CoBA: Counterbias Text Augmentation for Mitigating Various Spurious Correlations via Semantic Triples

Anonymous ACL submission

Abstract

Deep learning models often learn and exploit spurious correlations in training data, using these non-target features to inform their predictions. Such reliance leads to performance degradation and poor generalization on unseen data. To address these limitations, we introduce a more general form of counterfactual data augmentation, termed *counterbias* data augmentation, which simultaneously tackles multiple biases (e.g., gender bias, simplicity bias) and enhances out-of-distribution robustness. We present CoBA, a unified framework that operates at the semantic triple level: first decomposing text into subject-predicate-object triples, then selectively modifying these triples to disrupt spurious correlations. By reconstructing the text from these adjusted triples, CoBA generates *counterbias* data that mitigates spurious patterns. Through extensive experiments, we demonstrate that CoBA not only improves downstream task performance, but also effectively reduces biases and strengthens out-of-distribution resilience, offering a versatile and robust solution to the challenges posed by spurious correlations.

1 Introduction

Despite deep learning's success across various domains, spurious correlations continue to pose significant challenges in training effective models (Ye et al., 2024). Spurious correlations are patterns that appear in datasets but do not represent genuine relationships, such as correlations with background or textures (Beery et al., 2018; Geirhos et al., 2019; Sagawa et al., 2020). This phenomenon is also prevalent in text data, where spurious correlations frequently emerge at the word-level. In such cases, certain words or phrases become associated with specific labels due to their co-occurrence in particular contexts. This association often fails to reflect the actual meaning or intent, resulting in performance degradation in models (Wang et al., 2022; Joshi et al., 2022; Chew et al., 2024). Furthermore, spurious correlations are linked to various biases, including gender bias, simplicity bias, and challenges related to out-of-distribution (OOD) robustness. Consequently, mitigating these correlations is crucial for enhancing deep learning models in a broader context (McMilin, 2022; Liusie et al., 2022; Ming et al., 2022).

While several methods have been proposed to mitigate spurious correlations from a model-centric perspective by identifying spurious features, recent studies have shifted the focus toward a data-centric approach, particularly in the field of natural language processing (Ye et al., 2024). Early approaches suggested reweighting data samples to mitigate spurious features; however, this strategy can inadvertently introduce new biases by overemphasizing irrelevant features (Han and Tsvetkov, 2021; Shi et al., 2023). Subsequently, researchers have been exploring data manipulation techniques aimed at enhancing the generality and diversity of data distribution. These methods seek to improve model capabilities by reducing the impact of spurious correlations present in the original data (Ye et al., 2024). Recent studies suggest that augmenting datasets with counterfactual data—entailing minimal modifications to the original sentences—can effectively mitigate spurious correlations (Kaushik et al., 2020; Udomcharoenchaikit et al., 2022; Chan et al., 2023). While early studies relied on human-annotated counterfactuals, more recent works propose automatically generating them through data augmentation, demonstrating their effectiveness in reducing spurious correlations (Zeng et al., 2020; Wang and Culotta, 2021; Wen et al., 2022; Treviso et al., 2023; Sachdeva et al., 2024). However, due to the minimal modifications, this approach may lack diversity, potentially leading to issues such as overfitting and subsequent performance degradation (Qiu et al., 2024).

In this study, we extend current research on counterfactual data augmentation to *counterbias data augmentation*, which simultaneously addresses various biases and challenges, such as gender bias, simplicity bias, and out-of-distribution robustness. Although counterfactual data has been effective in mitigating spurious correlations, there remains significant potential for a unified approach that can concurrently tackle these diverse challenges. To explore this, we propose transforming the given text into a set of semantic triples using a large language model (LLM), with each triple encapsulating compressed information from the sentences. By generating counterfactual triples through modifications of the original triples and reconstructing text from these debiased triples using an LLM, we can create augmented *counterbias* data. This triple-level modification simplifies the generation of counterfactuals, as triples naturally contain the key elements of sentences. Additionally, with the support of LLMs in reconstructing text from triples, our framework can effectively diversify augmented text. *Counterbias data augmentation* differs from previous counterfactual data augmentation approaches, which aim to make minimal changes while flipping the original data's label.

Additionally, we conducted an analysis to identify principal words in various models using word importance measurements, revealing that each model has a distinct set of principal words. This finding suggests that counterbias data generated for a single model may not be effective for other models. To address this finding, we employ a majority-voting-based ensemble method to identify words that may contribute to spurious correlations. This approach is effective for augmenting counterbias data that can be universally applied across various models. Through experiments validating the effectiveness of our proposed framework, CoBA, we observed that it effectively alleviates various biases and challenges while also augmenting counterbias data applicable across different models.

Our main contributions are as follows:

- **A Unified Framework for Counterbias Augmentation:** We introduce CoBA, a novel approach that extends counterfactual data augmentation to *counterbias* data augmentation. Unlike prior methods that primarily focus on minimal label-flipping modifications, CoBA targets a broader range of biases and spurious correlations, improving both in-distribution performance and out-of-distribution robustness.

- **Insights into Spurious Correlations Across Models:** Through a detailed analysis of word importance, we reveal how spurious correlations vary significantly across different model architectures, underscoring the limitations of relying on a single model. This insight motivates our ensemble-based strategy to identify and mitigate problematic correlations more reliably.

- **Empirical Validation and Practical Benefits:** Extensive experiments across tasks like sentiment analysis, natural language inference, and text style transfer show that CoBA consistently alleviates multiple biases and enhances model resilience to distribution shifts. These results highlight CoBA's versatility and its potential to inform more robust, fair, and generalizable deep learning solutions.

2 Related Work

Counterfactual data augmentation has been shown to effectively mitigate spurious correlations. An early study introduced the concept of counterfactual data by manipulating existing data to alter the label with minimal modifications (Kaushik et al., 2020). These counterfactual data have been demonstrated to be useful for mitigating spurious patterns and precisely evaluating deep learning models, particularly with regard to local decision boundaries (Gardner et al., 2020).

Since these studies relied on human annotators to generate counterfactual data, producing such data for various datasets was challenging. As a result, researchers began exploring automated methods for generating counterfactual data, particularly in data augmentation contexts. In early explorations, predefined rules were applied to augment counterfactual data (Zmigrod et al., 2019; Wang and Culotta, 2021).

Beyond rule-based techniques, deep learning models have been employed to augment counterfactual data. For example, several studies have proposed leveraging well-trained classifiers to identify principal words (Wang et al., 2022; Wen et al., 2022; Bhan et al., 2023). Additionally, generating counterfactual data using deep learning models has proven effective in diversifying the generated data (Wu et al., 2021; Treviso et al., 2023; Sun et al.,

2024). Recently, researchers have also begun exploring the use of LLMs for counterfactual data augmentation (Sachdeva et al., 2024; Chang et al., 2024; Li et al., 2024).

3 Methodology

3.1 Overview

In this paper, we aim to alleviate various biases and obstacles by mitigating spurious correlations through counterbias data augmentation. Specifically, given an original dataset $\mathcal{D}_{ori}$, which consists of (x_i, y_i) where x_i and y_i denote the input text and its corresponding label, we aim to generate $(\hat{x}_i, \hat{y}_i)$, where $\hat{x}_i$ represents the augmented counterbias text and $\hat{y}_i$ denotes a different label from y_i. We define counterbias text as text that shares spurious words and semantics with the original text but is assigned a different label to mitigate spurious correlations, similar to counterfactual text. This represents a broader concept of counterfactual text, which refers to text with minimal differences from the original data but with different labels (Molnar, 2020). Unlike counterfactual data, counterbias data are not restricted to minimal differences; they can exhibit different syntactic structures and expressions compared to the original data, as long as they retain the spurious words and semantics of the original text. This distinction between counterbias data and counterfactual data allows counterbias data to introduce a wider variety of patterns, thereby amplifying the augmentation effect on the model.

To accomplish this, we first decompose x_i into a set of semantic triples, denoted as T_{x_i}. This T_{x_i} consists of semantic triples $t^j_{x_i} \in T_{x_i}$, each representing a triple of a sentence in x_i. A single semantic triple $t^j_{x_i}$ has the structure of (subject, predicate, object). This procedure is performed by an LLM using a designated prompt.

Next, we modify the decomposed $t^j_{x_i}$ to mitigate spurious correlations at the triple-level, resulting in modified triples $\hat{t}^j_{x_i} \in \hat{T}_{x_i}$. Specifically, we follow a step-by-step procedure as follows: **1)** We first identify sets of spurious words and principal words, W_s and W_p, representing the set of words that causes spurious correlations and the set of words that plays a crucial role in determining the label of x_i. We use multiple well-trained classifiers with different backbones and word importance measurement techniques to recognize $w_s \in W_s$ and $w_p \in W_p$. **2)** We then obtain $\hat{T}_{x_i}$ by modifying $t^j_{x_i}$ that includes w_p while maintaining $t^j_{x_i}$ involving w_s. This configuration allows us to make minimal changes that differentiate the label but maintain the spurious words, resulting in the generation of counterfactual triples. **3)** To introduce diverse patterns into the augmented data, thereby enhancing OOD robustness, we randomly permute the order and delete several triples in $\hat{T}_{x_i}$.

Finally, we augment $\hat{x}_i$ by reconstructing counterbias text from the modified $\hat{T}_{x_i}$ using the LLM with a designated prompt. Since we have modified T_{x_i} to $\hat{T}_{x_i}$ to retain $\hat{y}_i$ instead of y_i by modifying $t^j_{x_i}$ with w_p, the reconstructed $\hat{x}_i$ receives the label $\hat{y}_i$, which differs from y_i. This results in a counterbias-augmented dataset $\mathcal{D}_{cb}$, which is used to train a downstream task model in combination with the original dataset, $\mathcal{D}_{ori} \cup \mathcal{D}_{cb}$. Figure 1 illustrates this overall procedure.

	LIME	IG	SV
SST-2	26.72% (83.9%)	18.53% (83.2%)	14.21% (85.3%)
IMDB	8.64% (81.4%)	7.00% (74.6%)	7.99% (81.3%)

Table 1: The ratio of duplication among the top-5 most principal words for each model. The number in parentheses indicates the degree of overlap between two or more models, but not every models.

3.2 Analysis on Important Words

Before introducing CoBA in detail, we first present an analysis to investigate the differences in principal words across various models from two different perspectives. Implementation details are provided in the technical appendix A.1.

3.2.1 Word-level Importance Analysis

We used the three techniques mentioned above to measure the top-5 important words for each model on the SST-2 and IMDB datasets. Afterward, we evaluated the ratio of duplicated words among the important words identified by each model. Specifically, we counted the instances where all four models contained at least one duplicated word. Table 1 presents the result of this analysis. The findings suggest that the number of words consistently regarded as important across all models is small. Notably, this ratio was less than 10% of the total words in the IMDB, which contains relatively longer text compared to SST-2. Although the models used in this analysis share BERT-family architecture, they focus on different words within the input text when making predictions. However, when examining the overlap in important words between just two models at a time, we found that the majority

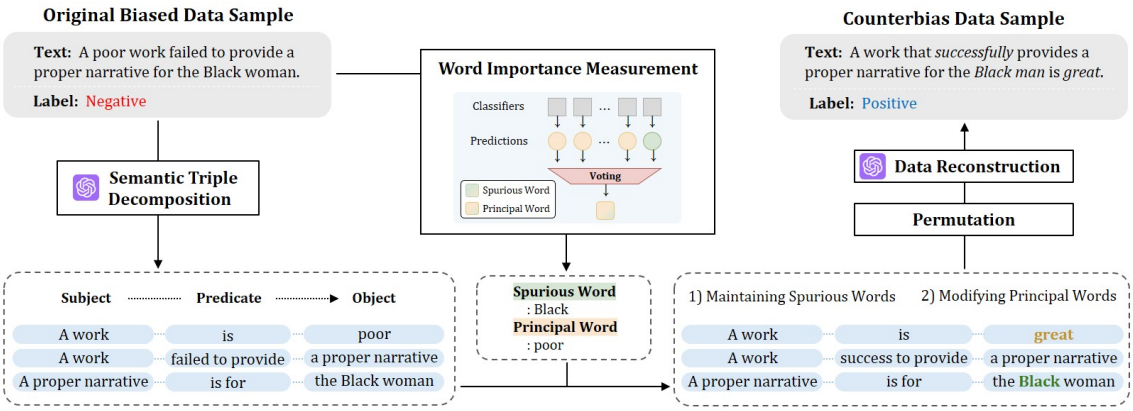

Figure 1: Overall procedure of CoBA.

SST-2	Noun	Verb	Adjective & Adverb	Others
BERT-base	**50.9**	10.1	18.1	*20.9*
BERT-large	**48.6**	2.3	*35.1*	14.0
RoBERTa-large	**67.9**	9.8	*15.6*	6.7
BART-base	**35.3**	*22.0*	21.0	21.6
IMDB	Noun	Verb	Adjective & Adverb	Others
BERT-base	*22.6*	12.2	16.0	**49.2**
BERT-large	*28.4*	12.4	8.3	**50.8**
RoBERTa-large	**46.7**	16.0	*29.1*	8.2
BART-base	*26.8*	11.3	24.9	**37.0**

Table 2: The ratio of POS tags among top-5 most important words for each model on SST-2 and IMDB. Bolded values represent the most frequent POS tag for each model and dataset, while italicized values represent the second most frequent POS tag.

of cases exceeded 80%. This indicates that while each model has its own tendencies, there is still a meaningful overlap in the patterns they recognize, suggesting that they focus on the semantics of the sentence in distinct yet related ways.

3.2.2 POS Tagging Analysis

To support the findings of the previous analysis, we conducted an additional analysis by performing POS tagging on top-5 important words identified from the analysis above. Table 2 presents the result of this analysis. The findings indicate that the important words identified by each model have different POS tags, revealing that each model focuses on different aspects of the given text. A qualitative evaluation of this tendency is provided in the technical appendix.

These two analyses suggest that counterfactual data augmented by previous methods, which leveraged a single model to identify important words from input text, may not be adequate for other models, diminishing the efficiency to be applied universally across various models. Inspired by this finding, we propose leveraging multiple models and using a majority-voting-based ensemble method to identify important words, including spurious and principal words.

3.3 CoBA

In this section, we introduce the detailed procedure of our CoBA; Counterbias Augmentation framework. CoBA consists of three major components: semantic triples decomposition, triple-level manipulation, and reconstruction of counterbias text. Each of these components will be explained in detail. The overall procedure is illustrated in Figure 1.

3.3.1 Semantic Triple Decomposition

To augment the given x_i into $\hat{x}_i$, we first decompose x_i into T_{x_i}, where T_{x_i} denotes a set of semantic triples $t^j_{x_i} \in T_{x_i}$. Each $t^j_{x_i}$ represents a sentence or phrase from x_i, and follows the structure of (subject, predicate, object). For instance, given x_i as "I love In-N-Out. Their burger feels incredibly fresh", the desired T_{x_i} is {(I, love, In-N-Out), (Their burger, feels, incredibly fresh)}. While various techniques exist for triple decomposition, they primarily focus on decomposing a *single* sentence into semantic triples, which differs from our purpose (Tan et al., 2019; Ye et al., 2021; Chen et al., 2021).

To effectively decompose text containing multiple sentences into semantic triples, we utilize LLMs, which can perform various tasks when given proper instructions through prompts (Brown et al., 2020; Ouyang et al., 2022). We achieve this by designing a dedicated prompt p_{ext} for an

LLM $\mathcal{L}$[1]. Consequently, the desired set of semantic triples T_{x_i} is obtained by $T_{x_i} = \mathcal{L}(x_i, p_{ext})$.

3.3.2 Triple-level Manipulation

Since T_{x_i} contains compressed information about the original x_i, we aim to modify this compressed T_{x_i} to mitigate the underlying spurious correlations. The procedure is detailed as follows:

First, we employ a set of multiple well-trained classifiers $\mathcal{M}$, where each $m_i \in \mathcal{M}$ represents an individual classifier trained on $\mathcal{D}_{ori}$. After training $\mathcal{M}$, we perform word importance measurement on x_i using each m_i and extract K important words, denoted as W_{m_i}. We then count the frequency of each word's appearance in the W_{m_i}. If a certain word appears in W_{m_i} more than the threshold τ[2], indicating its importance across various models, we include it in W_p, the set of principal words crucial for determining the label of x_i. Words in W_{m_i} that are not included in W_p are categorized into W_s, as they are important only for certain models and not universally significant, implying that such words may induce spurious correlations in the model. Additionally, arbitrary words that are known to introduce spurious correlations and biases can also be included in W_s if needed.

Second, we modify T_{x_i} to mitigate spurious correlations at the triple-level. Specifically, we first categorize each $t^j_{x_i}$ in T_{x_i} as a spurious triple if $t^j_{x_i}$ contains a word from W_s. Other triples that contain a word from W_p are categorized as principal triples. After categorization, we obtain $\hat{T}_{x_i}$ by modifying only the principal triples while maintaining the spurious triples. In particular, we use $\mathcal{L}$ to alter the label of x_i by modifying the principal triples, which play a crucial role in determining the label. This process results in the generation of modified principal triple, $\hat{t}^j_{x_i} = \mathcal{L}(t^j_{x_i}, \hat{y}_i, p_{mod})$, where $\hat{y}_i$ denotes the desired label different from the original y_i. This targeted manipulation preserves the spurious words and semantics of the original data while flipping the label, leading to the augmentation of counterbias data.

Finally, to effectively leverage the flexibility of counterbias data, which allows for various changes compared to the original data, such as different syntactic structures, we randomly permute the order of normal triples that are not categorized as spurious or principal triples within $\hat{T}_{x_i}$. Additionally, gender bias-inducing words are replaced with words of the opposite gender at the triple-level to mitigate gender bias. We used the WinoBias dataset (Zhao et al., 2018) to replace gender-related words. Furthermore, we randomly delete some of the normal triples with a small, predefined probability. Restricting the shuffling and deletion to normal triples helps introduce diverse patterns into the augmented data while preserving the core semantics. The completion of this process produces the final candidate set of triples for reconstruction, $\hat{T}_{x_i}$.

3.3.3 Reconstruction of Counterbias Text

Finally, we augment counterbias text $\hat{x}_i$ by reconstructing text given the processed $\hat{T}_{x_i}$. Specifically, we utilize $\mathcal{L}$ to achieve this, which is formulated as $\hat{x}_i = \mathcal{L}(\hat{T}_{x_i}, p_{rec})$. As a consequence, we obtain the counterbias data $(\hat{x}_i, \hat{y}_i)$. Note that we can easily generate multiple $\hat{x}_i$ using different configurations of decoding strategies for $\mathcal{L}$ or even different arrangements of $\hat{T}_{x_i}$. This is different from conventional counterfactual data augmentation, which is difficult to augment multiple data as they require minimal changes compared to original data.

4 Experiments

4.1 Improvement on Task Performance

We evaluated performance improvements in downstream tasks to determine if CoBA effectively mitigates spurious correlations and outperforms conventional data augmentation methods, including counterfactual data augmentation. For this purpose, we primarily used natural language inference (NLI) and sentiment analysis tasks. We used SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) for NLI tasks and SST-2 (Socher et al., 2013) and IMDB (Maas et al., 2011) for sentiment analysis tasks. Implementation details and baseline methods are provided in the appendix A.2 and appendix A.3.

Table 3 demonstrates the result of the experiment. A key finding is that CoBA outperformed other baselines, including counterfactual data augmentation methods, in most cases. Although counterfactual data augmentation methods effectively mitigate spurious patterns, they limit data diversity by introducing minimal modifications when converting labels. On the other hand, conventional data augmentation methods, particularly LLM-based methods such as GPT3Mix and AugGPT exhibit the variation in augmented data; they do not take

[1] Please refer to our technical appendix for full details on p_{ext} and other prompts.

[2] We simply set τ as $\frac{|\mathcal{M}|+1}{2}$, where $|\mathcal{M}|$ denotes the number of classifier models used. Note that $|\mathcal{M}|$ is an odd number.

	SST-2	IMDB	SNLI	MNLI
Baseline w/o Augmentation	92.8 / 94.0 / 94.5	91.5 / 91.6 / 92.3	86.2 / 86.6 / 85.4	82.4 / 84.5 / 83.8
EDA	93.1 / 93.1 / 92.9	90.8 / 91.6 / 91.6	86.8 / 86.2 / 88.8	80.6 / 81.5 / 82.6
Back-translation	93.2 / 93.5 / 89.3	91.4 / 92.2 / 88.2	87.7 / 84.1 / 88.1	83.1 / 82.8 / 83.4
C-BERT	91.9 / 94.0 / 93.2	92.1 / 91.0 / 90.8	84.4 / 89.0 / 91.2	82.1 / 84.7 / 85.4
Human-CAD	-	93.2 / 93.8 / 95.1	88.0 / *89.9* / 89.9	-
AutoCAD	**94.9** / 96.4 / 95.2	92.8 / 93.3 / 93.4	88.0 / **90.1** / 89.1	89.8 / *91.3* / **92.0**
GPT3Mix	93.2 / 95.2 / 95.3	93.9 / **94.1** / 93.9	-	-
AugGPT	94.2 / 95.4 / 95.7	92.2 / 94.0 / *94.2*	**90.3** / 87.5 / 88.9	88.7 / 87.6 / 85.1
CoBA (LLM-Identification)	*94.6* / **96.7** / *95.9*	*94.4* / *94.0* / 93.8	89.9 / 88.2 / **90.5**	*90.6* / 90.6 / 89.2
CoBA	**94.9** / *96.5* / **96.2**	**95.4** / **94.1** / **95.3**	*90.1* / **90.1** / **90.5**	**91.1** / **91.7** / *91.3*

Table 3: The comparison of downstream task performance of the model trained with each data augmentation strategy. The best performance in each group is boldfaced, and the second-best performance is italicized. The performance is presented in the form of "BERT / DeBERTaV3 / T5-Base". Note that Human-CAD only provides counterfactual datasets for IMDB and SNLI, and the official source code of GPT3Mix is limited to processing large datasets such as SNLI and MNLI.

	IMDB	SNLI
w/o Augmentation	52.3	70.2
AutoCAD	86.1	75.6
CoBA	**87.2**	**75.8**

Table 4: The comparison of models on Human-CAD test set. For this experiment, we trained BERT-base model.

spurious correlations into account. By combining the advantages of mitigating spurious correlations and generating diverse augmented data, CoBA was able to outperform other baseline methods.

Additionally, we conducted a small ablation study, where LLMs were used to identify W_s and W_p instead of using $\mathcal{M}$, a set of multiple well-trained classifiers. The results of this study are presented as "CoBA (LLM-Identification)" in Table 3. While this approach showed remarkable performance compared to other baselines, its performance improvement was smaller than that of the original CoBA. This suggests that the identification of spurious words and principal words using LLM may be less effective than our majority-voting-based ensemble method using downstream task models. We hypothesize this phenomenon arises because a single LLM may not effectively capture W_s and W_p, given the difference in important words across models, as evidenced by the analysis in Section 3.2.

4.2 Mitigation of Spurious Correlation

To verify that the effectiveness of the proposed method comes from mitigating spurious correlations rather than just data augmentation, we evaluated the Human-CAD test set, which provides human-annotated examples for assessing spurious correlation mitigation. For this evaluation, we trained BERT-base using a combination of original datasets and augmented data generated by CoBA and AutoCAD.

As shown in Table 4, the model trained without augmented data exhibited significantly lower performance on the Human-CAD test set, indicating that the baseline model is vulnerable to spurious correlations. In contrast, the models trained with AutoCAD and CoBA demonstrated more robust performance compared to the baseline model. Furthermore, CoBA, our proposed method based on counterbias augmentation, outperformed existing counterfactual data augmentation methods, underscoring its effectiveness in mitigating spurious correlations.

4.3 Alleviation of Gender Bias

To verify CoBA's effectiveness in reducing biases by mitigating related spurious correlations, we conducted an experiment focused on gender bias reduction. For this experiment, we adopted the list of gender bias-related words from a previous study (Zhao et al., 2018). By incorporating words from this list into W_s, we aim to mitigate the underlying spurious correlations, thereby alleviating gender bias in the model. To quantify the gender bias, we used two benchmarks: StereoSet (SS) (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020). For comparison, we established various baselines. BERT (Raw) refers to the original BERT model without any additional training, and BERT (IMDB) refers to the BERT model with additional pretraining on the IMDB training dataset. SentenceDe-

	SS	CrowS
BERT (Raw)	57.8	59.0
BERT (IMDB)	58.6	59.7
SentenceDebias (Liang et al., 2020)	53.8	58.1
Naive-masking (Thakur et al., 2023)	56.5	60.8
Random-phrase-masking (Thakur et al., 2023)	54.5	58.0
CoBA-based	**51.4**	**52.0**

Table 5: Comparison of gender bias across methods, as measured by SS and CrowS. A score close to 50 indicates that the model has less gender bias. Note that all models are based on `BERT-base`.

EDA	AutoCAD	AugGPT	CoBA
0.9957	0.9641	0.9658	0.9531

Table 6: Cosine similarity between the embedding vectors of the original and augmented texts for each method.

	IMDB → SST-2	IMDB → Yelp
Baseline	63.2	61.2
EDA	66.2	58.2
AutoCAD	80.1	73.6
AugGPT	83.0	71.6
ReAct (Sun et al., 2021)	84.5	75.3
CoBA	83.2	74.0

Table 7: AUROC(%) performance of the models in out-of-distribution scenario. "IMDB → SST-2" indicates a scenario where a model trained on IMDB is tested on SST-2, and "IMDB → Yelp" means the model trained on IMDB is tested on Yelp. For ReAct, we follow the reported performance from previous work (Baran et al., 2023).

Baseline	AugGPT	CoBA-based
72.10	74.50	75.70

Table 8: BLEU-4 scores (Papineni et al., 2002) for the informal-to-formal text style transfer task on the GYAFC dataset. The baseline model for this experiment is `T5-Base` (Raffel et al., 2020).

bias is a baseline method that achieves debiasing at the embedding level (Liang et al., 2020). Naive-masking and Random-phrase-masking are methods based on word-level substitution (Thakur et al., 2023). Lastly, CoBA-based refers to the BERT model with additional pretraining on the IMDB training dataset, combined with the augmented data generated by CoBA.

Table 5 shows the result of the experiment. The model trained with a combination of original and CoBA-augmented counterbias data achieved a score closest to the ideal 50, outperforming other baselines. Unlike other strategies, such as masking gender-related pronouns to neutral pronouns, our CoBA focuses on augmenting data with representations of the opposite gender, leading to a more balanced introduction of gender-related representations. As a result, CoBA contributed to mitigating spurious correlations, thereby alleviating gender bias in the model.

4.4 OOD Robustness with Diverse Augmented Data

Unlike counterfactual text augmentation, which introduces minimal modifications to alter labels, counterbias text augmentation has no such restrictions, allowing for a wider range of lexical and semantic expressions. This flexibility plays a crucial role in enhancing model performance through data augmentation (Cegin et al., 2024). To validate this, we randomly sampled 100 augmented data from IMDB using each method and measured the difference between the original and augmented data by calculating the cosine similarity produced by `BERT-base`. The results in Table 6 indicate that CoBA introduces meaningful differences in the data while preserving core semantics.

To further support the effectiveness of this diversification in augmented data, we conducted an evaluation in an OOD scenario. For this experiment, we trained a model on IMDB but tested it on SST-2 and Yelp (Zhang et al., 2015). The results of this evaluation are presented in Table 7. This evaluation suggests that our CoBA exhibits remarkable improvement in OOD robustness. While the performance gain is slightly lower than that of ReAct (Sun et al., 2021) baseline, it is important to note that ReAct is a strategy that solely focused on enhancing OOD robustness. In contrast, our CoBA offers various benefits such as mitigation of spurious correlation and other biases. In conclusion, we validated that CoBA jointly offers numerous benefits to the model, from the mitigation of spurious correlation to the improvement of OOD robustness.

4.5 Extension to Generation Tasks

In this paper, we proposed CoBA, which involves decomposing the given text into semantic triples, selecting spurious and principal triples, applying

	Original
Premise	A woman talks on a cellphone while sitting in front of blue railings that are in front of the ocean.
Hypothesis	She talks to her boyfriend about plans that night.
	Human-CAD
Premise	A woman talks on a cellphone while sitting in front of blue railings that are in front of the ocean.
Hypothesis	He has a conversation on her phone outdoors.
	CoBA
Premise	A man is sitting in front of blue railing while talking on a cellphone, with the railing positioned in front of the ocean.
Hypothesis	A man talks to his boyfriend while he is in a new car.

Table 9: The comparison of augmented data generated by Human-CAD and CoBA on SNLI.

bias-mitigation techniques, and then reconstructing the augmented text. This approach is applicable not only to classification tasks but also to text generation tasks. To verify CoBA's effectiveness in text generation, we conducted an experiment applying CoBA to a text style transfer task using the GYAFC dataset (Rao and Tetreault, 2018).

For applying CoBA to the text style transfer task, we first decomposed the given text into semantic triples. Next, we utilized $\mathcal{L}$ to identify principal triples in both formal and informal sentences.[3]. Additionally, we included the gender bias alleviation scheme introduced in Section 4.3. After permuting the order of normal triples that are not principal, we reconstructed the augmented text.

The results of this experiment, displayed in Table 8, show that CoBA, the triple-based augmentation method, exhibited a remarkable performance improvement compared to the AugGPT baseline. This underscores CoBA's extensibility to text generation tasks. We plan to investigate the strategies for identifying spurious patterns in text generation tasks in future work.

4.6 Qualitative Analysis

Table 9 compares Human-CAD and CoBA for data augmentation in the SNLI dataset. In this example, the relationship between the original premise and hypothesis is neutral, while the augmented pairs exhibit a contradiction. In the Human-CAD example, the premise remains unchanged, and the label change is introduced by altering the gender in the hypothesis. Conversely, CoBA modifies "outdoor" to "car" to change the label to a contradiction. Notably, CoBA effectively generated a contradiction pair without any human annotation. Additionally, CoBA's modification of the premise contributes to the diversity of augmented data, which likely contributes to the performance improvements shown in Table 3. Additional qualitative analysis results can be found in the appendix.

[3]Note that our purpose in this experiment is to verify the usefulness of CoBA, a triple-based augmentation method, in text generation tasks, rather than to mitigate spurious correlations in these tasks. Accordingly, we did not identify spurious triples. We leave the identification and mitigation of spurious correlations in text generation tasks as future work.

5 Conclusion

We introduced *counterbias data augmentation* as a more general and flexible extension of counterfactual data augmentation, capable of addressing multiple forms of bias and improving out-of-distribution robustness simultaneously. Through an analysis of word importance across different models, we highlighted the limitations of using a single model to identify spurious correlations. Building on these insights, we developed CoBA, a framework that leverages large language models to decompose text into semantic triples and apply triple-level modifications guided by a majority-voting-based ensemble. This approach enabled us to effectively mitigate spurious correlations, alleviate biases (such as gender and simplicity bias), and enhance OOD robustness.

Our extensive experiments demonstrated CoBA's versatility and effectiveness across various tasks, including sentiment analysis, natural language inference, and text style transfer. Unlike conventional counterfactual methods that emphasize minimal label-flipping modifications, CoBA allows for more diverse and semantically rich augmentations, leading to broader improvements in both accuracy and resilience.

Looking ahead, we plan to extend CoBA to more complex text generation scenarios, further exploring the framework's potential to mitigate spurious patterns in generated text. We envision future work refining the decomposition and reconstruction steps, optimizing the balance between information preservation and bias mitigation, and generalizing our approach to a wider range of application domains and model architectures.

Limitation

While our study demonstrates promising results in mitigating various biases and spurious correlations, several important limitations should be acknowl-

edged. First, our triple-based approach to text representation, although effective in capturing core semantic relationships, inherently entails some degree of information loss. Decomposing text into subject-predicate-object triples simplifies complex linguistic structures and potentially discards contextual nuances that could be relevant for certain tasks. This simplification, while advantageous for text manipulation and reconstruction, may inadvertently introduce new patterns that manifest as alternative forms of spurious correlations.

Another limitation arises from the trade-off between information preservation and the flexibility required for effective bias mitigation. Our method intentionally sacrifices some semantic granularity to facilitate diverse text generation and broader bias mitigation. While this approach has proven effective in our experiments, it's important to recognize that this trade-off might not be optimal for all applications or domains. Furthermore, the robustness of our majority-voting ensemble method for identifying spurious correlations, while validated in our tested scenarios, may vary when applied to different types of biases or domains beyond the scope of our current evaluation. Our analysis primarily focuses on specific biases and spurious correlations within the datasets examined. Although our framework has demonstrated success in these contexts, its generalizability to other types of biases or more complex spurious correlations remains to be fully explored.

Additionally, while we observed improvements in out-of-distribution robustness, the long-term stability of these improvements and their consistency across different application domains requires further investigation.

These limitations highlight important directions for future research. Subsequent work could explore more sophisticated methods for preserving semantic information while maintaining the flexibility necessary for bias mitigation, as well as extend the framework's applicability to a broader range of biases and domains.

References

Mateusz Baran, Joanna Baran, Mateusz Wójcik, Maciej Zięba, and Adam Gonczarek. 2023. Classical out-of-distribution detection methods benchmark in text classification tasks. In *Proceedings of ACL 2023 Student Research Workshop*, pages 119–129.

Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in terra incognita. In *Proceedings of ECCV*, pages 456–473.

Milan Bhan, Jean-Noël Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. 2023. Tigtec: Token importance guided text counterfactuals. In *Proceedings of ECML-PKDD (Part 3)*, pages 496–512.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*, pages 1877–1901.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2024. Effects of diversity incentives on sample diversity and downstream model performance in llm-based text augmentation. In *Proceedings of ACL*, pages 13148–13171.

Robin Chan, Afra Amini, and Mennatallah El-Assady. 2023. Which spurious correlations impact reasoning in nli models? a visual interactive diagnosis through data-constrained counterfactuals. In *Proceedings of ACL (Demo Track)*, pages 463–470.

Mingshan Chang, Min Yang, Qingshan Jiang, and Ruifeng Xu. 2024. Counterfactual-enhanced information bottleneck for aspect-based sentiment analysis. In *Proceedings of AAAI*, pages 17736–17744.

Zhexue Chen, Hong Huang, Bang Liu, Xuanhua Shi, and Hai Jin. 2021. Semantic and syntactic enhanced aspect sentiment triplet extraction. In *Findings of ACL*, pages 1474–1483.

Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, and Kuan-Hao Huang. 2024. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In *Findings of EACL*, pages 1013–1025.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of EMNLP*, pages 1307–1323.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2019. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of ICLR*.

Xiaochuang Han and Yulia Tsvetkov. 2021. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *Findings of EMNLP*, pages 4398–4409.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *Proceedings of ICLR*.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. In *Proceedings of EMNLP*, pages 9804–9817.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of ICLR*.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2024. Prompting large language models for counterfactual generation: An empirical study. In *Proceedings of LREC-COLING*, pages 13201–13221.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of ACL*, pages 5502–5515.

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. *arXiv preprint arXiv:1907.11692*.

Adian Liusie, Vatsal Raina, Vyas Raina, and Mark Gales. 2022. Analyzing biases to spurious correlations in text classification tasks. In *Proceedings of AACL*, pages 78–84.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of ACL 2002 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR*.

Andrew 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 ACL*, pages 142–150.

Emily McMilin. 2022. Selection bias induced spurious correlations in large language models. In *Proceedings of ICML 2022 Workshop on Spurious Correlations, Invariance and Stability*.

Yifei Ming, Hang Yin, and Yixuan Li. 2022. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of AAAI*, pages 10051–10059.

Christoph Molnar. 2020. *Interpretable machine learning*. Lulu.com.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of ACL*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of EMNLP*, pages 1953–1967.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-08-09.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS*, pages 27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Xiaoqi Qiu, Yongjie Wang, Xu Guo, Zhiwei Zeng, Yue Yu, Yuhong Feng, and Chunyan Miao. 2024. Paircfr: Enhancing model training on paired counterfactually augmented data through contrastive learning. *arXiv preprint arXiv:2406.06633*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of NAACL*, pages 129–140.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of KDD*, pages 1135–1144.

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The shapley value in machine learning. In *Proceedings of IJCAI*, pages 5572–5579.

Rachneet Sachdeva, Martin Tutek, and Iryna Gurevych. 2024. Catfood: Counterfactual augmented training for improving out-of-domain performance and calibration. In *Proceedings of EACL*, pages 1876–1898.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *Proceedings of ICLR*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of NeurIPS 2019 Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*, pages 86–96.

Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. 2023. How re-sampling helps for long-tail learning? In *Proceedings of NeurIPS*, pages 75669–75687.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642.

Yiyou Sun, Chuan Guo, and Yixuan Li. 2021. React: Out-of-distribution detection with rectified activations. In *Proceedings of NeurIPS*, pages 144–157.

Yuewen Sun, Erli Wang, Biwei Huang, Chaochao Lu, Lu Feng, Changyin Sun, and Kun Zhang. 2024. Acamda: Improving data efficiency in reinforcement learning through guided counterfactual data augmentation. In *Proceedings of AAAI*, pages 15193–15201.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of ICML*, pages 3319–3328.

Zhen Tan, Xiang Zhao, Wei Wang, and Weidong Xiao. 2019. Jointly extracting multiple triplets with multilayer translation constraints. In *Proceedings of AAAI*, pages 7080–7087.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of ACL*, pages 340–351.

Marcos Treviso, Alexis Ross, Nuno M Guerreiro, and André FT Martins. 2023. Crest: A joint framework for rationalization and counterfactual text generation. In *Proceedings of ACL*, pages 15109–15126.

Can Udomcharoenchaikit, Wuttikorn Ponwitayarat, Patomporn Payoungkhamdee, Kanruethai Masuk, Weerayut Buaphet, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Mitigating spurious correlation in natural language understanding with counterfactual inference. In *Proceedings of EMNLP*, pages 11308–11321.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in nlp models. In *Findings of NAACL*, pages 1719–1729.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of AAAI*, volume 35, pages 14024–14031.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of EMNLP*, pages 6382–6388.

Jiaxin Wen, Yeshuang Zhu, Jinchao Zhang, Jie Zhou, and Minlie Huang. 2022. Autocad: Automatically generate counterfactuals for mitigating shortcut learning. In *Findings of EMNLP*, pages 2302–2317.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of ACL*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP (Demo Track)*, pages 38–45.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of ACL*, pages 6707–6723.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Proceedings of ICCS*, pages 84–95.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Contrastive triple extraction with generative transformer. In *Proceedings of AAAI*, pages 14257–14265.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

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 EMNLP*, pages 2225–2239.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of EMNLP*, pages 7270–7280.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of NeurIPS*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL*, pages 15–20.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of ACL*, pages 1651–1661.

A Appendix

A.1 Experimental Setup for Important Words Analysis

In experiments about important word analysis in Section 3.2, we employed four models—`BERT-base`, `BERT-large` (Devlin et al., 2019), `RoBERTa-large` (Liu et al., 2019), and `DistilBERT`—to identify the top-importance words (Sanh et al., 2019). We measure word-level importance for each model using three different word importance measurement techniques: local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016), integrated gradient (IG) (Sundararajan et al., 2017), and Shapley value (SV) (Rozemberczki et al., 2022). In this experiment, we set the number of top-importance words to 5. For training the classifier, we set the batch size to 32, the initial learning rate of the AdamW optimizer (Loshchilov and Hutter, 2019) to 5e-5, the maximum token length to 300, and the maximum training epochs to 15. We selected the best checkpoint based on the accuracy of the validation set. Also, the second analysis involves performing part-of-speech (POS) tagging using NLTK (Loper and Bird, 2002) and comparing the tendencies of each model concerning the POS tags of important words. For this analysis, we employed four different models: `BERT-base`, `BERT-large` (Devlin et al., 2019), `RoBERTa-large` (Liu et al., 2019), and `DistilBERT` (Sanh et al., 2019). We trained these models on SST-2 (Socher et al., 2013) and IMDB (Maas et al., 2011) datasets.

A.2 Experimental Setup for Task Performance

We employed five models: `BERT-base`, `BERT-large` (Devlin et al., 2019), `RoBERTa-large` (Liu et al., 2019), `DistilBERT` (Sanh et al., 2019), and `BART-base` to identify w_s and w_p. In this experiment, we set the number of top-k words to 5. We used `GPT-4o-mini` (OpenAI, 2024) for triple generation and text reconstructing. The generated counterfactual data was combined with the original dataset and used to train `BERT-base`, `DeBERTaV3-base` (He et al., 2023) and `T5-base` (Raffel et al., 2020) classifiers. For training the classifiers, we set the batch size to 32, the initial learning rate of the AdamW optimizer (Loshchilov and Hutter, 2019) to 5e-5, and the maximum training epochs to 10. The best checkpoint was selected based on validation set accuracy. All experiments were conducted using the Transformers library (Wolf et al., 2020).

A.3 Baseline Methods for Task Performance

The baseline methods for comparison are as follows:

- EDA (Wei and Zou, 2019): A rule-based augmentation technique that modifies sentences through word-level modification. In this study, the modification ratio was set to 20%.
- Back-translation (Sennrich et al., 2016): An augmentation technique that translates the original sentence into a pivot language and then back-translates it into the source language.
- C-BERT (Wu et al., 2019): A strategy that leverages the contextual capabilities of the BERT model by filling in masked tokens.
- Human-CAD (Kaushik et al., 2020): This baseline uses the Human-CAD dataset, which was created by employing human annotators to generate counterfactual data from a subset of the SNLI and IMDB datasets. Specifically, we trained a model using a combination of the Human-CAD dataset and the original dataset.
- AutoCAD (Wen et al., 2022): A counterfactual data augmentation method that uses a text-infilling model.
- GPT3Mix (Yoo et al., 2021): An LLM-based augmentation technique using few-shot examples and the assignment of soft label predicted by the LLM. We used `GPT-4o-mini` for a fair comparison.
- AugGPT (Dai et al., 2023): An augmentation approach based on ChatGPT, where LLMs are prompted to generate paraphrases of original sentences. We used `GPT-4o-mini` for a fair comparison.

A.4 Prompt Setting

A.4.1 Prompt for Semantic Triple Decomposition (Sentiment Analysis)

`System:` You are a chatbot used for data augmentation. I will provide two paragraphs or internet comments for natural language understanding (NLU) tasks or sentiment analysis tasks.
`User:` Please create semantic triples for the following sentence.
Triple consists of three elements: subject, predicate, and object.

Here is an example of a sentence and its corresponding Semantic Triplet: A few people in a restaurant setting, one of them is drinking orange juice.

1. A few people | are in | a restaurant setting
2. One person | is drinking | orange juice

Here is another example of a sentence and its corresponding Semantic Triplet: A poor work that failed to provide a proper narrative for the black woman.
1. A work | is | poor
2. A work | failed to provide | a proper narrative
3. A proper narrative | is for | the black woman

Please provide no answers other than the semantic triplet. Output only the semantic triplet.

Here is an paragraph you should make a semantic triplet:
Content

A.4.2 Prompt for Reconstructing Triples into Sentences (Sentiment Analysis)

`System:` You are a chatbot used for data augmentation. Your job is reconstructing the selected triples into a sentence or paragraph.
`User:` Please create sentences for the following Triples.
Here is an example of a Semantic Triples and its corresponding reconstructed text:

1. A few people | are in | a restaurant setting
2. One person | is drinking | orange juice
Output format:
A few people in a restaurant setting, one of them is drinking orange juice.

Here is another example of a Semantic Triples and its corresponding reconstructed text:
2. I | am | a student
1. I | am | a professor
Output format:
I am a student and also a professor.

Please provide no answers other than the reconstructed text. Output only the reconstructed text. And don't consider the number of sentences in the input text.

Please follow the order of the inputs strictly as they are written. Do not consider the numbers provided in the inputs. For example:
2. I | am | a student
1. I | am | a professor
Output format:
I am a student and also a professor.
In this case, even though the sequence numbered "2" comes first numerically, ignore the numbers and generate the output starting with "I | am | a student" as shown in the example.

Here is a Semantic Triples you should make a text:
Content

A.4.3 Prompt for Semantic Triple Decomposition (Natural Language Inference)

`System:` You are a chatbot used for data augmentation. I will provide two paragraphs or internet comments for natural language understanding tasks. This natural language understanding task has a label of entailment, contradiction, or neutral.
`User:` You should creating semantic triples from the following paragraph, and select the most important semantic triples. Your task is to receive two sentences along with the label for a natural language understanding task corresponding to those sentences. For each sentence, you need to create semantic triples.
Here is an example of two input sentence and label:

sent1: A woman is walking across the street eating a banana, while a man is following with his briefcase.
sent2: An actress and her favorite assistant talk a walk in the city.
label: neutral

Here is an output example of semantic triples:

sent1:
1-1. A woman | is walking | across the street
1-2. A woman | is eating | a banana
1-3. A man | is following | a woman
1-4. A man | is carrying | a briefcase

sent2:
2-1. An actress | is walking | in the city
2-2. An actress | is with | her favorite assistant
2-3. An actress and her favorite assistant | are talking | while walking

1. A few people | are in | a restaurant setting
2. One person | is drinking | orange juice

Here is an another example of two input sentence and label:

sent1: Two women, holding food carryout containers, hug.
sent2: Two groups of rival gang members flipped each other off.
label: contradiction

Here is an output example of above example:

sent1:
1-1. Two women | are holding | food carryout containers
1-2. Two women | hug | each other

sent2:
2-1. Two groups of rival gang members | flipped | each other off

Please provide no answers other than the semantic triplet. Output only the semantic triples.

Here is an paragraph you should make a semantic triplet:
Content

A.4.4 Prompt for Reconstructing Triples into Sentences (Natural Language Inference)

`System:` You are a chatbot used for data augmentation. I will provide triples for natural language understanding tasks. This natural language understanding task has a label of entailment, contradiction, or neutral.
`User:` You should reconstruct the semantic triples into a sentence or paragraph. Don't change other triplet. Then reconstruct the semantic triples into a sentence or paragraph.
Here is an example of two input triples and label:

sent1:
1-1. An older woman | sits | at a small table
1-2. An older woman | has | orange juice
1-3. Employees | are smiling | in the background
1-4. Employees | are wearing | bright colored shirts

sent2:
2-1. A girl | flips | a burger

label: contradiction

Here is example of output:

reconstructed sent1:
An older woman sits at a small table with a glass of orange juice, while employees in bright-colored shirts smile in the background.
reconstructed sent2:
A girl flips a burger.

Here is another example of two input triples and label:

sent1:
1-1. The school | is having | a special event
1-2. The special event | is to show | American culture
1-3. American culture | deals with | other cultures in parties

sent2:
2-1. A school | is hosting | an event

Here is example of output:

reconstructed sent1:
The school is having a special event in order to show the american culture on how other cultures are dealt with in parties.
reconstructed sent2:
A school is hosting an event.

Please follow the example format exactly and only output the necessary graph triplets. Do not start with conversational phrases like "Here's" or "Sure."

Here is an semantic triples you should reconstruct:
Content