CADCon: An Approach for Robust Learning of Counterfactually Augmented Datasets based on Contrastive Learning

Anonymous ACL submission

Abstract

During the fine-tuning process of Pre-trained Language Models (PLMs), they encounter relatively small datasets that may have spurious correlation patterns. Counterfactually Augmented Data (CAD) has emerged as a solution to make models less sensitive to such spurious patterns. While there has been progress in generating CAD due to advancements in generation models, the focus has primarily been on the quality of CAD, with limited attention given to training models for robustness. We introduce CADCon, a novel contrastive learning approach to enhance robustness by effectively utilizing CAD, rather than simply augmenting it. Firstly, we utilize an LLM-based generative model to generate counterfactual samples from original sentences. This is achieved by using a simple prompt, without human intervention or additional models. Secondly, we propose a tagging-based noise infusion method, which infuses noise into sentences without altering genuine tokens that have causal relationships with labels. Lastly, we perform contrastive learning so that counterfactual samples are distant from the original sentences and noise-infused samples are close. Our method effectively mitigates spurious correlations and improves robustness. We demonstrate that our method outperforms in both counterfactual task and domain generalization task.

1 Introduction

Pre-trained language models (PLMs) (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019) trained on a large amount of unlabeled data have shown superior performance through fine-tuning in various tasks. PLMs can perform well with less data, but they are easily exposed to spurious correlation (Tu et al., 2020) between text and label, which is called *shortcuts*, by biasing the distribution within the training data. For example, when a model is trained on a majority of positive reviews for Spielberg movies, the word "*Spielberg*" will have a spurious correlation with the positive label (Wang and Culotta, 2020; Wang et al., 2022). Even if a negative comment about the movie is provided, a model trained on positive reviews would still predict a positive review for *Spielberg*. This phenomenon of shortcuts can lead to overfitting and a lack of generalization, resulting in challenges when dealing with out-of-domain (OOD) data. To mitigate these spurious correlations, research has explored two main approaches: 1) generating counterfactual augmented data (CAD), and 2) distinguishing causal features.

The first approach mainly focused on generating CAD (Kaushik et al., 2020; Samory et al., 2021), which is generated by minimally perturbing examples to flip the label. CAD was initially performed through manual annotation by humans. But due to the high cost, the approach shifted towards automatic generation methods. Yang et al. (2021) utilized a sentiment dictionary and Wang and Culotta (2021) used a statistical matching approach and pre-defined antonyms to automatically generate CAD. But, these methods were limited in generating high-quality CAD by using dictionary or statistics. Recent works tried to utilize PLMs for CAD generation, such as T5 (Zhou et al., 2022; Wen et al., 2022), GPT-2 (Madaan et al., 2021; Wu et al., 2021) and GPT-3 (Dixit et al., 2022; Liu et al., 2022; Chen et al., 2023). However, these previous works only concentrate on the generation of high-quality CAD. So, they result in subsequent costs associated with human intervention or the utilization of additional models for post-generation filtering. Moreover, they focus on augmenting CAD for training purposes only, without addressing the crucial issue of enhancing robustness through a model training perspective.

The second approach aims to mitigate spurious correlation and enhance robustness by distinguishing causal features through a classifier without requiring additional augmented data. This allows the model to ignore shortcut tokens and focus on genuine tokens during the learning process. To define shortcut tokens, Wang and Culotta (2020) utilized magnitude coefficients through a classifier and Wang et al. (2022) used the attention scores of the model and the frequency of domain-specific word. Choi et al. (2022) identified genuine tokens using the gradient from a fine-tuned model and output values from a masked language model. These studies distinguished these tokens using models trained on the train dataset. However, relying on such models, which are already biased due to spurious correlations, leads to inaccurate discrimination of these tokens.

In this paper, we propose a novel approach to effectively address the limitations of previous studies, aiming to resolve the spurious correlation problem and enhance model robustness. Our approach takes into account both data generation methods and model training strategies to offer a comprehensive and effective learning strategy. Our contributions can be summarized as follows:

- We leverage the knowledge of Large Language Models (LLMs) to generate counterfactual samples effectively using a simple-prompt approach. We analyze the datasets generated based on different prompts and demonstrate their excellence through experimental results.
- We introduce a Tagging-based Noise Infusion (TNI) technique and contrastive learning approach to effectively grasp the patterns of both original and counterfactual samples. This approach contributes to efficient representation learning, leading to improved robustness.
- To demonstrate the superiority of the proposed method, we show that it is superior in terms of robustness with improved generalization ability in both conventional fine-tuning and prompt-based fine-tuning with some interesting ablation.

2 Related Work

2.1 Counterfactually Augmented Dataset

A counterfactual text sample is a sentence that is generated by making minimal changes to the original text in order to flip its label. Prior studies (Kaushik et al., 2020; Samory et al., 2021) employed human annotators to create CAD. This augmented data was combined with the original dataset to train models, with the goal of improving the robustness and generalization of text classification models. However, manually annotating by humans is time-consuming and costly, so recent research has been focusing on automatically generating CAD. Yang et al. (2021) proposed an approach to automatically generate CAD by utilizing a sentiment dictionary. Madaan et al. (2021) utilized pre-trained GPT-2 to generate counterfactual samples based on conditions such as named-entity tags, semantic role labels, or sentiment. Wen et al. (2022) handled specific rationales as masked spans and employed a controllable text generation model to create CAD.

Recent studies have explored the use of Large Language Models (LLMs), such as GPT-3, for CAD generation. Dixit et al. (2022) proposed a CAD generation framework by combining a Counterfactual retrieval model with the GPT-3 model. Liu et al. (2022) proposed an effective dataset creation method through collaboration between human workers and LLMs, where human filtering was applied to the NLI dataset generated by LLM. Chen et al. (2023) constructed high-quality CAD without relying on human workers, utilizing GPT-3 generated CAD filtered by a teacher model. While the advancement of generation models has led to a surge in CAD generation, many of the mentioned studies focus on how to generate CAD accurately and effectively. As a result, they often require human validation or additional complex models for data verification. Furthermore, these sophisticatedly generated datasets are only used for straightforward augmentation in training, without introducing practical training methods aimed at effectively improving robustness.

2.2 Robust for Text classification

Gunel et al. (2021) jointly optimized cross-entropy loss and Supcon (Khosla et al., 2020) loss during the fine-tuning stage, demonstrating improved performance not only in general text classification but also enhanced robustness in few-shot and noisy environments. However, this approach has limitations in directly addressing the spurious correlation problem. The following studies aim to tackle the problem of spurious correlations, often referred to as shortcut issues, to enhance robustness in text classification. Wang and Culotta (2020) utilized features derived from matched samples to distinguish

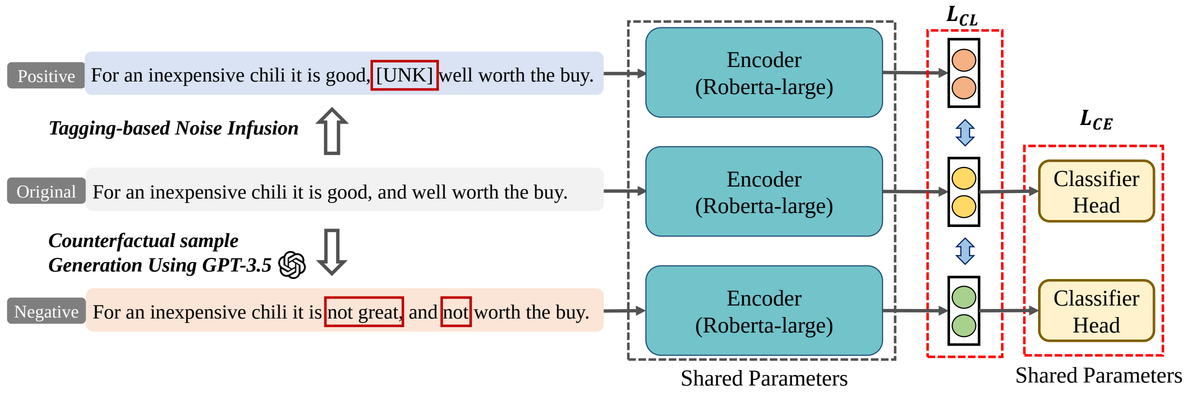

Figure 1: Overview of our proposed model. L_{CL} is the triplet loss for learning the representation for the original sentence, negative sentence, and positive sentence, and L_{CE} is the cross entropy loss for each label of the original sentence and the counterfactual sentence.

shortcuts from genuine ones. By removing words predicted to be shortcuts, they enhanced robustness in text classification. Wang et al. (2022) ensured robustness by distinguishing spurious tokens from important ones through cross-domain dataset analysis and knowledge-aware perturbation. Choi et al. (2022) proposed causally contrastive learning, training models to distinguish causal features. However, these methods all rely on gradient-based techniques to extract crucial words or utilize fine-tuned classifiers for consistency filtering in generated CAD. A notable drawback of these methods is their dependence on already-biased classifiers, which have encountered the spurious correlation problem.

3 Method

We propose a novel approach, CADCon, which utilizes simple prompts to generate counterfactual samples and effectively tackles the issue of spurious correlation through contrastive learning with CAD. Firstly, we utilize the GPT-3.5 generative model to generate counterfactual sentences from the original sentence by altering only minimally genuine tokens, those tokens that have an impact on the label. Next, we use the tagging information extracted from the original sentence to distinguish non-causal words that do not affect the sentence label. We then infuse noise into tokens associated with the identified non-causal tagging information to create positive sentences. We aim to learn the underlying patterns of the generated counterfactual and positive data in the representation space, focusing on capturing the key patterns associated with their influence on labels. Figure 1 provides an overview of CADCon, consisting of the following three processes: 1) Counterfactual Sample Generation Using GPT-3.5, 2) Tagging-based Noise Infusion, and 3) Contrastive Learning with Triplets.

3.1 Counterfactual Sample Generation Using GPT-3.5

Given a collection of sentences $\{A_i\}_{i=1}^{m}$, we construct a collection of counterfactual samples $\{N_i\}_{i=1}^{m}$ using GPT-3.5. In contrast to recent studies that use GPT-3.5 to generate CAD (Dixit et al., 2022; Liu et al., 2022; Chen et al., 2023), we concentrate on generating counterfactual samples using simple prompts, without the need for human intervention or additional models. We constructed the dataset by conducting experiments with the following three prompt instructions. Instruction 1 contains the "*Please make it a negative sentence.*" which outlines the intended behavior of the model. Instruction 2 provides the current task and label information for the sentence. In instruction 3, we offer specific guidance with phrases "*Just change a few words*" and "*while preserving the original text as much as possible.*" We use a similarly-designed prompt with instruction 3 corresponding to each task and label. Please refer to the Appendix B for a detailed description of the prompt instructions used for this purpose.

3.2 Tagging-based Noise Infusion (TNI)

In this paragraph, we introduce a method for constructing a collection of positive sentences $\{P_i\}_{i=1}^{m}$ aimed at addressing the fundamental issue of spurious correlation by preventing bias towards non-causal words that are not directly associated with the label. Previous data augmentation methods in contrastive learning (Gao et al., 2021b; Gunel et al., 2021), techniques such as dropout noise, EDA (Wei and Zou, 2019), and back translation (Sugiyama and Yoshinaga, 2019) have been used to generate positive samples. However, these augmentation techniques do not effectively address the aim of reducing spurious correlation, because they might destroy existing semantics. To address the issue of spurious correlation, distinguishing whether a token is a shortcut token or not is crucial. However, identifying tokens learned as shortcuts in the fine-tuned model is highly challenging. Therefore, we use a universal Part-of-speech (POS) tag set (Petrov et al., 2012) which is widely utilized across various NLP tasks to enhance performance. And, we utilize the logit output from the fine-tuned Model f to define the tagging that is not relevant to the labels. We iteratively removed tokens with specific tagging information to calculate the significance of their influence as described in Equation 1. Suppose that there is an input text $S = [w_0, w_1, ..., w_n]$ and universal POS tag set T = [`VERB`, `NOUN`, ..., `DET` ...]. The degree of accuracy reduction for the original model when removing all tokens belonging to each POS tag set is denoted as the importance I_{T_i}. It is represented by the following equation:

$$I_{T_i} = f(S) - f(S_{\backslash w_i \in T_i}) \tag{1}$$

We consider cases where the accuracy reduction is less than θ as POS tagging information that does not influence the label. We define this as the non-causal tag set G. And then, when given input S, we propose a noise infusion method to generate new positive samples by extracting k word tokens belonging to the set G and replacing these words with the `[UNK]` token. This approach allows us to maintain genuine tokens that influence the label while infusing noise for tokens that do not have an impact, thereby reducing bias towards shortcut tokens. Here, k is determined by multiplying a scaling factor α that reflects the average number of non-casual words in the train dataset. See the Appendix A for details on the k used for each dataset. In the example shown in Figure 1, words such as *'For'*, *'and'*, ... etc. can be decided for non-causal words. Especially, key point is that considering different non-causal words with varying noise infusion at each epoch allows us to consider multi-views. This helps reduce the tendency to become biased towards spurious correlation as the training progress.

3.3 Contrastive Learning with Triplets

We introduce contrastive learning for the effective training of models on the generated counterfactual and positive samples. First, the counterfactual sentences generated by altering only genuine tokens are considered not only to be a loss for direct label prediction but also to be a loss that encourages them to move further away from the original sentence in the latent space. Next, by bringing the positive samples generated through tagging-based noise infusion closer to the original samples in the representation space, we effectively mitigate the bias towards non-causal words and enhance the model's generalization ability. In summary, we aim to emphasize important features and eliminate unnecessary shortcuts through the generated triplets.

In conventional fine-tuning models, the `[CLS]` hidden representations from PLM M pass through a classifier head to produce the probability distribution on the label set y. As a result, the parameters θ of the entire model are trained in the direction of minimizing the cross-entropy loss between the predicted label $\hat{y}$ and the ground-truth label y:

$$L_{CE} = \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \cdot \log \hat{y}_{i,c} \tag{2}$$

where N denotes a batch of training examples of size and C denotes classes.

Recently, in order to narrow the gap between pre-training and downstream tasks prompt-based Fine-tuning models are attracting attention and few-shot setting (Brown et al., 2020; Gao et al., 2021a) Most prompt-based learning approach (Shin et al., 2020; Schick and Schütze, 2021; Gao et al., 2021a) utilize task-specific templates consisting of discrete prompts alongside input sentences. These prompts contain a `[MASK]` token and are designed to construct an objective that is similar to MLM training, where the goal is to map the `[MASK]` token to the right label (a specific word) with a pre-defined verbalizer. The probability distribution over the label is shown below:

$$P_M(\texttt{[MASK]} = v | T(x)) | v \in V_y) \tag{3}$$

where $T(\cdot)$ is a task-specific template and V_y is the label words of y.

In the standard (conventional) FT approach, representation learning was conducted using the hidden states of the `[CLS]` token as the representations of sentences. However, in the prompt-based FT approach, as demonstrated in (Jian et al., 2022), the final classification is performed using the [MASK] token. Therefore, representation learning was intuitively and effectively carried out using the representations of the [MASK] token, rather than the [CLS] token. We utilized a loss function similar to the training approach in C2L (Choi et al., 2022), which applied a margin-based ranking loss. The specific calculation of the triplet loss is as follows:

$$L_{CL} = \max(0, \frac{1}{M}\sum_{i=1}^{M} d(A_i, P_i) - \frac{1}{M} d(A_i, N_i) + \alpha) \tag{4}$$

where M is the number of sentences, A_i represents the i-th original sentence, N_i is the negative sentence generated from the i-th original sentence by the GPT model, P_i is the positive sentence generated from the i-th original sentence by TNI, α is a margin value enforced between positive and negative pairs, and $d(\cdot)$ computes the distance between the hidden states at [CLS] tokens or [MASK] tokens as the representations of two sentences. The final loss is as follows:

$$L = (1 - \lambda)L_{CE} + \lambda L_{CL} \tag{5}$$

λ is a scalar weighting hyperparameter that we tune for each downstream task.

4 Experiments

4.1 Datasets

Counterfactul Task Datasets To identify and address the phenomenon of being biased by spurious correlation in training data, we use two datasets (Kaushik et al., 2020; Samory et al., 2021), where the counterfactually-revised dataset (CF) is paired with the original dataset (O). Following Kaushik et al. (2020), we use the same train/valid/test datasets in sentiment analysis. In the sexism dataset (Samory et al., 2021), unlike sentiment analysis, there are pairs annotated by the crowdworkers only to make the sexist sentence a non-sexist sentence. Therefore, we used the original-counterfactual pairs from the dataset and ensured label balance by constructing a non-sexist dataset sampled from non-pairs within the dataset. Further, in the standard FT experiment, the dataset was split in a 9:1 ratio for training and testing, respectively. And we use 10% of the train dataset for validation. In both tasks, we also utilized the CF train dataset, which was not utilized during training, as a test dataset to demonstrate the impact of spurious correlations. Appendix A shows the statistical details of the counterfactual task datasets. We used YELP (Asghar, 2016), SST2 (Socher et al., 2013), FineFood (McAuley and Leskovec, 2013), and Tweet[1] as test data for sentiment analysis and Tweet[2] for Sexism classification as Out-Of-Distribution (OOD) datasets to evaluate the generalization ability.

Cross-Domain Generalization Datasets For cross-domain experiments, we use sentiment analysis datasets on SST-2 (Socher et al., 2013), IMDb (Maas et al., 2011), FineFood (McAuley and Leskovec, 2013) datasets. In standard FT, we utilized official train, validation, and test sets if available. In cases where such datasets were not provided, we randomly split the data into training and validation sets with an 8:2 ratio for each seed.

4.2 Baselines

Supcon In Gunel et al. (2021), the joint optimization of cross-entropy loss and SupCon loss (Khosla et al., 2020) in PLM fine-tuning was applied, showing enhanced robustness and improved generalization performance in text classification tasks.

C2L To enhance robustness, Choi et al. (2022) relies on the classifier model to identify causal words that significantly influence the label. They treat the masking of causal words as negative examples, and the masking of less significant words as regular positive examples, thereby jointly optimizing triplet loss and cross-entropy. We used the publicly available code on our experimental setup.

EDA Easy Data Augmentation (EDA) (Wei and Zou, 2019) proposed a method of augmenting sentences by randomly applying four heuristic techniques: synonym replacement, word insertion, word deletion, and word swapping. We employed this method to augment our dataset by applying one augmentation per sentence.

SSMBA Ng et al. (2020) proposed a corrupt-and-reconstruct text data augmentation technique using the BERT pre-trained model, showing performance improvements on out-of-domain datasets. In our experiments, we adopted the approach of augmenting data while keeping the labels unchanged. We

[1]https://www.kaggle.com/c/tweet-sentiment-extraction. We use only positive and negative tweets, excluding neutral labels.

[2]https://www.kaggle.com/datasets/dgrosz/sexist-workplace-statements.

Methods	In-Domain Dataset			Out-Of-Distribtuion Dataset				Overall
	O-Test	CF-Test	CF-Train	YELP	SST2	Food	Tweet	
Standard Fine-Tuning (full-data)								
RoBERTa-large (Liu et al., 2019)	93.85	93.31	89.75	95.38	86.00	95.65	78.75	90.38
Robust Learning								
SupCon (Gunel et al., 2021)	93.85	88.11	84.20	95.26	86.20	95.32	74.90	88.18
C2L (Choi et al., 2022)	93.92	91.67	89.55	95.22	88.47	95.32	80.66	90.69
Data Augmentation								
EDA (Wei and Zou, 2019)	**94.33**	93.51	91.88	95.59	89.22	95.71	80.31	91.51
SSMBA (Ng et al., 2020)	93.60	92.69	89.06	**95.90**	89.40	**96.12**	78.75	90.79
AugGPT (Dai et al., 2023)	93.37	91.46	87.97	95.32	90.21	94.18	78.66	90.17
Counterfactually Augmented Dataset								
Human-CAD	93.17	**97.47**	**99.02**	92.16	88.65	94.26	80.66	92.20
CORE-CAD	90.64	95.42	92.35	90.32	87.86	92.18	87.39	90.88
CADCon	93.37	95.83	95.04	95.29	**91.07**	94.89	**88.62**	**93.44**

Table 1: The accuracy (%) of various approaches in sentiment analysis for the counterfactual task under standard fine-tuning setting.

Methods	O-Test	CF-Test	CF-Train	Tweet
Baseline	92.69	49.23	45.14	81.00
SupCon	91.79	22.56	20.21	76.28
C2L	**93.21**	37.69	30.76	77.92
EDA	91.67	37.69	28.99	81.59
SSMBA	92.82	25.64	19.18	79.36
AugGPT	92.31	29.23	23.39	78.83
Human-CAD	91.79	**91.80**	**98.04**	**83.11**
CADCon	90.13	88.97	88.10	82.82

Table 2: The accuracy (%) of various approaches in sexism task under standard fine-tuning setting

also employed this method to augment our dataset by applying one augmentation per sentence.

AugGPT Dai et al. (2023) used GPT-3 to augment data, enhancing the performance of text classification in a few-shot setting. In our experiments, we augment data using single-turn dialogues with the prompt "Please rephrase the following sentence."

Human-CAD This method, often compared in papers that predominantly explore the automated generation of CAD, involves augmenting CAD generated by human annotators (Kaushik et al., 2020) and training it alongside the original train dataset.

CORE-CAD Dixit et al. (2022) proposed a retrieval-augmented generation framework for generating CAD using a combination of a retrieval model and GPT-3. In our approach, we use the publicly available dataset on our experimental setup.

5 Results and Discussion

Firstly, we demonstrate the superior performance of our proposed approach over existing previous methods for robust text classification through two counterfactual tasks. Secondly, we conducted an 8-shot experiment with extremely low data volume and a cross-domain generalization experiment for typical dataset environments to illustrate the enhancement of robustness. Lastly, we validate the superiority of the proposed method through a comprehensive ablation study.

5.1 Main results

Spurious Correlation in Counterfactual task As shown in Table 1 and 2, especially in the sexism task, the Roberta-large model trained on the original train dataset using standard FT achieves an accuracy of 92.69% on the original test dataset (O-Test). However, its accuracy drops significantly to 49.23% on the CF test dataset (CF-Test). In the case of sentiment analysis, the performance drop on the CF-Test dataset is relatively small by 0.5%, whil implies that larger PLMs are less sensitive to spurious patterns, as also noted by Yang et al. (2021). Nevertheless, for demonstrating the issue of shortcuts in the train dataset, we report the performance of the CF train dataset (CF-Train), which was not used during training. This results in a considerable performance drop in both sentiment analysis and sexism datasets. Furthermore, the low performance in Out-Of-Distribution dataset (OOD) suggests that both datasets suffer from spurious correlation within the training data, leading to poor

Methods (8-shot)	In-Domain Dataset			Out-Of-Distribtuion Dataset				Overall
	O-Test	CF-Test	CF-Train	YELP	SST2	Food	Tweet	
Prompt-based Fine-Tuning								
RoBERTa-large (Liu et al., 2019)	92.21	90.33	90.95	93.54	82.61	94.85	72.41	88.13
Robust Learning								
SupCon (Gunel et al., 2021)	91.52	90.45	91.38	**95.31**	84.16	95.28	73.51	88.80
Data Augmentation								
EDA (Wei and Zou, 2019)	91.02	91.64	92.71	94.18	84.34	94.79	71.00	88.53
SSMBA (Ng et al., 2020)	**92.25**	92.13	92.55	93.91	84.70	95.28	74.63	89.35
AugGPT (Dai et al., 2023)	92.13	92.30	92.69	92.68	81.55	94.64	70.53	88.07
Counterfactually Augmented Dataset								
Human-CAD	91.19	**93.16**	**93.61**	94.01	85.13	94.96	78.45	90.07
CORE-CAD	91.76	92.95	93.09	93.36	88.30	93.72	81.50	90.67
CADCon	91.11	91.93	93.09	95.28	89.59	**95.37**	**82.23**	**91.23**

Table 3: The accuracy (%) of various approaches in sentiment analysis for the counterfactual task under the prompt-based fine-tuning setting.

Methods	S → I	S → F	I → S	I → F	F → S	F → I	Overall
Standard Fine-Tuning (full-data)							
RoBERTa-large (Liu et al., 2019)	**91.67**	93.08	89.16	91.13	82.48	90.22	89.62
Robust Learning							
SupCon (Gunel et al., 2021)	90.82	89.64	91.21	94.95	73.40	89.68	88.28
C2L (Choi et al., 2022)	90.52	91.61	89.90	94.64	81.18	**90.50**	89.72
Data Augmentation							
EDA (Wei and Zou, 2019)	91.64	93.51	90.76	94.12	80.18	89.29	89.92
SSMBA (Ng et al., 2020)	90.71	90.78	**94.21**	93.96	78.75	89.31	89.62
CADCon	89.58	**93.75**	90.88	**94.96**	**87.30**	89.76	**91.04**

Table 4: The accuracy (%) of cross-domain generalization task. We denote each sentiment dataset as follows: SST-2 (S), IMDB (I), and FineFood (F).

generalization capabilities. We can also see that existing methods that do not utilize CAD still fail to catch spurious correlations.

Robustness in Counterfactual Task Table 1 and 3 shows that the proposed method outperforms various baselines in both settings (full-data, 8-shot) on the In-Domain Dataset (IDD) and OOD. Also, in the case of Human-CAD, which is directly generated by human, the performance on IDD is the highest since CF-Train was used for training. However, the performance on OOD is consistently lower compared to CADCon across all four datasets. This highlights that the proposed method demonstrates a remarkable performance by enhancing the generalization capabilities and ensuring model robustness, dramatically improving overall performance. While previous methods might exhibit better performance on the O-Test, this advantage can be attributed to their incorporation of biases from the spurious correlations present in the train dataset. However, their lack of adaptation to CF-Test and OOD becomes evident. In contrast, CADCon shows mostly dramatic performance improvements on IDD and OOD. Furthermore, in Table 1 and 2 considering the CF-Train, which demonstrated performance of 95.04% and 88.10% for the two tasks, it can be observed that the proposed approach is suitable for mitigating spurious correlations and enhancing robustness, which is the main aim of this paper.

Robustness in Domain Generalization Task In an environment with relatively abundant training data, we report the performance of domain generalization task to demonstrate that our proposed method is effective in securing robustness and enhancing generalization capabilities. As evident from Table 4, there is a substantial increase in performance, particularly in IMDB → FineFood and FineFood → SST2. This indicates that the efforts to address spurious correlations in CADCon can potentially contribute to improving generalization abilities even when the domain undergoes a shift.

Models	Data Augmentation		Loss			Datasets	
	Neg	Pos	CE	Triplet-Neg	Triplet-Pos	IDD	ODD
Human-CAD	Human	X	O	X	X	96.55	88.93
CORE-CAD	GPT	X	O	X	X	92.8	89.44
GPT-CAD	Our GPT	X	O	X	X	94.85	89.63
Human-**CADCon**	Human	TNI	O	O	O	**96.60**	90.08
CORE-**CADCon**	GPT	TNI	O	O	O	93.39	89.33
CADCon-Chat	Our GPT	Chat-Aug	O	O	O	94.19	91.46
CADCon-EDA	Our GPT	EDA	O	O	O	94.64	91.41
CADCon-Variant	Our GPT+ TNI	TNI	O	O	O	95.12	90.61
CADCon	Our GPT	TNI	O	O	O	94.75	**92.47**

Table 5: The accuracy (%) based on variations in CADCon. Our GPT refers to counterfactual samples generated by GPT-3.5 using instruction3 as a prompt, and TNI stands for Tagging-based Noise Infusion to generate positives from the original sentences. IDD represents the average accuracy on the In-Domain Dataset, and ODD represents the average accuracy on the Out-Of-Distribution Dataset.

5.2 Ablation Study

Analysis on generated CAD We evaluate our generated GPT-CAD in three metrics, as shown in Table 11. First, we measure the number of new corpora that did not appear in the original train dataset to evaluate diversity. Second, we calculate the overlap as a metric for the ratio of corpora that overlap with the original train dataset's corpora. Lastly, to examine how well the generated counterfactual sentences maintain the existing context, we use BERTScore (Zhang* et al., 2020), which computes cosine similarity between the original sentences and the generated counterfactual sentences using BERT encodings. Through these three metrics, we observe that our GPT-CAD exhibits similarity to Human-CAD, where humans manually generate counterfactual sentences. This suggests its suitability to preserve the original context while altering keywords. This tendency is evident in Table 5, where Human-CADCon and CADCon show significant performance improvement, indicating the effective application of our framework.

CAD	Diversity	Overlap (%)	BERTScore
Human-CAD	1392	92.68	0.969
CORE-CAD	498	60.15	0.914
GPT-CAD	1218	83.28	0.955

Table 6: Analysis of CAD on sentiment analysis. GPT-CAD is a counterfactually augmented dataset created by utilizing Instruction 3 in Table 10.

Analysis on CADCon As indicated in Table 5, we perform ablation studies on CADCon in a sentiment analysis task, focusing on two aspects. Firstly, CADCon demonstrates an improvement of approximately 2.84% over GPT-CAD trained by simply augmenting counterfactual sentences. This indicates that the proposed representation learning critically enhances the model's generalization ability. Secondly, to show the effectiveness of the proposed Tagging-based Noise Infusion (TNI) for generating positive samples, we compare the performance of Chat-Aug and EDA for augmenting positive samples. Of course, the performance is better than simply augmenting the data, but the proposed CADCon has the largest performance improvement, suggesting that the proposed TNI method is more effective than semantic diversity for the operation of CADCon.

6 Conclusion

We proposed CADCon, a novel approach for generating and effectively training counterfactually-augmented data (CAD). It took into account both data and model aspects to enhance robustness and addressed the problem of spurious correlation. We employed straightforward prompts to make minimal changes in the original data to create counterfactual samples, without the need for human annotators or extra models. By focusing on representation learning between the generated CAD and the original dataset, we aimed to effectively train genuine token embeddings. Additionally, we introduced the tagging-based noise infusion technique to produce positive samples which helps mitigate bias towards non-causal tokens, thus enhancing generalization capability. We demonstrated the superiority of CADCon through experiments and ablation studies.

Limitations

In this work, we utilized the GPT-3.5 model to generate the dataset. GPT-CAD for CADCon is data that flips the label of sentences without the need for human intervention or additional models. If the CAD we generate is re-labeled by humans or generated by humans, it may perform better. However, our focus is not on meticulously generating CADs but rather on verifying and analyzing how effective learning with CADs can be. Therefore, in future work, if various high-quality CADs become available, we believe that our proposed framework could be utilized, much like the performance improvement observed in Human-CADCon.

Ethics Statement

Our work will not lead to any ethical concerns. The data we used in the experiment is publicly accessible, and the dataset created directly using the GPT-3.5 model was also used only for experimental research purposes.

References

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *CoRR*, abs/1605.05362.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. 2022. C2l: Causally contrastive learning for robust text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10526–10534.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Contrastive learning for prompt-based few-shot language learners. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5577–5587, Seattle, United States. Association for Computational Linguistics.

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*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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, Portland, Oregon, USA. Association for Computational Linguistics.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13516–13524.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 573–584.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

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, Seattle, Washington, USA. Association for Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14024–14031.

Jason Wei and Kai 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, Hong Kong, China. Association for Computational Linguistics.

Jiaxin Wen, Yeshuang Zhu, Jinchao Zhang, Jie Zhou, and Minlie Huang. 2022. AutoCAD: Automatically generate counterfactuals for mitigating shortcut learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2302–2317, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony 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, Online. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Linyi Yang, Jiazheng Li, Padraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 306–316, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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, Dublin, Ireland. Association for Computational Linguistics.

A Implementation Details

All our models are implemented with Pytorch framework (Paszke et al., 2019), Huggingface trasnformers (Wolf et al., 2020), NLTK library (Bird and Loper, 2004), OpenPrompt toolkit (Ding et al., 2021). We use RoBERTa-large (Liu et al., 2019) as our PLM backbone and the batch size is 8 and the maximum sequence length is 256. Also, we run all experiments three times with different random seeds and report the mean performances. In few-shot experiments, we train only K=8 examples per class. For each number of 8-shots, we randomly sample 5 times from the training set with different random seeds and report the mean performances. For each experiment that includes a contrastive objective, we conduct a grid-based hyperparameter sweep for coefficient $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

A.1 Statistics of Counterfactual Task Dataset

Table 9 shows the statistics of the dataset used in the counterfactual task.

Task	Type	pos/sexist	neg/non-sexist
Sentiment	O-Train	856	851
	O-Test	245	243
	CF-Train	851	856
	CF-Test	243	245
Sexism	O-Train	1036	1036
	O-Test	130	130
	CF-Train	-	1036
	CF-Test	-	130

Table 7: Statistics of counterfactual task datasets.

A.2 Hyper-parameters

We set the environment for all experiments as follows: one NVIDIA 3090 GPU with 24GB graphic memory, Ubuntu 22.04, Python 3.8, and CUDA 11.7 version. As mentioned in the paper, we employ different hyperparameters, denoted as k and λ, for each dataset. Especially in the Tagging-based Noise Infusion method, the parameter k, determining the number of word tokens to which noise is added, showed significant performance improvement with a value of 8 for the CF-IMDB dataset, particularly on the out-of-distribution (OOD) dataset. Therefore, using CF-IMDB as a reference, the scaling factor α was calculated. This calculation is determined by dividing the average number of non-causal tokens, which is 45 for CF-IMDB, resulting in a value of 0.18. Consequently, we calculate the value of k for each dataset by multiplying its respective average non-causal token count with the scaling factor. Summarizing the relevant hyperparameters, they are presented in Table 8.

Dataset	k	λ
CF-IMDB (Kaushik et al., 2020)	8	0.9
Sexism (Samory et al., 2021)	1	0.3
SST2 (Socher et al., 2013)	1	0.1
IMDB (Maas et al., 2011)	8	0.9
FineFood (McAuley and Leskovec, 2013)	5	0.1

Table 8: Hyper-parameters of CADCon.

A.3 Prompt Templates for Prompt-based Fine-tuning

Table 9 shows all the pre-defined prompt templates and verbalizers used in few-shot setting.

Dataset	Template	Verbalizer
CF-IMDB	It was <mask>. <S_1>	negative/positive
Sexism	It was <mask>. <S_1>	nonsexism/sexism
SST2	It was <mask>. <S_1>	negative/positive
IMDB	It was <mask>. <S_1>	negative/positive
FineFood	It was <mask>. <S_1>	negative/positive

Table 9: Templates and verbalizer in our experiments.

B Analysis of Prompt Instructions

As mentioned in 3.1, we utilized the GPT-3.5 model to create three instructions, obtaining counterfactual sentences from the original sentences through prompts. A specific example of this is identical to Table 10. In this section, we aim to compare and analyze the performance and quality associated with each prompt instruction.

Num	Instructions
1	Please make it a negative sentence.
2	The following sentence is a positive sentence in sentiment analysis. Please make it a negative sentence.
3	The following sentence is a positive sentence in sentiment analysis. Just change a few words to make it a negative sentence while preserving the original text as much as possible.

Table 10: Example of instructions for positive samples in a sentiment analysis task.

B.1 Evaluations on CAD by Prompt Instructions

We evaluate the generated CAD using three metrics, as described in the ablation study. Additionally, we assess the performance of our CAD based on three prompt instructions. Instruction1, which simply flips labels, shows a very low word overlap of 55.26% with the original sentence. Particularly in instruction3, by incorporating the phrase "while preserving the original text as much as possible," we identify preservation of up to 83.28% of the original sentence while flipping the label. Moreover, with a diversity count of 1218, indicating the number of corpora not used in the original sentence, it can be considered the most superior CAD among the three instructions. The CAD generated with instruction3 exhibits similarity to Human-CAD, as indicated by the BERTScore metric.

CAD	Diversity	Overlap (%)	BERTScore
Human	1392	92.68	0.969
Instruction1	758	55.26	0.895
Instruction2	1183	76.91	0.934
Instruction3	1218	83.28	0.955

Table 11: Analysis of CAD with different prompt instructions on sentiment analysis. The number following "Instruction" corresponds to the instructions associated with each number used in Table 10.

Also, we conducted an ablation study on datasets generated by three different prompts. Table 2 reports the performance of applying CADCon to the datasets generated through instructions for the three different scenarios. Interestingly, we find that even in instructions where task-related information is limited, such as in CADCon1, there is a significant improvement in the ability to generalize to OOD data compared to the baseline model Roberta-large. Furthermore, the addition of task-related information in CADCon2 and the inclusion of the instruction "while preserving the original text as much as possible" in CADCon3 gradually lead to performance improvements. Particularly, CADCon3, which generates CAD with the aim of minimally flipping the label by changing only genuine tokens, proves to be the most effective in achieving robustness through representation learning. Consequently, we utilized the GPT-CAD generated with Instruction3 in all final experiments.

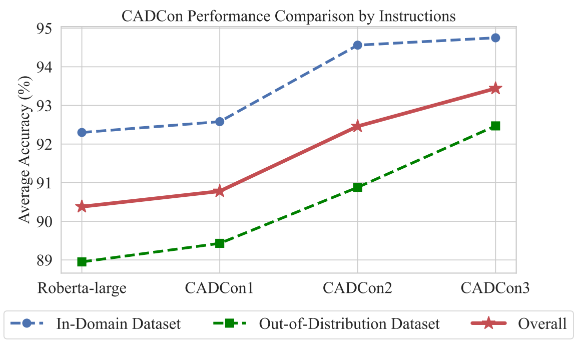

Figure 2: The performance variations of CADCon on datasets generated for each instruction. The number following "CADCon" corresponds to the instructions associated with each number used in Table 10.

C More Detail about Tagging-based Noise Infusion

In the Tagging-based Noise Infusion method, we defined the non-causal tag set G by iteratively removing each POS tag set for each dataset and calculating the importance. The following Figure 3 is an ablation study on the results of calculating importance for each dataset. We estimated θ to be 1%, defining the non-causal tag set as the part-of-speech tagging information for which the accuracy drop is less than 1%.

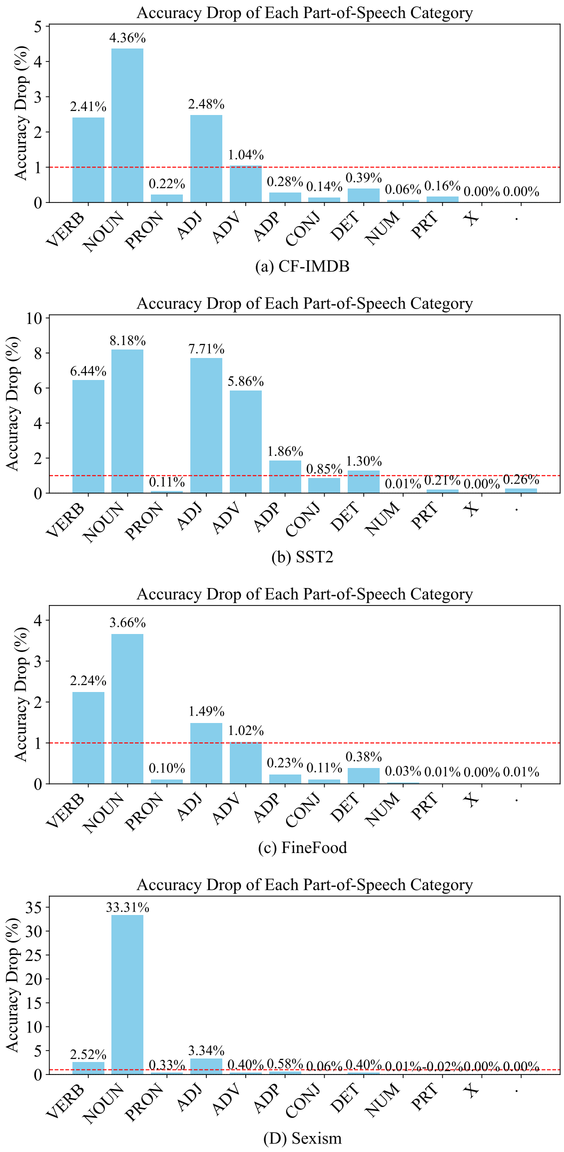

Figure 3: The accuracy drop of each part-of-speech category across datasets