Enhancing the Aspect Robustness Score of the HAABSA++ Model Using Adversarial Training

Milad Agha[1], Flavius Frasincar[1][0000−0002−8031−758X], Beilly Zhu[1], and Tarmo Robal[2][0000−0002−7396−8843]

[1] Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, The Netherlands
532622ma@student.eur.nl, frasincar@ese.eur.nl, 611700bz@eur.nl
[2] Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia
tarmo.robal@taltech.ee

Abstract. Sentiment analysis is an important tool in understanding users of the Web through the cues they leave while communicating and providing feedback. Sentiment classification models may have lower robustness because they detect irrelevant patterns instead of sentiment-bearing words related to the target aspects. We evaluated the robustness of the Hybrid Approach for the Aspect-Based Sentiment Analysis++ (HAABSA++) model. We first generated an Aspect Robustness Test Set (ARTS) through the augmentation of the original test set with different augmentation techniques: RevTgt, RevNon, and AddDiff. These techniques modify the target and non-target aspects to test whether the model still makes correct predictions. Aspect robustness is evaluated using the Aspect Robustness Score (ARS). In addition, we investigated the improvement of ARS through adversarial training by applying the three data augmentation methods to the training set. We find that the robustness of the HAABSA++ model as measured by ARS is average when we compare the HAABSA++ model with other models found in the literature. We also find that performing adversarial training improves the robustness of the model as measured by ARS. However, this improvement comes at the cost of a lower accuracy of the HAABSA++ model for the original test set instances.

Keywords: Aspect-based sentiment analysis · Aspect robustness · Attention-based model · Neural network.

1 Introduction

The field of sentiment analysis has increased in importance with the growth of various Web platforms, including for social media and online reviews. On these Web platforms, vast amount of information is generated daily. Sentiment analysis enables the extraction of opinions and emotions from content, providing valuable insights for businesses, academics, and consumers from brand reputation and customer sentiment to social insights, trend identification, and interdisciplinary studies. This in turn can be utilised for Web Engineering (WE), for instance to

enhance user experience, develop better user-centred interfaces, optimize search, and automate processes on Web platforms (e.g., virtual assistants). However, manual analysis of this data is exhaustive and inefficient. As a result, sentiment analysis has become an essential tool [9]. In particular Aspect Based Sentiment Analysis (ABSA) is noteworthy for its approach to evaluate individual aspects or characteristics, rather than assigning an overall sentiment score to each sentence or to the entire text [17].

ABSA can be broken down into three sub-tasks: aspect identification, sentiment classification, and sentiment aggregation [17]. The identification process is about pinpointing the specific aspects mentioned, while the classification process involves labeling these aspects with a sentiment value. Lastly, aggregation is about compiling these sentiment values to reflect the collective sentiment. This work aims to optimize solutions for the sentiment classification task, also known as Aspect-Based Sentiment Classification (ABSC).

ABSC techniques are typically grouped into three main types: knowledge-based approaches, machine learning models, and hybrid models [1]. Studies indicate that the hybrid approach yields superior results to the individual methods [18]. However, it is still an open question which specific mix of these techniques achieves the optimal balance of accuracy and operational effectiveness.

Recent research by [20] has developed the Hybrid Approach for Aspect-Based Sentiment Analysis (HAABSA), which achieves state-of-the-art results in sentiment classification. This approach uses a lexicalized domain-specific ontology to assess the sentiment towards the target. When the results of the ontology are ambiguous, the sentences are processed by a specialized Neural Network (NN) with a Left-Center-Right structure and Rotatory attention (LCR-Rot) [23] mechanism serving as a secondary model.

Two additional extensions to the HAABSA model to improve the accuracy of sentiment classification results were introduced in [19], resulting a model called HAABSA++. These extensions involve the use of deep contextual word embeddings (i.e., ELMO [13] and BERT [2] designed to grasp the semantic information of text by considering the surrounding context) to better account for the semantics of words, and the introduction of hierarchical attention to distinguish the importance of high-level input sentence representations.

Although HAABSA++ achieves cutting-edge results in the sentiment classification task, recent studies have shown that the apparent high performance of most models is due to their ability to detect spurious patterns [10]. As a result, the robustness of these models remains uncertain. Models must respond only to the sentiment-related words associated with the target, without being influenced by the sentiment expressed about unrelated aspects. [21] explored the robustness problem by developing a framework designed to generate a comprehensive Aspect Robustness Test Set (ARTS). Combined with a new measure, the Aspect Robustness Score (ARS), ARTS can be used to assess whether a model can reliably detect the intended sentiment in the presence of spurious patterns. [21] also proposes to improve the ARS with adversarial training by using the automatic generation framework on the training set.

In this paper, we evaluate the robustness of the HAABSA++ model, and investigate the improvement of the ARS through adversarial training. The main novelty of this paper is the generation of ARTS for the SemEval-2015 [14] and SemEval-2016 [15] datasets. The generated ARTS are used to compute the ARS for the HAABSA++ model. We also evaluate whether adversarial training improves the ARS of the HAABSA++ model and how well HAABSA++ compares to other models found in the literature in terms of robustness measured by ARS.

The remaining sections of this paper are organized as follows. Section 2 reviews relevant studies related to ABSC. Section 3 outlines the data used in this study. Section 4 explains the research methodology, with the results reported in detail in Section 5. The final section, Section 6, summarizes the conclusions and suggests directions for future research. Access to the source code is freely available at `https://github.com/MiladAgha/rHAABSA-pp`.

2 Related Work

Most approaches to ABSC can be grouped into three types [1]. Knowledge-based methods rely on part-of-speech tags and lexicons. This type of method was the first to tackle the ABSC [8]. Later, machine learning methods emerged as a convenient alternative with high performance rates [16]. These methods offer more flexibility, while knowledge-based solutions are better for sentiment classification within a specific domain because they are based on human labour. These two strategies are complementary [22], giving birth to hybrid approaches.

Recently, a growing number of studies have explored the hybrid models. [3] proposes a hybrid model called BBLSTM, which includes a bitmask layer to focus attention on specific aspects within the text and incorporates sentiment lexicons into the domain-specific word embeddings. [5] enhances conventional attention-based LSTM models by introducing a method that better captures the meaning of the opinion target and by integrating syntactic information into the attention mechanism. [11] proposes a hybrid model called ALDONAr, which merges a lexicalized domain ontology with a regularized neural attention model. [20] introduced a hybrid model called HAABSA, which also uses a lexicalized domain ontology as [11] but combines it with a neural attention model with a rotatory attention mechanism. [19] presented a hybrid model called HAABSA++ that extends the HAABSA model with deep contextual word embeddings and hierarchical attention. These sentiment classification models produce impressive results. However, the robustness of these models must yet be tested [10].

[6] shows that the majority of sentences in existing ABSA datasets contain either a single aspect or multiple aspects of the same polarity, simplifying the ABSC task to sentence-level sentiment classification, thereby causing low robustness. [6] presented a new large-scale dataset called Multi-Aspect Multi-Sentiment (MAMS). Unlike existing ABSA datasets, MAMS sentences contain at least two different aspects with different sentiment polarities. This protects ABSC models from collapsing into sentence-level sentiment analysis, thereby increasing the

robustness of the ABSC model. However, it can be very expensive to annotate the robustly driven MAMS data, as it is fully obtained through human labor.

[21] explored the robustness problem by developing an automatic generation framework designed to generate a comprehensive ARTS used to evaluate aspect robustness. It employs three strategies (RevTgt, RevNon, and AddDiff) for generating variations that preserve the original content and aspect terms while separating the sentiment polarity of non-target aspects from the intended target. Combined with the new measure ARS, ARTS can be used to assess whether a model can reliably detect the intended sentiment in the presence of spurious patterns.

3 Specification of Used Datasets

To train and test our model for aspect-based sentiment classification, we use the datasets from SemEval-2015 Task 12 [14] and SemEval-2016 Task 5 Subtask 1 [15]. SemEval is a series of workshops that are commonly used for this type of evaluation task. This is an advantage as our results can be easily compared to other methods using the same datasets.

The SemEval datasets contain restaurant reviews, with each review holding at least one sentence. Each sentence has zero or more opinions, where an opinion is made up of the target aspect, the category of the aspect being evaluated, and a polarity that expresses whether the reviewer is positive, negative, or neutral about the specific aspect. An illustration of a sentence from the SemEval-2016 dataset in XML format is shown in Listing 1.1.

```
<sentence id="404464:2">
  <text>The entree was bland and small, dessert was not inspired</text>
  <Opinions>
    <Opinion target="entree" category="FOOD#QUALITY" polarity="negative"
        from="4" to="10"/>
    <Opinion target="entree" category="FOOD#STYLE_OPTION" polarity="
        negative" from="4" to="10"/>
    <Opinion target="dessert" category="FOOD#QUALITY" polarity="negative"
         from="32" to="39"/>
  </Opinions>
</sentence>
```

Listing 1.1. Example sentence of the SemEval-2016 dataset in XML format.

We clean the data of implicit aspects because our model requires explicit aspects for the aspect-based sentiment classification task. This step removes 25% of the data. We also eliminate repeated annotation of the same target aspects in the data. This happens because aspects belong to more than one aspect category. An example of this is shown in Listing 1.1. This step removes another 5% of the data. After cleaning, the SemEval-2015 data has a training set of 254 reviews with 1210 aspects and a test set of 96 reviews with 559 aspects.

The SemEval-2016 data has a training set of 350 reviews with 1769 aspects and a test set of 90 reviews with 623 aspects. For the training set and the test set of the SemEval 2015 and SemEval 2016 data sets, some statistics are shown. Fig. 1 shows the division of the aspects according to the sentiments they convey, indicating that positive polarity is the most common among the aspects and that the datasets have comparable proportions of polarity. Fig. 2 shows the frequency of the number of aspects per sentence. We can see that sentences with a single aspect are the most common.

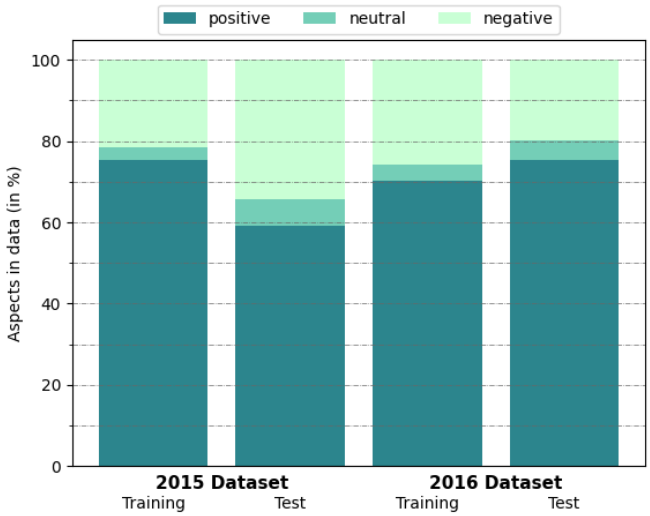

Fig. 1. Relative frequencies of sentiment labels.

4 Methodology

This section outlines the approach to check HAABSA++ robustness. First, the sentiment of the inputs is assessed using a domain sentiment ontology. If this method fails, a neural network is used as a fallback option. The neural network receives augmented data through a proposed data generation method to determine the Aspect Robustness Score (ARS). Furthermore, the study investigates the improvement of the ARS through adversarial training, which involves augmenting the training data.

In this study, we focus on the aspect robustness of the HAABSA++ model. To evaluate the aspect robustness of ABSA models, [21] introduced a novel metric called ARS and developed ARTS to augment the test set with three types of data augmentation techniques: RevTgt, which introduces tokens that invert

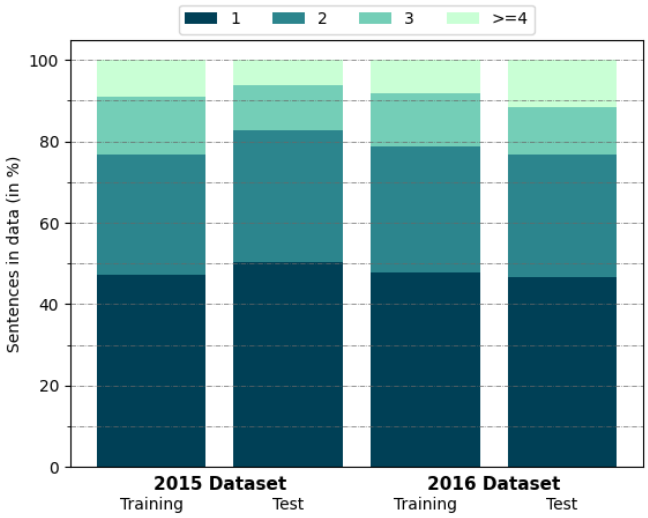

Fig. 2. Relative frequencies of aspects per sentence.

the sentiment associated with the target aspect, RevNon, which inserts tokens that maintain the sentiment of the target aspect while changing the sentiments of other aspects, and AddDiff, which appends new segments containing different aspects than the target with different sentiment from the target. The heuristics of these augmentation techniques are explained in more detail below.

RevTgt. The first augmentation technique aims to generate sentences that reverse the sentiment of the target aspect. This is achieved by two primary methods: opinion reversal and negation addition. When reversing opinion words, the approach starts with antonym replacement using WordNet, a lexical database [12]. However, to ensure context compatibility, only antonyms with the same Part-of-Speech tag as the original word are considered. When multiple options are available, random selection is applied, giving priority to antonyms already in the vocabulary. For target-oriented extraction of opinion words, manually annotated data provided by [4] is used. An example of the flip opinion strategy is shown in Table 1. In this example, the opinion word “best" for the target aspect “meal" in the sentence “It is simply the best meal in NYC" is flipped to the opinion word “worst" to produce the new sentence “It is simply the worst meal in NYC". In cases where no suitable antonyms can be found, or for longer phrases, negation is introduced based on linguistic features. Typically, this involves adding “not" in front of adjectives or verbs that express sentiment. If the sentiment term is not in the form of an adjective or verb, negation is applied to the closest verb.

Table 1 also provides an example of the add negation strategy, where the opinion word “authentic" for the target aspect “food" in the sentence “The food was authentic" is negated to produce the new sentence “The food was not au-

Table 1. Strategies and examples of RevTgt. The target aspects are underlined and all of the relevant opinion words and conjunctions are in bold type.

Strategy	**Example**
Flip Opinion	It's simply the **best** meal in NYC. → It's simply the **worst** meal in NYC.
Add Negation	The food was **authentic**. → The food was **not authentic**.
Adjust Conjunctions	The food is so **cheap and** the waiters are nice. → The food is so **expensive but** the waiters are nice.

thentic". Adjustments are also made to conjunctions to ensure fluency. Reversing the sentiment of one aspect can result in conflicting sentiments between aspects. Conjunctions are adjusted accordingly. Cumulative conjunctions such as "and" are used when adjacent sentiments have the same polarity, while adversative conjunctions such as "but" are used when sentiments differ. Also, an example of the adjust conjunctions strategy is shown in Table 1. In this example, because the opinion word "cheap" for the target aspect "food" in the sentence "The food is so cheap and the waiters are nice" is flipped to the opinion word "expensive", the conjunction word "and" is changed to the conjunction word "but" to produce the new sentence "The food is so expensive, but the waiters are nice".

RevNon. The second augmentation technique reverses sentiments for non-target aspects that have the same sentiment as the target aspect, using the RevTgt method. An example of the flip same-sentiment non-target aspects strategy is shown in Table 2. In this example the opinion word "clean" for the non-target aspect "restaurant" in the sentence "The service is good and the restaurant is clean" is flipped to the opinion word "dirty" and the conjunction word "and" is changed to the conjunction word "but" to produce the new sentence "The service is good but the restaurant is dirty". In addition, for non-target aspects that already exhibit sentiments opposite to the target sentiment, the approach amplifies that opposition. This is achieved by randomly selecting adverbs from a specialized dictionary of degree adverbs compiled from the training data. This deliberate exaggeration serves to emphasize the existing contrast. Also, an example of the exaggerated opposite-sentiment non-target aspects strategy is shown in Table 2. In this example the opinion word "amazing" for the non-target aspect "food" in the sentence "Decor needs to be upgraded but the food is amazing!" is exaggerated to "greatly amazing" to produce the new sentence "Decor needs to be upgraded but the food is greatly amazing!".

AddDiff. The last augmentation technique explores the impact of appending additional non-target aspects on model performance. The first step of this technique is the creation of the AspectSet, where all aspect expressions from the dataset are extracted. This is achieved by identifying sentiment terms in each instance (e.g., "bad" in "Plain and simple it's bad thai food.") and then retrieving their linguistic branches (e.g., "bad thai food") using pretrained con-

Table 2. Strategies and examples of RevNon. The target aspects are underlined, the non-target aspects are in italics, and all of the relevant opinion words and conjunctions are in bold type.

Strategy	Example
Flip same-sentiment non-target aspects	The service is good **and** the *restaurant* is **clean**. → The service is good **but** the *restaurant* is **dirty**.
Exaggerate opposite-sentiment non-target aspects	Decor needs to be upgraded but the *food* is **amazing**! → Decor needs to be upgraded but the *food* is **greatly amazing**!

stituency parsing [7]. Some examples of found aspect expressions contained in the AspectSet are shown in Table 3.

Then, one to three aspects not mentioned in the original sample and with sentiments different from the target aspect are randomly selected from AspectSet and added to the original instance. For example, the aspect expressions "a lot of food" and "a nice pizza place" are added to the sentence "The kitchen however, is almost always slow" to obtain the new sentence "The kitchen however, is almost always slow, *but* a lot of food *and* a nice pizza place." This augmentation technique allows us to assess whether the introduction of more unrelated aspects with opposite sentiments confuses the model.

The trio of data augmentation techniques are applied to extend the test sets for the SemEval 2015 and 2016 restaurant domain datasets, as [21] did for the SemEval 2014 restaurant domain dataset. Aspect robustness is then evaluated using ARS [21], a measure that considers the accurate classification of the original example and its modified versions (RevTgt, RevNon, and AddDiff) as a single correct instance. Additionally, improving ARS through adversarial training is explored by applying these three data augmentation methods to the training set.

Fig. 3 provides a high-level overview of how our method is structured to help to visualise where the augmented data is fed into our method.

Table 3. Examples of aspect expressions in the AspectSet.

Sentiment	Aspect Expression
Positive	decor is nice and minimalist the service is always outstanding every pie is ultra fresh
Negative	the atmosphere is noisy lobster ravioli was very salty their deliveries take for ever

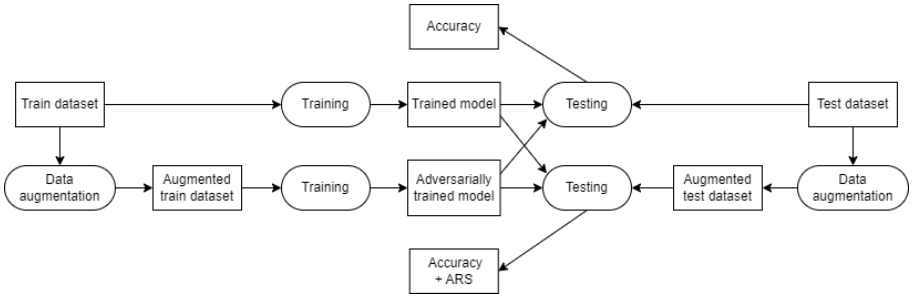

Fig. 3. A high-level overview of the structure of the proposed method, showing where the augmented data is fed into the method.

5 Results

In this section, we present the obtained results. In Section 5.1, some statistics for the generated ARTS of SemEval 2015 and SemEval 2016 are discussed. In Section 5.2, the results for the adversarial training are presented. Last, in Section 5.3, we compare the HAABSA++ model with other models in the literature to see how it performs in terms of robustness as measured by ARS.

5.1 Data Augmentation

In Table 4 the number of instances in the tests and trains set for the original and augmented SemEval 2015 and SemEval 2016 data sets are presented. We see that for the test sets the original data is enriched from 559 and 623 instances to 1725 and 1868 instances of the augmented data for the SemEval 2015 and SemEval 2016 datasets, respectively. This is an increase in size of 209% and 200%, respectively. For the train sets the original data is enriched from 1210 and 1769 instances to 3938 and 5640 instances of the augmented data for the SemEval 2015 and SemEval 2016 datasets, respectively. An increase in size of 225% and 219%, respectively. Ideally, the desired increase in size is 300%, reflecting the three augmentation techniques. However, not all original instances qualify for each strategy, leading to this gap. AddDiff is applicable in all cases, but RevTgt and RevNon require explicit opinion words. The main obstacle for RevTgt is the lack of opinion words. Consequently, the number of new test instances generated by RevTgt matches the number of available opinion word instances. RevNon imposes additional constraints. First instances with only one aspect are filtered out. These instances have no non-target aspect and therefore RevNon cannot be applied. Furthermore, instances with overlapping opinion words between target and non-target aspects are also filtered out. This is because in these cases RevNon cannot be used to reverse the sentiment of the non-target aspect without also changing the sentiment of the target aspect. Last, instances with neutral sentiments for all non-target aspects are excluded. There is no point in reversing or exaggerating the neutral sentiment of non-target

Table 4. The number of instances in the test set and train set for the original and augmented SemEval 2015 and SemEval 2016 datasets. Also, the instance count for each augmentation technique is shown.

	Original	Augmented	RevTgt	RevNon	AddDiff
Dataset	Train Test	Train Test	Train Test	Train Test	Train Test
SemEval 2015	1210 559	3938 1725	1068 442	450 165	1210 559
SemEval 2016	1769 623	5640 1868	1498 454	604 168	1769 623

aspects. Therefore, no new instance is created with RevNon if all non-target aspects have neutral sentiment. The instance counts for each used augmentation technique are also shown in Table 4.

5.2 Model Performance

Because the ontology model is not trainable, we do not use the augmented test data on the ontology model and the ontology part of the HAABSA++ model. To evaluate the robustness of the neural network part of the HAABSA++ model, we feed the original test data into the ontology reasoner and augment the inconclusive instances of the original test data.

In Table 5, we see that the accuracy of the HAABSA++ model with regular training for the original test set of the SemEval 2015 and SemEval 2016 datasets is 74.6% and 83.0%, respectively. After adversarial training, we see that the accuracy for the original test set of the SemEval 2015 and SemEval 2016 datasets drops to 73.2% and 81.5%, respectively. A possible reason for this drop is that by adding the new instances to the training set, the distribution of the training set is changed in a way that makes it more difficult for the HAABSA++ model to correctly predict the sentiment of the original test set. The HAABSA++ model performs better when compared to the stand-alone ontology model, which has an accuracy of 65.7% and 79.0% for the original test set of the SemEval 2015 and SemEval 2016 datasets, respectively. Surprisingly, however, and in contrast to the results found in the paper of [19], the HAABSA++ model performs worse than the standalone neural network model, which has an accuracy of 77.1% and 86.7% with regular training, and 76.6% and 83.3% with adversarial training for the original test set of the SemEval 2015 and SemEval 2016 datasets, respectively. We see again that after adversarial training the accuracy of the original test sets drop.

To better understand the performance of the HAABSA++ model, we look at the performance of the individual parts of the hybrid model. We see that the accuracy of the ontology part of the HAABSA++ model is quite high, with values of 80.1% and 86.2% for the original test set of the SemEval 2015 and SemEval 2016 datasets, respectively. However, we see that the accuracy of the ontology part of the HAABSA++ model is even higher when done by the neural network, with values of 85.1% and 92.2% for the original test set of the SemEval 2015 and

Table 5. Accuracy of the regularly and adversarially trained HAABSA++ model for the original and augmented test sets of the SemEval 2015 and SemEval 2016 datasets. The accuracy of the ontology and the neural network within the HAABSA++ model as well as stand-alone models are also shown. For the augmented test sets ARS is given in parentheses.

	SemEval 2015		SemEval 2016	
	Original	Augmented	Original	Augmented
HAABSA++				
Regular training	74.6%	-	83.0%	-
Adversarial training	73.2%	-	81.5%	-
Ontology part	80.1%	-	86.2%	-
(If done by Neural Network)	(85.1%)		(92.2%)	
Neural Network part				
Regular training	69.3%	56.7% (24.7%)	77.8%	61.0% (30.1%)
Adversarial training	66.4%	68.5% (47.3%)	74.1%	77.7% (60.7%)
Ontology	65.7%	-	79.0%	-
Neural Network				
Regular training	77.1%	62.0% (24.2%)	86.7%	67.3% (30.5%)
Adversarial training	76.6%	76.9% (57.6%)	83.3%	84.9% (70.9%)

SemEval 2016 datasets, respectively. This explains why the HAABSA++ model performs worse than the standalone neural network model. The neural network part of the HAABSA++ model performs rather poorly with an accuracy of 69.3% and 77.8% with regular training and 66.4% and 74.1% with adversarial training for the original test set of the SemEval 2015 and SemEval 2016 datasets, respectively. A possible explanation for this is that the ontology reasoner picks out all of the easy instances of the test set and leaves the neural network with all the difficult instances.

To find the benefits of the adversarial training we have to look at the accuracy and ARS of the models for the augmented test sets. The stand-alone neural network model has an accuracy (ARS) of 62.0% (24.2) and 67.3% (30.5) with regular training but an accuracy (ARS) of 76.9% (57.6) and 84.9% (70.9) with adversarial training for the augmented test set of the SemEval 2015 and SemEval 2016 datasets, respectively. We see a big increase in accuracy and ARS for the adversarially trained model compared to the regularly trained model. This increase in accuracy and ARS is a direct result of the adversarial training. The adversarial training changes the distribution of the training set in a way that allows the model to better recognize the new instances of the augmented test set, increasing accuracy and ARS. We observe a similar but smaller increase in both the accuracy and ARS when we look at the outcomes of the neural network part of the HAABSA++ model. The neural network part of the HAABSA++

model has an accuracy (ARS) of 56.6% (24.7%) and 61.0% (30.1%) with regular training but an accuracy (ARS) of 68.5% (47.3%) and 77.7% (60.7%) with adversarial training for the augmented test set of the SemEval 2015 and SemEval 2016 datasets, respectively. A possible reason why the increase for both accuracy and ARS is smaller is again that the neural network part of HAABSA++ is initially left with all the difficult instances because the ontology reasoner has picked out all the easy instances of the test set. The augmentation of these difficult instances creates even more difficult new instances, which causes the difference in the increase of both accuracy and ARS.

In all results of the augmented test sets, we see that the ARS is lower than the accuracy when we compare the ARS with the corresponding accuracy of the augmented test sets. This is expected due to the strict definition of ARS. ARS considers the correct classification of the original example and its modified versions as a single correct instance. Thus, if only one is incorrectly classified, ARS will be incorrect for this instance.

If we compare the results of the original test sets with the augmented test sets (Table 5), we see a decrease in accuracy for regular training. This decrease can be explained by the fact that the new instances of the augmented test set are in some way different from the instances on which the model is trained with regular training. When we compare the results of the original test sets with the augmented test sets for adversarial training, this decrease in accuracy turns into a slight increase in accuracy. This is because adversarial training causes the model to better recognize the new instances of the augmented test set.

5.3 Robustness Comparison

Table 6 lists for several models the accuracy for the original test set of the SemEval 2014 Restaurant dataset and the ARS after regular training and after adversarial training for the augmented test set of the SemEval 2014 Restaurant dataset. Except for the HAABSA++ model, all models, accuracies, and ARS values in Table 6 are taken over from the paper of [21]. The accuracy for the augmented test set is not shown because they are not reported in the paper of [21]. To conclude how well the HAABSA++ models compare to these other models in terms of robustness as measured by ARS, we also evaluated our models on the SemEval 2014 Restaurant dataset. From Table 6, we conclude that ARS increases after adversarial training for all models and that the HAABSA++ model without ontology reasoner performs averagely in terms of accuracy and robustness as measured by ARS. The best results for ARS are obtained by the BERT models. It appears that capturing the semantic information of text by considering the surrounding context in addition to the words themselves, as BERT contextual word embeddings do, increases the robustness of these models as measured by ARS. Although the HAABSA++ model also makes use of BERT word embeddings, the performance of the HAABSA++ model without ontology reasoner is more similar to the TD-LSTM model. A possible explanation for this is that the HAABSA++ model uses pre-trained BERT word embeddings, rather

Table 6. The table reports the accuracy for the original test set of the SemEval 2014 Restaurant dataset, and the ARS after regular training and after adversarial training for the augmented test set of the SemEval 2014 Restaurant dataset for several models.

		ARS	
	Accuracy	Regular training	Adversarial training
BERT-PT	86.7%	59.3%	74.6%
CapsBERT	83.5%	55.4%	75.8%
BERT	83.0%	54.8%	74.8%
HAABSA++ (w/o ont)	77.8%	31.5%	66.2%
TD-LSTM	78.1%	30.2%	62.8%
HAABSA++	72.2%	29.6%	55.1%
GCN	77.9%	24.7%	61.5%
MemNet	75.2%	21.5%	38.0%
attLSTM	76.0%	14.6%	48.7%
GatedCNNN	77.0%	13.1%	37.5%
Average	**78.4%**	**31.5%**	**58.4%**

than training the word embeddings on the data, as is the case with the other BERT models.

6 Conclusion

In this paper, we inspected the robustness of the HAABSA++ model. This model uses a domain-specific ontology with a specialized neural network as a fail-safe to evaluate the sentiment toward aspect targets. To evaluate the aspect robustness of the HAABSA++ model, we augmented the test set with three types of data augmentation techniques: RevTgt, which inverts the sentiment associated with the target aspect, RevNon, which maintains the sentiment of the target aspect while changing the sentiments of other aspects, and AddDiff, which appends new segments containing different aspects than the target with different sentiment from the target. Aspect robustness is evaluated using ARS. Additionally, we explored improving ARS through adversarial training by applying the three data augmentation methods to the training set.

Surprisingly, we find that the HAABSA++ model performs better without using the ontology part of the model. Furthermore, we find that performing adversarial training indeed improves the robustness of the model as measured by ARS. However, this improvement comes at the cost of a lower accuracy of the HAABSA++ model for the original test set instances. When comparing the robustness as measured by ARS of the HAABSA++ model with other models found in the literature, we find that the HAABSA++ score is average.

A suggestion for future work is to improve the data augmentation method. In particular, the need for annotation of opinion words is a major drawback. This

limits the use of this method for other datasets where this annotation is not present. Furthermore, the data augmentation techniques could be modified to apply to the entire dataset. When an instance of the original data set has fewer newly created instances in the augmented data set, the ARS is less stringent. A more uniform ARTS prevents this. Also, it would be interesting to test how the change of order of model components in the classification pipeline from ontology first to neural network first, would affect the outcome.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Brauwers, G., Frasincar, F.: A survey on aspect-based sentiment classification. ACM Computing Surveys **55**(4), 1–37 (2023)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Pre-training of deep bidirectional transformers for language understanding. In: 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HCL 2019). pp. 4171–4186. ACL (2019)
3. Do, B.T.: Aspect-based sentiment analysis using bitmask bidirectional long short term memory networks. In: 31st International Florida Artificial Intelligence Research Society Conference (FLAIRS 2018). pp. 259–264. AAAI Press (2018)
4. Fan, Z., Wu, Z., Dai, X., Huang, S., Chen, J.: Target-oriented opinion words extraction with target-fused neural sequence labeling. In: 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HCL 2019). pp. 2509–2518. ACL (2019)
5. He, R., Lee, W.S., Ng, H.T., Dahlmeier, D.: Effective attention modeling for aspect-level sentiment classification. In: 27th International Conference on Computational Linguistics (COLING 2018). pp. 1121–1131. ACL (2018)
6. Jiang, Q., Chen, L., Xu, R., Ao, X., Yang, M.: A challenge dataset and effective models for aspect-based sentiment analysis. In: 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). pp. 6280–6285. ACL (2019)
7. Joshi, V., Peters, M., Hopkins, M.: Extending a parser to distant domains using a few dozen partially annotated examples. In: 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). vol. 1, pp. 1190–1199. ACL (2018)
8. Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S.: NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In: 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 437–442. ACL (2014)
9. Liu, B.: Sentiment analysis: Mining opinions, Sentiments, and Emotions. Cambridge university press, second edn. (2020)
10. Liu, X., Ding, Y., An, K., Xiao, C., Madhyastha, P., Xiao, T., Zhu, J.: Towards robust aspect-based sentiment analysis through non-counterfactual augmentations. arXiv preprint arXiv:2306.13971 (2023)
11. Meškelė, D., Frasincar, F.: ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. Information Processing & Management **57**(3), 102211 (2020)

12. Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
13. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2227–2237. ACL (2018)
14. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 task 12: Aspect based sentiment analysis. In: 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 486–495. ACL (2015)
15. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al.: SemEval-2016 task 5: Aspect based sentiment analysis. In: 10th International Workshop on Semantic Evaluation (SemEval 2016). pp. 19–30. ACL (2016)
16. Ren, Y., Zhang, Y., Zhang, M., Ji, D.: Context-sensitive twitter sentiment classification using neural network. In: 13th AAAI Conference on Artificial Intelligence (AAAI 2016). vol. 30, pp. 215–221. AAAI Press (2016)
17. Schouten, K., Frasincar, F.: Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering **28**(3), 813–830 (2016)
18. Schouten, K., Frasincar, F.: Ontology-driven sentiment analysis of product and service aspects. In: 15th Extended Semantic Web Conference (ESWC 2018), LNCS. vol. 10843, pp. 608–623. Springer (2018)
19. Truşcă, M.M., Wassenberg, D., Frasincar, F., Dekker, R.: A hybrid approach for aspect-based sentiment analysis using deep contextual word embeddings and hierarchical attention. In: 20th International Conference on Web Engineering (ICWE 2020), LNCS. vol. 12128, pp. 365–380. Springer (2020)
20. Wallaart, O., Frasincar, F.: A hybrid approach for aspect-based sentiment analysis using a lexicalized domain ontology and attentional neural models. In: 16th Extended Semantic Web Conference (ESWC 2019), LNCS. vol. 11503, pp. 363–378. Springer (2019)
21. Xing, X., Jin, Z., Jin, D., Wang, B., Zhang, Q., Huang, X.: Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. In: 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). pp. 3594–3605. ACL (2020)
22. Yanase, T., Yanai, K., Sato, M., Miyoshi, T., Niwa, Y.: bunji at SemEval-2016 task 5: Neural and syntactic models of entity-attribute relationship for aspect-based sentiment analysis. In: 10th International Workshop on Semantic Evaluation (SemEval 2016). pp. 289–295. ACL (2016)
23. Zheng, S., Xia, R.: Left-center-right separated neural network for aspect-based sentiment analysis with rotatory attention. arXiv preprint arXiv:1802.00892 (2018)