A Details on Experimental Datasets and Baselines

To verify the validity of the proposed method, a variety of text classification tasks are explored on 8 different dataset as shown in Table 1. For sentiment classification task, the following data sets are considered: Movie Review (**mr**) dataset containing binary categories [Pang and Lee, 2005]. FineFood (**foods**) [McAuley and Leskovec, 2013] for food reviews scored on a scale from 1 to 5. Following [Moon *et al.*, 2021], ratings 5 are regarded as positive and ratings 1 are regarded as negative. Stanford Sentiment Treebank (**sst2**) [Socher *et al.*, 2013] with sentence binary classification task containing human annotations in movie reviews and their emotions. Kindle reviews (**kindle**) [He and McAuley, 2016] from the Kindle Store, where each review is rated from 1 to 5. Following [Wang and Culotta, 2021], reviews with ratings $4, 5$ are positive and reviews with ratings $1, 2$ are negative.

For toxic detection, the following datasets are included: **Davidson** [Davidson *et al.*, 2017] collected from Twitter which contains three categories, hate speech, offensive or not. **OffEval** [Zampieri *et al.*, 2019] collected from Twitter which is divided into offensive and non-offensive. **ToxicTweet**[3] from Twetter, where toxic, severe toxic, obscene, threat, insult, and identity hate are marked. We choose toxic or not as our dichotomous task and obtain balanced categories by downsampling the non-toxic samples. **Abusive** from Kaggle[4] for binary abusive language detection. We treat offensive, hateful, abusive speech as toxic, and we convert toxic language detection as a binary text classification task. For all the toxic detection datasets, we delete non-English characters, web links, dates, and convert all the words to lowercase.

Baselines and Details

Baselines. The cross-domain generalization is verified by training on the source domain and testing on the target domain. Several different shortcut mitigation or automatic counterfactual agumentation approaches are compared. **Automatically Generated Counterfactuals (AGC)** [Wang and Culotta, 2021], which augments the training data with automatically generated counterfactual data by substituting causal features with the antonyms and assigning the opposite labels. Then, the augmented samples are added to the training dataset to train a robust model. **Masker** [Moon *et al.*, 2021], which improves the cross-domain generalization of language models through the keyword shortcuts reconstruction and entropy regularization. It uses tokens with high PLMs attention scores as possible shortcuts. **C2L** [Choi *et al.*, 2022], which monitors the causality of each word collectively through a set of automatically generated counterfactual samples and uses contrastive learning to improve the robustness of the model. **Drop**$_{tfidf}$ [Chao *et al.*, 2023], which performs word dropout on the original sample based on the TF-IDF score to obtain an augmented sample of semantic retention and protect important keywords.

Details. As our main experiment, we train PLMs on the training set of the source domain $\mathcal{X}_{source}^{train}$, and save the optimal models which have the best results on $\mathcal{X}_{source}^{test}$. Then, the optimal models are used to perform text attack testing and fairness testing. The batch of all datasets and all baselines is uniformly set to 64, and the learning rate is $1e-5$. We set epoch to 5 and use Adam as the optimizer. All the codes are written using pytorch and trained on four NVIDIA A40 GPUs. For the baselines, officially published codes are used to replicate the experimental results. For AGC, we identify the causal features by picking the closest opposite matches which have scores greater than 0.95 as suggested in the original paper. For MASKER, we set the weights of the two regularization terms to 0.001 and 0.0001 for cross-domain generalization. For C2L, we set the number of positive/negative pairs for comparison learning to 1, and search for the optimal weight of contrastive learning loss in $[0.1, 0.7, 1.0]$. For Drop$_{tfidf}$, we use TF-IDF Word Dropout approach which is better method according to the original paper. All the key parameters of the baselines are consistent with those reported in the original paper.

B Results on RoBERTa

RoBERTa shows results similar to BERT's in Table 4, in which our proposed ACWG outperforms various different baseline methods, and C2L can achieve suboptimal results in most cases. This shows the positive contributions of the new semantic diversity brought about by reasonable counterfactual methods to robust models.

C Gender Fairness Test

Furthermore, although our method does not specifically study fairness on minority groups, such as gender and race, robust feature learning still helps to alleviate the bias of the model [Yao *et al.*, 2021]. To verify this idea, we explore the gender bias that has been extensively studied by a set of gender attribute terms given by [Nadeem *et al.*, 2021]. If a sample contains any of the keywords in the gender attribute terms, then we assume that the sample is likely to have gender unfairness. We screen potential gender bias samples in the test sets of Davidson and ToxicTweets since they have more samples than Abusive and OffEval.

Fairness Metrics. Subsequently, the trained model is used to test fairness on the above subsets, using the following metrics. **Perturbation Consistency Rate (PCR)**. PCR is used to assess the robustness of the model to the gender perturbation of the sample, which measures the percentage of predicted results that have not changed if a gender attribute term in a sample is replaced with the opposite gender-related word. For example, if a sample '*She is a good girl*' is predicted by the model as positive, then its gender perturbation sample '*He is a good boy*' should have the same prediction result. If the results are different, the model may be gender-sensitive and make unfair judgments about *She, girl* and *He, boy*. **False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED)** [Zhang *et al.*, 2020a]. They are relaxations of Equalized Odds (also known as Error Rate

[3] https://huggingface.co/datasets/mc7232/toxictweets

[4] https://www.kaggle.com/datasets/hiungtrung/abusive-language-detection

Table 4: **RoBERTa**'s results (accuracy %) on cross-domain text classification. Bold indicates the optimal result, green indicates the average of the test results on different target domains with a fixed source domain.

Datasets		Models						Datasets		Models					
Source	Target	RoBERTa	AGC	Masker	C2L	Drop$_{tfidf}$	ACWG	Source	Target	BERT	AGC	Masker	C2L	Drop$_{tfidf}$	ACWG
mr	mr	87.96	88.89	88.60	89.39	88.74	89.05	foods	mr	65.53	72.40	71.53	77.57	78.53	80.39
	foods	78.13	86.29	75.26	85.44	81.33	85.90		foods	94.25	97.20	93.93	97.32	96.18	97.21
	sst2	93.23	93.12	93.35	92.89	93.22	93.69		sst2	75.00	74.66	77.88	81.77	76.71	84.52
	kindle	87.90	88.00	89.05	88.95	88.37	89.02		kindle	77.78	77.30	79.68	81.81	80.82	84.29
	Average	86.81	89.08	86.57	89.17	87.92	**89.42**		Average	78.14	80.39	80.76	84.62	83.06	**86.60**
sst2	mr	89.59	89.11	89.05	89.76	89.16	89.95	kinde	mr	83.03	82.89	82.36	82.40	81.62	85.31
	foods	79.73	90.23	80.00	86.78	87.30	90.16		foods	79.98	84.72	82.51	83.02	82.55	89.08
	sst2	94.15	94.04	93.69	94.15	93.62	95.30		sst2	87.50	87.65	87.27	87.67	87.39	88.53
	kindle	88.00	87.02	87.56	87.40	87.24	88.62		kindle	90.38	90.89	91.02	91.86	90.80	90.67
	Average	87.87	90.10	87.56	89.52	89.33	**91.01**		Average	85.22	86.54	85.79	86.24	85.59	**88.40**
Davidson	Davidson	95.75	96.26	96.30	96.32	96.42	96.02	OffEval	Davidson	85.80	83.93	84.75	82.19	83.81	84.44
	OffEval	79.88	80.12	78.84	79.65	79.25	80.47		OffEval	81.98	83.14	83.26	84.77	82.68	84.88
	Abusive	78.94	80.50	79.78	80.62	80.38	81.72		Abusive	79.44	81.80	80.54	84.76	82.52	84.84
	ToxicTweet	82.37	83.60	80.74	82.40	81.96	84.51		ToxicTweet	88.60	87.79	88.15	88.03	87.49	88.80
	Average	84.24	85.12	83.92	84.75	84.50	**85.68**		Average	83.96	84.17	84.18	84.94	84.13	**85.74**
Abusive	Davidson	82.44	83.23	81.68	82.15	81.83	83.39	ToxicTweet	Davidson	87.21	87.13	86.75	87.06	86.77	87.52
	OffEval	76.51	80.12	81.51	79.93	80.22	80.58		OffEval	78.95	80.00	78.60	80.70	81.54	82.09
	Abusive	93.60	92.78	92.75	91.56	92.08	95.11		Abusive	78.52	78.85	81.30	78.43	79.18	81.37
	ToxicTweet	81.76	84.31	82.53	88.22	83.55	89.69		ToxicTweet	93.40	94.09	93.15	94.49	93.95	94.59
	Average	83.58	85.11	84.62	85.47	84.42	**87.19**		Avurage	84.52	85.02	84.95	85.17	85.36	**86.39**

Balance) defined by [Borkan *et al.*, 2019] as follows:

$$FPED = \sum_z |FPR_z - FPR_{all}|, \\ FNED = \sum_z |FNR_z - FNR_{all}|, \tag{7}$$

where FPR_{all} and FNR_{all} denotes False Positive Rate and False Negative Rate in the whole test set, FPR_z and FNR_z represents the results in the corresponding gender group z, where $z = \{male, female\}$. The lower their values, the more fair the model is. **Fairness Results**. The measurement of fairness is reported in Table 5. For PCR, ACWG outperforms BERT on both Davidson and ToxicTweets, indicating that the proposed method is more stable when flipping the attributes, without misjudgment due to differences between male and female. In addition, the lower FPED and FNED also indicate that ACWG made more balanced predictions for the male/female samples, further verifying its fairness. ACWG's fairness also stems from a more explicit causal feature reflected in word-groups, since gender is not the actual cause of the model's predictions, and it is easy for ACWG to exclude the influence of such non-causal features.

Table 5: Measurement results of gender fairness compared with BERT on Davidson and ToxicTweets. ↑ indicates that the smaller the value, the higher the fairness, while ↓ is opposite.

	Davidson		ToxicTweets	
	BERT	ACWG	BERT	ACWG
PCR (% ↑)	99.10	99.51	99.04	99.22
FPED (↓)	0.0201	0.0116	0.0228	0.0181
FNED (↓)	0.1441	0.0949	0.0406	0.0389