Diffusion based Counterfactual Augmentation for Dual Sentiment Classification

Dancheng Xin, Jiawei Yuan, and Yang Li [*]
College of Computer and Control Engineering,
Northeast Forestry University, Harbin 150040, China
{dcxin, jwyuan, yli}@nefu.edu.cn

Abstract

State-of-the-art NLP models have demonstrated exceptional performance across various tasks, including sentiment analysis. However, concerns have been raised about their robustness and susceptibility to systematic biases in both training and test data, which may lead to performance challenges when these models encounter out-of-distribution data in real-world applications. Although various data augmentation and adversarial perturbation techniques have shown promise in tackling these issues, prior methods such as word embedding perturbation or synonymous sentence expansion have failed to mitigate the spurious association problem inherent in the original data. Recent counterfactual augmentation methods have attempted to tackle this issue, but they have been limited by rigid rules, resulting in inconsistent context and disrupted semantics. In response to these challenges, we introduce a diffusion-based counterfactual data augmentation (DCA) framework. It utilizes an antonymous paradigm to guide the continuous diffusion model and employs reinforcement learning in combination with contrastive learning to optimize algorithms for generating counterfactual samples with high diversity and quality. Furthermore, we use a dual sentiment classifier to validate the generated antonymous samples and subsequently perform sentiment classification. Our experiments on four benchmark datasets demonstrate that DCA achieves state-of-the-art performance in sentiment classification tasks.

Keywords: counterfactual data augmentation, spurious association, continuous diffusion

1. Introduction

Sentiment analysis aims to identify and understand the emotion or sentiment expressed in text, which could include documents, sentences, or tweets. It is a foundational task in natural language processing (NLP) and has surged in popularity in recent years, due to its wide-ranging practical applications (Kertkeidkachorn and Shirai, 2023; Nzeyimana, 2023). In recent years, deep learning technology has experienced significant growth and achieved remarkable success in the field of sentiment analysis (Zhang et al., 2015; Yadav and Vishwakarma, 2020). However, the inherent complexity of human sentiments (Bravo-Marquez et al., 2014) presents an ongoing challenge in practical sentiment analysis, resulting in issues such as overfitting. This challenge leads to model failures when confronted with minor modifications in real-world examples (Zhang et al., 2020; Xing et al., 2020). Researchers have attempted to tackle these challenges by employing data augmentation and adversarial perturbation, such as augmenting training data by generating synonymous sentences (Xu et al., 2019) and introducing random noise during the word embedding phase (Croce et al., 2020). However, these methods still cannot fully resolve the problem of spurious associations (Gardner et al., 2020).

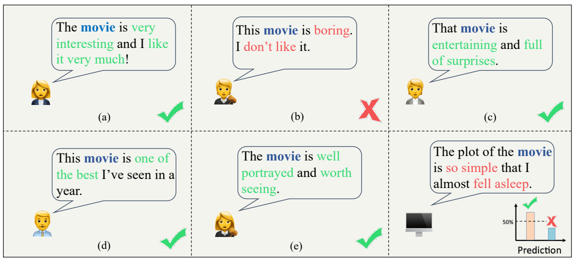

The movie is very interesting and I like it very much!
(a)
✓
This movie is boring. I don't like it.
(b)
✗
That movie is entertaining and full of surprises.
(c)
✓
This movie is one of the best I've seen in a year.
(d)
✓
The movie is well portrayed and worth seeing.
(e)
✓
The plot of the movie is so simple that I almost fell asleep.
50%
✓
✗
Prediction

Figure 1: An Illustration of spurious association in SA Tasks, where ✓ represents positive sentiment, and ✗ represents negative sentiment. If the model is trained using the first five examples, it will learn the spurious association between neural word "*movie*" and positive sentiment. Consequently, it might inaccurately classify the sentiment of the last example as positive.

Taking the word "*movie*" in Figure 1 as an example, the frequent co-occurrence of the neutral word "*movie*" with positive sentiment in the data distribution leads the model to learn a spurious association between "*movie*" and positive sentiment, resulting in inaccurate predictions. The issue of spurious patterns impacts the out-of-domain (OOD) generalization of models trained on independent and identically distributed (IID) data, causing a decline in performance when there is a shift in distribution (Sugiyama and Kawanabe, 2012; Snoek et al., 2019).

[*] Corresponding Author.

LREC-COLING 2024, pages 4901–4911
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

To address the issue of spurious association in sentiment analysis (SA), researchers have explored counterfactual data augmentation to aid models in understanding the true causal relationships between sentiment data samples and their corresponding labels (Kaushik et al., 2020; Xing et al., 2020; Wang and Culotta, 2021; Raedt et al., 2022; Wu et al., 2021; Ou et al., 2022). An essential aspect of counterfactual data augmentation is generating samples with sentiments opposite to the original ones. Nevertheless, current counterfactual methods have several limitations: (i) Most studies rely on fixed rules or known negation words, synonyms, and antonyms from WordNet to create antonymous samples, limiting the diversity and semantic consistency of the generated samples. (ii) The generated antonymous samples are merged with the original samples during training without considering their correspondence. (iii) Data generation and classification are often handled as distinct tasks, trained sequentially, leading to issues such as error accumulation.

In this paper, we present a Diffusion-based Counterfactual Augmentation (DCA) framework aimed at mitigating spurious association in sentiment analysis (SA). Our approach excels in generating counterfactual samples with exceptional diversity and fluency, distinguishing it from previous research. The framework combines a generator, a discriminator, and a dual sentiment classification model, all integrated into a framework that leverages reinforcement learning and contrastive learning techniques. In the generation phase, we start by using multi-label learning to craft optimal antonymous paradigms for the original samples. Next, we utilize contrastive learning between antonymous paradigms and original samples to instruct the diffusion model sampling process. By incorporating the reward from the discriminator, we generate diverse antonymous samples with controllable sentiment label and coherent semantics. In the sentiment classification phase, we establish a dual sentiment classifier, comprising both an original sample predictor and an antonymous sample predictor. It leverages training from the discriminator and ultimately serves as the final sentiment classifier.

The main contributions of this study can be summarized as follows:

- We propose a novel diffusion-based counterfactual data augmentation framework for sentiment generation. It guides the diffusion model in generating antonymous samples from an antonymous paradigm and collaborates with the dual sentiment classifier for sentiment analysis.
- We leverage reinforcement learning and contrastive learning to jointly optimize both the generator and the discriminator, improving the quality and diversity of the generated samples.
- We conduct experiments on four sentiment analysis (SA) benchmark datasets, demonstrating that our model outperforms state-of-the-art methods. Additionally, qualitative analysis shows that the counterfactual samples we generate exhibit high fluency and diversity.

2. Related Work

In recent years, the advent of deep learning has significantly enhanced the performance of natural language processing, including sentiment analysis (Devlin et al., 2019; Qian et al., 2023; Fan et al., 2022). Nevertheless, challenges in generalization persist when dealing with out-of-distribution (OOD) data (Xing et al., 2020; Huang et al., 2017; Ren et al., 2019; Zhang et al., 2020). To address these challenges,some research is dedicated to enhancing the robustness of neural networks. These studies primarily fall into three categories: adversarial training, causal inference, and data augmentation. The adversarial training approach introduces adversarial noise during the training process to assist the model in producing consistent output, even when faced with data perturbations (Croce et al., 2020; Miyato et al., 2017). Causal inference methods aim to improve a model's robustness by identifying the factors in the text that truly impact the sentiment. Paul (2017) emphasized establishing causal relationships between word features and sentiment labels. The primary idea behind data augmentation is to improve a model's generalization by generating more training data. In Zhang et al. (2015); Kobayashi (2018); Xu et al. (2019), synonym-based augmentation is achieved by randomly replacing synonyms, hypernyms, or hyponyms in the original samples. However, synonym augmentation methods cannot address the problem of spurious association. To this end, Kaushik et al. (2020); Srivastava et al. (2020) introduced minimal modifications to the training data, employing human annotators to perform label inversion. However, this manual approach is costly and time-consuming. Furthermore, Wang and Culotta (2021); Chen et al. (2021); Yang et al. (2021) utilized automated techniques, such as rules for antonym replacement, to generate antonymous samples.

Recently, diffusion models have been applied to controlled text generation (Li et al., 2022; Lin et al., 2023; Gong et al., 2023). Compared to traditional methods like GANs, diffusion models often exhibit a more stable training process and the

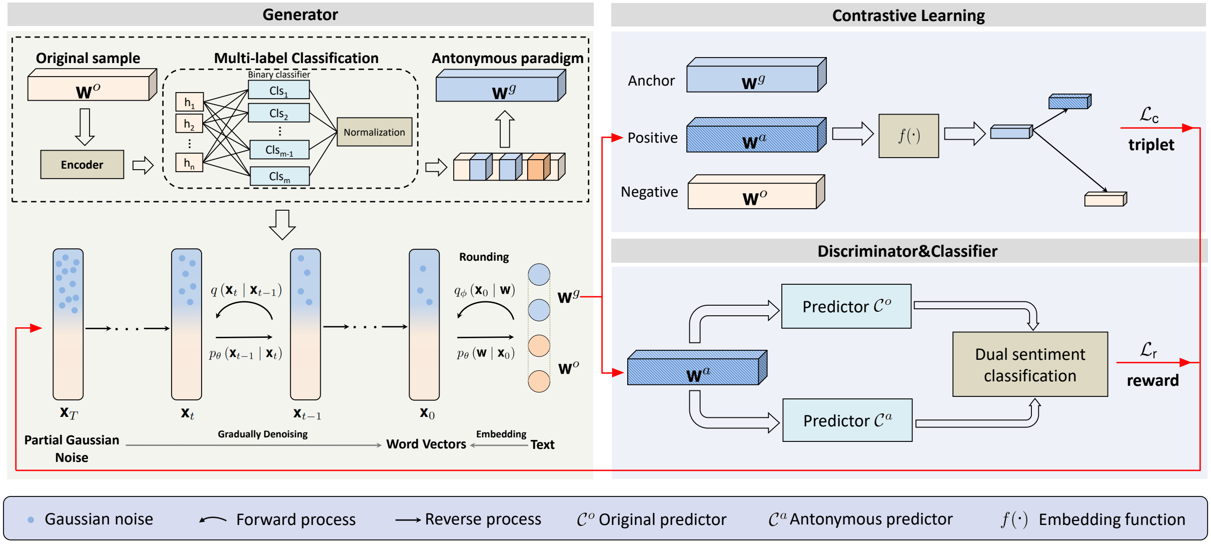


Figure 2: The architecture of diffusion-based counterfactual augmentation (DCA) framework.

ability to generate diverse content without requiring retraining. This feature is particularly valuable for content control and enhancing diversity. Building upon prior research, we propose a diffusion-based counterfactual augmentation (DCA) framework for sentiment analysis. It's essential to emphasize the significant differences between our model and the one proposed by Chen et al. (2021). The rule-based substitution they used may restrict the diversity and fluency of generated samples. Hence, we have chosen to utilize a counterfactual enhancement framework based on the diffusion model to encourage the generation of antonymous samples.

3. Methodology

The overall structure of our proposed Diffusion-based Counterfactual Augmentation (DCA) framework is shown in Figure 2, which comprises four parts: (i) diffusion based counterfactual generator, (ii) antonymous discriminator, (iii) contrastive learning, and (iv) dual sentiment prediction.

3.1. Diffusion based Counterfactual Generator

Antonymous Paradigm In this study, we focus on the sentence-level sentiment analysis (SA) task. Each sentence from the dataset is designated as the original sample. Taking inspiration from previous methods, we apply three rules to build the antonymous sentence through a combination of word substitution and a multi-label classification algorithm, designating it as the antonymous paradigm. These learned paradigms serve as guides for the diffusion model in generating semantic antonymous samples. First, we retain stop words unrelated to sentiment to prevent the intrusion of irrelevant information. Next, we replace adjectives, adverbs, and verbs that impact the sentiment of sentences with their corresponding antonyms from WordNet. Lastly, for words that do not adhere to the aforementioned rules, we substitute them with their corresponding synonyms.

Given a sequence of original samples with a length of n, denoted as $\mathbf{w}^o = \{w_1, w_2, \ldots, w_n\}$, and a vocabulary with size m, for each token w_k, we perform a supervised multi-label classification to determine whether the j-th word in the vocabulary can be a substitution for it. The substitution probabilities can be defined as:

$$p\left(y_k^j \mid w_k\right) = \frac{1}{1+\exp\left(\mathbf{W}_j \mathbf{h}_k + \mathbf{b}_j\right)}, \quad j \in [1, m] \tag{1}$$

where $\mathbf{h}_k$ is the hidden representation for token w_k, $\mathbf{W}_j$ is weight matrix and $\mathbf{b}_j$ is the bias.

Subsequently, we re-normalize the obtained probabilities by setting the probabilities of words with low likelihood and those not included in WordNet to zero.

$$P_k = \text{normalize}\left(\{p(y_k^1) = 1, \ldots, p(y_k^m) = 1\}\right) \tag{2}$$

Consequently, we have a multinomial normalized distribution denoted as $w_k \sim \text{Multinomial}(P_k)$.

Finally, we derive the antonymous paradigm $(\mathbf{w}^g, \bar{s})$ from the original sample $(\mathbf{w}^o, s)$ to guide the diffusion generation, with s representing the sentiment label of the original sample $\mathbf{w}^o$, $\bar{s}$ is the opposite sentiment label.

Generator Utilizing the reference antonymous paradigm, we train a generator to generate additional antonymous samples. When compared to traditional generative models, such as Generative

Adversarial Networks (GANs) (Goodfellow et al., 2014), diffusion models have emerged as a novel paradigm for generative models. They come with several potential advantages, particularly in the generation of high-quality text and images. Typically, a diffusion model includes both a forward and a reverse diffusion process. Let $\mathbf{x}$ represent the latent representations of the original sample ($\mathbf{w}^o$). At the initial step of the forward noise-adding process, we follow the Diffusion-LM proposed by Li et al. (2022) to map the discrete sample $\mathbf{w}^o$ into a continuous space. Specifically, we concatenate the original sample $\mathbf{w}^o$ and the antonymous paradigm $\mathbf{w}^g$ to embed them into a continuous feature space, denoted as Emb($\mathbf{w}^{o \oplus g}$).

$$q_\phi\left(\mathbf{x}_0 \mid \mathbf{w}^{o \oplus g}\right) = \mathcal{N}\left(\text{Emb}\left(\mathbf{w}^{o \oplus g}\right), \beta_0 \mathbf{I}\right) \tag{3}$$

where β_0 refers to the of noise added in the time step, and $\mathbf{I}$ is an identity matrix.

Beginning with $\mathbf{x}_0$ drawn from $q_\phi(\mathbf{x})$, we obtain a sequence of latent variables, $\mathbf{x}_1, \ldots, \mathbf{x}_T$, representing the intermediate steps in the process of diffusion. These variables are generated through a forward process, progressively introducing small amounts of Gaussian noise into the sample:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \tag{4}$$

In contrast to conventional diffusion models, which perturb $\mathbf{x}_t$ in its entirety, we introduce partial noise solely to $\mathbf{w}_t^g$. This is a crucial aspect for enabling the diffusion model to conduct conditional language modeling.

In the reverse conditional denoising process, the objective is to recover the initial $\mathbf{x}_0$ from the partially Gaussian-noised $\mathbf{x}_T$:

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta\left(\mathbf{x}_t, t\right), \Sigma_\theta\left(\mathbf{x}_t, t\right)\right) \tag{5}$$

where $\mu_\theta(\cdot)$ and $\Sigma_\theta(\cdot)$ are parameters of Gaussian distribution, learned during the training process. We replace $\mathbf{w}_{t-1}^o$ in $\mathbf{x}_{t-1}$ with $\mathbf{w}_0^o$. In each diffusion sampling step, we apply a rounding operation to the reparameterized $\mathbf{x}_t$ to project it back into the word embedding space. It can be inferred that the core of our generation process is learning the data distribution between the original samples and antonymous paradigms, rather than controlling them through numerous classifiers. Building on this strategy, our model autonomously simulates the semantic relationship between $\mathbf{w}^o$ and $\mathbf{w}^g$ by connecting the embeddings of the original sequence and the paradigm sequence, thereby jointly training over two distinct feature spaces. We compute the variational lower bound following the original diffusion process:

$$\begin{aligned} \mathcal{L}_{\text{vlb}} = \mathbb{E}_q &\left[D_{\text{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_T\right)\right)\right. \\ &+ \sum_{t=2}^{T} D_{\text{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, t\right)\right) \\ &+ D_{\text{KL}}\left(q_\phi\left(\mathbf{x}_0 \mid \mathbf{w}^{o \oplus g}\right) \| p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)\right) \\ &\left.- \log p_\theta\left(\mathbf{w}^{o \oplus g} \mid \mathbf{x}_0\right)\right] \end{aligned} \tag{6}$$

where $\mathbb{E}q$ represents the expectation over the joint distribution $q(\mathbf{x}_{0:T})$, and the last term, denoted as $\mathcal{L}_{\text{round}}$, corresponds to the rounding operation illustrated in Figure 2. Utilizing the aforementioned method, we generate the final antonymous sample ($\mathbf{w}^a, \bar{s}$) based on ($\mathbf{w}^o, s$) and ($\mathbf{w}^g, \bar{s}$).

3.2. Antonymous Discriminator

In order to evaluate the quality of the generated antonymous sample, we construct a dual discriminator, in which the original sample set $\mathcal{D}^o$ is used to train the original predictor $\mathcal{C}^o$, and the antonymous sample set $\mathcal{D}^a$ is to train the antonymous predictor $\mathcal{C}^a$. The antonymous sample $\mathbf{w}^a$ is input to both $\mathcal{C}^o$ and $\mathcal{C}^a$. They generate their respective representations through softmax layers, denoted as $\mathbf{h}^o$ and $\mathbf{h}^a$, to make dual sentiment predictions. We utilize LSTM, Bert-base, and Bert-large(Devlin et al., 2019) as our text encoders.

$$\begin{aligned} p_{ori}(s \mid \mathbf{w}^a) &= \text{softmax}\left(\mathbf{W}_{ori}\mathbf{h}^o + \mathbf{b}_{ori}\right) \\ p_{ant}(s \mid \mathbf{w}^a) &= \text{softmax}\left(\mathbf{W}_{ant}\mathbf{h}^a + \mathbf{b}_{ant}\right) \end{aligned} \tag{7}$$

where $\mathbf{W}_{ori}$ and $\mathbf{b}_{ori}$ represent the parameters of $\mathcal{C}^o$, while $\mathbf{W}_{ant}$ and $\mathbf{b}_{ant}$ correspond to $\mathcal{C}^a$. The parameters of $\mathcal{C}_a$ change during the generation process of antonymous samples, however, we prevent updates to the parameters of the original predictor by blocking gradient backpropagation.

We optimize the generator by providing reward to the generated antonymous samples through the discriminator. The predictive reward is derived from both $\mathcal{C}^o$ and $\mathcal{C}^a$:

$$r(\mathbf{w}^a) = \alpha p_{ant}(\bar{s} \mid \mathbf{w}^a) + \gamma\left(s - p_{ori}(s \mid \mathbf{w}^a)\right) \tag{8}$$

where α and γ are two inversely related trade-off parameters, and the sum of them is 1. In the cold-start phase of $\mathcal{C}^a$, the value of α incrementally increases from 0 to 1 as the performance of $\mathcal{C}^a$ improves. To ensure the quality of the generated samples, we employ a policy gradient-based approach similar to that of Chen et al. (2021). We sample M antonymous sentences for each original sample and calculate the average reward as the baseline reward, denoted as r_{avg}. We encourage the generator $\mathcal{G}$ to produce antonymous samples when $r(\mathbf{w}^a)$ exceeds the average reward r_{avg}, the loss of the discriminator can be computed as follows:

$$\mathcal{L}_{\text{r}} = -\log(r(\mathbf{w}^a) - r_{avg}) P_{\mathcal{G}}(\mathbf{w}^a \mid \mathbf{w}^o) \tag{9}$$

3.3. Contrastive Learning

To encourage the generator to produce antonymous samples that are closer to the antonymous paradigm and ensure antonymous semantics between the original sample and the antonymous paradigm, we employ a contrastive learning strategy. The antonymous paradigm $\mathbf{w}^g$ serves as the anchor, $\mathbf{w}^a$ is the generated antonymous sample, and $\mathbf{w}^o$ is the original sample.

$$\mathcal{L}_{\mathrm{c}}\left(\mathbf{w}^g, \mathbf{w}^a, \mathbf{w}^o\right) = \sum_{\mathbf{w}\in\mathcal{W}} \max\Big(0, \|f(\mathbf{w}^g) - f\left(\mathbf{w}^a\right)\|_2^2 - \|f(\mathbf{w}^g) - f\left(\mathbf{w}^o\right)\|_2^2 + \epsilon\Big) \tag{10}$$

where the parameter ϵ serves as the safety margin.

In conclusion, we derive the overall objective function by summing up the three components:

$$\mathcal{L} = \mathcal{L}_{\mathrm{vlb}} + \mathcal{L}_{\mathrm{r}} + \mathcal{L}_{\mathrm{c}} \tag{11}$$

3.4. Dual Sentiment Classification

We combine antonymous sample generation and end-to-end dual sentiment classification with a reinforcement learning and contrastive learning framework. Specifically, using the original sample $\mathbf{w}^o$ and the generated antonymous sample $\mathbf{w}^a$, we employ $\mathcal{C}^o$ and $\mathcal{C}^a$ respectively to perform dual sentiment prediction.

$$p(s\,|\,\mathbf{w}^o) = \begin{cases} p_{ori}(s\,|\,\mathbf{w}^o), & \text{if} > \min\left(\mu, p_{ant}(s\,|\,\mathbf{w}^a)\right) \\ p_{ant}(s\,|\,\mathbf{w}^a), & \text{otherwise} \end{cases} \tag{12}$$

where μ represents the confidence threshold. When the confidence of the original predictor exceeds either that of the antonymous predictor or the threshold μ, we adhere to the results of the original predictor. In the opposite case, the sentiment of the original sample $\mathbf{w}^o$ should be regarded as the sentiment opposite to $\mathbf{w}^a$.

4. Experiments

In this section, we conduct experiments to explore the following research questions: (i) Does our proposed data augmentation approach have the capability to substantially improve the sentiment analysis (SA) performance of the model? If so, how does the enhancement achieved by our approach compare to other baseline methods? (ii) Do the individual components of our framework contribute positively to the overall effectiveness of the model? (iii) Is the proposed DCA framework effective in addressing the problem of spurious association and generating samples that exhibit both high quality and diversity?

4.1. Datasets

We evaluate the proposed DCA model on four benchmark datasets for sentence-level sentiment analysis (SA). **SST-2 & SST-5** is Originating from the Stanford Sentiment Treebank (Socher et al., 2013), they categorize sentiments in movie reviews into two sets of labels for binary and 5-class classification tasks, respectively. **RT** is generated from online movie reviews (Pang and Lee, 2005), where sentiments are classified into two classes for SA tasks. **Yelp-5** is collected from the Yelp website platform and classifies sentiments into five levels, ranging from 0 to 4. We follow the data split provided by Xu et al. (2019), using 100K samples for the training set, 10K for the validation set, and 10K for the test set. We utilize the traditional classification accuracy (%) as our evaluation metric following previous study.

4.2. Experimental Settings

For the **SST-2 & SST-5** and **RT** datasets, we establish a maximum sequence length of 50. However, for the **Yelp-5** dataset, the maximum sequence length is extended to 200. For the diffusion generator, we set the embedding dimension d to 300. We set the diffusion step T to 2000, following a square-root noise schedule. Regarding the encoder LSTM, the batch size is configured as 16, the sentence sampling times M as 32, and the learning rate as $1e^{-3}$. We set the confidence threshold μ to 0.80 and 0.41 for binary and 5-class datasets, respectively. When Bert is employed as the text encoder, we modify the learning rate and batch size, setting them to $2e^{-6}$ and 32, respectively. Accordingly, μ is set to 0.52 and 0.22. The experiments are executed using NVIDIA A100 Tensor Core GPUs, and all parameters within our experiments are optimized using the Adam optimizer (Kingma and Ba, 2015).

4.3. Baselines

To conduct a comprehensive evaluation of the DCA framework, we compare it against the following state-of-the-art methods:

SynDA (Zhang et al., 2015) is a data augmentation method that is based on synonymy. It utilizes an English thesaurus from WordNet to create new samples by selectively replacing words in the original samples with their corresponding synonyms.

Back-tran (Sennrich et al., 2016) leverages a translation model to translate the original sample into various languages and then back to the source language to obtain synonymous samples.

ConDA (Kobayashi, 2018) is a contextual data augmentation technique that employs a bidirectional language model to identify synonyms for individual words. It generates adversarial samples by randomly replacing words in original samples with these model-predicted counterparts.

VAT (Miyato et al., 2017) is an adversarial training mechanism designed to enhance robustness in text classification tasks. It achieves this by strategically perturbing embeddings within recurrent neural structures to generate adversarial examples.

AGC (Wang and Culotta, 2021) utilizes WordNet to identify antonyms of the top N critical words in a corpus and generates antonymous samples through a word substitution strategy.

LexicalAT (Xu et al., 2019) employs a generator to randomly replace words with their corresponding synonyms, hyponyms, or hypernyms, generating new samples. Then, it utilizes adversarial learning to jointly optimize both the generator and discriminator.

RCDA (Chen et al., 2021) introduces an end-to-end reinforcement learning framework for generating counterfactual data. It employs the antonymous word substitution strategy to generate counterfactual data, addressing the problem of spurious association.

KATG (Shen et al., 2022) utilizes a generator-discriminator architecture that leverages prior sentences and keyword-bias sampling to generate adversarial samples.

5. Results

5.1. Main Experimental Results

In table 1, we compare our model DCA using LSTM, Bert$_{\text{base}}$, and Bert$_{\text{large}}$ as text encoders with the baseline methods on four benchmark datasets. As expected, DCA demonstrates a significant performance advantage over the competitive baselines. Moreover, compared to the current most advanced counterfactual data augmentation method, RCDA, our method has achieved comprehensive leadership on four datasets. The increase in performance is statistically significant according to the paired t-test. Specifically, (i) for the LSTM text encoder, DCA outperforms the LSTM baseline by 2.89% on SST-2, 2.66% on SST-5, 3.13% on RT, and 0.93% on Yelp-5. DCA improves the baseline methods SynDA, Back-tran, ConDA, VAT, AGC, LexicalAT, KATG, and RCDA with significant accuracy improvements on all datasets. (ii) DCA consistently outperforms all baseline methods when using the Bert$_{\text{base}}$ and Bert$_{\text{large}}$ text encoders in combination with Back-tran, AGC, LexicalAT, and RCDA across all datasets. The results presented in Table 1 clearly illustrate the consistent superiority of our DCA approach over baseline methods across different text encoders and datasets, highlighting its effectiveness in enhancing sentiment analysis performance.

Method	SST-2	SST-5	RT	Yelp-5
LSTM	80.28	39.97	76.03	61.79
+SynDA	80.30	40.20	/	/
+Back-tran	80.77	39.59	76.32	61.76
+ConDA	80.10	40.55	/	/
+VAT	81.16	37.38	75.94	59.69
+AGC	76.00	32.03	71.80	60.53
+LexicalAT	81.60	41.99	76.22	61.18
+KATG	81.90	/	/	/
+RCDA	82.97	42.35	78.87	62.44
+DCA	**83.17**	**42.63**	**79.16**	**62.72**
Method	**SST-2**	**SST-5**	**RT**	**Yelp-5**
Bert$_{\text{base}}$	91.52	53.66	87.14	66.17
+Back-tran	91.81	53.93	87.41	65.54
+AGC	89.51	52.76	85.30	65.54
+RCDA	91.18	54.02	88.23	66.57
+DCA	**92.26**	**54.34**	**88.34**	**66.73**
Method	**SST-2**	**SST-5**	**RT**	**Yelp-5**
Bert$_{\text{large}}$	92.86	55.25	88.33	66.93
+Back-tran	92.96	54.70	88.21	66.84
+AGC	93.02	53.24	87.69	66.17
+LexicalAT	93.03	53.38	88.68	67.50
+RCDA	93.30	55.62	89.07	67.41
+DCA	**93.68**	**56.33**	**89.16**	**67.76**

Table 1: Comparison of accuracy (%) results between our DCA model and the baselines on four SA benchmark datasets.

5.2. Ablation Study

We conduct ablation studies to validate the effectiveness of each component in DCA. These studies cover **C**andidate words **S**election (CS) for the antonymous paradigm, the architecture of the **D**ual Sentiment **C**lassifier (DC), the embedding of the **Contr**astive Learning (Cntr) strategy, and the Generator based on **Diff**usion (Diff). As presented in Table 2, the "w/o CS" and " w/o DC" indicate that we randomly select candidate words to generate the antonymous paradigm and only use the antonymous classifier for sentiment prediction, respectively. We find that high-quality antonymous paradigm plays a crucial role in the performance of the diffusion model, and the removal of the dual sentiment prediction structure results in a significant decline in accuracy. In the case of "w/o Cntr", it means that we omit the contrastive learning mechanism during diffusion generation. The experimental results highlight that contrastive learning effectively enhances sample quality. Furthermore, "w/o Diff" denotes that we do not utilize the diffusion-based generator and instead generate samples directly using a word substitution strat-

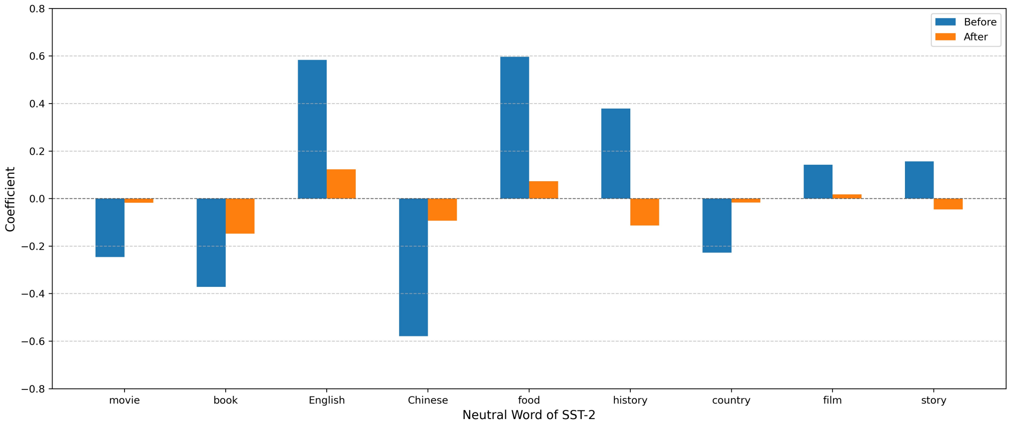


Figure 3: Sentiment polarity coefficients before and after integrating antonymous samples. "Before" refers to the initial word coefficients, and "After" indicates the updated coefficients.

LSTM	**SST-2**	**SST-5**	**RT**	**Yelp-5**
w/o CS	76.86	38.23	74.32	60.94
w/o DC	81.31	40.46	77.34	61.52
w/o Cntr	82.96	42.17	78.92	62.44
w/o Diff	82.97	42.35	78.87	62.44
DCA	**83.17**	**42.63**	**79.16**	**62.74**
Bert$_{\text{base}}$	**SST-2**	**SST-5**	**RT**	**Yelp-5**
w/o CS	79.83	39.72	77.18	63.24
w/o DC	82.13	42.36	80.42	64.39
w/o Cntr	92.14	53.98	88.26	66.19
w/o Diff	91.98	54.02	88.23	66.57
DCA	**92.26**	**54.34**	**88.34**	**66.73**
Bert$_{\text{large}}$	**SST-2**	**SST-5**	**RT**	**Yelp-5**
w/o CS	81.24	41.85	79.12	63.86
w/o DC	83.46	43.37	81.21	64.96
w/o Cntr	93.32	55.36	89.11	67.02
w/o Diff	93.30	55.62	89.07	67.41
DCA	**93.68**	**56.33**	**89.16**	**67.76**

Table 2: Ablation experimental results on four SA benchmark datasets.

Method	**SST-2**		**SST-5**	
	dis-1↑	dis-2↑	dis-1↑	dis-2↑
DSA	0.1134	0.5432	0.0937	0.5246
AGC	0.1215	0.5615	0.0984	0.1043
RCDA	0.1251	0.5677	0.1043	0.5541
DCA	**0.1285**	**0.5812**	**0.1097**	**0.5628**

Method	**RT**		**Yelp-5**	
	dis-1↑	dis-2↑	dis-1↑	dis-2↑
DSA	0.0741	0.5202	0.0098	0.1342
AGC	0.0935	0.5437	0.0106	0.1386
RCDA	0.1052	0.5553	**0.0114**	0.1431
DCA	**0.1130**	**0.5621**	0.0109	**0.1449**

Table 3: Evaluation of antonymous sentence diversity, where ↑ symbol signifies that a higher value denotes greater diversity.

egy. This leads to a comprehensive drop in experimental results. Based on this analysis, it's clear that the absence of any individual component results in a decline in DCA's performance. When considering the "w/o Cntr" scenario, it means the exclusion of the contrastive learning mechanism during the generation of diffused samples. The experimental results emphasize the effectiveness of contrastive learning in improving the quality of generated samples. On the other hand, "w/o Diff" indicates that the diffusion-based generator is not employed, and samples are generated directly using a word substitution strategy. This omission results in a significant reduction in the experimental results. Based on this analysis, it is clear that the absence of any individual component results in a decline in DCA's performance. This shows the importance of each component within the DCA framework in enhancing sentiment analysis accuracy.

5.3. Qualitative Analysis

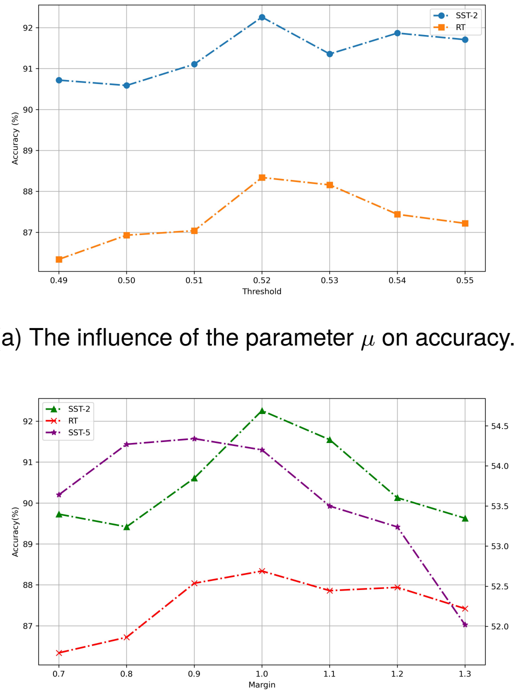
We employ distinct-N(Li et al., 2016), with N set to 1 and 2, to further assess the diversity of antonymous samples generated by different methods. As shown in Table 3, our DCA method outperforms the baseline methods, except for distinct-1 on Yelp-5 compared to RCDA. It is demonstrated that the diversity of antonymous samples produced by DCA significantly outperforms most of the baseline methods. Furthermore, both the antonymous paradigm and the generated antonymous samples are not unique, which are crucial for ensuring the diversity of generated texts.

To further investigate whether our proposed DCA method can identify intricate consistencies in the text and produce contextually coherent antony-

	Example 1	**Example 2**
Original Sample	The ring left me cold and wet like I was out in the Seattle drizzle unprotected. *Negative*	The dragons are the real stars of reign of fire and you will not be disappointed. *Positive*
Generated Sample (word substitution)	The ring left me warm and dry like I was out in the Seattle mizzle protected. *Positive*	The dragons are the unreal stars of sovereignty of firing and you will be disappointed. *Negative*
Generated Sample (ours)	The ring made me warm and happy like I was in the Seattle basking in the sunshine with joy. *Positive*	The dragons are the fake roles of the fire and you will become unsatisfied. *Negative*

Table 4: Antonymous sentences generated by word substitution strategies and DCA model.

mous samples, we provide two training samples with different sentiments in Table 4. In the first negative sample, "wet" and "drizzle" are replaced with "dry" and "mizzle" respectively. In the second positive sample, phrases like "sovereignty of firing" appear. While they do encompass the basic antonymous operations, there are also unreasonable replacements and non-standard expressions. However, samples generated by the DCA method not only capture sentiment-related features of sentences but also exhibit superior expressiveness.



(a) The influence of the parameter μ on accuracy.

(b) The influence of the parameter ϵ on accuracy.

Figure 4: Results of parameter sensitivity analysis.

5.4. Data Visualization

To further investigate the effectiveness of our method in addressing the issue of spurious association, we selected nine words with neutral connotations, as identified on Wikipedia. We conducted this analysis using the SST-2 dataset and trained logistic regression models before and after integrating antonymous samples, respectively. The sentiment polarity coefficient of each word in a sentence ranges from -1 to 1. A coefficient close to 0 indicates a likely neutral word, around 1 conveys a positive sentiment, and near -1 implies a negative sentiment. Our observations revealed a significant correction in the sentiment biases of these representative words. As shown in Figure 3, the sentiment polarity of these representative words was noticeably adjusted after data augmentation with our method, which demonstrates the effectiveness of our approach in mitigating the issue of spurious association present in the original dataset.

5.5. Parameter Sensitivity Analysis

In Figure 4, we conduct a sensitivity analysis of model parameters. We examine the variations in classification accuracy under different parameter values of μ in the dual sentiment classifier. The results presented in Figure 4a indicate that configuring the confidence threshold μ at 0.22 for binary tasks using Bert$_{\text{base}}$ improves the model's ability to control prediction selection. Additionally, we conducted an analysis on the parameter ϵ which represents the safety margin introduced in Section 3.3. As shown in Figure 4b, the model exhibits optimal classification performance when ϵ is set to 1.0. Furthermore, for tasks that demand fine-grained discrimination, a smaller value of ϵ helps the model maintain more rigorous control over each category, facilitating more precise classification.

6. Conclusion

In this article, we introduce a Diffusion-based Counterfactual Augmentation (DCA) framework to address spurious association in sentiment analysis. DCA excels in generating diverse and fluent counterfactual samples by combining a generator, a discriminator, and a dual sentiment classification model. We conducted experiments on four benchmark datasets and evaluated our approach against several state-of-the-art models. Experimental results indicate that DCA outperforms the baseline methods. Through qualitative analysis and visualization, we demonstrate that DCA improves the quality and diversity of generated counterfactual samples, effectively alleviating spurious association.

7. Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work has been supported by the National Natural Science Foundation of China (NSFC) via Grant 62276059 and the Heilongjiang Provincial Natural Science Foundation of China via Grant YQ2023F001. Corresponding author: Yang Li, E-mail: yli@nefu.edu.cn.

8. References

Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. 2014. Meta-level sentiment models for big social data analysis. *Knowledge-based systems*, 69:86–99.

Hao Chen, Rui Xia, and Jianfei Yu. 2021. Reinforced counterfactual data augmentation for dual sentiment classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 269–278. Association for Computational Linguistics.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

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.

Shuai Fan, Chen Lin, Haonan Li, Zhenghao Lin, Jinsong Su, Hang Zhang, Yeyun Gong, Jian Guo, and Nan Duan. 2022. Sentiment-aware word and sentence level pre-training for sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4984–4994. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1307–1323. Association for Computational Linguistics.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial attacks on neural network policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

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*.

Natthawut Kertkeidkachorn and Kiyoaki Shirai. 2023. Sentiment analysis using the relationship between users and products. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8611–8618. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. 2023. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 21051–21064. PMLR.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Antoine Nzeyimana. 2023. KINLP at semeval-2023 task 12: Kinyarwanda tweet sentiment analysis. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 718–723. Association for Computational Linguistics.

Jiao Ou, Jinchao Zhang, Yang Feng, and Jie Zhou. 2022. Counterfactual data augmentation via perspective transition for open-domain dialogues. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1635–1648. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124. The Association for Computer Linguistics.

Michael J. Paul. 2017. Feature selection as causal inference: Experiments with text classification. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 163–172, Vancouver, Canada. Association for Computational Linguistics.

Fan Qian, Jiqing Han, Yongjun He, Tieran Zheng, and Guibin Zheng. 2023. Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12966–12978. Association for Computational Linguistics.

Maarten De Raedt, Fréderic Godin, Chris Develder, and Thomas Demeester. 2022. Robustifying sentiment classification by maximally exploiting few counterfactuals. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11386–11400. Association for Computational Linguistics.

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

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

Lingfeng Shen, Shoushan Li, and Ying Chen. 2022. KATG: keyword-bias-aware adversarial text generation for text classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence,* *AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11294–11302. AAAI Press.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D. Sculley, Joshua V. Dillon, Jie Ren, and Zachary Nado. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13969–13980.

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.

Megha Srivastava, Tatsunori B. Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9109–9119. PMLR.

Masashi Sugiyama and Motoaki Kawanabe. 2012. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14024–14031. AAAI Press.

Tongshuang Wu, Marco Túlio Ribeiro, Jeffrey Heer, and Daniel S. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6707–6723. Association for Computational Linguistics.

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuanjing Huang. 2020. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3594–3605, Online. Association for Computational Linguistics.

Jingjing Xu, Liang Zhao, Hanqi Yan, Qi Zeng, Yun Liang, and Xu Sun. 2019. LexicalAT: Lexical-based adversarial reinforcement training for robust sentiment classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5518–5527, Hong Kong, China. Association for Computational Linguistics.

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artif. Intell. Rev.*, 53(6):4335–4385.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 306–316. Association for Computational Linguistics.

Ruiyi Zhang, Changyou Chen, Zhe Gan, Wenlin Wang, Dinghan Shen, Guoyin Wang, Zheng Wen, and Lawrence Carin. 2020. Improving adversarial text generation by modeling the distant future. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2516–2531, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.