Tasty Burgers, Soggy Fries: Probing Aspect Robustness in Aspect-Based Sentiment Analysis

Xiaoyu Xing[1*] **Zhijing Jin**[2*] **Di Jin**[3] **Bingning Wang**[4] **Qi Zhang**[1] **Xuanjing Huang**[1]

[1]Fudan University [2]Max Planck Institute, [3]CSAIL, MIT, [4]Sogou Inc.

{xyxing18,qz,xjhuang}@fudan.edu.cn
zjin@tuebingen.mpg.de
jindi15@mit.edu, wangbingning@sogou-inc.com

Abstract

Aspect-based sentiment analysis (ABSA) aims to predict the sentiment towards a specific aspect in the text. However, existing ABSA test sets cannot be used to probe whether a model can distinguish the sentiment of the target aspect from the non-target aspects. To solve this problem, we develop a simple but effective approach to enrich ABSA test sets. Specifically, we generate new examples to disentangle the confounding sentiments of the non-target aspects from the target aspect's sentiment. Based on the SemEval 2014 dataset, we construct the Aspect Robustness Test Set (ARTS) as a comprehensive probe of the aspect robustness of ABSA models. Over 92% data of ARTS show high fluency and desired sentiment on all aspects by human evaluation. Using ARTS, we analyze the robustness of nine ABSA models, and observe, surprisingly, that their accuracy drops by up to 69.73%. We explore several ways to improve aspect robustness, and find that adversarial training can improve models' performance on ARTS by up to 32.85%.[1]

1 Introduction

Aspect-based sentiment analysis (ABSA) is an advanced sentiment analysis task that aims to classify the sentiment towards a specific aspect (e.g., *burgers* or *fries* in the review "Tasty *burgers*, and crispy *fries*."). The key to a strong ABSA model is it is sensitive to *only* the sentiment words of the target aspect, and therefore not be interfered by the sentiment of any non-target aspect. Although state-of-the-art models have shown high accuracy on existing test sets, we still question their robustness. Specifically, given the *prerequisite* that a model outputs correct sentiment polarity for the test example, we have the following questions:

Equal Contribution. Please email (or cc') both first authors for correspondence.

[1]Our code and new test set are available at https://github.com/zhijing-jin/ARTS_TestSet.

(Q1) If we reverse the sentiment polarity of the target aspect, can the model change its prediction accordingly?

(Q2) If the sentiments of all non-target aspects become opposite to the target one, can the model still make the correct prediction?

(Q3) If we add more non-target aspects with sentiments opposite to the target one, can the model still make the correct prediction?

A robust ABSA model should both meet the prerequisite and have affirmative answers to all the questions above. For example, if a model makes the correct sentiment classification (i.e., positive) for *burgers* in the original sentence "Tasty *burgers*, and crispy fries", it should flip its prediction (to negative) when seeing the new context "Terrible *burgers*, but crispy fries". Hence, these questions together form a probe to verify if an ABSA model has high ***aspect robustness***.

Unfortunately, existing ABSA datasets have very limited capability to probe the *aspect robustness*. For example, the Twitter dataset (Dong et al., 2014) has only one aspect per sentence, so the model does not need to discriminate against non-target aspects. In the most widely used SemEval 2014 Laptop and Restaurant datasets (Pontiki et al., 2014), for 83.9% and 79.6% instances in the test sets, the sentiments of the target aspect, and all non-target aspects are all the same. Hence, we cannot decide whether models that make correct classifications attend only to the target aspect, because they may also wrongly look at the non-target aspects, which are *confounding factors*. Only a small portion of the test set can be used to answer our target questions proposed in the beginning. Moreover, when we test on the subset of the test set (59 instances in Laptop, and 122 instances in Restaurant) where the target aspect sentiment differs from all non-target aspect sentiments (so that the confounding

SubQ.	Generation Strategy	Example
Prereq.	SOURCE: The original sample from the test set	Tasty **burgers**, and crispy fries. (Tgt: burgers)
Q1	REVTGT: Reverse the sentiment of the *target* aspect	Terrible **burgers**, but crispy fries.
Q2	REVNON: Reverse the sentiment of the *non-target* aspects with originally the same sentiment as target	Tasty **burgers**, but soggy fries.
Q3	ADDDIFF: Add aspects with the *opposite* sentiment from the target aspect	Tasty **burgers**, crispy fries, but poorest service ever!

Table 1: The generation strategies and examples of the prerequisite (Prereq) and three questions (Q1)-(Q3). Each example are annotated with the **target aspect (Tgt)**, and altered sentence parts.

factor is disentangled), the best model (Xu et al., 2019a) drops from 78.53% to 59.32% on Laptop and from 86.70% to 63.93% on Restaurant. This implies that the success of previous models may over-rely on the confounding non-target aspects, but not necessarily on the target aspect only. However, *no* datasets can be used to analyze the aspect robustness more in depth.

We develop an automatic generation framework that takes as input the original test instances from SemEval 2014, and applies three generation strategies showed in Table 1. New test instances generated by REVTGT, REVNON, and ADDDIFF can be used to answer the questions (Q1)-(Q3), respectively. The generated new instances largely overlap with the content and aspect terms of the original instances, but manage to disentangle the confounding sentiment polarity of non-target aspects from the target, as showed in the examples in Table 1. In this way, we produce an "all-rounded" test set that can test whether a model robustly captures the target sentiment instead of other irrelevant clues.

We enriched the laptop dataset by 294% from 638 to 1,877 instances and the restaurant dataset by 315% from 1,120 to 3,530 instances. By human evaluation, more than 92% of the new aspect robustness test set (ARTS) shows high fluency, and desired sentiment on all aspects. Our ARTS test set is in line with other recent works on NLP challenge sets (McCoy et al., 2019; Gardner et al., 2020). Using our new test set, we analyze the aspect robustness of nine existing models. Experiment results show that their performance degrades by up to 55.64% on Laptop and 69.73% on Restaurant. We also use our generation strategy to conduct adversarial training and find it improves aspect robustness by at least 11.87% and at most 35.37% across various models.

The contributions of our paper are as follows:

1. We develop simple but effective automatic generation methods that generate new test instances (with over 92% accuracy by human evaluation) to challenge the aspect robustness.

2. We construct ARTS, a new test set targeting at aspect robustness for ABSA models, and propose a new metric, Aspect Robustness Score.

3. We probe the aspect robustness of nine models, and reveal up to 69.73% performance drop on ARTS compared with the original test set.

4. We provide several solutions to enhance aspect robustness for ABSA models, including adversarial training detailed in Section 5.4.

2 Data Generation

As shown in Table 1, we aim to build a systematic method to generate all possible aspect-related alternations, in order to remove the confounding factors in the existing ABSA data. In the following, we will introduce three different ways to disentangle the sentiment of the target aspect from sentiments of non-target aspects.

2.1 REVTGT

The first strategy is to generate sentences that reverse the original sentiment of the target aspect. The word spans of each aspect's sentiment of SemEval 2014 data are provided by (Fan et al., 2019a). We design two methods to reverse the sentiment, and one additional step of conjunction adjustment on top of the two methods to polish the resulting sentence.

Strategy	Example
Flip Opinion	It's **light** and **easy** to transport. → It's **heavy** and **difficult** to transport.
Add Negation	The menu **changes** seasonally. → The menu **does not change** seasonally.
Adjust Conjunctions	The food is good, **and** the decor is **nice**. → The food is good, **but** the decor is **nasty**.

Table 2: Three strategies and examples of REVTGT.

Strategy	Example
Original sentence & sentiment	It has great <u>food</u> and a reasonable price, but the service is poor. **(Tgt) <u>food</u>:+** **price:+** **service:−** **overall:○**
REVNON Flip same-sentiment non-target aspects (and adjust conjunctions)	It has great <u>food</u> **but** an **unreasonable** price, **and** the service is poor.
Exaggerate opposite-sentiment non-target aspects	It has great <u>food</u> **but** an **unreasonable** price, **and** the service is **extremely** poor. **(Tgt) <u>food</u>:+** **price:−** **service:−−** **overall:−**

Table 3: The generation process of REVNON. The <u>target aspect</u> (Tgt), and sentiments of all aspects are annotated.

Flip Opinion Words Suppose we have the sentence "*Tasty* **burgers** and crispy fries," where the sentiment term for the target aspect is *Tasty*. We aim to generate a new sentence that flips the sentiment *Tasty*. A baseline approach is antonym replacement by looking up WordNet (Miller, 1995). However, due to polysemy, the simple lookup is very likely to derive an inappropriate antonym and cause incompatibility with the context. Among the retrieved set of antonyms, we only keep words with the same Part-of-Speech (POS) tag as original, using the stanza package[2] which takes the context into account by the state-of-the-art neural network-based model.[3] Lastly, in the case of multiple antonyms, we prioritize the words that are already in the existing vocabulary, and then randomly select an antonym from the candidate set.

Add Negation As the above strategy of flipping by the antonym is constrained by whether appropriate antonyms are available. For those cases without suitable antonyms, including long phrases, we add negation according to the linguistic features. In most cases, the sentiment expression is an adjective or verb term, so we simply add negation (i.e., "not") in front of it. If the sentiment term is not an adjective or verb, we add negation to its closest verb. For example, in Table 2, there are no available antonyms for "change" in the original example "The <u>menu</u> **changes** seasonally.", so we simply negate it as "The <u>menu</u> **does not change** seasonally."

Adjust Conjunctions As pinpointed in Section 1, 79.6∼83.9% of the original test data of SemEval 2014 (Pontiki et al., 2014) have the same sentiment for all aspects. A possible result of reverting one aspect's sentiment is that the other aspects' sentiments will be opposite to the altered one. So

we need to adjust the conjunctions for language fluency. If the two closest surrounding sentiments of a conjunction word have the same polarity, then cumulative conjunctions such as "and" should be applied; otherwise, we should adopt adversative conjunctions such as "but." In the example in Table 2, after flipping the sentiment, we derive the example "The food is good, **and** the decor is nasty" which is very unnatural, so we replace the conjunction "and" with "but," and thus generate the example "The food is good, **but** the decor is nasty."

2.2 REVNON

Changing the target sentiment by REVTGT can test if a model is sensitive enough towards the target-aspect sentiment, but we need to further complement this probe by perturbing the sentiments of the non-target aspects (REVNON). As showed in Table 3, for all the non-target aspects with the same sentiment as the target aspect's, we reverse their sentiments using the same method as REVTGT. And for all the remaining non-target aspects, whose sentiments are already opposite from the target sentiment, we exaggerate the extent by randomly adding an adverb (e.g., "very", "really" and "extremely") from a dictionary of adverbs of degree that is collected based on the training set. The resulting test example will be a solid proof of the ABSA quality, because only the target aspect has the desired sentiment, and all non-target aspects have been flipped to or exaggerated with the opposite sentiment.

2.3 ADDDIFF

The first two strategies, REVTGT and REVNON, have explored how the sentiment changes of existing aspects will challenge an ABSA model, and ADDDIFF further investigate if adding more non-target aspects can confuse the model. Moreover, the existing SemEval 2014 test sets have only on average 2 aspects per sentence, but the real-world applications can have more aspects. With these

---

[2]https://stanfordnlp.github.io/stanza/

[3]For the candidate filter, we do not use GPT-2 perplexity because its low accuracy, e.g., 38.4% on a random sample set. And its output is also less interpretable than the POS filter.

motivations, we develop ADDDIFF as follows.

We first form a set of aspect expressions AspectSet[4] by extracting all aspect expressions from the entire dataset. Specifically, for each example in the dataset, we first identify each sentiment term (e.g., "reasonable" in "Food at a reasonable price") and then extract its linguistic branch as the aspect expression (e.g., "at a reasonable price") by pretrained constituency parsing (Joshi et al., 2018). Table 4 shows several examples of AspectSet in the restaurant domain.

Sentiment	Aspect Expression
Positive	staff is *friendly* and *knowledgeable* desserts are *out of this world* texture is a *velvety*
Negative	service is *severely slow* dining experience is *miserable* tables are *uncomfortably close*

Table 4: Example aspect expressions from AspectSet of the restaurant domain.

Using the AspectSet, we randomly sample 1-3 aspects that are not mentioned in the original test sample and whose sentiments are different from the target aspect's, and then append these to the end of the original example. For example, "Great food and best of all GREAT beer!" $\xrightarrow{\text{ADDDIFF}}$ "Great food and best of all GREAT beer, *but management is less than accommodating, music is too heavy, and service is severely slow*." In this way, ADDDIFF enables the advanced testing of whether the model will be confused when there are more irrelevant aspects with opposite sentiments.

3 ARTS Dataset

3.1 Overview

Our source data is the most[5] widely used ABSA dataset, SemEval 2014 Laptop and Restaurant Reviews (Pontiki et al., 2014).[6] We follow (Wang et al., 2016; Ma et al., 2017; Xu et al., 2019a) to remove instances with conflicting polarity and only keep positive, negative, and neutral labels. We use the train-dev split as in (Xu et al., 2019a). The resulting Laptop dataset has 2,163 training, 150

validation, and 638 test instances, and Restaurant has 3,452 training, 150 validation, and 1,120 test instances.

Building upon the original SemEval 2014 data, we generate enriched test sets of 1,877 instances (294% of the original size) in the laptop domain, and 3,530 instances (315%) in the restaurant domain using generation method introduced in Section 2. The statistics of our ARTS test set are in Table 5. (A more detailed explanations of the number of instances generated by each strategy is in Appendix A.1.)

	Laptop	Restaurant
Original Test Set	638	1,120
Enriched Test Set	1,877	3,530
Relative Size	294.20%	315.17%
Fluency Check		
Accepted Instances	1,732	3,260
Fixed Instances	145	270
Acceptance Rate	92.27%	92.35 %
Inter-Agreement	91.10%	92.69%
Sentiment Check		
Accepted Instances	1,763	3,362
Fixed Instances	114	168
Acceptance Rate	93.93%	95.24%
Inter-Agreement	94.14%	95.61%

Table 5: Overall statistics of the ARTS test set and results of fluency and sentiment checks.

3.2 Quality Inspection

We conduct human evaluation to validate the generation quality of our ARTS dataset on two criteria:

1. **Fluency**: Does the generated sentence maintain the fluency of the source sentence?

2. **Sentiment Correctness**: Does the sentiment of each aspect have the desired polarity?

- REVTGT: Is the target sentiment reversed?
- REVNON: For non-target aspects with originally the same sentiment as the target, is it reversed? For the rest, are they exaggerated?
- ADDDIFF: Is the target sentiment unchanged?

Each task is completed by two native-speaker judges. We first calculate the inter-agreement rate of the human annotators, and then resolve the divergent opinions on samples that they disagree with. We accept the samples that both judges considered as correct or are resolved to be correct after our check. Finally, we ask the annotators to fix the rejected samples by minimal edit which does not

---

change the aspect term or the sentence meaning, but satisfies both criteria.

Fluency Check The evaluation results on fluency are showed in Table 5. Most samples (92.27% of Laptop and 92.35% of Restaurant test sets) are accepted as fluent text. The inter-agreement rate between the two human judges is also high, 91.10% and 92.69% on the two datasets.

Sentiment Check We also evaluate the sentiment correctness of the generated text. Note that for REVNON, we count the samples with all "yes" answers as accepted samples. Overall, the acceptance rate of the generated samples is 93.93% on Laptop and 95.24% on Restaurant, along with inter-agreement rates of over 94.14% on both datasets.

3.3 Dataset Analysis

After checking the quality of our enriched ARTS test set, we analyze the dataset characteristics and make comparisons with the original test sets.

	Laptop		Restaurant	
	Ori	ARTS	Ori	ARTS
#Words/Sent	18.56	22.27	19.37	23.15
Vocab Size	1565	1746	2197	2451
Labels				
Positive	341	883	728	1953
Negative	128	587	196	1104
Neutral	169	407	196	473
#Positive/#Negative	2.66	1.5	3.71	1.77
Aspect-Related Challenge				
#Aspects/Sent	2.05	2.75	2.57	3.28
Opp. Nontgt ≥ 1	16%	59%	20%	67%
Opp. Nontgt $=$ All	9%	38%	11%	42%
#Opp. Nontgt/Sent	0.23	1.16	0.27	1.39

Table 6: Characteristics of the ARTS test sets in comparison to the Original ("Ori") Laptop and Restaurant test sets.

For *general statistics*, we can see from Table 6 that the sentence length in the new test set is on average 4 words more than the original, and the vocabulary size is also larger by around two hundred. *For the label distribution*, we can see that the new test set has an increasing number of all labels, and especially balances the ratio of positive-to-negative labels from the original 2.66 to 1.5 on Laptop, and from 3.71 to 1.77 on Restaurant.

For the *aspect-related challenge* in the test set, the new test set, first of all, has a larger number of aspects per sentence than the original. Our test set also features the higher disentanglement of the target aspect from the non-target aspects that have

the same sentiment as the target: the portion of samples with at least one non-target aspects of sentiments different from the target is 59∼67%, and on average 45% higher than the original test sets. And the portion of the most challenging samples where all non-target aspects have sentiments different from the target one on the new test set is on average 30% more than that of the original test set. The average number of non-target aspects with opposite sentiments per sample in the new test set is on average 5 times that of the original set.

3.4 Aspect Robustness Score (ARS)

As mentioned in Section 1, a model is considered to have high aspect robustness if it satisfies both the prerequisite and all three questions (Q1)-(Q3). So we propose a novel metric, Aspect Robustness Score (ARS), that counts the correct classification of the source example and all its variations (REVTGT, REVNON, and ADDDIFF) as one unit of correctness. Then we apply the standard calculation of accuracy. Note that the three variations correspond to questions (Q1)-(Q3), respectively.

4 Evaluating ABSA Models

We use our enriched test set as a comprehensive test on the aspect robustness of ABSA models.

4.1 Models

For a comprehensive overview of the ABSA field, we conduct extensive experiments on models with a variety of neural network architectures.

TD-LSTM: (Tang et al., 2016a) uses two Long Short-Term Memory Networks (LSTM) to encode the preceding and following contexts of the target aspect (inclusive) and concatenate the last hidden states of the two LSTMs to make the sentiment classification.

AttLSTM: Wang et al. (2016) apply an Attention-based LSTM on the concatenatation of the aspect and word embeddings of each token.

GatedCNN: Xue and Li (2018) use a Gated Convolutional Neural Networks (CNN) that applies a Tanh-ReLU gating mechanism to the CNN-encoded text with aspect embeddings.

MemNet: Tang et al. (2016b) use memory networks to store the sentence as external memory and calculate the attention with the target aspect.

GCN: Aspect-specific Graph Convolutional Networks (GCN) (Zhang et al., 2019a) first applies

GCN over the syntax tree of the sentence and then imposes an aspect-specific masking layer on its top.

BERT: Xu et al. (2019a) uses a BERT-based baseline (Devlin et al., 2019) and takes as input the concatenation of the aspect term and the sentence.

BERT-PT: Xu et al. (2019a) post-train BERT on other review datasets such as Amazon laptop reviews (He and McAuley, 2016) and Yelp Dataset Challenge reviews, and finetune on ABSA tasks.

CapsBERT: (Jiang et al., 2019) encode the sentence and the aspect term with BERT, and then feed it into Capsule Networks to predict the polarity.

BERT-Sent: For more in-depth analysis, we also implement a sentence classification baseline. BERT-Sent takes as input sentences without aspect information, and predicts the "global" sentiment. We use it because if other ABSA models fails to pay attention to aspects, they will degenerate to a sentence classifier. If so, they will resemble BERT-Sent, which performs well on original tests and badly on ARTS. So BERT-Sent is a reference to check degenerated aspect-level models.

4.2 Implementation Details

For all existing models, we use the authors' official implementation. For our self-proposed BERT-Sent, we use Adam optimizer with a learning rate of 5e-5, weight decay of 0.01, and batch size of 32. We apply the l_2 regularization with $\lambda = 10^{-4}$, and train 50 epochs. Note that the tokenization of the ASTS dataset is the same as the original SemEval 2014, as we prepared the new test set by inverting the NLTK tokenization rules we used when applying the generation strategies.[7]

4.3 Results on ARTS

We list the accuracy[8] of the nine models on the Laptop and Restaurant test sets in Table 7. For *Entire Test-New* in Table 7, accuracy is calculated using ARS. Supplementary to ARS, Table 7 also decomposes ARS into single-strategy scores (the right three columns) by splitting the entire ARTS test set into three subsets according to the corresponding data generation techniques. Each of the single-strategy scores explains from one perspective the reason for large performance drop in ARS, which will be elaborated later.

---

[7]We used the metanl package to detokenize: https://github.com/commonsense/metanl.

[8]For ABSA, accuracy is the standard metric to be reported (Wang et al., 2016; Xue and Li, 2018; Tang et al., 2016b).

[9]CapsBERT hand-crafted the Capsule Guided Routing specifically for the restaurant domain, so it fails significantly.

Overall Performance On the entire test set, we can see that the accuracy of all models on the original test set is very high, achieving up to 78.53% on Laptop and 86.70% on Restaurant, but it drops drastically (↓69%∼↓25%) on our new test sets.

Performance of Different Models From the overall performance on our new test set, we can see that BERT models on average are more robust to the aspect-targeted challenges that our new test set poses. The most effective model BERT-PT scores the best on both original accuracy and robustness. It has 53.29% ARS on Laptop and 59.29% on Restaurant. However, the accuracy of non-BERT models on average drops drastically to under 30% by over ↓50%.

Performance on Different Subsets We list in detail the performance of each model on the three subsets of our new test set: REVTGT, REVNON, and ADDDIFF. They correspond to the three questions (Q1)-(Q3). REVTGT on average induces the most performance drop, as it requires the model to pay precise attention to the target sentiment words. REVNON makes the performance of the sentence classifier BERT-Sent drops the most by up to ↓45.93%, and the model CapsBERT also drops by up to ↓39.26%. The last subset ADDDIFF causes most non-BERT models to drop significantly, indicating that these models are not robust enough against an increased number of non-target aspects, which should have been irrelevant.

Laptop vs. Restaurant The performance drop on Restaurant is higher than that on Laptop. There are two possible reasons: (1) the original performance on restaurant is higher, and (2) the new test set is more challenging in the Restaurant domain. We verify this by calculating the relative drop ($\frac{new-old}{old}$) in addition to the reported absolute values of the change. The relative drop on Laptop is 64.76%, which is higher than 60.36% on Restaurant. For the laptop dataset, both the lower original performance and the larger relative decrease of performance might be due to the nature of the dataset. For example, Laptop restaurant has far fewer training data than Restaurant, which makes the models less accurate originally and weaker on ARTS.

5 Analysis

5.1 Variations of Generation Strategies

Combining Multiple Strategies Each sample in the ARTS test set is generated by one of the three

Model	Entire Test Ori → New (Change)	RevTgt Subset Ori → New (Change)	RevNon Subset Ori → New (Change)	AddDiff Subset Ori → New (Change)
Laptop Dataset				
MemNet	64.42 → 16.93 (↓47.49)*	72.10 → 28.33 (↓43.77)*	82.22 → 79.26 (↓02.96)	64.42 → 56.58 (↓07.84)*
GatedCNN	65.67 → 10.34 (↓55.33)*	75.11 → 24.03 (↓51.08)*	83.70 → 78.52 (↓05.18)	65.67 → 45.14 (↓20.53)*
AttLSTM	67.55 → 09.87 (↓57.68)*	72.96 → 27.04 (↓45.92)*	85.93 → 75.56 (↓10.37)*	67.55 → 39.66 (↓27.89)*
TD-LSTM	68.03 → 22.57 (↓45.46)*	73.39 → 29.83 (↓43.56)*	83.70 → 77.04 (↓06.66)	68.03 → 60.66 (↓07.37)*
GCN	72.41 → 19.91 (↓52.50)*	78.33 → 35.62 (↓42.71)*	88.89 → 74.81 (↓14.08)*	72.41 → 52.51 (↓19.90)*
BERT-Sent	73.04 → 17.40 (↓55.64)*	78.76 → 59.44 (↓19.32)*	88.15 → 42.22 (↓45.93)*	73.04 → 34.64 (↓38.40)*
CapsBERT	77.12 → 25.86[9] (↓51.26)*	80.69 → 57.73 (↓22.96)*	88.89 → 49.63 (↓39.26)*	77.12 → 45.14 (↓31.98)*
BERT	77.59 → 50.94 (↓26.65)*	83.05 → 65.02 (↓18.03)*	93.33 → 71.85 (↓21.48)*	77.59 → 71.00 (↓06.59)*
BERT-PT	78.53 → 53.29 (↓25.24)*	82.40 → 60.09 (↓22.31)*	93.33 → 83.70 (↓09.63)*	78.53 → 75.71 (↓02.82)
Average	71.60 → 25.23 (↓46.37)*	77.42 → 43.01 (↓34.41)*	87.57 → 70.29 (↓17.28)*	71.60 → 53.45 (↓18.15)*
Restaurant Dataset				
MemNet	75.18 → 21.52 (↓53.66)*	80.73 → 27.54 (↓53.19)*	84.46 → 73.65 (↓10.81)*	75.18 → 60.71 (↓14.47)*
GatedCNN	76.96 → 13.12 (↓63.84)*	85.11 → 23.17 (↓61.94)*	88.06 → 72.97 (↓15.09)*	76.96 → 54.91 (↓22.05)*
AttLSTM	75.98 → 14.64 (↓61.34)*	82.98 → 28.96 (↓54.02)*	86.26 → 61.26 (↓25.00)*	75.98 → 52.32 (↓23.66)*
TD-LSTM	78.12 → 30.18 (↓47.94)*	85.34 → 34.99 (↓50.35)*	88.51 → 75.68 (↓12.83)*	78.12 → 70.18 (↓07.94)*
GCN	77.86 → 24.73 (↓53.13)*	86.76 → 35.58 (↓51.18)*	88.51 → 79.50 (↓09.01)*	77.86 → 65.00 (↓12.86)*
BERT-Sent	80.62 → 10.89 (↓69.73)*	89.60 → 44.80 (↓44.80)*	89.86 → 57.21 (↓32.65)*	80.62 → 30.89 (↓49.73)*
CapsBERT	83.48 → 55.36 (↓28.12)*	89.48 → 71.87 (↓17.61)*	90.99 → 74.55 (↓16.44)*	83.48 → 77.86 (↓05.62)*
BERT	83.04 → 54.82 (↓28.22)*	90.07 → 63.00 (↓27.07)*	91.44 → 83.33 (↓08.11)*	83.04 → 79.20 (↓03.84)*
BERT-PT	86.70 → 59.29 (↓27.41)*	92.20 → 72.81 (↓19.39)*	92.57 → 81.76 (↓10.81)*	86.70 → 80.27 (↓06.43)*
Average	79.77 → 31.62 (↓48.15)*	86.92 → 44.75 (↓42.17)*	88.96 → 73.32 (↓15.64)*	79.77 → 63.48 (↓16.29)*

Table 7: Model accuracy on Laptop and Restaurant data. We compare the accuracy on the **Ori**ginal and our **New** test sets (*Ori → New*), and calculate the *change* of accuracy. Besides the Entire Test Set, we also list accuracy on subsets where the generation strategies RevTgt, RevNon and AddDiff can be applied. The accuracy of *Entire Test-New* is calculated using ARS. ⋆ indicates whether the performance drop is statistically significant (with p-value ≤ 0.05 by Welch's t-test).

strategies. However, it is also worth exploring whether combining several strategies can make a more challenging probe on the aspect robustness of ABSA models. As a case study, we analyze the model robustness against test samples generated by the combination of RevNon+AddDiff. By comparing the performance decrease caused

Model	Laptop Ori → New (Change)	Restaurant Ori → New (Change)
MemNet	82.22 → 72.59 (↓09.63)	84.46 → 50.90 (↓33.56)*
GatedCNN	84.44 → 59.26 (↓25.18)*	87.84 → 53.83 (↓34.01)*
AttLSTM	85.93 → 51.85 (↓34.08)*	86.26 → 38.06 (↓48.20)*
TD-LSTM	83.70 → 68.89 (↓14.81)*	88.51 → 65.99 (↓22.52)*
GCN	88.89 → 60.74 (↓28.15)*	88.51 → 72.52 (↓15.99)*
BERT-Sent	88.15 → 11.85 (↓76.30)*	89.86 → 11.94 (↓77.92)*
CapsBERT	90.37 → 24.44 (↓65.93)*	90.99 → 66.89 (↓24.10)*
BERT	93.33 → 68.15 (↓25.18)*	91.44 → 76.58 (↓14.86)*
BERT-PT	93.33 → 78.52 (↓14.81)*	92.57 → 78.60 (↓13.97)*
Average	87.57 → 55.14 (↓32.43)*	88.96 → 57.26 (↓31.70)*

Table 8: The accuracy of each model on the original test set and the new test set generated by RevNon+AddDiff in laptop and restaurant domains.

by RevNon+AddDiff in Table 8 and by only RevNon and AddDiff in Table 7, we can see that the accuracy of each model decreases by a much larger extent on RevNon+AddDiff than either of RevNon and AddDiff. The performance drop by the RevNon+AddDiff subset is almost the sum of RevNon and AddDiff in most cases, and sometimes larger than the sum, for example, in

the case of BERT-PT on Laptop where the sum of the performance drop by the two separate strategies 9.63%+2.82% is smaller than the combined strategy's 14.81%.

AddDiff with More Aspects Some strategies such as AddDiff can be parameterized by k, where k is the number of additional non-target aspects to be added. We select three models (the best, the worst, and an average-performing one), and plot their accuracy on test samples generated by AddDiff(k) on Laptop in Figure 1. As k gets larger, the test samples become more difficult. The sentence classification baseline BERT-Sent drops drastically, BERT-PT remains high, and GCN lies in the middle.

5.2 How to Effectively Model the Aspect?

An important usage of our ARTS is to understand what model components are key to aspect robustness. We list the aspect-specific mechanisms of all models according to the ascending order of their ARS on Laptop dataset in Table 9. We can see that for BERT-based models, BERT-PT, which is further trained on large review corpora, gets the best accuracy and aspect robustness. More complicated structures like CapsBERT underperforms the basic BERT by 25.08%.
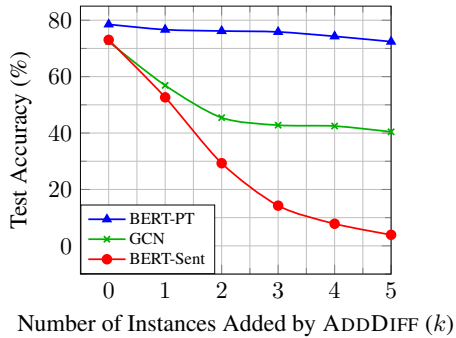
Figure 1: Accuracy of BERT-PT, GCN, and BERT-Sent on the test samples in the laptop domain generated by ADDDIFF(k) where k varies from 1 to 5.

Model	ARS	Asp+W Emb	Posi-Aware	Asp Att
AttLSTM	9.87	✓	✗	✓
GatedCNN	10.34	✓	✗	✓
MemNet	16.93	✗	✗	✓
GCN	19.91	✗	✓	✓
TD-LSTM	22.57	✗	✓	✗
CapsBERT	25.86	✗	✗	✓
BERT	50.94	✗	✗	✗
BERT-PT	53.29	✗	✗	✗

Table 9: Models in the ascending order of their ARS on Laptop. We list their aspect-specific mechanisms, including concatenating the aspect and word embeddings (Asp+W Emb), position-aware mechanism for aspects (Posi-Aware), and attention using the aspect (Asp Att). We highlight ✓ for Posi-Aware as it is the most related to aspect robustness for non-BERT models.

Among the non-BERT models, the aspect position-aware models TD-LSTM and GCN are the most robust, as they have a stronger sense of the location of the target aspect in a sentence. On the contrary, the other models with poorer robustness (9.87%~16.93% in Table 9) only use mechanisms such as aspect-based attention, or concatenating the aspect embedding to the word embedding.

To summarize, the main takeaways are

- For BERT models, additional pretraining is the most effective.
- For non-BERT models, explicit position-aware designs lead to more aspect robustness.

5.3 Does a More Diverse Training Set Help?

A recent dataset, Multi-Aspect Multi-Sentiment (MAMS) (Jiang et al., 2019), is collected from the same data source as the SemEval 2014 Restaurant dataset (Ganu et al., 2009). However, its sentences are more complicated, each having at least two aspects with different sentiment polarities.

We use this dataset to inspect whether a stronger training set can help improve aspect robustness.

Training and Testing on MAMS Table 10a checks the aspect robustness of models trained on MAMS using the original MAMS test set (O→O) and the new test set that we produced by applying the same generation strategies to its test set (O→N). Models trained and tested on MAMS have a smaller decrease rate than those on the Restaurant dataset. This shows that a more challenging training set can make models more robust.

Training on MAMS and Testing on Restaurant As MAMS and Restaurant are collected from the same source data, we test whether MAMS-trained models perform well on the new test set of Restaurant (in the column "MAMS→N" of Table 10b). We can see that all models trained on MAMS are more robust than those trained on the Restaurant dataset. For example, the accuracy of BERT and BERT-PT on the new test set is lifted up to 62.77%.

5.4 Does Adversarial Training Help?

Although the MAMS described in Section 5.3 provides a training set with diversity, it remains difficult to improve aspect robustness for other domains, or future new datasets. Therefore, we propose a flexible method, adversarial training, for aspect robustness, which is applicable to any given dataset.

We conducted adversarial training on the Laptop and Restaurant datasets, and analyze its effect in Table 10b. Specifically, for the column "Adv→N", we generated an additional training set by applying the three proposed strategies on training data, then trained models on the augmented data obtained by combining the original training data and this newly generated data, and finally evaluated on the ARTS test data. This practice follows Table 7 of (Zhang et al., 2019b) which is a similar stream of work as ours for the paraphrasing domain.

In both Restaurant and Laptop domains, adversarial training (Adv→N) leads to significant performance improvement than only training on the original datasets (O→N). On the Restaurant datasets, adversarial training is even more effective than training on MAMS, because our generated data instances comprehensively covered all possible perturbations of the non-target aspects, and naturally collected datasets might not be comparable.

6 Error Analysis for Data Generation

We analyze the error types in the subset of ARTS that was fixed by human judges. Two most significant error types are wrong antonyms (~2%), such

Model	MAMS		Restaurant				Laptop		
	O→O	O→N	O→O	O→N	MAMS→N	Adv→N	O→O	O→N	Adv→N
MemNet	70.51	37.80	75.18	21.52	24.02	37.95	64.42	16.93	31.82
GatedCNN	66.02	32.93	76.96	13.13	18.48	37.50	65.67	10.34	41.85
AttLSTM	67.14	39.67	75.98	14.64	22.32	48.66	67.55	9.87	42.63
TD-LSTM	77.62	49.25	78.12	30.18	41.60	62.76	68.03	22.57	54.86
GCN	76.95	47.98	77.86	24.73	46.51	61.52	72.41	19.91	56.43
BERT-Sent	49.25	10.48	80.62	10.89	12.95	45.80	73.04	17.40	53.92
CapsBERT	83.38	60.18	83.66	55.36	61.43	75.80	76.80	25.86	61.23
BERT	84.51	61.38	83.04	54.82	62.77	74.82	77.59	50.94	65.67
BERT-PT	85.10	64.37	86.70	59.29	62.77	74.64	78.53	53.29	66.93

(a) Accuracy of each model trained on the MAMS **O**riginal training data and evaluated on the **O**riginal test data (O→O), as well as the **N**ew test set generated by our models (O→N).

(b) Accuracy of each model trained on the **O**riginal data and evaluated on the **O**riginal test set (O→O), and the **N**ew test set (O→N), as well as that trained on the **Adv**ersarial data and evaluated on the **N**ew test set (Adv→N). For Restaurant, we also test models trained on **MAMS** dataset and tested on the **N**ew test set of Restaurant (MAMS→N).

Table 10: Improvements on the new test set using different training data.

as "the weight of the laptop is light→dark", and negation which causes grammatical errors (∼1.1%). In future work, we can fix the latter by applying a grammatical error correction system on top of our generation. Also, REVTGT and REVNON cannot be applied to 1.4∼6.6% instances with complicated sentiment expressions which rely on commonsense. For example, "a 2-hour wait" is negative bust too difficult to alter in our current generation framework. It needs more advanced models such as text style transfer (Shen et al., 2017; Jin et al., 2019b).

7 Related Work

Robustness in NLP Robustness in NLP has attracted extensive attention in recent works (Hsieh et al., 2019; Li et al., 2016). As a popular method to probe the robustness of models, adversarial text generation becomes an emerging research field in NLP. Techniques include adding extraneous text to the input (Jia and Liang, 2016), character-level noise (Belinkov and Bisk, 2018; Ebrahimi et al., 2018), and word replacement (Alzantot et al., 2018; Jin et al., 2019a). Using the adversarial generation techniques, new adversarial test sets are proposed for several tasks such as paraphrasing (Zhang et al., 2019b) and entailment (Glockner et al., 2018; McCoy et al., 2019).

Aspect-Based Sentiment Analysis ABSA has emerged as an active research area recently. Early works hand-craft sentiment lexicons and syntactic features for rule-based classifiers (Vo and Zhang, 2015; Kiritchenko et al., 2014). Recent neural network-based models use architectures such as LSTM (Tang et al., 2016a), CNN (Xue and Li, 2018), Attention mechanisms (Wang et al., 2016),

Capsule Network (Jiang et al., 2019), and the pre-trained model BERT (Xu et al., 2019a). Similar to the motivation in our paper, some work shows preliminary speculation that the current ABSA datasets might be downgraded to sentence-level sentiment classification (Xu et al., 2019b).

8 Conclusion

In this paper, we proposed a simple but effective mechanism to generate test instances to probe the aspect robustness of the models. We enhanced the original SemEval 2014 test sets by 294% and 315% in laptop and restaurant domains. Using our new test set, we probed the aspect robustness of nine ABSA models, and discussed model designs and better training that can improve the robustness.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2890–2896. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36. Association for Computational Linguistics.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019a. Target-oriented opinion words extraction with target-fused neural sequence labeling. 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 2509–2518. Association for Computational Linguistics.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019b. Target-oriented opinion words extraction with target-fused neural sequence labeling. 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 2509–2518. Association for Computational Linguistics.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *12th International Workshop on the Web and Databases, WebDB 2009, Providence, Rhode Island, USA, June 28, 2009*.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna 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 NLP models via contrast sets. *CoRR*, abs/2004.02709.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 650–655. Association for Computational Linguistics.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 507–517. ACM.

Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. 2019. On the robustness of self-attentive models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1520–1529. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 6279–6284. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019a. Is BERT really robust? natural

language attack on text classification and entailment. *CoRR*, abs/1907.11932.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019b. Imat: Unsupervised text attribute transfer via iterative matching and translation. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3095–3107. Association for Computational Linguistics.

Vidur Joshi, Matthew E. Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1190–1199. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 437–442. The Association for Computer Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4068–4074. ijcai.org.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

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.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6830–6841.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3298–3307. ACL.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 214–224. The Association for Computational Linguistics.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1347–1353. AAAI Press.

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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 606–615. The 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. 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.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019b. A failure of aspect sentiment classifiers and an adaptive re-weighting solution. *CoRR*, abs/1911.01460.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2514–2523. Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019a. Aspect-based sentiment classification with aspect-specific graph convolutional 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 2019, Hong Kong, China, November 3-7, 2019, pages 4567–4577. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. PAWS: paraphrase adversaries from word scrambling. 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 1298–1308. Association for Computational Linguistics.

A Appendices

A.1 Test Set Generation Details

In our main paper, we mentioned that the size of our enriched ARTS test set is 294% of the original Laptop data size, and 315% of the original Restaurant test data size. These two ratios should ideally be both 400%, because there are three generation strategies, plus one original sentence. However, this gap is because not every original test sentence can qualify for every generation strategy. The number of instances generated by each strategy is shown in Table **??**.

Although AddDiff can apply to all test cases, the other two strategies, REVTGT and REVNON, cannot apply to all instances. For REVTGT and REVNON, we need to flip opinion words, so we can only apply these two strategies on instances with explicit opinion words. The number of opinion words is the main bottleneck for REVTGT. Specifically, Fan et al. (2019b) provide opinion words for 466 instances of Laptop, and 846 instances of Restaurant. Since REVTGT is applicable to these instances, the number of new test instances in ARTS generated by REVTGT are the same as the number of opinion word-available instances.

In addition to the opinion word constraint, REVNON has some further requirements. We filter out instances with only one aspect (224 in Laptop, and 263 in Restaurant), as well as instances where the opinion words of the target aspect are overlapped with the opinion words of the non-target aspect (102 in Laptop and 132 in Restaurant). Also we did not consider instances of which all the sentiment of non-target aspects are neutral (4 in Laptop, and 7 in Restaurant).