Counterfactual-Enhanced Information Bottleneck for Aspect-Based Sentiment Analysis

Mingshan Chang[1,2], **Min Yang**[1*], **Qingshan Jiang**[1], **Ruifeng Xu**[3]
[1]Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Harbin Institute of Technology (Shenzhen)
{ms.chang, min.yang, qs.jiang}@siat.ac.cn, xuruifeng@hit.edu.cn

Abstract

Despite having achieved notable success for aspect-based sentiment analysis (ABSA), deep neural networks are susceptible to spurious correlations between input features and output labels, leading to poor robustness. In this paper, we propose a novel **C**ounterfactual-**E**nhanced **I**nformation **B**ottleneck framework (called CEIB) to reduce spurious correlations for ABSA. CEIB extends the information bottleneck (IB) principle to a factual-counterfactual balancing setting by integrating augmented counterfactual data, with the goal of learning a robust ABSA model. Concretely, we first devise a multi-pattern prompting method, which utilizes the large language model (LLM) to generate high-quality counterfactual samples from the original samples. Then, we employ the information bottleneck principle and separate the mutual information into factual and counterfactual parts. In this way, we can learn effective and robust representations for the ABSA task by balancing the predictive information of these two parts. Extensive experiments on five benchmark ABSA datasets show that our CEIB approach achieves superior prediction performance and robustness over the state-of-the-art baselines. Code and data to reproduce the results in this paper is available at: https://github.com/shesshan/CEIB.

Introduction

Aspect-based sentiment analysis (ABSA), which aims to identify the sentiment of a specific aspect in a sentence, has raised increasing interest in both academic and industrial communities (Zhang et al. 2022). For accurate and stable sentiment prediction in this fine-grained sentiment analysis task, it is essential to capture the context words expressing opinions towards the target aspect.

So far, deep learning techniques have been predominant in the ABSA task. Deep neural networks can automatically and efficiently learn discriminative contextual representations of both the context and aspect without time-consuming human annotation (Negi and Buitelaar 2014). To model the semantic relationship between the target aspect and its context, various attention mechanisms have been proposed to learn interactive features of the context and aspect (Tang, Qin, and Liu 2016; Ma et al. 2017; Lei et al. 2019). In another line, several works explicitly capture syntax-aware features for the aspect by incorporating the syntactic knowledge with graph neural networks (Huang and Carley 2019; Wang et al. 2020; Liang et al. 2022). More recently, pre-trained language models (PLMs), such as BERT (Devlin et al. 2019), have been applied to the ABSA task, yielding state-of-the-art results (Song et al. 2019; Xu et al. 2019; Jiang et al. 2019; Zhang, Zhou, and Wang 2022). Extensive linguistic knowledge learned from the large-scale textual corpus can be utilized to improve the performance for ABSA.

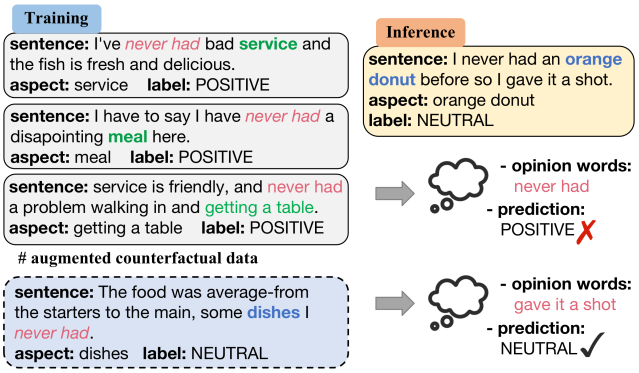

Figure 1: Spurious correlation between the context words "*never had*" and the sentiment label "POSITIVE" in the restaurant dataset. We use counterfactual data featuring identical spurious context words while different sentiment labels to encourage the model to capture vital opinion words.

Despite the effectiveness of prior studies, few efforts are devoted to mitigating the spurious correlation problem for ABSA. Specifically, deep ABSA models appear to associate superficial patterns with predicted labels, which are held by most training samples but not intrinsic to the ABSA task. For example, as shown in Figure 1, due to the high co-occurrence of the context words "*never had*" and the sentiment label "POSITIVE" in the training corpus, deep models tend to learn the strong correlation between the context words "never had" and the label "POSITIVE", rather than capturing the semantically crucial opinion expressions. As a result, models would fail to infer the ground-truth label "NEUTRAL" for the testing instance, which contains the

*Corresponding author.

words "*never had*" without holding this spurious correlation. Under such an inductive bias, models that have achieved promising performance on the in-domain data would suffer from poor robustness against the out-of-distribution or more challenging data. One possible solution to tackle this challenge is to introduce counterfactual data with the similar spurious context words while opposite sentiment labels to motivate the *counterfactual thinking* (Wang et al. 2022) ability of the ABSA model. In this way, the model can pay more attention to semantically relevant opinion words for the target aspect. In addition, incorporating the original data with the augmented counterfactual data without considering their interactions would even exacerbate the model performance. Thus, it poses a non-trivial challenge to devise a strategy to effectively exploit the interactions between factual and counterfactual data for improving the robustness of the deep ABSA model.

In light of this, we propose a **C**ounterfactual-**E**nhanced **I**nformation **B**ottleneck framework (called CEIB) to reduce spurious correlations for ABSA, aiming to improve the robustness of the deep ABSA model. The proposed CEIB framework learns a more robust model by taking benefits of both the large language model (LLM) to generate counterfactual data from the original training data and the information bottleneck (IB) principle to model interactions between the original data and augmented data. Specifically, we first devise a multi-pattern prompting method, utilizing LLM to generate high-quality counterfactual samples from the original training samples. Then, we employ the IB principle to discard spurious features from the input while retaining essential predictive information for the sentiment label. To enhance the capacity of CEIB in characterizing adversarial and out-of-distribution data, we separate the mutual information in the original IB objective into factual and counterfactual parts by leveraging the original factual sample and the generated counterfactual sample. By balancing the predictive information of these two parts, we can learn more robust and balanced representations for the ABSA task. The main contributions in this paper can be summarized as follows:

- We propose a novel CEIB framework for robust aspect-based sentiment analysis, which reduces spurious correlations by taking advantage of both the IB principle and counterfactual data augmentation, with the aim of learning a more robust ABSA model.
- We devise a multi-pattern prompting-based method, utilizing LLM to generate high-quality counterfactual data, which are then leveraged to balance the predictive information of the original training data to learn effective and robust representations.
- We conduct extensive experiments on five widely utilized benchmark ABSA datasets. Experimental results show that CEIB achieves better prediction and robustness performance compared to the strong competitors.

Related Work

Aspect-Based Sentiment Analysis

As an essential task in natural language processing, sentiment analysis is commonly studied at document-level or sentence-level, which makes distinguishing sentiment polarities of different aspects in a single document or sentence difficult. To address this limitation, aspect-based sentiment analysis (ABSA), a fine-grained sentiment analysis task, is proposed to identify the sentiment polarity towards a specific aspect within a sentence or document.

So far, deep neural networks have dominated the literature on ABSA. Earlier approaches centered around devising diverse attention mechanisms to learn attention-based representations of the context and the target aspect, which implicitly captured the semantic relationship between the given aspect and its context (Wang et al. 2016; Ma et al. 2017; Lei et al. 2019). For example, Wang et al. (2016) first proposed attention-based LSTMs to capture relevant sentiment information from the context given the target aspect. Ma et al. (2017) introduced an interactive attention to interactively learn the attention-aware representations of the target aspect and its context.

In another trend, several studies focus on explicitly capturing syntax-aware features for the target aspect by leveraging syntactic knowledge and graph neural networks (Huang and Carley 2019; Wang et al. 2020; Tian, Chen, and Song 2021; Liang et al. 2022). The key idea of these methods involves exploiting the syntactic structures, such as syntax dependency trees, to build graphs. Then, graph convolutional networks (GCNs) (Tian, Chen, and Song 2021; Liang et al. 2022) or graph attention networks (GATs) (Huang and Carley 2019; Wang et al. 2020) can be utilized to aggregate sentiment information from the syntactically adjacent nodes to the target aspect node.

More recently, the pre-trained language models (PLMs), such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019), have been applied to ABSA and yielded state-of-the-art results (Song et al. 2019; Jiang et al. 2019; Wang et al. 2020; Zhang, Zhou, and Wang 2022). These methods either employed BERT/RoBERTa as an embedding layer to acquire better initial embeddings (Wang et al. 2020; Jiang et al. 2019) or fine-tuned BERT/RoBERTa-based models by incorporating a task-specific classification layer (Xu et al. 2019). They absorbed the merit of rich linguistic and world knowledge contained in PLMs.

Spurious Correlation Reduction in NLP

Despite the preliminary success, deep neural networks are notoriously prone to learning spurious correlations between superficial feature patterns and the predicted label. For instance, models can achieve promising results in natural language inference (NLI) without capturing the semantic correlations between hypothesis and premises, due to the reliance on specific linguistic patterns in hypothesis (Gururangan et al. 2018) or superficial heuristics between the input text pairs (McCoy, Pavlick, and Linzen 2019). Similar biases have also been revealed in other tasks, including question answering (Jia and Liang 2017) and reading comprehension (Kaushik and Lipton 2018). These models are "right for the wrong reasons", which results in poor robustness when the data distribution shifts.

Existing solutions to mitigate the spurious correlation problem can be roughly grouped into two categories: (1)

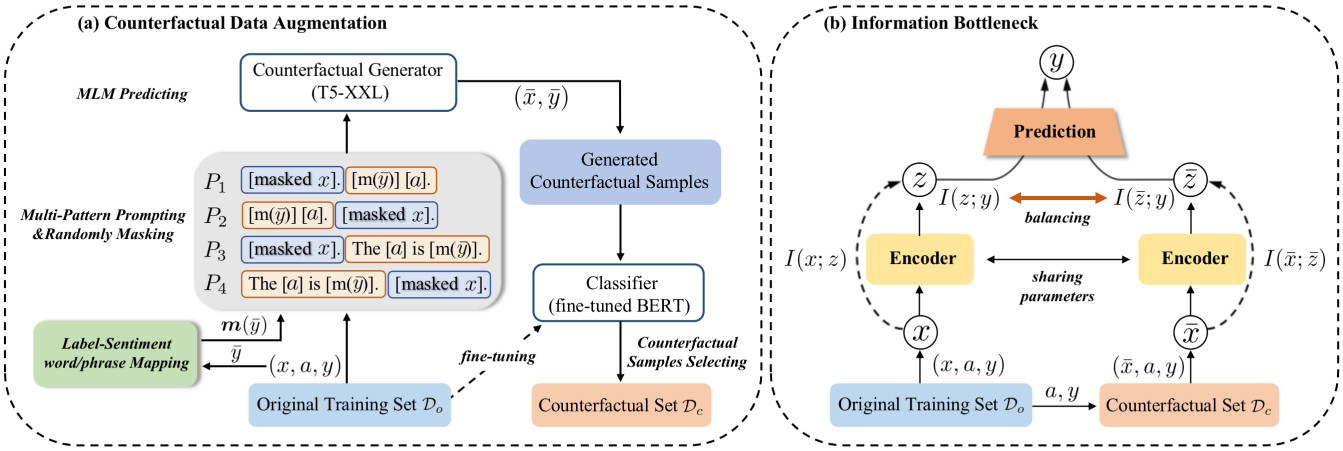

Figure 2: The overview of our CEIB, encompassing two primary modules: (a) counterfactual data augmentation module that employed LLM to generate the counterfactual data and (b) information bottleneck module with a factual-counterfactual balance setting to learn a more robust ABSA model.

data augmentation (Zellers et al. 2018; Nie et al. 2020; Wang and Culotta 2021; Wu et al. 2022), and (2) ensemble learning (Clark, Yatskar, and Zettlemoyer 2020; Stacey et al. 2020; Sanh et al. 2021; Tian et al. 2022).

The key idea of the data augmentation-based methods is to generate adversarial samples without the superficial patterns or spurious associations to alleviate the dataset bias, thus training more robust models. For instance, Zellers et al. (2018) proposed an adversarial filtering method to generate counterfactual samples and filter them in an adversarial way, which reduced spurious stylistic artifacts in the original dataset. Nie et al. (2020) augmented the original training dataset with human-written samples which exposed the model's brittleness on spurious correlations in an iterative human-in-the-loop manner. Wang and Culotta (2021) introduced a dataset de-biasing paradigm from the causal-theoretic perspective, which generated causally counterfactual data to train debiased models.

Ensemble learning-based methods proposed to leverage bias-only models to capture superficial features or shallow patterns presented in the training data, and then train a debiased model with the detected spurious correlations. For instance, Stacey et al. (2020) designed a classifier to learn the biases and discouraged the hypothesis encoder from learning them, which in turn updated the biased classifier in an adversarial learning way. Clark, Yatskar, and Zettlemoyer (2020) leveraged a low-capacity model as the bias-only model to capture simple patterns and down-weighted the corresponding loss to train a more robust model via ensemble learning. Tian et al. (2022) detected the spurious correlations in the training dataset based on the causal inference theories and incorporated a new counterfactual model with the factual model to mitigate the bias.

In this paper, we reduce spurious correlations for robust ABSA by taking benefits of both the data augmentation-based and ensemble learning-based approaches. We first generate counterfactual data where the spurious correlations do not hold in order to encourage the trained model to capture semantically relevant opinion words for the target aspect. Then, we employ the IB principle to balance the predictive information of the original factual data and the augmented counterfactual data to learn a more robust ABSA model in an ensemble manner.

Methodology

Task Description Let $\mathcal{D}_o = \{(x_i, y_i)\}_{i=1}^N$ be our original training dataset of N instances, where each instance x_i contains a context text $s = \{w_j^s\}_{j=1}^n$ with n words and a target aspect $a = \{w_j^a\}_{j=1}^m$ with m words. w_i^s (or w_i^a) denotes the i-th word in the context (or target aspect). Each instance x_i has a sentiment label y_i, where $i \in \{1, \ldots, C\}$ and C stands for the number of sentiment categories. The goal of ABSA is to predict the sentiment polarity $\hat{y}_i$ towards the aspect a given the input instance x_i.

Model Overview As illustrated in Figure 2, the proposed CEIB framework mainly consists of two modules: (a) a counterfactual data augmentation module that employs multi-pattern prompting on the large language model to generate high-quality counterfactual samples from the original training samples; (b) an information bottleneck module that leverages the augmented counterfactual data to balance the predictive information of the original factual data to learn a more robust ABSA model. Next, we will introduce these two modules in detail.

Multi-Pattern Prompting-Based Counterfactual Data Augmentation

To mitigate the spurious correlations between input features and output labels in the original dataset, we adopt counterfactual data augmentation to generate counterfactual samples for the original training samples. Inspired by (Zhou et al. 2022; Schick and Schütze 2021), we formulate the task of counterfactual data augmentation as a cloze task with

multi-pattern prompting. Then, we employ large language model to generate high-quality counterfactual data.

Counterfactual Samples Generation To enhance the counterfactual data generation, we design multiple patterns $\boldsymbol{P} = \{\boldsymbol{P}_j\}_{j=1}^{K}$ to create task-specific prompts. We first define a mapping function $\boldsymbol{m}$, which maps the sentiment label y to specific sentiment word/phrase which is consistent with the label and compatible with all aspects. Then, each pattern $\boldsymbol{P}_j$ utilizes both the sentiment word/phrase of other labels $\bar{y}$ (i.e. $\boldsymbol{m}(\bar{y})$) and the target aspect (i.e. a) to form the aspect-aware prompt. In addition, we inject the domain-specific information into the prompt text. Further, we randomly mask several tokens of the sentence x, which is then combined with the prompt to obtain the candidate.

We employ the large language model T5-XXL[1] as our counterfactual data generator by taking advantage of the consistency of our cloze-style text generation task and the pre-training "fill-in-the-blank" task of T5 (Raffel et al. 2020). Concretely, for each training sample (x_i, y_i), we feed the candidates of each pattern $\boldsymbol{P}_j$ into T5-XXL to predict the masked tokens and obtain the counterfactual samples $C_{i,j} = \{\bar{x}_{i,j}^k\}_{k=1}^{N_c}$ with other different labels $\bar{y}$.

Counterfactual Samples Selection To obtain a counterfactual set with diversity and high-quality, we first fine-tune BERT (Devlin et al. 2019) with the original training set $\mathcal{D}_o$. Then, we use the fine-tuned BERT as the vanilla ABSA classifier f to select the augmented counterfactual samples. Specifically, for each training sample (x_i, y_i), given the generated counterfactual samples $C_{i,j}$, we use f to predict the probability for each counterfactual sample and select one sample with the highest prediction score for each label $\bar{y} \neq y_i$ and add it into our counterfactual set $\mathcal{D}_c$.[2]

Information Bottleneck with Counterfactual Set

To better exploit the interactions between the original data and the augmented counterfactual data to learn a more robust model, we employ the information bottleneck (IB) principle and leverage the augmented counterfactual data to balance the predictive information of the original factual data. In this way, the learned model can capture crucial features related to the target aspect rather than the spurious features that can not generalize well.

The Encoder Given an instance x with a sequence of context words $s = \{w_i^s\}_{i=1}^{n}$ and the corresponding target aspect $a = \{w_i^a\}_{i=1}^{m}$, we adopt BERT (Devlin et al. 2019) as the text encoder and take the formatted sequence $x = (\texttt{[CLS]}s\texttt{[SEP]}a\texttt{[SEP]})$ as the input, where the special tokens [CLS] and [SEP] represent the classification token and the separation token respectively. Then, we use the representation of the [CLS] token from all output token representations H for sentiment classification:

$$H = \text{BERT}(x), \quad h = h_{\texttt{[CLS]}} \tag{1}$$

where $H = \{h_1, \ldots, h_{m+n+3}]$ represents the output hidden representations of BERT, and $h_{\texttt{[CLS]}}$ represents the representation of the [CLS] token.

Information Bottleneck The goal of the information bottleneck (IB) principle (Tishby, Pereira, and Bialek 2000) is to learn a representation z from the hidden representation (i.e. h) which is minimal to the input x while maximally informative to the label y, which can be defined as minimizing the following loss:

$$\mathcal{L}_{IB} = \beta I(x, z) - I(z, y) \tag{2}$$

where $I(\cdot)$ indicates the mutual information between two variables. The hyper-parameter β controls the compression-accuracy trade-off.

However, merely learning compressed yet informative representations for each training sample can not guarantee that the learned representations for different sentiment classes are equally informative to the target label and the model would struggle to characterize the adversarial or out-of-distribution data. Thus, we use the generated counterfactual data to learn more robust and balanced representations against the dataset bias.

Specifically, given a training sample (x, a, y) from the original set $\mathcal{D}_o$, we first randomly select a counterfactual sample $(\bar{x}, a, y)$ with the same aspect a and label y from the counterfactual set $\mathcal{D}_c$, which is generated from the other original training samples with different sentiment labels (i.e. with the counterfactual contexts). Then, we sequentially feed the factual sample and the counterfactual sample into the encoder sharing the same parameters to obtain the representations z and $\bar{z}$, respectively. Assuming that z and $\bar{z}$ are independent, the minimization and maximization term in the original IB objective can be factorized into the factual and counterfactual part:

$$\mathcal{L}_{IB} = \beta \left[I(x; z) + I(\bar{x}; \bar{z})\right] - \left[I(z; y) + I(\bar{z}; y)\right] \tag{3}$$

To balance the factual-label mutual information and counterfactual-label mutual information, the maximization term can be further refined as:

$$I(z; y) + I(\bar{z}; y) = \left[I(\bar{z}; y) - I(z; y)\right] + 2I(z; y) \tag{4}$$

Here, we derive a difference term between the factual and counterfactual part, which can regularize the confidence of the model on the factual part. That is, we use the counterfactual representation to balance the predictive information of the factual representation, to implicitly remove the spurious features from the learned representations. Then, we introduce a hyper-parameter α to control the balance-confidence trade-off and our training objective can be specified as:

$$\begin{aligned}\mathcal{L}_{CEIB} =& \beta \left[I(x; z) + I(\bar{x}; \bar{z})\right] \\ &+ \alpha \left[I(z; y) - I(\bar{z}; y)\right] - I(z; y)\end{aligned} \tag{5}$$

However, mutual information is computationally intractable for general deep neural networks, making the optimization of Eq.(5) difficult. To tackle this challenge, we apply approximation techniques to find tractable solutions for the mutual information terms of $\mathcal{L}_{\text{CEIB}}$.

[1] https://huggingface.co/t5-11b

[2] Figure 1 provides an exemplary counterfactual sample. For more details, please refer to our released full augmented data.

Minimization Term Similar to (Alemi et al. 2016), we employ the Variational Information Bottleneck (VIB) principle to derive a variational approximation for $I(x;z)$, which can be upper bounded as:

$$I(x;z) = \mathbb{E}_x[\mathbb{KL}[p(z|x)|q(z)]] \tag{6}$$

where $q(z)$ is the parametric variational approximation to $p(z)$, which is assumed to follow a normal distribution $\mathcal{N}(0,I)$. Suppose that $z = e^{\mu}(x) + \epsilon \odot e^{\sigma}(x)$, where $\epsilon \sim \mathcal{N}(0,I)$, $e^{\mu}(x)$ is the encoding of $x \in \mathbb{R}^d$ and $e^{\sigma}(x)$ is a diagonal matrix with elements $\{\sigma_i\}_{i=1}^{d}$, which is set as the zero value without using the re-parameterization trick (Kingma and Welling 2013). Then, Eq.(6) can be simplified as applying the l_2-norm regularization on the encoding z:

$$I(x;z) = \|z\|_2^2 \tag{7}$$

Maximization Term Since $I(z;y)$ is equal to $H_p(y) - H_p(y|z)$, where $H_p(\cdot)$ is the entropy of $p(\cdot)$ and $H_p(y)$ is a constant, maximizing $I(z;y)$ is equivalent to minimizing $H_p(y|z)$. Thus, we further derive a variational lower bound for $-H_p(y|z)$:

$$\begin{aligned} -H_p(y|z) &= \int_y \int_z p(y,z) \log p(y|z) dz dy \\ &\geq \int_y p(y|z) \log q(y|z) dy \\ &= -H_{p,q}(y|z) \end{aligned} \tag{8}$$

where $p(y|z)$ is variationally approximated by $q(y|z)$, with a parameterized classifier, which can be regarded as the classification process of the models.

Then, the factual-counterfactual balancing term $I(z;y) - I(\bar{z};y)$ can be further formulated as:

$$\begin{aligned} & H_{p,q}(y|\bar{z}) - H_{p,q}(y|z) \\ &= \int_y p(y|z) \log q(y|z) dy - \int_y p(y|\bar{z}) \log q(y|\bar{z}) dy \\ &= \int_{\bar{z}} \int_z \int_y p(y,z,\bar{z}) \left[\log q(y|z) - \log q(y|\bar{z})\right] dy dz d\bar{z} \\ &= \int_y p(y|z,\bar{z}) \left[\log q(y|z) - \log q(y|\bar{z})\right] dy \\ &\leq \int_y q(y|z) \left[\log q(y|z) - \log q(y|\bar{z})\right] dy \\ &= H_q(y|z;y|\bar{z}) - H_q(y|z) \end{aligned} \tag{9}$$

where we assume z and $\bar{z}$ are independent, and $p(y,z,\bar{z})$ is equal to $p(y|z,\bar{z})p(z)p(\bar{z})$. To make the term tractable, we use $q(y|z)$ as the variational estimation of $p(y|z,\bar{z})$ and the derived $H_q(y|z;y|\bar{z})$ is consistent with the goal of factual-counterfactual balancing.

Consequently, the final training objective can be formulated as:

$$\begin{aligned} \mathcal{L}_{CEIB} =& \alpha \left(H_q(y|z;y|\bar{z}) - H_q(y|z)\right) \\ &+ H_{p,q}(y|z) + \beta(\|z\|_2^2 + \|\bar{z}\|_2^2) \end{aligned} \tag{10}$$

Datasets	Division	# Pos.	# Neu.	# Neg.
REST14	Train	2164	637	807
	Test	728	196	196
LAP14	Train	994	464	870
	Test	341	169	128
REST15	Train	912	36	256
	Test	326	34	182
REST16	Train	1240	69	439
	Test	469	30	117
MAMS	Train	3380	5042	2764
	Dev	403	604	325
	Test	400	607	329
REST14-ARTS	Test	1953	473	1104
LAP14-ARTS	Test	883	407	587

Table 1: Statistics on the ABSA datasets.

Experimental Setup

Datasets

We conduct our experiments on five benchmark ABSA datasets: **REST14** and **LAP14** from (Pontiki et al. 2014), **REST15** from (Pontiki et al. 2015), **REST16** from (Pontiki et al. 2016), and **MAMS** from (Jiang et al. 2019). We adopt the official data splits, which keep the same as in the original papers. We also use **ARTS** dataset (Xing et al. 2020), including REST14-ARTS and LAP14-ARTS, to test the robustness of the ABSA models. Each instance in these datasets consists of a review sentence, a target aspect, and the sentiment polarity (i.e., POSITIVE, NEGATIVE, NEUTRAL) towards the target aspect. The detailed statistics of the utilized datasets are shown in Table 1.

Baselines and Evaluation Metrics

We compare our CEIB approach with several state-of-the-art ABSA methods based on BERT, including **BERT-SPC** (Song et al. 2019) that takes sentence-aspect pair as the input sequence of BERT to learn aspect-aware representation; **BERT-PT** (Xu et al. 2019) that post-trains BERT with external domain-specific corpus to improve the model performance; **CapsNet-BERT** (Jiang et al. 2019) that adopts the capsule network to capture semantic interactions of the aspect and its context; **DGEDT-BERT** (Tang et al. 2020) that proposes a dual-transformer network which jointly learns the flat and graph-based representations; **RGAT-BERT** (Wang et al. 2020) that adopts a relational graph attention network to encode the aspect-oriented syntactic dependency; **DualGCN-BERT** (Li et al. 2021) that designs a SynGCN module with rich syntactic knowledge and a SemGCN module to capture semantic correlations; **TGCN-BERT** (Tian, Chen, and Song 2021) that introduces syntactic dependency types into GCN and adopts attentive layer ensemble to fully exploit the type information; **SenticGCN-BERT** (Liang et al. 2022) that integrates affective knowledge into the dependency graph to learn the sentiment associations between the context and aspect; **SSEGCN-BERT** (Zhang, Zhou, and Wang 2022) that leverages both the syn-

Models	REST14 (%)		LAP14 (%)		REST15 (%)		REST16 (%)		MAMS (%)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
BERT-SPC (Song et al. 2019)[♭]	84.11	76.68	77.59	73.28	83.48	66.18	90.10	74.16	<u>83.98</u>	<u>83.41</u>
BERT-PT (Xu et al. 2019)[♯]	84.95	76.96	78.07	75.08	-	-	-	-	-	-
CapsNet-BERT (Jiang et al. 2019)[♭]	85.36	78.41	78.97	75.66	82.10	65.57	90.10	75.15	83.76	83.15
DGEDT-BERT (Tang et al. 2020)[♯]	86.30	80.00	79.80	75.60	84.00	71.00	91.90	79.00	-	-
RGAT-BERT (Wang et al. 2020)[♭]	86.60	<u>81.35</u>	78.21	74.07	83.22	69.73	89.71	76.62	82.71	82.21
DualGCN-BERT (Li et al. 2021)[♯]	87.13	81.16	81.80	78.10	-	-	-	-	-	-
TGCN-BERT (Tian et al. 2021)[♯]	86.16	79.95	80.88	77.03	85.26	<u>71.69</u>	<u>92.32</u>	77.29	83.38	82.77
SenticGCN-BERT (Liang et al. 2022)[♯]	86.92	81.03	<u>82.12</u>	<u>79.05</u>	<u>85.32</u>	71.28	91.97	<u>79.56</u>	-	-
SSEGCN-BERT (Zhang et al. 2022)[♯]	<u>87.31</u>	81.09	81.01	77.96	-	-	-	-	-	-
CEIB (Ours)	**87.77**	**82.08**	**82.92**	**79.50**	**86.16**	**72.97**	**92.86**	**81.08**	**84.95**	**84.41**
w/o IB	85.54	78.46	77.43	75.96	84.76	69.44	91.40	76.96	83.28	83.81
w/o CDA	86.25	80.00	80.94	76.62	85.54	71.80	92.00	77.74	84.06	83.63

Table 2: Main experimental results on five ABSA benchmark datasets. The results with [♯] are retrieved from the corresponding original papers and [♭] indicates our reproduced results. Best scores are in bold and the second best ones are underlined. All models are based on BERT$_{base}$.

tactic and semantic information by aspect-aware attention mechanism and syntactic mask matrices.

To evaluate the performance of the ABSA models, we adopt two widely-used metrics: Accuracy (**Acc.**) and macro-averaged F1 score (**F1**). We report the averaged scores of 5 runs with random initialization to ensure the reliability and stability our experiments.

Implementation Details

The proposed CEIB framework encompasses two parts: counterfactual data augmentation and information bottleneck. For the counterfactual data augmentation, we design 4 prompting patterns to improve the diversity of generated texts and alleviate the sensitivity to prompt templates simultaneously. The masking ratio dynamically ranges from 0.3 to 0.8 based on the length of the input sequence. That is, we set a smaller mask ratio for longer sentences. Employing these settings, we generate 10 counterfactual samples for each original training sample.

For the training stage based on the information bottleneck, we adopt the PyTorch implemenetation of uncased BERT$_{base}$[3] (12 layers and 768 hidden dimensions) as the base text encoder. We train all our models for 30 epochs. Adam is used as the optimizer with the initial learning rate as $5e^{-5}$ and the weight decay as $1e^{-4}$. The hyper-parameter α is set in range 0.5 to 1.0 and β for l_2-norm regularization is adaptively set by the optimizer.

Experimental Results and Analysis

Main Results

The experimental results of the ABSA methods on five benchmark datasets (i.e. REST14, LAP14, REST15, REST16, MAMS) are reported in Table 2. We can observe that CEIB substantially and consistently outperforms all compared baselines on the overall datasets in terms of both accuracy (Acc.) and macro-averaged F1 (F1) score, which verifies the effectiveness of our proposed approach. In particular, CEIB achieves the best improvement of 1.52% macro-F1 score compared with the best baseline (i.e. SenticGCN-BERT) on the REST16 dataset.

Compared with the competitive baselines SenticGCN-BERT and SSEGCN-BERT that take advantages of both the rich semantic knowledge from BERT and syntactic information from syntax dependency structures, our CEIB achieves better performance. The advancement is benefited from both the counterfactual data augmentation and the information bottleneck which can reduce spurious correlations while capturing the crucial contexts from the training data, thus improving the performance on the testing data.

Ablation Study

To verify the effectiveness of the counterfactual data augmentation and information bottleneck methods for robust ABSA, we conduct an ablation study on the five ABSA datasets. Specifically, we remove the counterfactual data augmentation module (denoted as "w/o CDA") from CEIB by selecting samples featuring the same aspect while different sentiment label from the original training dataset rather than the augmented counterfactual dataset. In addition, we also remove the information bottleneck module (denoted as "w/o IB") from CEIB by combining the generated counterfactual data with the original data to train the model without considering their interactions.

The ablation test results are summarized in Table 2. The performance of CEIB suffers from a sharp degradation when discarding the IB module. This is because IB enables CEIB to utilize the interactions between the factual and counterfactual data to learn effective yet robust representations. Also, CDA makes a notable contribution to CEIB, which is within our expectation since CDA can generate counterfactual data with satisfying quality and diversity to facilitate the learning

[3]https://huggingface.co/bert-base-uncased

Models	REST14-ARTS				LAP14-ARTS			
	Acc. (Ori→New)	Drop	F1 (Ori→New)	Drop	Acc. (Ori→New)	Drop	F1 (Ori→New)	Drop
BERT-SPC	84.11 → 57.81	-26.30	76.68 → 48.08	-28.60	77.59 → 58.02	-19.57	73.28 → 54.58	-18.70
CapsNet-BERT	85.36 → 69.24	-16.12	78.41 → 55.25	-23.16	78.97 → 60.31	-18.66	75.66 → 51.50	-24.16
RGAT-BERT	86.60 → 71.64	-14.96	81.35 → 60.10	-21.25	78.21 → 66.30	-11.91	74.07 → 55.68	-18.39
CEIB (Ours)	**87.40 → 80.00**	**-7.40**	**82.03 → 73.97**	**-8.06**	**81.35 → 69.18**	**-12.17**	**77.53 → 65.51**	**-12.02**

Table 3: Robustness results on ARTS test set. We compare accuracy (Acc.) and macro-averaged F1 (F1) on the original and the adversarial test sets. We also calculate the performance drop (Drop) from the original to the new sets.

of a more robust ABSA model. It is no surprise that combining both IB and CDA modules contribute a significant improvement to CEIB.

Robustness Analysis

We evaluate the robustness of our CEIB method on Aspect Robustness Test Sets (**ARTS**) (Xing et al. 2020) (statistics are shown in Table 1), which are built to test whether a ABSA model can robustly capture the aspect-relevant opinion words to distinguish the sentiment of the target aspect from the non-target aspects. ARTS augmented the original REST14 and LAP14 corpus by applying three adversarial strategies: (1) REVTGT that reverses the original sentiment of the target aspect, (2) REVNON that reverses the sentiment of the non-target aspects and (3) ADDDIFF that generates more non-target aspects with opposite sentiment polarities from the target aspect. Since CEIB can reduce spurious correlations from the non-target aspects and encourage the model to capture the crucial opinion words related to the target aspect, we assume that CEIB will show strong robustness in adversarial scenarios.

The results are shown in Table 3. We observe that CEIB substantially outperforms the compared methods when injecting adversarial perturbations, verifying the robustness of our proposed method. In particular, on the ARTS-REST14 dataset, the drop in accuracy and macro-averaged F1 scores are 7.40% and 8.06% respectively, which are much better than that produced by the baseline methods.

Long-tail Evaluation

We further evaluate the robustness of our CEIB method in the long-tail scenario. As shown in Table 1, for both REST15 and REST16 datasets, the class size of the POSITIVE class (the largest class) divided by the NEUTRAL class (the smallest class) is more than 10. Thus, we conduct experiments on these two datasets presenting an imbalanced data distribution. In Table 4, we report the averaged prediction results of three sentiment classes (i.e., POSITIVE, NEGATIVE and NEUTRAL), separately. It can be observed that CEIB achieves substantially better performance on the minority class (i.e., NEUTRAL) than the compared baselines, verifying the robustness of CEIB against the dataset imbalance and long-tail bias. The proposed CEIB can reduce spurious correlations between input features and certain class label, to learn more robust and generalizable representations for all sentiment classes, thus showing impressive generalization capability in the long-tail scenario.

Models	Pos.	Neg.	Neu.
REST15 (%)			
BERT-SPC	92.10	81.52	11.76
CapsNet-BERT	91.12	79.36	11.76
RGAT-BERT	91.03	81.50	17.65
CEIB (Ours)	**92.80**	**85.16**	**26.47**
REST16 (%)			
BERT-SPC	96.10	80.25	33.33
CapsNet-BERT	95.74	83.76	30.00
RGAT-BERT	95.95	84.62	13.33
CEIB (Ours)	**96.59**	**88.03**	**53.33**

Table 4: The performance of the ABSA models on two datasets with long-tail data distributions. Here, POSITIVE (Pos.) is the largest sentiment class and NEUTRAL (Neu.) is the smallest one.

Conclusion

In this paper, we proposed a counterfactual-enhanced information bottleneck framework (CEIB) to mitigate the spurious correlation problem for the ABSA task. CEIB learned a more robust model by taking benefits of both the counterfactual data augmentation and the IB principle. First, we devised a multi-pattern prompting method, which leveraged LLM to generate counterfactual data for the original training data. Then, we separate the mutual information in the original IB objective into factual and counterfactual parts. We can learn robust and generalizable representations by balancing the predictive information of these two parts. Consequently, the trained model could reduce spurious correlations while capturing semantically relevant opinion words for the target aspect, thus improving the robustness for ABSA. We conducted extensive experiments on five benchmark ABSA datasets. The experimental results demonstrated the effectiveness and robustness of our proposed CEIB approach.

Acknowledgments

Min Yang was supported by National Key Research and Development Program of China (2022YFF0902100), National Natural Science Foundation of China (62376262), Shenzhen Science and Technology Innovation Program (KQTD20190929172835662), Shenzhen Basic Research Foundation (JCYJ20210324115614039 and JCYJ20200109113441941). Qingshan Jiang was supported by National Key Research and Development Program of

China (2021YFF1200100 and 2021YFF1200104).

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2020. Learning to Model and Ignore Dataset Bias with Mixed Capacity Ensembles. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3031–3045. Online: Association for Computational Linguistics.

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

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 107–112. New Orleans, Louisiana: Association for Computational Linguistics.

Huang, B.; and Carley, K. 2019. Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks. 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)*, 5469–5477. Hong Kong, China: Association for Computational Linguistics.

Jia, R.; and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2021–2031. Copenhagen, Denmark: Association for Computational Linguistics.

Jiang, Q.; Chen, L.; Xu, R.; Ao, X.; and Yang, M. 2019. A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. 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)*, 6280–6285. Hong Kong, China: Association for Computational Linguistics.

Kaushik, D.; and Lipton, Z. C. 2018. How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5010–5015. Brussels, Belgium: Association for Computational Linguistics.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lei, Z.; Yang, Y.; Yang, M.; Zhao, W.; Guo, J.; and Liu, Y. 2019. A Human-Like Semantic Cognition Network for Aspect-Level Sentiment Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 6650–6657.

Li, R.; Chen, H.; Feng, F.; Ma, Z.; Wang, X.; and Hovy, E. 2021. Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6319–6329. Online: Association for Computational Linguistics.

Liang, B.; Su, H.; Gui, L.; Cambria, E.; and Xu, R. 2022. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235: 107643.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ma, D.; Li, S.; Zhang, X.; and Wang, H. 2017. Interactive Attention Networks for Aspect-Level Sentiment Classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, 4068–4074. Melbourne, Australia: AAAI Press.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. Florence, Italy: Association for Computational Linguistics.

Negi, S.; and Buitelaar, P. 2014. INSIGHT Galway: Syntactic and Lexical Features for Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 346–350.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901. Online: Association for Computational Linguistics.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; AL-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; Hoste, V.; Apidianaki, M.; Tannier, X.; Loukachevitch, N.; Kotelnikov, E.; Bel, N.; Jiménez-Zafra, S. M.; and Eryiğit, G. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 19–30. San Diego, California: Association for Computational Linguistics.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; and Androutsopoulos, I. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 486–495. Denver, Colorado: Association for Computational Linguistics.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evalu-*

ation (SemEval 2014), 27–35. Dublin, Ireland: Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Sanh, V.; Wolf, T.; Belinkov, Y.; and Rush, A. M. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

Schick, T.; and Schütze, H. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2339–2352. Online: Association for Computational Linguistics.

Song, Y.; Wang, J.; Jiang, T.; Liu, Z.; and Rao, Y. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Stacey, J.; Minervini, P.; Dubossarsky, H.; Riedel, S.; and Rocktäschel, T. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8281–8291. Online: Association for Computational Linguistics.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect Level Sentiment Classification with Deep Memory Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 214–224. Austin, Texas: Association for Computational Linguistics.

Tang, H.; Ji, D.; Li, C.; and Zhou, Q. 2020. Dependency Graph Enhanced Dual-transformer Structure for Aspect-based Sentiment Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6578–6588. Online: Association for Computational Linguistics.

Tian, B.; Cao, Y.; Zhang, Y.; and Xing, C. 2022. Debiasing NLU Models via Causal Intervention and Counterfactual Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10): 11376–11384.

Tian, Y.; Chen, G.; and Song, Y. 2021. Aspect-based Sentiment Analysis with Type-aware Graph Convolutional Networks and Layer Ensemble. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2910–2922. Online: Association for Computational Linguistics.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Wang, H.; Liang, W.; Shen, J.; Van Gool, L.; and Wang, W. 2022. Counterfactual Cycle-Consistent Learning for Instruction Following and Generation in Vision-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15471–15481.

Wang, K.; Shen, W.; Yang, Y.; Quan, X.; and Wang, R. 2020. Relational Graph Attention Network for Aspect-based Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3229–3238. Online: Association for Computational Linguistics.

Wang, Y.; Huang, M.; Zhu, X.; and Zhao, L. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 606–615. Austin, Texas: Association for Computational Linguistics.

Wang, Z.; and Culotta, A. 2021. Robustness to Spurious Correlations in Text Classification via Automatically Generated Counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16): 14024–14031.

Wu, Y.; Gardner, M.; Stenetorp, P.; and Dasigi, P. 2022. Generating Data to Mitigate Spurious Correlations in Natural Language Inference Datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2660–2676. Dublin, Ireland: Association for Computational Linguistics.

Xing, X.; Jin, Z.; Jin, D.; Wang, B.; Zhang, Q.; and Huang, X. 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)*, 3594–3605. Online: Association for Computational Linguistics.

Xu, H.; Liu, B.; Shu, L.; and Yu, P. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. 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)*, 2324–2335. Minneapolis, Minnesota: Association for Computational Linguistics.

Zellers, R.; Bisk, Y.; Schwartz, R.; and Choi, Y. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 93–104. Brussels, Belgium: Association for Computational Linguistics.

Zhang, W.; Li, X.; Deng, Y.; Bing, L.; and Lam, W. 2022. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 1–20.

Zhang, Z.; Zhou, Z.; and Wang, Y. 2022. SSEGCN: Syntactic and Semantic Enhanced Graph Convolutional Network for Aspect-based Sentiment Analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4916–4925. Seattle, United States: Association for Computational Linguistics.

Zhou, J.; Zheng, Y.; Tang, J.; Jian, L.; and Yang, Z. 2022. FlipDA: Effective and Robust Data Augmentation for Few-Shot Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8646–8665. Dublin, Ireland: Association for Computational Linguistics.