies have proposed data augmentation methods that transform the data while preserving their attributes as much as possible, such as rotation and Cutout (DeVries and Taylor, 2017). For textual data, fundamental textual operations, such as replacement, insertion, deletion, and shuffling, have been widely accepted in various augmentation frameworks (Wei, 2019). This straightforward data augmentation strategy enhances model robustness by focusing the optimization process on strengthening its ability to handle noise (Neelakantan et al., 2015; Piedboeuf and Langlais, 2022). The introduction of noise increases robustness, enabling the model to maintain its performance against intentional text corruption or modifications (Karpukhin et al., 2019). However, these techniques face the challenge of not guaranteeing the preservation of attributes and readability between the original and augmented sentences (Hsu et al., 2021). Methods using pretrained language models trained on diverse data to learn language representations have been proposed for data augmentation to address the preservation of attributes and readability problems. Compared to the noise addition, these models offer better preservation of readability and attributes (Wiechmann et al., 2022). Nevertheless, they have limited ability to provide various sentences as they generate sentences based on the existing data distribution (Ott et al., 2018; Vanmassenhove et al., 2019). Furthermore, as these sentences are generated using language models pretrained on massive data, they tend to generate typical expressions. Moreover, the data is generated without consideration of decision boundaries, leading to a lack of robustness (Dong et al., 2021).

On the other hand, One of the popular data augmentation techniques is mixup, which involves combining two or more pieces of information from different images to create a new image. Mixup configures the vicinal risk minimization learning method using soft labels instead of one-hot encoding labels in the learning process. This approach helps prevent overconfidence, enhances robustness against adversarial attacks, and preserves the content of each attribute (Zhang et al., 2018). It offers several benefits, such as enhancing model robustness against adversarial attacks by training the decision boundary using soft labels, as found in previous studies (Chen et al., 2023). However, directly apply-

[*]Equal contribution
[†]Corresponding author

LREC-COLING 2024, pages 5930–5943
20-25 May, 2024.

ing mixup-based approaches to the text domain is limited. Unlike images, where attributes can be interpreted as continuous signals, sentences consist of a discrete set of words. Consequently, the modification of words in equal ratios does not guarantee an equal influence on the sentence label (Kim et al., 2021; Chen et al., 2022).

From a geometric perspective, the robustness of a deep learning model is expected to be influenced by its decision boundary (Goodfellow et al., 2015a). Decision boundaries extend across the entire feature space used for training and are not limited to the provided data points. Therefore, investigating the decision boundaries is a crucial aspect of understanding the decision-making behavior of deep neural network classifiers.

Accordingly, we propose a data augmentation technique that leverages the advantages of each method. Similar to mixup, the proposed approach involves shifting the latent features toward the decision boundary, leveraging soft labels while effectively preserving existing attributes. We define ambiguous data as values close to the decision boundary. Furthermore, we employ a pretrained language model to ensure readability and attribute consistency between the original and augmented sentences. Additionally, we introduce variability to sentences through mid-K sampling and enhance robustness through decision-boundary-aware gradient modification. Experiments demonstrate that the proposed method improves the performance of the model by constructing a more robust decision boundary by shifting it using augmented data. Additionally, the experimental result showcased the superiority of our method compared to previous baselines in terms of performance enhancement, statistical durability, and robustness against adversarial attacks.

Our contributions are summarized as follows.

- We propose a novel and intuitive data augmentation technique considering the decision boundary in latent space. In addition, we exploit a pretrained language model to preserve the readability and attribute consistency in the generated sentences, enhancing the effectiveness of the data augmentation.
- We introduce mid-K sampling to generate diverse augmented data. This method generates diverse data by selecting the top k words while considering the middle K words, generating data while preserving essential information.
- We demonstrate the effectiveness and robustness of the proposed method through comparative experiments including soft labels, curriculum learning, and various decoding strategies.

2. Related Work

2.1. Text Augmentation

Text data augmentation is a training strategy designed to enhance the robustness of a model by generating new sentences using various techniques. One common approach is the manipulation of words according to predefined rules. Easy data augmentation (EDA) is a well-known method that employs rule-based techniques, including synonym replacement, random insertion, swap, and deletion, to introduce diverse types of noise (Wei, 2019). However, these techniques might alter the original meaning of the sentences by randomly deleting words or changing their order without considering the context. In contrast to word-level rule-based augmentation, other methods have been proposed to consider the context of sentences using deep learning models.

Early research in this area employed language models to replace words with their alternatives, considering the context of the sentence (Kobayashi, 2018; Wu et al., 2019; Zhou et al., 2022). These model-based augmentation techniques preserve the semantics of the original data by modifying only a portion of the original sentence. However, they may have limited diversity compared to the original data. Further developing the existing research, a study proposed using a pretrained large language model (LLM) (Yoo et al., 2021). However, this method requires existing LLM knowledge and may result in bias.

On the other hand, several approaches have been proposed to introduce mixup augmentation into the natural language processing field (Zhang et al., 2018). Early studies suggested performing mixup interpolation at the word embedding level or the level of encoded sentence representations (Guo, 2020). An end-to-end scheme to perform mixup and train the model was also introduced (Sun et al., 2020; Chen et al., 2022). These mixup-based approaches complement the sparsity of the data distributions through interpolation. However, they have limited interpretability because the augmented results cannot be directly observed.

2.2. Analysis of Decision-Boundary Neighbored Data

In training and evaluating deep learning models, various studies have suggested the importance of counterfactual data, which have minimal differences in the input space but different label values (Teney et al., 2020; Gardner et al., 2020; Kaushik et al., 2020). Specifically, the concept of a contrast set has been introduced to explain this phenomenon from the viewpoint of the decision boundary (Gardner et al., 2020). Ambiguous data that

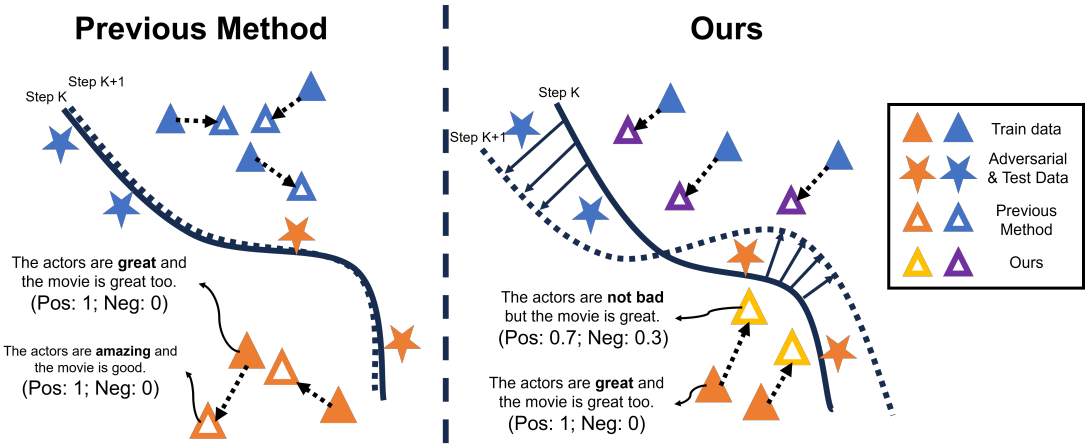

Figure 1: The figure illustrates the concept of the decision-boundary-aware gradient modification. In the previous method, augmentation was performed without the consideration of decision boundaries. However, in the proposed method, augmentation is performed in decision-boundary-aware manner.

are close to the decision boundary and challenging to distinguish are essential in forming a robust decision boundary (Ding et al., 2020; Zhu et al., 2022) and in training (Margatina et al., 2021; Zhang et al., 2022a). Data points close to the decision boundary are also significant from the perspective of adversarial attacks because vulnerability to adversarial assaults increases with the distance from the decision boundary (Zhang et al., 2021). Considering this geometric characteristic, strategies to improve the generalizability of a model by adjusting the perturbation have been studied (Zhang et al., 2022b; Holtz et al., 2022; Cheng et al., 2022; Yang and Xu, 2022; Zhang et al., 2023).

Previous studies have demonstrated the potential to enhance model robustness through data augmentation (Rebuffi et al., 2021; Shorten et al., 2021; Gowal et al., 2021). Based on these findings, the concept of decision-boundary-aware data augmentation (Zhu et al., 2022; Chen et al., 2023) has been emerged. Decision-Boundary-aware augmentation method aims to establish a more robust decision boundary by leveraging data augmentation techniques that can shift the current decision boundary in a targeted direction (Park et al., 2022; Peng et al., 2023). A previous study suggested a method to identify adversarial samples that lie close to the decision boundary without crossing it, improving adversarial robustness without sacrificing performance (Zhu et al., 2022).

Furthermore, studies have explored the utilization of mixup to generate data points close to the decision boundary. However, existing mixup augmentation methods are claimed to be ineffective in improving adversarial robustness when interpolation samples are randomly selected (Chen et al., 2023). To address this problem, researchers have proposed a method that explicitly selects the current attackable data as mixup interpolation samples to generate attackable data points. However, in the field of natural language processing, the generation of meaningful sentences is difficult because of the discrete representation of text and variations in length (Yoon et al., 2021). Thus, previous studies that applied mixup augmentation in the natural language processing domain have primarily focused on the feature level (Guo, 2020; Sun et al., 2020). To overcome this limitation, SSMix (Yoon et al., 2021) suggested generating sentences with soft labels by modifying words that influence the label. However, these studies involve word-level modifications and differ from the direct generation of sentences that combine two sets of label information. Inspired by these studies, we construct sentences with soft labels by reconstructing them from ambiguous representations after shifting the encoded representations to the decision boundary.

3. Methodology

3.1. Task Definition

In this paper, the decision boundary refers to the region where the probability of each class is equal (Karimi and Derr, 2022). We defined ambiguous data as values close to the decision boundary. A task is defined as obtaining $\hat{x}$, corresponding to generating ambiguous data from the given source data x, and each data x is paired with an attribute vector y. For example, in sentiment analysis, sentiments such as 'positive' and 'negative' become attributes y. The data augmentation process consists of four steps. First, we create a well-trained attribute classifier C_π with source data x and source attribute y. Second, the source data x is encoded based on the encoder from the first step to obtain z, which is the latent representation of x. To examine the

similarity of z to the decision boundary in the latent space, we pass z to C_π to obtain the classification $\tilde{y}$. Third, based on the gradient of $\tilde{y}$ for the decision boundary, we apply n times the iterative gradient modification to the value of z to obtain an ambiguous representation z'^n. The Transformer-based decoder reconstructs x from this z'^n, resulting in $\hat{x}$, the augmented data of x. Finally, $\hat{x}$ is input into C_π for scoring, and the result is assigned as a soft label to create an augmented data pair $D' = \{\hat{x}, \hat{y}\}$ in the result. Figure 1 describes the overall proposed task.

3.2. Model Training

The proposed model consists of three subcomponents: the encoder, decoder, and attribute classifier. The entire procedure to augment data is divided into four steps. First, the encoder E_θ encodes a given sentence into its latent representation z. Then, the attribute classifier is trained by z, resulting in a well-trained attribute classifier. In addition, the encoder learns how to precisely distinguish each attribute in the latent space. The training object is defined as follows:

$$\mathcal{L}_{cls}(C_\pi(E_\theta(x), y; \theta, \pi)) = \varepsilon_{cls} \sum_{i}^{|C|} u_i log(q_i) - (1 - \varepsilon_{cls}) \sum_{i}^{|C|} \bar{q}_i log(q_i) \quad (1)$$

Second, the frozen E_θ generates z, which is used by the Transformer-decoder D_γ to conduct reconstruction from the provided z. Through this process, D_γ is trained to reconstruct a sentence from z. The training object is defined as follows:

$$\mathcal{L}_{recon}(D_\gamma(E_\theta(x), x; \gamma)) = -\sum_{k=1}^{|N|} \sum^{|x_k|} \Big((1 - \varepsilon_{recon}) \sum_{i}^{|V|} \bar{p}_i log(p_i) + \varepsilon_{recon} \sum_{i}^{|V|} \bar{u}_i log(p_i) \Big) \quad (2)$$

where $|N|$ represents the size of the training data, $|x_k|$ represents the length of x_k, and $|V|$ and $|C|$ denote the size of the vocabulary and number of classes, respectively. In addition, $\bar{p}_i$ and $\bar{q}_i$ represent the probability distributions predicted by the decoder and classifier, respectively. Further, p_i and q_i represent the true distribution of reconstruction and classification, respectively. Moreover, ε_{recon} and ε_{cls} are label smoothing parameters for each loss term, and u_i represents a uniform noise distribution for label smoothing, defined as $1/|V|$ and $1/|C|$, respectively.

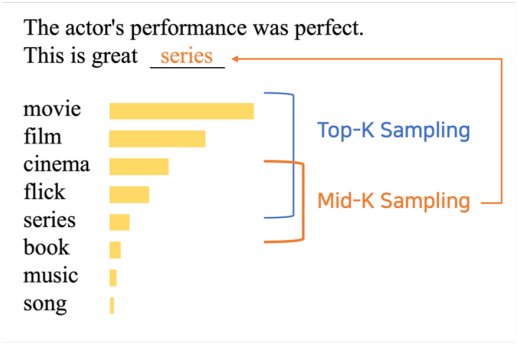

Figure 2: The figure illustrates the concept of the mid-K Sampling. Mid-K Sampling is a method to increase the diversity of generated sentences by sampling the middle K sentences instead of selecting the K sentences with the highest probability values (Top-K Sampling).

During training, we first trained E_θ and C_π using $\mathcal{L}_{cls}$ and then trained D_γ using $\mathcal{L}_{recon}$ while fixing E_θ. Thus, C_π and D_γ are trained separately, independent of each other. We found that separate training of C_π and D_γ yields superior results compared to joint training. Following this approach to model training, the decision-boundary-aware gradient is modified to provide enhanced data $\hat{x}$.

3.3. Decision Boundary-aware Gradient Modification

During the inference time (i.e., augmenting a given sentence), D_γ takes $z'^{(n)}$, the ambiguous representation of $z' = E_\theta(x)$ as input. To acquire the transformed representation, we passed z to C_π and computed the gradient. The direction of modification is determined by back-propagating the gradient of the attribute classification loss, inspired by previous work (Goodfellow et al., 2015b). During the iterative modification process, instead of using z, we adapted the previous step $z'^{(n-1)}$. In addition, λ is a hyperparameter for modification:

$$z'^{(n)} = z'^{(n-1)} - \lambda \bigtriangledown_{z'^{(n-1)}} \mathcal{L}_{cls}(C_\pi(z'^{(n-1)}), \bar{y}) \quad where\ z'^{(0)} := z,\ n >= 1 \quad (3)$$

where $\bar{y}$ indicates the decision boundary of the model. The decision boundary is defined as every class with equal likelihood (e.g., $\{0.5, 0.5\}$ for a binary classification task).

3.4. Mid-K Sampling

Although reconstruction was performed from a modified representation, the difference between generated and original sentences may not be noticeable,

depending on the decoding strategy such as beam search (Freitag and Al-Onaizan, 2017). Existing decoding strategies, such as top-K (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2020), aim to generate sentences close to the correct sentence. Although this characteristic is useful for general tasks, such as machine translation, which requires an optimal solution, it may be improper for data augmentation, which requires diverse generated sentences (Feng et al., 2021). We propose a novel technique called mid-K sampling to achieve the goal of the augmentation method, which generates sentences that differ from the original while preserving the core semantics.

Figure 2 illustrates the concept of mid-K sampling. Similar to top-K sampling, mid-K sampling selects the k most probable words at the current time step to exclude inappropriate words that are too different from the original sentence. Next, we further selected the word k' with the highest probability among the k words. To consider the significance of the selected k' words, we determine d whether the cumulative sum of the probabilities of these words exceeds the predefined threshold p.

If the cumulative sum is below the threshold, it indicates that the word distribution is relatively flat at this time step. Therefore, the words generated in this step are less crucial in determining the meaning of the sentence, allowing a variety of word selections. By explicitly excluding the most likely k' words from the distribution, it is possible to generate an ambiguous sentence that differs from the original and prevents the uniformity of the generated sentences.

Conversely, if the cumulative sum exceeds the threshold, the word distribution has a high skewness. In this case, the k' words can be expected to be crucial in maintaining the meaning of the sentence. Therefore, sampling is performed among all k words, including k' words, which is the same as top-K sampling. By allowing critical words to be generated, the core semantics of the original sentence can be preserved.

Algorithm 1 The Procedure of Mid-K Sampling

Require: Word probability distribution p, Vocabulary V, k, k', threshold t
Ensure: Next word token
1: Get top k token from distribution p
2: Normalize word probability with respect to k tokens and get p_k
3: Get the cumulative sum of the probability of top k' tokens from p_k
4: **if** cumsum $\geq t$ **then**
5: Get $p_{k'}$ by excluding k' tokens from p_k and normalize
6: Sample next token from $p_{k'}$
7: **else**
8: Sample next token from p_k
9: **end if**

4. Experiments

4.1. Experimental Setup

We conducted the experiments on various text classification datasets (Socher et al., 2013; Warstadt et al., 2019; Pang et al., 2002; Li, 2002; Maas et al., 2011) and baselines (Wei, 2019; Wu et al., 2019; Sennrich et al., 2016; Anaby-Tavor et al., 2020; Yoo et al., 2021). Additionally, to simulate more challenging scenarios in learning the model, we evaluated the proposed method under low-resource conditions using only a subset of the training set for each dataset. We repeated the same experiment five times with different random seeds and reported the mean value and standard deviation of the results. We adopted *t5-large* from Transformers library (Wolf et al., 2020) as the encoder and decoder of our model. The evaluation of downstream tasks was performed using the *bert-base-cased* model and the *microsoft/deberta-v3-base* model from Transformers.

4.2. Effectiveness

Table 1 presents the experimental results. This study conducts experiments with BERT, which is commonly used, and DeBERTa to evaluate the effectiveness of the proposed method on larger models. The results reveal that the proposed method demonstrates a high average performance gain compared to other augmentation methods in the majority of datasets.

Both EDA (Wei, 2019) and back-translation (Sennrich et al., 2016), which are widely used as existing data augmentation techniques, performed similarly or even led to performance degradation compared to the baseline in BERT and DeBERTa. This phenomenon may arise from two factors: the potential semantic damage in EDA and the lack of diversity in the augmented data generated by back-translation.

The use of pretrained language models and the incorporation of soft labels proved to be effective in data augmentation. For DeBERTa, both GPT3-MIX (Yoo et al., 2021) and the proposed method employing soft labels performed significantly better than other methods. Although GPT3-MIX and the proposed method performed better on most datasets, the application of GPT3-MIX was limited for long sentences, such as in IMDb, due to the maximum token limit of the GPT3 API. The difference between the method proposed in this paper and GPT3-MIX is whether soft labels are assigned to randomly generated sentences or intentionally

BERT		Baseline	EDA	BT	C-BERT	LAMBADA	GPT3-MIX	Ours
SST2	1%	82.0 (2.8)	79.6 (1.9)	80.7 (3.1)	83.1 (2.7)	84.7 (4.4)	*87.7 (0.6)*	**88.9 (0.6)**
	10%	89.4 (6.1)	88.2 (2.4)	86.9 (4.5)	88.9 (4.1)	*90.1 (3.6)*	86.2 (0.5)	**90.5 (1.6)**
CoLA	1%	61.7 (3.0)	63.2 (4.7)	56.6 (6.4)	62.9 (7.3)	**68.9 (3.9)**	*68.5 (0.3)*	67.5 (2.0)
	10%	64.7 (6.3)	68.3 (4.1)	69.2 (4.1)	63.6 (7.1)	67.6 (4.0)	*69.4 (2.2)*	**69.7 (0.6)**
SUBJ	1%	73.4 (11.8)	72.7 (4.3)	73.5 (4.2)	72.3 (4.3)	85.5 (1.5)	**90.6 (1.1)**	*86.7 (1.1)*
	10%	77.7 (10.2)	80.2 (5.7)	76.5 (7.7)	80.2 (14.4)	87.4 (6.7)	*91.1 (1.2)*	**91.7 (3.9)**
SST5	1%	26.5 (4.1)	25.8 (3.6)	25.8 (2.4)	26.4 (2.9)	29.4 (2.6)	**33.3 (0.1)**	*30.4 (1.6)*
	10%	44.4 (4.5)	46.6 (1.5)	46.5 (2.9)	47.6 (5.1)	*48.5 (4.5)*	43.0 (1.8)	**49.5 (1.7)**
TREC6	1%	67.0 (7.5)	65.9 (7.1)	**69.3 (6.3)**	66.6 (5.9)	63.2 (4.9)	60.5 (6.1)	*68.2 (4.1)*
	10%	83.4 (3.9)	84.1 (5.4)	86.1 (7.7)	86.3 (7.5)	71.3 (6.9)	*86.6 (2.7)*	**88.3 (1.6)**
IMDB	1%	70.0 (4.1)	65.6 (4.8)	66.9 (3.3)	78.2 (0.1)	80.7 (6.7)	**86.2 (1.8)**	*84.6 (5.7)*
	10%	81.6 (5.3)	67.3 (3.0)	77.5 (5.9)	75.2 (5.3)	*82.1 (1.5)*	-	**87.7 (1.9)**
DeBERTa		Baseline	EDA	BT	C-BERT	LAMBADA	GPT3-MIX	Ours
SST2	1%	81.4 (5.6)	85.2 (4.9)	85.5 (6.6)	84.5 (6.6)	87.5 (5.3)	*88.2 (1.3)*	**88.6 (1.1)**
	10%	86.5 (6.7)	89.2 (3.2)	85.5 (6.6)	91.6 (1.4)	92.5 (1.8)	*93.2 (0.8)*	**93.4 (0.9)**
CoLA	1%	66.5 (3.7)	67.4 (3.7)	69.2 (5.2)	69.2 (5.2)	69.0 (0.5)	*75.4 (4.4)*	**79.9 (3.1)**
	10%	79.0 (2.7)	65.6 (4.5)	66.4 (4.7)	73.9 (5.2)	**80.5 (6.5)**	79.0 (1.3)	*79.3 (2.5)*
SUBJ	1%	77.7 (7.8)	78.0 (3.2)	78.7 (7.9)	75.8 (4.3)	80.2 (2.4)	**88.6 (2.1)**	*83.7 (1.8)*
	10%	83.1 (6.5)	86.3 (7.1)	85.8 (7.1)	83.7 (8.9)	*88.2 (5.2)*	82.9 (14.7)	**89.6 (2.2)**
SST5	1%	26.4 (2.4)	26.7 (6.7)	24.7 (4.5)	26.7 (2.8)	26.9 (1.5)	**37.5 (4.0)**	*29.2 (0.3)*
	10%	25.8 (4.0)	25.6 (3.2)	26.2 (2.9)	*28.6 (4.8)*	**29.9 (5.4)**	27.4 (11.0)	**29.9 (0.7)**
TREC6	1%	61.0 (3.8)	64.4 (4.4)	65.5 (3.4)	66.0 (2.9)	*67.0 (2.7)*	63.6 (11.8)	**68.1 (4.1)**
	10%	90.1 (6.1)	86.7 (7.0)	90.4 (6.3)	93.6 (2.6)	93.4 (5.7)	*94.2 (0.9)*	**95.6 (1.2)**
IMDB	1%	77.8 (2.1)	77.0 (6.3)	74.8 (5.5)	87.4 (7.3)	70.3 (3.5)	*87.7 (1.8)*	**89.9 (2.2)**
	10%	86.0 (3.5)	84.8 (4.9)	85.1 (4.3)	*88.0 (3.0)*	83.9 (3.2)	-	**91.9 (2.1)**

Table 1: Performance (%) of baseline and proposed method on the BERT (Devlin et al., 2019) and DeBERTa (He et al., 2021) models. The statistics are presented in the mean (standard deviation) format, and 1% and 10% indicate what percentage of the original was used. The best results are boldfaced, and the second-best results are italicized. Lower scores than the baseline are in gray. For the Internet Movie Database (IMDB) dataset, we could not experiment with the IMDB-10% environment due to the limitation of the official source code.

set soft labels. Additionally, if sentences are augmented through an LLM, the model relies heavily on the knowledge of the LLM. In other words, the unique characteristics of the dataset may become less prominent. However, as the proposed method focuses more on the given dataset, it is more effective in terms of this perspective because we can create soft labels without losing the characteristics of the dataset.

The proposed method exhibited a relatively small standard deviation of performance compared to other techniques, indicating that the proposed approach is statistically stable against changes in a random seed, unlike other augmentation methods. The construction of the low-resource dataset may vary with random seeds when randomly extracting data, leading to statistical instability. Nevertheless, the proposed method consistently increased performance through soft labels and mid-K sampling.

4.3. Robustness

We experimented to evaluate the robustness against adversarial attacks, which induce changes in the predicted value of the model by altering several words. We employed the TextFooler (Jin et al., 2020) and probability-weighted word saliency (Ren et al., 2019) strategies provided by the TextAttack (Morris et al., 2020) library. Following the previous approach (Si et al., 2021), we evaluated the approach by selecting 10% of the testing set from the SUBJ and IMDb datasets. We report accuracy under attack (AUA) and attack success rate (ASR) as metrics to assess the results. Table 2 presents the results.

The experimental results demonstrate that the model trained using the proposed method exhibits higher robustness against adversarial attacks compared to other methods. The proposed method reduces overconfidence by generating data with a soft label that is close to the decision boundary for training (Müller et al., 2019; Thulasidasan et al., 2019). Alleviating the overconfidence in this manner enhances the robustness of the model against adversarial attacks (Grabinski et al., 2022). These findings are consistent with a previous study (Zhu et al., 2022) demonstrating robustness improvement using data close to the decision boundary.

TextFooler	SUBJ	IMDB
EDA	AUA: 22.0% ASR: 76.09%	AUA: 0.0% ASR: 100.0%
GPT3-MIX	AUA: 8.0% ASR: 91.11%	AUA: 2.0% ASR: 97.14%
Ours	AUA: **36.0%** ASR: **60.0%**	AUA: **6.0%** ASR: **92.68%**
PWWS	SUBJ	IMDB
EDA	AUA: 30.0% ASR: 67.39%	AUA: 0.0% ASR: 100.0%
GPT3-MIX	AUA: 22.0% ASR: 75.56%	AUA: 0.0% ASR: 100.0%
Ours	AUA: **42.0%** ASR: **53.33**%	AUA: **4.0%** ASR: **95.12%**

Table 2: Robustness against TextFooler and probability-weighted word saliency (PWWS) attacks on the baseline and proposed methods. AUA denotes accuracy under attack and ASR denotes attack success rate.

	Baseline (Soft-Label)	Hard-Label
SST2	**88.9 (0.6)**	87.2 (1.2)
SUBJ	**86.7 (1.1)**	86.4 (5.0)
TREC6	**68.2 (4.1)**	51.3 (5.6)

Table 3: An ablation study on comparison between the soft-label and hard-label. We denote performance (%) in mean (std) format.

4.4. Ablation Study

4.4.1. Effectiveness of Soft Labels

We performed an ablation study to investigate the effectiveness of soft labels. This study compares the performance of forming augmented data pairs by assigning hard labels from the original sentences instead of using soft labels. Table 3 lists the results, confirming that the proposed method based on soft labels outperformed the approach based on assigning hard labels from the original data. These results suggest that soft labels play a crucial role in decision-boundary recognition augmentation methods. Additionally, using hard labels exhibited a higher standard deviation of performance than soft labels, indicating that this is more unstable than the soft label approach.

4.4.2. Effectiveness of Curriculum Data Augmentation

Recently, studies have explored the combination of textual data augmentation methods with curriculum learning (Bengio et al., 2009), suggesting the concept of curriculum data augmentation (Wei et al., 2021; Lu and Lam, 2023). The augmentation method proposed in this paper can also be extended to curriculum data augmentation by adjusting the number of gradient modifications. In this study, we employed a curriculum data augmentation approach, where data close to the decision boundary were gradually generated by involving data obtained through moving lambda (λ) one more time every two epochs. The performance compared to the baseline is presented in Table 4. For SST2, SUBJ, and CoLA, which consist of relatively short sentences, the curriculum augmentation showed a slight improvement at about 0.1 to 0.3. Additionally, it was not effective for dataset with long sentences such as IMDb. This result suggests that even though performance can be further improved with curriculum learning, its effectiveness is limited to the dataset with short sentences. However, since there was an improvement in performance in short sentences, it is expected that the improved approach will lead to performance enhancements in longer sentences as well. We leave this to future work.

	Baseline (w/o Curr. Aug.)	w/ Curr. Aug.
SST2	**88.9 (0.6)**	**88.9 (1.7)**
SUBJ	86.7 (1.1)	**87.0 (4.8)**
CoLA	67.5 (2.0)	**67.8 (1.1)**
IMDB	**84.6 (5.7)**	80.6 (2.8)

Table 4: An ablation study on curriculum augmentation. We denote performance (%) in mean (std) format.

	Greedy	Beam Search	Top-K	Mid-K
SST2	87.6 (2.5)	86.5 (2.7)	88.6 (0.7)	**88.9 (0.6)**
SUBJ	73.6 (12.3)	81.3 (6.0)	86.3 (1.0)	**86.7 (1.1)**
CoLA	63.9 (4.3)	67.4 (1.8)	66.4 (0.8)	**67.5 (2.0)**

Table 5: An ablation study on decoding strategy. We denote the performance (%) of our proposed Mid-K sampling and other decoding methods in mean (std) format.

4.4.3. Effectiveness of Different Decoding Strategies

We conducted an ablation study to validate the effectiveness of mid-K sampling compared to other conventional decoding strategies. Table 5 presents the results of applying greedy decoding, the beam search, and top-K sampling to the proposed method, compared to the baseline method that uses mid-K sampling. Specifically, applying greedy decoding or the beam search resulted in lower performance compared to mid-K sampling. Greedy decoding and beam search are optimization methods that aim to determine an optimal solution, making them effective in such tasks as translation. Top-K sampling introduces relatively more

Original	Augmented
a dark, quirky road movie that constantly **defies expectation.** (Positive: 75% / Negative 25%)	a has all but **no sense of story but fascinating humor.** (Positive: 100% / Negative 0%)
This film contains more action before the opening credits than are in entire Hollywood films of this sort. This film is produced by Tsui Hark and stars Jet Li. This team has brought you many worthy **Hong Kong** cinema productions, including the Once Upon A Time in China series. **The action was fast and furious** with amazing wire work. (Positive: 100% / Negative 0%)	This film begins more action before the opening credits than are in entire Hollywood films of this sort. This film is based byeki Hol and stars Michael Ryan. This team has made you many wonderful **Hollywood** cinema productions, including Most Lear Aot in Friday The Wellle car series. **The action, the story actually was dry.** (Positive: 74% / Negative 26%)
These thoughts are **hugely entertaining** and women will also enjoy this movie I'm sure! All cast members perform well, and this film could have been a tremendous hit all over the world if it was made in **England or the US.** But for those of you who are fortunate enough to understand Swedish, **you are in for a treat!** (Positive: 100% / Negative 0%)	Some brains are **terrific** and women will also enjoy this movie I'm sure! First cast members plays well, and this film could have been a particularly look all over the world if it was made in **Hollywood or the era.** But for those of you who are complicit enough to Aotta, **you are in for hatred!** (Positive: 75% / Negative 25%)

Table 6: Augmented samples with its soft-label through our proposed method. The left column of each row denotes the original data, and the right column represents the augmented data through our method. Important differences between original data and augmented data are boldfaced.

diversity through sampling compared to the greedy method; however, the produced sentences still frequently resemble the original because this method prefers words in the original sentences with the highest probability. In contrast, the mid-K sampling proposed in this paper achieves the highest performance because it provides diversity while effectively preserving semantic coherence.

4.5. Case Study

Table 6 presents augmented sentences through the proposed method. Three main characteristics of the proposed method are observed by comparing the original and augmented data.

First, the method proposed in this paper introduces different proper nouns compared to the original data. For instance, it changed "Hong Kong cinema productions" to "Hollywood cinema productions." This transformation allows the model to learn from augmented data that are distinct from the original, serving as new and diverse training examples. This outcome distinguishes the proposed approach from such methods as word-level modification or back-translation, which only replace a few words in each sentence, resulting in augmented sentences that are similar to the original.

Second, the proposed method generates data close to the decision boundary while maintaining the core attribute. For example, through the example of "no sense of story but fascinating humor," the proposed method produces sentences close to the decision boundary by adding the aspect of "no sense of story" while preserving the core attribute of "fascinating humor." This characteristic allows us to leverage data with soft labels that are close to the decision boundary, preventing overconfidence in the model and improving performance and adversarial robustness.

Finally, the suggested approach allows for phrase-level modification of the expression and the generation of various expressions. For instance, applying a rule-based synonym replacement to "hugely entertaining" may result in such expressions as "very entertaining" or "hugely humorous." However, the proposed approach can produce expressions like "terrific." This characteristic facilitates the training of the model with a variety of expressions by allowing for a broader range of expressions beyond simple word-level modifications. Furthermore, the word modifications take place at a level that is uncommon but may still be inferred from their meaning, such as changing "thoughts" to "brains." By employing such exceptional modifications, as opposed to relying on predefined dictionaries, the proposed model can improve its robustness against an adversarial attack that confuses the model with unexpected expressions or expressions the model might not have learned well.

5. Conclusion

This paper introduces a novel text augmentation method aimed at enhancing model robustness by shifting features closer to the decision boundary and increasing data diversity through mid-K sampling. Experimental results demonstrate the efficacy of decision-boundary-aware soft labels and mid-K sampling in augmenting data diversity and robustness. While the proposed approach shows promise, there are potential limitations, particularly related to covariate shift between augmented and real-world data distributions. Future research should focus on quantifying and mitigating this shift to ensure the augmented data's representativeness.

Limitations

While our paper introduces a novel augmentation technique with various benefits, we acknowledge several limitations:

- **Concerns on Covariate Shift**: The proposed augmentation method assumes alignment between augmented and real-world data distributions, potentially introducing covariate shift. Future research would address strategies to quantify and mitigate this shift, especially in low-resource settings.

- **Linguistic Correctness**: Generated sentences may lack linguistic correctness, albeit previous studies suggest performance enhancement despite linguistic inaccuracies. Our mid-K sampling approach primarily modifies proper nouns, preserving linguistic structure and semantics while introducing diversity. Additionally, adjusting hyperparameters and incorporating soft-label adaptation aids to mitigate linguistic damage.

- **Curriculum Data Augmentations**: While curriculum augmentation was not significantly effective in this study, its potential benefits, especially in specific datasets, suggest further exploration. Future work will focus on refining curriculum augmentation strategies to improve effectiveness and robustness.

Ethics Statement

The proposed method relies on pretrained language models to generate sentences. This dependence on specific models leads to potential bias in the augmented sentences. However, as the proposed method aims to move the given representation close to the decision boundary in the feature space, resulting in weakening strong expressions, it may help neutralize biased expressions in the original data.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1C1C1008534), and Institute for Information & communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program, Chung-Ang University).

Bibliographical References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Chen Chen, Jingfeng Zhang, Xilie Xu, Lingjuan Lyu, Chaochao Chen, Tianlei Hu, and Gang Chen. 2023. Decision boundary-aware data augmentation for adversarial training. *IEEE Transactions on Dependable and Secure Computing*, 20(3):1882–1894.

Hui Chen, Wei Han, Diyi Yang, and Soujanya Poria. 2022. DoubleMix: Simple interpolation-based data augmentation for text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4622–4632.

Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, and Cho-Jui Hsieh. 2022. Cat: Customized adversarial training for improved robustness. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 673–679.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. 2020. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*.

Xinshuai Dong, Anh Tuan Luu, Min Lin, Shuicheng Yan, and Hanwang Zhang. 2021. How should pre-trained language models be fine-tuned towards

adversarial robustness? *Advances in Neural Information Processing Systems*, 34:4356–4369.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015a. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015b. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. 2021. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233.

Julia Grabinski, Paul Gavrikov, Janis Keuper, and Margret Keuper. 2022. Robust models are less over-confident. *Advances in Neural Information Processing Systems*, 35.

Hongyu Guo. 2020. Nonlinear mixup: Out-of-manifold data augmentation for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4044–4051.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Chester Holtz, Tsui-Wei Weng, and Gal Mishne. 2022. Learning sample reweighting for accuracy and adversarial robustness. *arXiv preprint arXiv:2210.11513*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Semantics-preserved data augmentation for aspect-based sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4417–4422.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 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, pages 8018–8025.

Hamid Karimi and Tyler Derr. 2022. Decision boundaries of deep neural networks. In *2022 21st IEEE international conference on machine learning and applications (ICMLA)*, pages 1085–1092.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Yekyung Kim, Seohyeong Jeong, and Kyunghyun Cho. 2021. Linda: Unsupervised learning to interpolate in natural language processing. *arXiv preprint arXiv:2112.13969*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457.

Byeong Tak Lee, Yong-Yeon Jo, Seon-Yu Lim, Youngjae Song, and Joon-myoung Kwon. 2022. Efficient data augmentation policy for electrocardiograms. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4153–4157.

Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheoneum Park, and Kyomin Jung. 2021. Crossaug: A contrastive data augmentation method for debiasing fact verification models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3181–3185.

Dan Li, Xinand Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Hongyuan Lu and Wai Lam. 2023. PCC: Paraphrasing with bottom-k sampling and cyclic learning for curriculum data augmentation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 68–82.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in Neural Information Processing Systems*, 32.

Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. 2015. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*.

Sejoon Oh, Sungchul Kim, Ryan A Rossi, and Srijan Kumar. 2021. Influence-guided data augmentation for neural tensor completion. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1386–1395.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3956–3965.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86.

Jungsoo Park, Gyuwan Kim, and Jaewoo Kang. 2022. Consistency training with virtual adversarial discrete perturbation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5646–5656.

Letian Peng, Yuwei Zhang, and Jingbo Shang. 2023. Generating efficient training data via llm-based attribute manipulation. *arXiv preprint arXiv:2307.07099*.

Frédéric Piedboeuf and Philippe Langlais. 2022. Effective data augmentation for sentence classification using one VAE per class. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3454–3464.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. 2021. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: short-text

data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 991–1000.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1569–1576.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for NLP tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. In *European Conference on Computer Vision*, pages 580–599.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Jason Wei, Chengyu Huang, Soroush Vosoughi, Yu Cheng, and Shiqi Xu. 2021. Few-shot text classification with triplet networks, data augmentation, and curriculum learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5493–5500.

Kai Wei, Jasonand Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6382–6388.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-)linguistic and readability features and their spill over effects on the prediction of eye movement patterns. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5276–5290.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Anthony Delangue, Clementand Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95.

Shuo Yang and Chang Xu. 2022. One size does not fit all: Data-adaptive adversarial training. In *European Conference on Computer Vision*, pages 70–85.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239.

Soyoung Yoon, Gyuwan Kim, and Kyumin Park. 2021. SSMix: Saliency-based span mixup for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3225–3234.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Jie Zhang, Bo Li, Chen Chen, Lingjuan Lyu, Shuang Wu, Shouhong Ding, and Chao Wu. 2023. Delving into the adversarial robustness of federated learning. *arXiv preprint arXiv:2302.09479*.

Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. 2021. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022a. ALLSH: Active learning guided by local sensitivity and hardness. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1328–1342.

Wang Zhang, Lam M. Nguyen, Subhro Das, Pin-Yu Chen, Sijia Liu, Alexandre Megretski, Luca Daniel, and Tsui-Wei Weng. 2022b. Tactics on refining decision boundary for improving certification-based robust training.

Jing Zhou, Yanan Zheng, Jie Tang, Li Jian, and Zhilin Yang. 2022. FlipDA: Effective and robust data augmentation for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8646–8665.

Bin Zhu, Zhaoquan Gu, Le Wang, Jinyin Chen, and Qi Xuan. 2022. Improving robustness of language models from a geometry-aware perspective. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3115–3125.

Original Sentence	This movie followed movies within a movie, much like Scream 3 and Urban Legend 2. This was pure crap! The whole movie within a movie crap.
Fine-tuned BERT	This movie followed movies within a movie, much like Scream 3 and Urban Legend 2. This was **pure** crap! The whole movie within a movie crap.
Ours	This movie followed movies within a movie, much like Scream 3 and Urban Legend 2. This was pure **crap**! The whole movie within a movie **crap**.

Table 7: Examples for case study in robustness. The bolded words represent important terms selected via Lime (Ribeiro et al., 2016) for the given model.

A. Case Study in Robustness

We investigated which words in the sentence contained the adversarial attack in order to evaluate the model's robustness using examples. The sentence chosen from IMDB for investigation was, "This movie followed movies within a movie, much like Scream 3 and Urban Legend 2." This was really terrible! The entire film within a film is crap.". This sentence has negative sentiment owing to expressions such as "crap". Table 7 presents augmented sentences through the proposed method.

However, when we applied adversarial attack to the fine-tuned BERT model trained on original data using this sentence, TextFooler attack algorithm changed the word "pure" instead of the word "crap", which actually contributes the negative sentiment of the sentence. For further investigation, we examined the word importance of the model on this sentences, and discovered that "pure" had a high level of importance according to the fine-tuned BERT model. Nonetheless, in the BERT model trained on original data and augmented data with our method, the word "crap" received a high level of word importance. Moreover, the model successfully defended the adversarial attack on this example sentence, different from the model trained only on original data. We concluded that this correction of word importance enhances the robustness of the model.