Enhancing Robustness in Aspect-based Sentiment Analysis by Better Exploiting Data Augmentation

Anonymous ACL submission

Abstract

In this paper, we propose to leverage data augmentation to improve the robustness of aspect-based sentiment analysis models. Our method not only exploits augmented data but also makes models focus more on predictive features. We show in experiments that our method compares favorably against strong baselines on both robustness and standard datasets. In the contrary, the widely used adversarial training that only leverages the augmented data fails to improve performance due to the distribution shift caused by the augmented data.

1 Introduction

Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis task with the aim of identifying the sentiment polarity (i.e., positive, negative or neutral) for a specified aspect in a sentence. While the state-of-the-art of ABSA has been advanced significantly, typically such systems are developed and tested on those well-defined, clean corpora. More recently, there has been considerable interest in using these systems in a more practical environment. For example, Xing et al. (2020) enrich the SemEval 14 test data by introducing utterances with irrelevant aspects into each sample. Such a change to data is trivial to humans but is catastrophic to most ABSA systems. In Xing et al. (2020)'s work, even the best performing system degrades in aspect robustness score (ARS)[1] by 24% and degrades in accuracy by 6% on the new test data.

The robustness problem with ABSA is partially because of the small-sized data available to training. A simple solution to this is to leverage automatically generated samples. However, data augmentation is difficult for robust ASBA because machine-made data is noisy and does not align well with human utterances. Table 1 shows two data augmentation examples together with the original data. One example is from the ARTS benchmark inside which all data is auto-generated and is followed by manual revision; the other is from our fully auto-generated data. The ARTS data is obviously much more fluent and natural than ours with no checks or revisions from humans. Consequently, there would be some distribution shift between training and test if one learns an ABSA model using auto-generated data but tests it on natural language-like data (as in ARTS).

Data source	Instance
Original	3D rendering slows it down considerably.
ARTS	3D rendering slows it down considerably, but keyboard is a love, battery life is amazing and quality is a superlative.
Ours	3D rendering slows it down considerably, but for the price, I was very pleased with the condition and the overall product and my new Toshiba works great on both.

Table 1: A sample from the SemEval 14 Laptop testset, its AddDiff + manual revision counterpart from ARTS and a sample generated by our reimplementation of AddDiff (aspects are underlined).

We note that, despite significant development effort, we were not able to consistently improve our ABSA system on either the ARTS data or the standard ABSA data by using adversarial training on both original and auto-augmented data. This result agrees with a previous finding that adversarial samples occasionally harms NLP systems when one collects them in different annotation schemas (Huang et al., 2020).

In this paper, we investigate how to better exploit data augmentation for robust ABSA. Our work is motivated by an intuition: the auto-generated utterances will not change the prediction if they are irrelevant to the target aspect. In response, we take the difference in predictions as a regularization factor when switching from the original data to the

[1] ARS is a strict measure for robustness: a model is considered handling one question type correctly only if all the variations of that question type are predicted correctly.

augmented data. This forces an ABSA system to concentrate more on learning predictive features and to pay less attention on irrelevant features. Our method significantly improves upon a strong baseline and an adversarial learning counterpart on both the ARTS and the SemEval 14 original datasets.

2 Method

2.1 Background

To measure robustness performance of ABSA models, Xing et al. (2020) propose to extend the SemEval 2014 datasets (Pontiki et al., 2014) with three data augmentation operations: (1) **REVTGT** reverses the sentiment of the target aspect. (2) **REVNON** retains the target aspect's sentiment, but changes all the non-target aspects' sentiments.[2] (3) **ADDDIFF** continues the sentence with new segments involving aspects different from the target aspect.[3]

In this work, we focus on using the ADDDIFF operation to perform data augmentation in ABSA. ADDDIFF does not modify the original sentence and thus is less likely to generate erroneous data. Most importantly, ADDDIFF does not require annotations for sentiment words' positions. Rather, it just needs sentiment polarity annotations, making it cost effective.

While the state-of-the-art of ABSA has been advanced significantly

2.2 Inspection for augmented data

Such cost effective generation has its own issues. In our experiments, we use the tools provided in (Xing et al., 2020) to generate our own augmentation data on the training set.[4] However, probably because we do not use manual quality inspection and manual modifications as what have been used in ARTS to build the test dataset, the generated data is clearly of less good quality, as illustrated by Table 1 where we performed our own ADDDIFF operation on the same test instance as in ARTS.

We believe that this distribution shift between augmented data and the real test data corresponds closely to what happens in real-life scenario when applying data augmentation. We show in our experiments that applying adversarial training with such augmented data does not consistently improve model performance (see Section 4.2), contrary to when the augmented data aligns perfectly with the test data (see Appendix A.2).

2.3 The KL-Regular Model

We notice that adversarial training only leverages generated data but not the prior knowledge about the generation process. Specifically, the relationship between the original sentence and the generated sentence has not been exploited. While such relationship is not always available for all data augmentation techniques, we propose in this work a simple way to leverage this prior knowledge for all *predictive feature invariant* data augmentation, which includes ADDDIFF operation that we apply here.

Take for example the ADDDIFF operation that we apply in Table 1. Since we have controlled in the augmentation process that the appended text says nothing about the main aspect, it does not imply any predictive features for the target label *a priori*. In other words, the predictive features remain unchanged when we switch between the original sentence and the generated one, and so does the predicted probability. We propose to take into account such prior knowledge to guide the model to learn predictive features and thus achieve better generalization over all distributions (Arjovsky et al., 2020).[5]

To incorporate the prior knowledge that the operation is predictive feature invariant, we thus propose to make the two probabilities closer. More formally, for each instance X_i, let $p(Y_i|X_i)$ be the label probability of the original sentence; $p(Y_i|X_i^a)$ be the counterpart probability where X_i^a denotes the sentence after applying our ADDDIFF operation; over each sentence, the cross entropy loss and the KL regularization loss are:

$$\begin{aligned}\mathcal{L}_{NLL}^i &= -\log p(Y_i|X_i) - \log p(Y_i|X_i^a)\\ \mathcal{L}_{KL}^i &= KL(p(Y_i|X_i), p(Y_i|X_i^a))\end{aligned} \quad (148)$$

that sums up to the loss function below where KL regularization loss is α-weighted:

$$\mathcal{L} = \sum_i (\mathcal{L}_{NLL}^i + \alpha \mathcal{L}_{KL}^i)$$

We have also tried the KL regularizer in the other direction and the JS divergence, but preliminary

[2]The operation also exaggerates the extent for certain aspects' sentiments already opposite to the target one.

[3]REVTGT and REVNON could only apply to the instances with explicit opinion words, while ADDDIFF could operate on all the instances.

[4]https://github.com/zhijing-jin/ARTS_TestSet

[5]By assuming that the sentence can be decoupled into predictive features and irrelevant features, we can draw causal graphs to show that $p(Y|X)$ equals to $p(Y|X^a)$ (see A.1).

results suggest that KL divergence with the proposed direction may perform slightly better; the probability is calculated based on the softmax of a RoBERTa based model (Dai et al., 2021).

3 Experiment Settings

Data & Processing. We conduct experiments on the SemEval 2014 Laptop and Restaurant Reviews (*Laptop* and *Restaurant*) (Pontiki et al., 2014) and the ARTS (Xing et al., 2020) extension. We follow previous stuides to remove instances with conflicting polarity (Wang et al., 2016; Ma et al., 2017; Xu et al., 2019a) and use the train-dev split as in (Xu et al., 2019b). For compairison, we report the accuracy, aspect robustness scores (ARS) and macro F1 scores that are averaged over 5 experiments.

Baselines. Previous works show strong robustness performance when using pretrained models (Radford et al., 2021; Hendrycks et al., 2020; Xing et al., 2020). Inspired by this, we use the same RoBERTa based model as in Dai et al. (2021)'s work and fine tune the model on the original SemEval data (Ori) as our baseline in this work. We find that it significantly outperforms the best results reported in (Xing et al., 2020) (i.e., the result given by the *BERT-PT* model). For completeness, we compare our method with the other two following methods:

1. *BERT-PT* which is the best performing model in (Xing et al., 2020). Xu et al. (2019b) propose this method which first post-trains a BERT based model on other review datasets and then fine tune it on ABSA task.

2. *Adversarial* which trains the RoBERTa baseline with both the original training data and the AddDiff data that we generate as described in Section 2.1.

Parameter Setting. We use fastNLP[6] to implement our models. We fine tune the RoBERTa-large model with a batch size $b = 64$, a dropout rate $d = 0.3$, and an AdamW optimizer (Loshchilov and Hutter, 2019) for both *Laptop* and *Restaurant* datasets. We perform grid search over learning rate $\{5e^{-6}, 1e^{-5}, 2e^{-5}\}$ for both datasets in all experiments; for *KL-Regular* that we propose in this work, we also grid search over the regularization weights $\{1, 3, 5\}$. We train the model up to 40 epochs and select the best model according to the result on the validation set, which we set to the Ori validation set. [7]

[6]https://github.com/fastnlp/fastNLP

4 Results and Analysis

4.1 Main Results

We show our main results in Table 2. For all datasets, we report accuracy and Macro F1; for ARTS we also consider ARS an evaluation metric. We observe that:

RoBERTa baseline outperforms BERT-PT on all testing scenarios. For example, on the Laptop dataset, our RoBERTa baseline outperforms BERT-PT by 4.1% in accuracy on the original test set and by 5.77% in ARS on the ARTS test set respectively. In consequence, we choose RoBERTa as our baseline to compare in the following.

Adversarial training does not improve consistently. Training on our noisy data in addition, the adversarial models have worse performance in ARS compared to the RoBERTa baseline on both the Laptop and the Restaurant datasets; the result for accuracy is mixed.

KL-Regular achieves the best performance overall. With the same noisy augmented data, our proposed KL-Regular model shows improvements in ARTS on both the Laptop and the Restaurant datasets, which outperforms the RoBERTa baseline by 1.72% in accuracy (3.64% in ars) and by 1.65% in accuracy (3.57% in ars) respectively. Our model also improves over baseline on the original datasets, making our model bring improvements over all testing cases. This makes our approach particularly promising since robustness focuses on all potentially encountered distributions.

4.2 Model Analysis

How do different methods behave on ARTS AddDiff subset? By comparing the performance change between our RoBERTa baseline and the adversarial training-based system in Table 3, we see that leveraging noisy AddDiff augmented data can still improve the performance on ARTS AddDiff subset. This might be because the generated data still share sentence structure similarity with the AddDiff subset in ARTS. However, this improvement might hinder its performance on

[7]We are aware of the limitations of such choices as pointed out in (Csordás et al., 2021); however, given that our objective is to generalize to all *unknown* O.O.D settings, we consider the original validation set a sensible choice.

Model	Ori		ARTS		
	F1	**Acc.**	**F1**	**Acc.**	**ARS**
Laptop					
BERT-PT	75.08	78.07	–	71.82	53.29
RoBERTa	79.22	82.63	73.90	77.32	59.06
Adversarial	**80.15**	**83.26**	74.11	78.34	58.06
KL-Regular	80.04	**83.26**	**75.66**	**79.04**	**62.70**
Restaurant					
BERT-PT	76.96	84.95	–	80.99	59.29
RoBERTa	79.11	86.73	74.62	81.32	59.48
Adversarial	78.61	86.23	73.72	81.51	58.50
KL-Regular	**80.86**	**87.59**	**77.22**	**82.97**	**63.05**

Table 2: Model accuracy on Laptop and Restaurant reviews from SemEval 14. **Ori** setting tests on the original test set and **ARTS** setting tests on its ARTS counterpart. Texts in bold indicate the best results.

Model	ADDDIFF Subset Ori->New(Change)	
	Laptop	**Restaurant**
RoBERTa	82.63->80.47(02.16)	86.73->87.46(00.73)
Adversarial	83.26->81.91(01.35)	86.23->87.79(01.56)
KL-Regular	83.26->83.51(00.25)	87.59->89.64(02.05)

Table 3: The model accuracy change on the AddDiff subset. We report the accuracy on Ori and on ARTS ADDDIFF subset (New), as well as their difference.

other datasets, as on the Restaurant original dataset, adversarial training underperforms the RoBERTa baseline.

Compared to adversarial training, our proposed KL-Regular method not only leads to best performance on the ADDDIFF subset, with more than 3% ARS improvements on both datasets, but also performs the best without performance degradation on the original dataset. Our result is related to the distribution shift described in section 2.2; adversarial training can be most effective when augmented data distribution aligns perfectly with the test distribution, see Appendix A.2.

Is our approach sensitive to the regularization weight? To answer this question, we conduct experiments over different regularization weights $\{1, 2, 3, 4, 5\}$ for the same model with the same hyperparameters. The result in Figure 1 shows that different weights result in quite similar improvements on the model performance. We also observe that the regularization indeed makes the predicted probabilities $p(Y_i|X_i)$ and $p(Y_i|X_i^a)$ closer, see Appendix A.3.

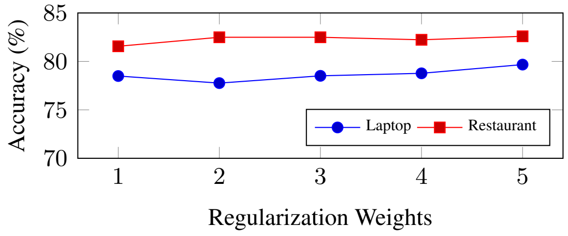

Figure 1: Accuracy and ARS for KL-Regular model with same hyperparameters on ARTS with the different weighs in {1,2,3,4,5}.

5 Related Works

Recent works improve ABSA robustness on ARTS by leveraging multiple dependency parses (Hou et al., 2021) or by leveraging external ABSA related data sources efficiently (Li et al., 2021). Our proposed method can be combined with theirs to further boost robustness performance on ARTS or other datasets (Jiang et al., 2019).

From technical perspectives, Liesting et al. (2021) try various data augmentation techniques on ABSA tasks; we not only leverage augemented data but also integrate prior knowledge about the generation. Our algorithm is similar to (Garg et al., 2018); however, our work considers leveraging general, automatic data augmentation tools with minimum cost. Such augmented data is noisy by nature and does not align well with the test distribution, leading to our observation that applying adversarial training does not lead to consistent improvements (Huang et al., 2020). Our work has theoretical foundation to bias the model focusing on features that have causal relationships with target labels for which we refer readers to (Mitrovic et al., 2021).

6 Conclusions and Future Work

For aspect-based sentiment analysis, we propose in this work a simple but effective method to improve aspect robustness by further exploiting the prior knowledge in data augmentation process. Experimental results show that our method can improve over the strong RoBERTa-based baseline on both original test and robustness test. We leverage noisy augmentation data, which corresponds closely to real-life scenario when applying data augmentation. In the future, we plan to apply our method to other NLP tasks and with other forms of data augmentation such as paraphrases.

7 Ethical Considerations

The experiment data we use are the most used datasets in ABSA studies and publicly released ARTS datasets and do not involve privacy disclosure. Our model architecture is based on open source releases. We do not anticipate any major ethical concerns.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization.

Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. 2021. The devil is in the detail: Simple tricks improve systematic generalization of transformers. 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 619–634. Association for Computational Linguistics.

Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. 2021. Does syntax matter? A strong baseline for aspect-based sentiment analysis with roberta. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1816–1829. Association for Computational Linguistics.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2018. Counterfactual fairness in text classification through robustness. *CoRR*, abs/1809.10610.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. *CoRR*, abs/2004.06100.

Xiaochen Hou, Peng Qi, Guangtao Wang, Rex Ying, Jing Huang, Xiaodong He, and Bowen Zhou. 2021. Graph ensemble learning over multiple dependency trees for aspect-level sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2884–2894. Association for Computational Linguistics.

William Huang, Haokun Liu, and Samuel R. Bowman. 2020. Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data. *CoRR*, abs/2010.04762.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 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)*, pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. *arXiv preprint arXiv:2111.02194*.

Tomas Liesting, Flavius Frasincar, and Maria Mihaela Trusca. 2021. Data augmentation in a hybrid approach for aspect-based sentiment analysis. *CoRR*, abs/2103.15912.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.

Jovana Mitrovic, Brian McWilliams, Jacob C. Walker, Lars Holger Buesing, and Charles Blundell. 2021. Representation learning via invariant causal mechanisms. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Towards causal representation learning. *CoRR*, abs/2102.11107.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

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 2020, Online, November 16-20, 2020*, pages 3594–3605. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019a. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019b. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2324–2335. Association for Computational Linguistics.

A Appendices

A.1 ABSA Causal Graph

Inspired by the recently works on causality (Arjovsky et al., 2020; Mitrovic et al., 2021; Schölkopf et al., 2021), we consider the ABSA task from a causal view.

Specifically for ABSA task assume that: a) Given a sentence-aspect pair, the sentence could be divided into key content K and irrelevant content I according to whether it contains the polarized description of the aspect. b) Only K contributes to the sentiment polarity classification.

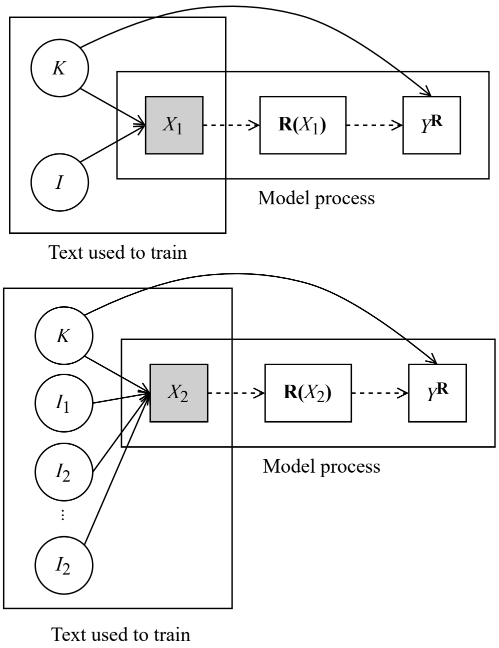

Figure 2: Causal graph and the learning process for ABSA task. Compared to the orginal sentence (above) AddDiff only adds irrelevant content.

We draw the causal graph based on the assumptions (solid lines in Figure 2). From the causal graph, it can be seen that the label Y only depends on the key content K of the target aspect which remains unchanged with AddDiff operation, thus $p(Y|X^1) = p(Y|K) = (Y|X^2)$; in other words, AddDiff doesn't change the label probability *a priori*. On dashed lines, we also show the learning process, the learning process only receives sentences X^1, X^2 and does not have the knowledge about the above prior knowledge (i.e, the two probabilities are equal). We show in our experiments that incorporating such prior knowledge can indeed improve model performance on various learning scenarios.

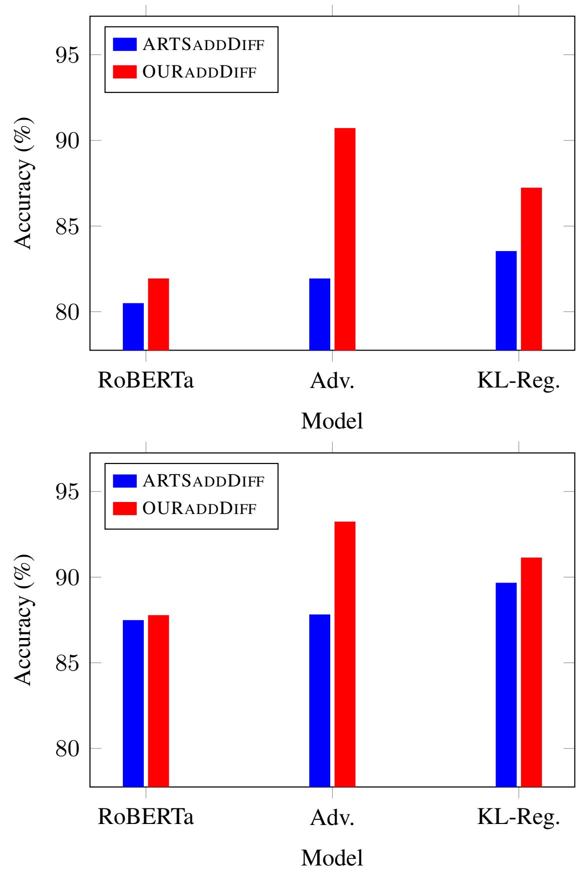

Figure 3: Comparison of model performance on ADDDIFF data from ARTS and its noisy counterpart we generated on the Laptop dataset (above) and the Restaurant dataset (below).

A.2 Different ADDDIFF Distribution

To understand distribution difference between ADDDIFF data from ARTS (ARTSADDDIFF) and its counterpart we generated (OURADDDIFF), we test models on these two test subsets and summarize the results in Figure 3. We observe that:

1. **Adversarial** performs very differently on ARTSADDDIFF and OURADDDIFF. Specifically, adversarial hardly improve the accuracy on ARTSADDDIFF, but reach the best performance on OURADDDIFF. It shows that adversarial training can be most effective when training data aligns perfectly with the test distribution, which we argue is a condition hard to obtain when applying data augmentation.

2. **KL-Regular** which also use our noisy ADDDIFF augmented data improves the performance on ARTSADDDIFF significantly and shows more similar improvements on the two test subsets. Since a model learning only predictive key features (i.e., K in Figure 2) will achieve exactly the same performance on both test subsets, the result might suggest that our model indeed focuses on predictive features to improve robustness over all tested datasets.

A.3 KL Divergence During Training

To verify that the KL divergence indeed decreases during training, we visualize its trend in Figure 4. The results show that the KL divergence is minimized through training despite some fluctuations in the first few epochs.

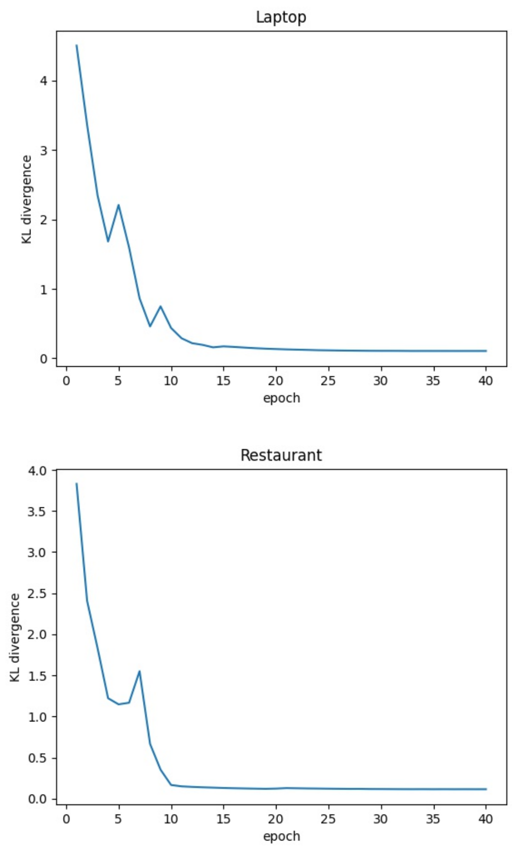

Figure 4: Trend of KL divergence while training our approach. We sum the kl divergence value of all the instances in training set at the end of each epoch.