On the current state of reproducibility and reporting of uncertainty for Aspect-based Sentiment Analysis

Anonymous ACL submission

Abstract

For the latter part of the past decade, Aspect-Based Sentiment Analysis has been a field of great interest within Natural Language Processing. Supported by the Semantic Evaluation Conferences in 2014 – 2016, a variety of methods has been developed competing in improving performances on benchmark data sets. Exploiting the transformer architecture behind BERT, results improved rapidly and efforts in this direction still continue today. Our contribution to this body of research is a holistic comparison of six different architectures which achieved (near) state-of-the-art results at some point in time. We utilize a broad spectrum of five benchmark data sets and introduce a fixed setting with respect to the preprocessing, the train/validation splits, the performance measures and the quantification of uncertainty. Overall, our findings are two-fold: First, we find that the results reported in the scientific articles are hardly reproducible, since in our experiments the observed performance (most of the time) fell short of the reported one. Second, the results are burdened with notable uncertainty (depending on the data splits) which is why a reporting of uncertainty measures is crucial.

1 Introduction

The field of Natural Language Processing (NLP) has profited a lot from technical and algorithmic improvements within the last years. Before the successful times of Machine Learning and Deep Learning, NLP was mainly based on what linguists knew about how languages work, i.e. grammar and syntax. Thus, primarily rule-based approaches were employed in the past. Nowadays, far more generalized models based on neural networks are able to learn the desired language features.

On the other hand, data in written form is available in huge amounts and thus might be an important source for valuable information. For instance, the internet is full of comparison portals, forums, blogs and social media posts where people state their opinions on a broad range of products, companies and other people. Product developers, politicians or other persons in charge could profit from this information and improve their products, decisions and behavior.

We specifically focus on *Aspect-Based Sentiment Analysis (ABSA)* in our work. ABSA is often used as a generic term for several unique tasks, which is caused by the inconsistency of terms in literature where many different names are widely used. To be as precise as possible, we explicitly use different terms than ABSA to refer to the exact tasks. The first one (Subtask 2, Pontiki et al., 2014) assumes that in each text, aspect terms are already marked and thus given exactly as written in the text (this differs from so-called aspect categories which do not necessarily appear in the text). Here, the task is to classify the sentiments for those aspect terms. This is why the term *Aspect Term Sentiment Classification (ATSC)* is most accurate.

When referring to ATSC methods, we usually think of *single-task* approaches. These methods are designed to carry out only aspect term sentiment classification as the aspect terms are already given. Whether these were identified manually or by an algorithm is not relevant in this setting. In practice, however, the aspect terms oftentimes are not already known. Thus, approaches dealing with the step of *Aspect Term Extraction (ATE)* have been developed. They can either work on their own or be combined with an ATSC method. For these combined methods, which we refer to as *ATE+ATSC*, one can further distinguish between *pipeline*, *joint* and *collapsed* models. In pipeline models, ATE and ATSC are simply stacked one after another, i.e. the output of the first model is used as input to the second model. The latter two are often also referred to as *multi-task* models, since both tasks are carried out simultaneously or in an alternating way. These models only differ in their labeling mechanisms: There are two label sets for joint models, one to indicate whether a word is part of an aspect term and the other one to state its polarity. For collapsed models, a unified labeling scheme indicates whether a word is part of a positive, negative or neutral aspect term or not.

We re-evaluate four different models for ATSC, covering a variety of different architectures (RNNs, Capsule networks, LCF-based, BERT-based), as well as two different ATE+ATSC models, one of which is a pipeline approach while the other one works in a collapsed fashion. All models are re-trained five times using five different (identical) train/validation splits and tested on the respective test sets in order to (i) compare them on a common ground and (ii) quantify the epistemic uncertainty associated with the architectures and the data.

2 Related work

Related experiments were conducted by Mukherjee et al. (2021), yet with a different focus. On the one hand, the authors also try to reproduce results on the benchmark data sets from SemEval-14 about Restaurants and Laptops. However, they selected six other models than we did for which the implementations are provided in one repository [1]. For these, the authors observed a consistent drop of 1-2 % with respect to both accuracy and macro-averaged F1-Score F_1^{macro}. Mukherjee et al. (2021) reported a doubling of this drop when using 15% of the training data as validation data. On the other hand, they executed additional tasks which included the set-up of two new data sets about Men's T-shirts and Television as well as the model evaluation on them. Furthermore, they also experimented with cross-domain training and testing. Yet, several important points are not addressed by their work which is why we investigate them in our work. First, while they mostly care about comparing different types of architectures (Memory Networks vs. BERT), we instead focus on comparing the best performing models for different *tasks* (ATSC vs. ATE+ATSC). Further, we cover a larger variety of types of architectures by selecting the best performing representatives of several different types. Second, we stick closer to the original implementations (by using them, when available) whereas they exclusively rely on community designed implementations, which adds a further potential source of errors. Third, and most important, we provide estimates for the epistemic uncertainty of performance values and are thus able to (at least tentatively) explain performance differences due to different reporting standards.

3 Materials and Methods

This section will introduce the data sets we utilized for training and evaluation as well as the selected model architectures. We start by briefly explaining the data, before the models are described, since (reported) performance values on these data sets partly motivate our choices regarding the models. Descriptive statistics for all used data sets can be found in Tab. 1. Note that the data sets we eventually use for training and testing the models are all based on the *original* train/test splits. Further we apply *small* modifications (as described below) which were (a) also applied by some of the authors whose models we re-evaluate and (b) we perceive as reasonable. This allows us to evaluate all of the architectures on a common ground, which is not possible by comparing the reported values from the original publications alone. Nevertheless, we are aware of the fact that this might limit comparability of our results to the original ones to some extent.

3.1 Data Sets

SemEval-14 Restaurants This data set contains reviews about restaurants in New York. Pontiki et al. (2014) chose a subset of the restaurant data from Ganu et al. (2009) as training data[2], while collecting test data[3] themselves. Both were labeled for several subtasks in the same way. These data sets were designed for ATSC as well as its equivalent on *Aspect-category* level (ACSC), but we stick to ATSC samples only. For each identified aspect term within a sentence, the polarity is given as *positive*, *negative*, *neutral* or *conflict*. We deleted the labels of the latter category (*conflict*) from the data sets due to their rare appearance. This is similar to previous work (Fan et al., 2018; Bai et al., 2020; Yang et al., 2020; Li et al., 2019a), yet, they do not all mention or explain the removing process explicitly. Rarely appearing duplicate sentences which occurred in the training set were also removed in our work. Due to their small amount, this proce-

[1] https://github.com/songyouwei/ABSA-PyTorch

[2] http://metashare.ilsp.gr:8080/repository/browse/semeval-2014-absa-restaurant-reviews-train-data/479d18c0625011e38685842b2b6a04d72cb57ba6c07743b9879d1a04e72185b8/

[3] http://metashare.ilsp.gr:8080/repository/browse/semeval-2014-absa-test-data-gold-annotations/b98d11cec18211e38229842b2b6a04d77591d40acd7542b7af823a54fb03a155/

dure should not cause severe problems concerning the over-estimation of metrics. This might be the reason why a similar preprocessing step was, to the best of our knowledge, only performed in one other work (Xue and Li, 2018).

SemEval-14 Laptops The second domain-specific subset of the SemEval-14 data is on Laptops. The data were collected and annotated by Pontiki et al. (2014) for the task of ATE and/or ATSC. The training data set is publicly available,[4] just like the test data (see Footnote 3). Again, there were duplicate sentences in the training data which we deleted Xue and Li (cf. 2018). Unlike other benchmark data sets, both SemEval-14 data sets come without an official train/validation split.

MAMS A *Multi-Aspect Multi-Sentiment (MAMS)* data set for the restaurant domain was introduced by Jiang et al. (2019) who criticized existing data sets for not being adequate for ATSC. Since the data sets described above mainly consist of sentences which exhibit (i) only one single aspect or (ii) several aspects with the same sentiment, they argued that the task would not be much more difficult than a sentiment prediction on the sentence-level. To circumvent this issue, they extracted sentences of Ganu et al. (2009) which comprise at least two aspects with differing sentiments.[5] The data sets have the same structure as the SemEval-14 data sets, with the difference that Jiang et al. (2019) provide a fixed validation set for MAMS. The size of the validation split comprises about ten percent of the whole training set, which also inspired our choice when it comes to creating train/validation splits from the two SemEval-14 training data sets.

ARTS Xing et al. (2020) questioned the suitability of existing data sets for testing the aspect robustness of a model, i.e. whether the model is able to correctly identify the words corresponding to the chosen aspect term and predict its sentiment only based on them. Thus, the authors created an automatic generation framework that takes SemEval-14 test data (Restaurants and Laptops) as input and creates an *Aspect Robustness Test Set (ARTS)*. They used three different strategies to enrich the existing test set: The first one, REVTGT ("*reverse target*"), aims to reverse the sentiment of the chosen aspect term (also called "*target aspect*"). This is reached by flipping the opinion using antonyms or adding negation words like "not". Additionally, conjunctions may be changed in order to make sentences sound more fluent. Another strategy to augment the test set is REVNON ("*reverse non-target*") for which the sentiment of non-target aspects are (i) changed if they have the same sentiment as the target aspect or (ii) exaggerated if the non-target aspect is of a differing polarity. The third strategy called ADDDIFF ("*add different sentiment*") adds non-target aspects with an opposite sentiment which is intended to confuse the model. These non-target aspects are selected from a set of aspects collected from the whole data set and appended to the end of the sentence. ARTS are only designed to be used as test sets after training an architecture on the respective SemEval-14 training sets. The test sets for both restaurants and laptops are publicly available.[6] During the preparation of the ARTS data for CapsNet-BERT, we noticed that the start and end positions of some aspect terms were not correct. We changed them in order to make the code work properly and we also deleted duplicates (cf. Xue and Li (2018)). For these specific test sets, the *Aspect Robustness Score (ARS)* was introduced by Xing et al. (2020) in order to measure how well models can deal with variations of sentences. Therefore, each sentence and all its variations are regarded as one unit for which the prediction is only considered to be correct if the predictions for *all* variations are correct. These units alongside with their corresponding predictions are then used to compute the regular accuracy on the unit-level.

More Data Sets Recently more data sets have been published in addition to the ones mentioned beforehand. Mukherjee et al. (2021) proposed two new data sets about *Men's T-Shirts* and *Television*. The YASO data set (Orbach et al., 2020) has a different structure as it is a multi-domain collection. This is an interesting approach, yet also the reason for not considering it for our experiments: This data set is far better suited for cross-domain analyses, which is out of the scope of this work.

3.2 Models

MGATN A *multi-grained attention network (MGATN)* was proposed by Fan et al. (2018). Its *multi-grained attention* as able to take into account

[4] http://metashare.ilsp.gr:8080/repository/browse/semeval-2014-absa-laptop-reviews-train-data/94748ff4624e11e38d18842b2b6a04d7ca9201ec33f34d74a8551626be122856

[5] https://github.com/siat-nlp/MAMS-for-ABSA

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

Data Set	**Subset**	**Original Sentences in total**	**Sentences without Duplicates**	**Sentences for 3-class ATSC**	**Multi-Sentiment Sentences**	**Aspect Terms in total**	**Positive Aspect Terms**	**Negative Aspect Terms**	**Neutral Aspect Terms**	**Removed Conflict Aspect Terms**
SemEval-14 Restaurants	Training	3,044	3,038	1,978	320	3,605	2,161	807	637	91
	Test	800	800	600	80	1,120	728	196	196	14
SemEval-14 Laptops	Training	3,048	3,036	1,460	166	2,317	988	866	463	45
	Test	800	800	411	38	638	341	128	169	16
ARTS Restaurants	Test	2,784	2,784	2,784	206	3,528	1,952	1,103	473	0
ARTS Laptops	Test	1,576	1,576	1,576	74	1,877	883	587	407	0
MAMS Restaurant	Training	4,297	4,297	4,297	4,297	11,186	3,380	2,764	5,042	0
	Validation	500	500	500	500	1,332	403	325	604	0
	Test	500	500	500	500	1,336	400	329	607	0

Table 1: Descriptive Statistics for the five utilized data sets. "*Multi-Sentiment sentences*" are those with at least two different polarities after removing "conflict" polarity. "*Aspect Terms in total*" also exclude "conflict".

the interaction between aspects. We chose MGATN since it is reported to be the best performing RNN-based model on SemEval-14 data sets.

CapsNet-BERT *Capsules Networks* were initially proposed for the field of Computer Vision (Hinton et al., 2011; Sabour et al., 2017), with the so-called *capsules* being responsible for recognizing certain implicit entities in images. Each capsule performs internal calculations and returns a probability that the corresponding entity appears in the image. A variation of Capsule Networks for ATSC and its combination with BERT was introduced by Jiang et al. (2019). It was reported to outperform all other capsule networks with respect to their accuracy on the SemEval-14 Restaurants data. Additionally, it performed second-best on MAMS, which is why we selected it for this study. Furthermore, we assumed their results on SemEval-14 Restaurants data to be for three-class classification, as all the other results they refer to are also three-class. Yet, it is not fully clear to us which makes this experiment even more interesting.

RGAT-BERT The *Relational Graph Attention Network (RGAT)* was introduced by Bai et al. (2020). It utilizes a dependency graph representing the syntactic relationships between words of a sentence as an additional input. The RGAT encoder creates syntax-aware aspect term embeddings following the representation update procedures from *Graph Attentional Networks (GATs)* (Velickovic et al., 2018). It exhibits the best performance among graph-based models and also performs best on the MAMS data in terms of both accuracy and F_1^{macro}.

LCF-ATEPC Yang et al. (2020) built upon the idea of the *Local Context Focus (LCF)* mechanism (Zeng et al., 2019). The local context of an aspect term is defined as a fixed-size window around it, words outside this window are taken into account with lower weights or not at all. For each input token two labels, for aspect and sentiment, are assigned according to the joint labeling scheme described in Sec. 1. We chose LCF-ATEPC to be part of this meta-study since it reached the highest F_1^{macro} and accuracy on SemEval-14 data of all approaches. Yet, this only holds for the variant that is trained using additional domain adaptation.

BERT+TFM The approach described by Li et al. (2019b) consists of a BERT model followed by a Transformer (TFM) layer (Vaswani et al., 2017) for classification. BERT+TFM was the best model on SemEval-14 Laptops among all collapsed models at the time point of its introduction. There were also models using other layers on top instead of the Transformer layer, but our variant of choice was TFM as it produced slightly better results than the rest.

GRACE GRACE, a *Gradient Harmonized and Cascaded Labeling* model introduced by Luo et al. (2020), belongs to the category of pipeline approaches. It includes a post-training step of the pre-trained BERT (Devlin et al., 2019) model using Yelp[7] and Amazon data (He and McAuley, 2016). The post-trained model then shares its first l layers between the ATE and the ATSC task. The remaining layers are only used for the former. They are followed by a classification layer for the detected aspect terms. These classification outputs are then used again as inputs for a Transformer decoder which performs sentiment classification. The principle of using the first set of labels as input for the second is called *Cascaded Labeling* here and is assumed to deal with interactions between different

[7] https://www.yelp.com/dataset

aspect terms. *Gradient Harmonization* is applied in order to cope with imbalanced labels during training. GRACE appears to be the best of the pipeline models according to the literature. Furthermore, it is reported to be the best ATE+ATSC model on both SemEval-14 data sets. However, these successes have to be taken into account with care, as their results are based on four-class classification. This means that in comparison to the other authors' settings they did not exclude conflicting reviews of SemEval-14 data. Thus, our analyis contributes to comparability even more since it has not been established yet for our model-data combinations.

4 Experiments[8]

We re-evaluate six models (cf. Sec. 3.2) on the five data sets presented in Sec. 3.1. Our overall goals are to establish comparability between the models, to examine whether reported performance can be reproduced and to quantify epistemic model uncertainty that might exist due to the lacking knowledge about the train/validation splits.

First, we re-use the implementations provided by the authors and try to reproduce their results on the data sets they used. Second, we adapt their code to the remaining data sets and conduct the necessary modifications, again sticking as closely as possible to the original hyperparameter settings (cf. Appendix A). The biggest change we made was increasing the number of training epochs drastically and adding an early stopping mechanism. For all ATSC models, we selected the optimal model during the training process based on the validation accuracy and/or F_1^{macro}. For performing the experiments, we had a *Tesla V100 PCIe 16GB* GPU at our disposal.

Data Preparation Unlike other data sets, both SemEval-14 data sets come *without* an official validation split. Thus, we created five different train/validation splits (90/10) for each of the two SemEval-14 training sets. For each split, five training runs with different random initializations were conducted per model. The resulting 25 different versions per model per data set were subsequently evaluated on the two official SemEval-14 test sets as well as on the ARTS test sets. In Sec. 5 we report overall means per model per test set as well as means and standard deviations per model and test set for each of the different splits. Since there is an official validation set for MAMS, we did not apply the splitting procedure from above when training on this data set. Consequently, the given means and standard deviations are based on five training runs with different random initializations only.

MGATN As there exists no publicly available implementation by its authors, we used the one from a collection of re-implemented ABSA methods from GitHub.[9] We slightly modified the early stopping mechanism from that repository and then implemented it into the other re-evaluated models.

CapsNet-BERT We used the implementation of CapsNet-BERT provided by its authors.[10]

RGAT-BERT We relied on the implementation of RGAT-BERT provided by its authors.[11] Since the authors manually created an accuracy score different to the one from `sklearn`,we substituted their metric to ensure comparability. For data transformation, we selected the stanza tokenizer (Qi et al., 2020) over the Deep Biaffine Parser,[12] which was used by Bai et al. (2020), since the former provides the necessary syntactic information, whereas the latter failed to produce the syntactic dependency relation tags and head IDs the model requires.

LCF-ATEPC We were not able to run the best-performing LCF-ATEPC variant based on domain adaptation due to missing pretrained models. Thus, we decided to go for the second best, LCF-ATEPC-Fusion, using the official implementation of LCF-ATEPC.[13] During our experiments, the authors of LCF-ATEPC started building a new repository[14] based on the existing code which we did not use as it was still subject to changes.

BERT+TFM We used the implementation of BERT+TFM provided by its authors.[15] Our model selection was based on F_1^{micro} and F_1^{macro}, which were calculated based on *(start position, end position, polarity)*-triples for each identified aspect. Due to the collapsed labeling scheme, these scores account for both ATE and ATSC.

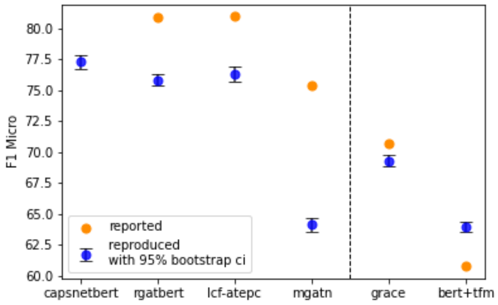
GRACE We used the post-trained BERT model provided by Luo et al. (2020).[16] Our model se-

[8]The complete source code (see appended zip-file) will be made available on GitHub upon publication.

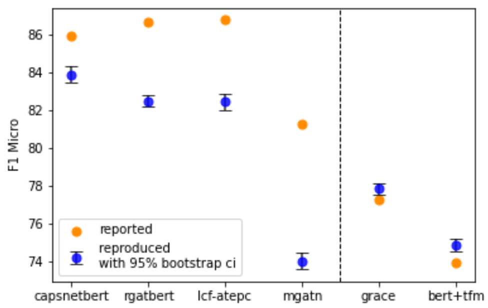
[9]https://github.com/songyouwei/ABSA-PyTorch
[10]https://github.com/siat-nlp/MAMS-for-ABSA
[11]https://github.com/muyeby/RGAT-ABSA
[12]https://github.com/yzhangcs/parser
[13]https://github.com/yangheng95/LCF-ATEPC
[14]https://github.com/yangheng95/pyabsa
[15]https://github.com/lixin4ever/BERT-E2E-ABSA
[16]https://github.com/ArrowLuo/GRACE

lection was based on ATSC-F_1^{micro} and -F_1^{macro} as well as on ATE-F_1^{micro}, with their calculations being slightly adjusted in order to match the calculation of those from BERT+TFM.

5 Results



(a) SemEval-14 Laptops



(b) SemEval-14 Restaurants

Figure 1: Comparison of reported and reproduced performance. The reproduced value is the mean of all 25 runs per model in total. Further, 95% bootstrap (n = 2000) confidence intervals are displayed. Note that absolute performance of GRACE (four classes) and BERT+TFM cannot be compared to the other models due to different tasks. No F_1^{micro} was reported for CapsNet-BERT on SemEval-14 Laptops.

In general, reported values were not reproducible. Fig. 1 shows a comparison of our average results to the reported results from the original publications on the SemEval-14 data sets. For all architectures there exists a notable gap between the blue (reproduced) and the orange (reported) values. In general, the gap tends to be larger for the ATSC models compared to the two ATE+ATSC models, where we could even reach a better performance for BERT+TFM within our replication study.[17]

It is also interesting to see how different runs can lead to rather broad ranges of results, although having done only five training runs per model and data split. An example for this phenomenon is the Accuracy of MGATN on SemEval-14 Laptops (cf. Fig. 2). For the first, the fourth and fifth split, all of the values lie very close together (within mean $\pm$ std), whereas the results of the other two splits show a rather high variance.

MGATN For MGATN, our reproduces results fell short of the reported values for accuracy, around five to ten percentage points for SemEval-2014 Laptops and Restaurants, respectively (cf. Tab. 4). Fig. 2 depicts the results on Laptops, the difference between reported and reproduced performance on the Restaurant data (not shown) looks similar. A reason for this behavior might be that we could not use the official implementation of the authors. In terms of ARS Accuracy on ARTS Restaurants, MGATN was the only model that reached only a single-digit value which means that it is not good at dealing with perturbed sentences.

CapsNet-BERT Comparing all the selected models on the ATSC task, CapsNet-BERT performed best on all data sets regarding all the metrics except for ARS Accuracy on ARTS Restaurant data (cf. Tab. 4). For ARTS, it seems as if the reported ARS accuracy for Laptops matched our result for Restaurants, and vice versa, as Fig. 3 illustrates. As far as we can tell, we did not mix up the data sets during our calculations which makes this look quite peculiar. The difference between the reported and reproduced values on SemEval-14 Restaurants data (as shown in Fig. 1b) may be explained by the fact that we did three-class classification and we only assumed so for the reported value.

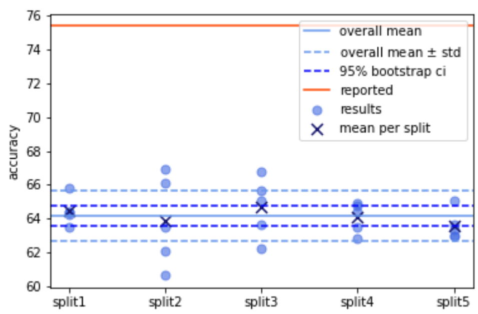
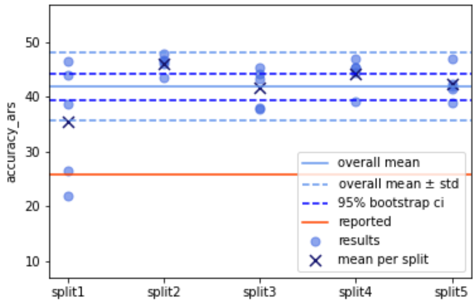
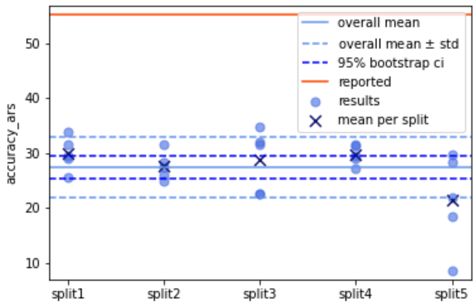


Figure 2: Example for high differences between data splits: Accuracy of MGATN on SemEval-14 Laptops.

[17] We do not give a similar figure for MAMS or ARTS as there are not enough reported values to form a good graph.



(a) ARTS Laptops



(b) ARTS Restaurants

Figure 3: Aspect Robustness Score (ARS) Accuracy of CapsNet-BERT.

RGAT-BERT For both SemEval-14 and MAMS we missed the reported values by around five percentage points (cf. Tab. 4). ARTS Restaurants is the only data set on which the best ARS Accuracy was not reached by CapsNet-BERT, but RGAT-BERT. Regarding MAMS, Bai et al. (2020) provided accuracy as well as F_1^{macro}, which is why we also compare these results here. Figure 4 shows the all five values of the four different measures as well as the average. For accuracy and F_1^{macro}, reported values from Bai et al. (2020) were added.

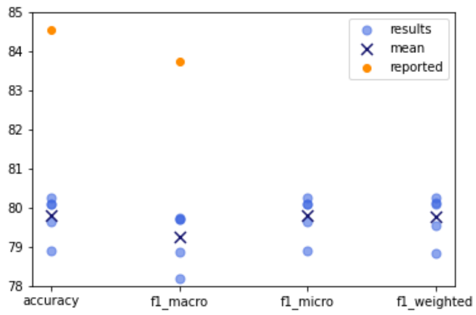
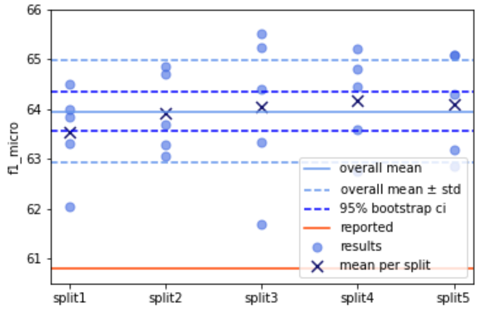
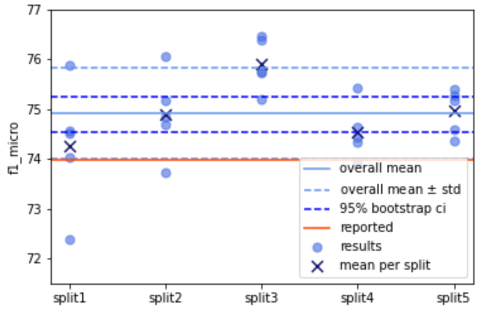


Figure 4: Performance of RGAT-BERT on MAMS.



(a) SemEval-14 Laptops



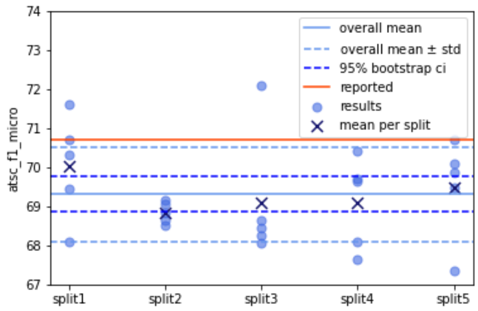
(b) SemEval-14 Restaurants

Figure 5: F_1^{micro} of BERT+TFM.

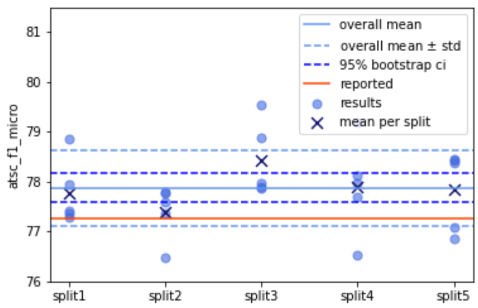
LCF-ATEPC Our experiments resulted in on average about five percentage points lower accuracies for LCF-ATEPC than were reported. Yet, LCF-ATEPC reached the best ARS Accuracy value on ARTS Restaurant data in our analysis.

BERT+TFM In contrast to the majority of the other models, for BERT+TFM the (average) performance of our runs surpassed the reported performance values on the SemEval-14 data. As Fig. 5 indicates, this holds for all runs (Laptop domain) and on average (Restaurant domain). The reasons for our improved values may lie in the chosen hyperparameters, yet we cannot tell for sure.

GRACE During our experiments with GRACE, we were able to produce results approximately in the same range as the reported values. Regarding SemEval-14 Restaurants our results on average were better than the reported ones (cf. Fig. 6b), while Laptops we could not quite reach the performance (cf. Fig. 6a). For the latter case, our results of single runs were better than (or at least equal to) the reported one, which is kind of a symptom of the problem. If we only reported the best of all runs, our conclusion would have been that we were able to outperform the original model. However, as



(a) SemEval-14 Laptops



(b) SemEval-14 Restaurants

Figure 6: ATSC F_1^{micro} of GRACE.

we have already mentioned, reported results were based on four-class classification, whereas our results were made for three-class. This might be the reason for different results. In the ATE+ATSC task, GRACE outperformed BERT+TFM on all data sets except for MAMS (cf. Tab. 5).

6 Discussion

Results differing from the reported values can be explained by various reasons. First, we often do not know how the reported values were created, i.e. whether the authors took the best or an average value of their runs. In Fig. 6a, it is clear to see that taking the best value compared the mean of the runs yields a difference of about almost three percentage points. Unfortunately there are also, to the best of our knowledge, no clear guidelines for how to properly report the uncertainty resulting from different data splits. One potential starting point could be to *always* perform multiple runs on multiple splits and use the different results to report variance values between and within splits. While the former gives an impression for the uncertainty induced by data heterogeneity, the latter rather reflects the model's share of the overall uncertainty. Second, our data usually are not identical to the data sets used for the original papers due to the preprocessing steps we explained beforehand. Also, training and validation splits are probably different from ours. Some models required additional syntactical information which we (potentially) inferred from other packages than indicated, because either none were given or because the ones that were given did not work as stated. Third, hyperparameter configurations are often not totally clear due to a lack of concise descriptions in the original work. In these cases we took those that were chosen by default in the implementations we used. Since those were not necessarily always provided by the authors of the models, we have no information about how close they are to the original configurations. What we could find out regarding hyperparameters can be found in Table 2 and 3 in Appendix A. Consequently, it is not surprising that we were not able to exactly reproduce given results, since hyperparameter tuning often has a large impact on the model performance. This insight is also shared by Mukherjee et al. (2021), although they tested other models in a different setup.

7 Conclusion & Future work

Our experiments revealed that reproducing reported results is hardly possible, given the current practice of performance reporting (at least for this subset of selected models). A tendency towards lower results is visible in our experiments, sometimes even five to ten percentage points lower than the original values. The only exception was BERT+TFM for which given values were surpassed. The reasons for these observations may lay in the data preprocessing step, in the hyperparameters or in the absence of a convention on which values to report (best or mean of several runs). This discovery of models hardly being comparable based on their performance measures is a very important one from our point of view. When new models are proposed, one of the main aspects during their evaluation is the improvement with respect to the state of the art. But when the performance of a single model can vary between single runs, the question is which results to take into account for model rankings.

A reporting convention indicating a common procedure combined with already prepared data sets with all possible labels could improve the comparability between models a lot. Also a huge practical meta-analysis of all models on several data sets would clarify the situation.

References

Xuefeng Bai, Pengbo Liu, and Yue Zhang. 2020. Investigating typed syntactic dependencies for targeted sentiment classification using graph attention neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:503–514.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, Brussels, Belgium. Association for Computational Linguistics.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. *Twelfth International Workshop on the Web and Databases (WebDB 2009)*.

Ruining He and Julian 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 '16, page 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. 2011. Transforming auto-encoders. In *Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I*, ICANN'11, page 44–51, Berlin, Heidelberg. Springer-Verlag.

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.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting BERT for end-to-end aspect-based sentiment analysis. *CoRR*, abs/1910.00883.

Huaishao Luo, Lei Ji, Tianrui Li, Nan Duan, and Daxin Jiang. 2020. GRACE: gradient harmonized and cascaded labeling for aspect-based sentiment analysis. *CoRR*, abs/2009.10557.

Rajdeep Mukherjee, Shreyas Shetty, Subrata Chattopadhyay, Subhadeep Maji, Samik Datta, and Pawan Goyal. 2021. Reproducibility, replicability and beyond: Assessing production readiness of aspect based sentiment analysis in the wild. *arXiv preprint arXiv:2101.09449*.

Matan Orbach, Orith Toledo-Ronen, Artem Spector, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Yaso: A new benchmark for targeted sentiment analysis. *arXiv preprint arXiv:2012.14541*.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. *CoRR*, abs/1710.09829.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Petar Velickovic, Guillem Cucurull, A. Casanova, Adriana Romero, P. Lio', and Yoshua Bengio. 2018. Graph attention networks. *ArXiv*, abs/1710.10903.

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. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3594–3605.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. *CoRR*, abs/1805.07043.

Heng Yang, Biqing Zeng, Jianhao Yang, Youwei Song, and Ruyang Xu. 2020. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *CoRR*, abs/1912.07976.

Biqing Zeng, Haishun Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9:3389.

Appendix

A Specifications and hyperparameters of the evaluated models

For upcoming tables, the following conventions will be used:

B	BERT Dimension
BS	Batch Size
CS	Capsule Size
E	Embedding Dimension
H	Hidden Dimension
# L	Number of Layers
LR	Learning Rate

		Training						Model				*Deviation*
Model	Data Set	Epochs	BS	LR	Dropout	Warmup %	L2	Max Len	# L	E	H	*(to Original)*
MGATN	all	50	16	1,00E-03	0.1	0	1,00E-02	85		300	300	Dropout: 0.5; L2: 1e-5
RGAT-BERT	Laptops	15	16	1,00E-05	BERT-specific	0	1,00E-05	90	2			
	Restaurants/MAMS	15	16	1,00E-05	BERT-specific	0	1,00E-05	90	2			
CapsNet-BERT	all	50	32	2,00E-05	0.1	0	0	-				LR: 3e-5
LCF-ATEPC	all	50	16	3,00E-05	0	0	1,00E-05	80				Epochs: 5
BERT+TFM	Laptops	50	32	2,00E-05		0	0	128				BS: 25; 1.5k steps, no epochs
	Restaurants/MAMS	50	16	2,00E-05		0	0	128				model selection after 1k steps
GRACE	ATE+GHL	5	32	3,00E-05		0.1		128				
	ATE+GHL+VAT	1	32	1,00E-05		0.1		128				
	ATE+ATSC+GHL+VAT	50	32	3,00E-06		0.1		128				Epochs: 10

Table 2: Model hyperparameters (Part I)

		BERT (Specific)			Other Specifications (Model specific)					*Deviation*
Model	Data Sets	H	Output Dimension	LR	Input Dropout	Attention Dropout	Dependency Dimension	SRD	Capsule Size	*(to Original)*
MGATN	all	768	-	-				3		
RGAT-BERT	Laptops		100	2,00E-05	0.1	0.1	100	-		
	Restaurants/MAMS		100	2,00E-05	0.1	0	80	-		
CapsNet-BERT	all	bert-base-uncased (huggingface)						-	300	
LCF-ATEPC	all	bert-base-uncased (huggingface)						5		
BERT+TFM	Laptops	bert-base-uncased (huggingface)						-		
	Restaurants/MAMS	bert-base-uncased (huggingface)						-		
GRACE	ATE+GHL	bert-base-uncased (huggingface)						-		
	ATE+GHL+VAT	bert-base-uncased (huggingface)						-		
	ATE+ATSC+GHL+VAT	bert-base-uncased (huggingface)						-		

Table 3: Model hyperparameters (Part II)

B Complete results

The following tables show the quantitative results of our experiments. For SemEval-14, five train-validation splits were created out of the original training set. On each split pair, five runs were performed which lead to split-specific means and standard deviations. In the overall mean and deviation, all runs of all splits are included. Consequently, they are based on 25 values for SemEval-14 and ARTS data and five values for MAMS data (as there were no splits applied).

Metric	Model	SemEval-14 Restaurant						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
Accuracy = F1 Micro	MGATN	74.32 (±1.24)	74.36 (±1.47)	74.70 (±0.73)	73.23 (±1.07)	73.66 (±0.81)	74.05 (±1.14)	81.25
	RGAT-BERT	82.52 (± 0.60)	83.21 (±0.88)	82.00 (±1.13)	82.70 (±0.67)	82.09 (±0.60)	82.50 (±0.86)	86.68
	CapsNetBERT	84.46 (±0.84)	84.07 (±0.92)	84.68 (±0.87)	83.46 (±0.63)	82.77 (±1.40)	83.89 (±1.13)	85.93
	LCF-ATEPC	82.56 (±0.89)	83.09 (±0.49)	82.87 (±1.28)	82.01 (±1.06)	81.78 (±1.52)	82.46 (±1.13)	86.77
F1 Macro	MGATN	62.04 (±2.37)	60.48 (±2.78)	61.34 (±0.99)	59.05 (±3.13)	57.15 (±3.70)	60.01 (±3.08)	71.94
	RGAT-BERT	72.88 (±0.68)	75.00 (±1.72)	72.86 (±2.21)	73.59 (±2.27)	72.39 (±0.81)	73.34 (±1.79)	80.92
	CapsNetBERT	76.21 (±1.59)	76.85 (±0.87)	77.02 (±1.66)	74.50 (±1.06)	72.43 (±4.07)	75.40 (±2.66)	-
	LCF-ATEPC	73.33 (±2.34)	75.17 (±0.38)	74.03 (±2.85)	73.22 (±1.58)	71.38 (±2.76)	73.43 (±2.36)	80.54
F1 Weighted	MGATN	72.83 (±1.56)	71.91 (±1.81)	72.53 (±0.48)	71.08 (±1.75)	70.03 (±2.23)	71.68 (±1.84)	-
	RGAT-BERT	81.03 (±0.54)	82.42 (±1.11)	81.09 (±1.37)	81.80 (±1.32)	80.76 (±0.67)	81.42 (±1.15)	-
	CapsNetBERT	83.50 (±1.00)	83.65 (±0.75)	83.98 (±1.09)	82.48 (±0.71)	81.02 (±2.44)	82.93 (±1.65)	-
	LCF-ATEPC	83.86 (±0.73)	83.80 (±0.70)	83.97 (±0.89)	82.88 (±1.09)	83.61 (±1.37)	83.63 (±0.99)	-
Metric	**Model**	**SemEval-14 Laptop**						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
Accuracy = F1 Micro	MGATN	64.48 (±0.85)	63.86 (±2.66)	64.67 (±1.78)	64.08 (±0.88)	63.61 (±0.85)	64.14 (±1.49)	75.39
	RGAT-BERT	76.14 (±1.05)	76.24 (±1.43)	75.27 (±0.63)	76.39 (±1.19)	75.20 (±1.02)	75.85 (±1.13)	80.94
	CapsNetBERT	76.21 (±1.01)	77.52 (±1.80)	77.49 (±1.13)	77.55 (±1.22)	77.84 (±1.70)	77.32 (±1.41)	-
	LCF-ATEPC	76.22 (±2.37)	76.93 (±1.24)	75.61 (±1.35)	77.58 (±1.16)	75.44 (±1.16)	76.36 (±1.62)	80.97
F1 Macro	MGATN	56.98 (±0.92)	56.36 (±3.09)	55.82 (±2.29)	56.81 (±2.87)	56.93 (±2.05)	56.58 (±2.21)	72.47
	RGAT-BERT	70.54 (±1.54)	70.86 (±2.51)	69.49 (±1.13)	71.94 (±1.62)	70.59 (±1.23)	70.68 (±1.73)	78.2
	CapsNetBERT	70.76 (±1.87)	72.92 (±2.45)	72.68 (±1.72)	72.56 (±2.43)	73.39 (±3.21)	72.46 (±2.37)	-
	LCF-ATEPC	70.23 (±3.60)	72.43 (±0.89)	70.20 (±1.58)	73.34 (±1.72)	70.63 (±2.07)	71.37 (±2.37)	77.86
F1 Weighted	MGATN	63.71 (±0.66)	63.20 (±2.63)	62.52 (±1.87)	63.22 (±2.30)	63.50 (±1.48)	63.23 (±1.79)	-
	RGAT-BERT	75.16 (±1.26)	75.37 (±1.87)	74.38 (±1.00)	76.14 (±1.32)	74.99 (±0.97)	75.21 (±1.34)	-
	CapsNetBERT	75.29 (±1.47)	77.20 (±2.09)	76.97 (±1.38)	76.73 (±2.00)	77.43 (±2.59)	76.72 (±1.95)	-
	LCF-ATEPC	77.33 (±1.93)	77.08 (±1.72)	76.43 (±1.37)	77.74 (±0.99)	75.59 (±1.23)	76.84 (±1.56)	-
Metric	**Model**	**MAMS**						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
Accuracy = F1 Micro	MGATN	-	-	-	-	-	61.95 (±3.17)	-
	RGAT-BERT	-	-	-	-	-	79.79 (±0.55)	84.52
	CapsNetBERT	-	-	-	-	-	83.04 (±0.70)	83.39
	LCF-ATEPC	-	-	-	-	-	78.94 (±0.56)	-
F1 Macro	MGATN	-	-	-	-	-	59.25 (±3.78)	-
	RGAT-BERT	-	-	-	-	-	79.24 (±0.69)	83.74
	CapsNetBERT	-	-	-	-	-	82.44 (±0.81)	-
	LCF-ATEPC	-	-	-	-	-	78.43 (±0.64)	-
F1 Weighted	MGATN	-	-	-	-	-	61.24 (±3.53)	-
	RGAT-BERT	-	-	-	-	-	79.77 (±0.59)	-
	CapsNetBERT	-	-	-	-	-	83.04 (±0.74)	-
	LCF-ATEPC	-	-	-	-	-	78.94 (±0.50)	-
Metric	**Model**	**ARTS Restaurant**						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
Accuracy = F1 Micro	MGATN	57.19 (±1.42)	57.61 (±2.47)	58.04 (±1.91)	57.74 (±1.01)	58.45 (±0.57)	57.81 (±1.54)	-
	RGAT-BERT	72.32 (±0.83)	73.20 (±1.52)	72.57 (±2.37)	71.38 (±1.54)	72.44 (±1.09)	72.38 (±1.54)	-
	CapsNetBERT	78.80 (±1.17)	78.38 (±0.75)	78.91 (±1.98)	78.80 (±0.77)	75.23 (±5.86)	78.02 (±2.98)	-
	LCF-ATEPC	73.59 (±0.55)	73.92 (±1.43)	74.88 (±1.58)	71.11 (±3.27)	73.13 (±0.90)	73.32 (±2.09)	-
F1 Macro	MGATN	47.03 (±0.76)	43.15 (±6.16)	43.17 (±7.18)	45.96 (±1.69)	43.13 (±2.40)	44.49 (±4.40)	-
	RGAT-BERT	63.53 (±2.11)	66.20 (±2.04)	64.77 (±3.19)	62.99 (±3.07)	63.70 (±1.27)	64.24 (±2.51)	-
	CapsNetBERT	71.22 (±1.36)	71.94 (±0.65)	71.63 (±2.65)	71.02 (±1.32)	65.87 (±7.49)	70.34 (±4.06)	-
	LCF-ATEPC	64.94 (±1.38)	66.82 (±1.76)	66.55 (±2.61)	62.91 (±2.71)	63.84 (±0.99)	65.01 (±2.39)	-
F1 Weighted	MGATN	54.89 (±0.81)	52.59 (±3.92)	52.79 (±5.22)	55.02 (±0.25)	52.96 (±1.44)	53.65 (± 2.96)	-
	RGAT-BERT	70.96 (±1.15)	72.65 (±1.66)	72.03 (±2.49)	70.61 (±2.07)	71.41 (±1.16)	71.53 (±1.79)	-
	CapsNetBERT	78.12 (±1.19)	78.29 (±0.48)	78.55 (±1.85)	78.19 (±0.84)	74.20 (±6.39)	77.47 (±3.25)	-
	LCF-ATEPC	74.74 (±0.37)	74.41 (±1.36)	75.83 (±1.34)	72.04 (±3.37)	74.70 (±0.91)	74.34 (±2.07)	-
ARS Accuracy	MGATN	9.13 (±1.42)	9.50 (±2.51)	10.00 (±3.03)	9.90 (±1.00)	9.57 (±0.67)	9.62 (±1.81)	-
	RGAT-BERT	35.17 (±3.16)	36.47 (±3.02)	35.47 (±4.52)	33.33 (±3.31)	35.73 (±3.14)	35.23 (±3.34)	-
	CapsNetBERT	29.96 (±3.11)	27.70 (±2.60)	28.75 (±5.70)	29.74 (±1.84)	21.43 (±8.50)	27.52 (±5.57)	55.36
	LCF-ATEPC	39.16 (±1.66)	40.30 (±3.24)	40.10 (±3.89)	34.02 (±6.20)	39.16 (±3.12)	38.55 (±4.28)	-
Metric	**Model**	**ARTS Laptop**						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
Accuracy = F1 Micro	MGATN	52.31 (±0.20)	52.14 (±1.56)	52.29 (±1.20)	52.19 (±0.83)	52.83 (±0.77)	52.35 (±0.96)	-
	RGAT-BERT	65.81 (±3.23)	64.66 (±5.33)	66.31 (±1.68)	68.25 (±1.35)	66.31 (±2.56)	66.27 (±3.12)	-
	CapsNetBERT	66.68 (±6.17)	72.51 (±0.73)	70.80 (±2.32)	71.97 (±1.48)	71.84 (±1.85)	79.77 (±3.60)	-
	LCF-ATEPC	69.38 (±1.78)	67.57 (±2.58)	68.99 (±0.74)	69.45 (±2.12)	67.50 (±1.56)	68.58 (±1.91)	-
F1 Macro	MGATN	46.58 (±0.76)	46.86 (±2.05)	44.91 (±1.69)	46.81 (±2.63)	48.41 (±1.57)	46.71 (±2.03)	-
	RGAT-BERT	60.30 (±4.14)	59.96 (±5.90)	61.46 (±1.73)	64.37 (±1.69)	62.75 (±2.62)	61.77 (±3.68)	-
	CapsNetBERT	61.61 (±6.59)	68.53 (±1.71)	66.57 (±3.09)	67.36 (±2.66)	68.29 (±3.51)	66.47 (±4.38)	-
	LCF-ATEPC	63.90 (±2.70)	63.79 (±3.44)	64.19 (±1.64)	66.02 (±2.87)	63.81 (±1.99)	64.34 (±2.53)	-
F1 Weighted	MGATN	50.54 (±0.45)	50.67 (±1.20)	49.60 (±1.30)	50.83 (±1.70)	52.10 (±1.00)	50.75 (±1.37)	-
	RGAT-BERT	64.30 (±3.69)	63.47 (±5.71)	65.23 (±1.58)	67.60 (±1.52)	65.73 (±2.70)	65.27 (±3.43)	-
	CapsNetBERT	65.34 (±6.43)	71.89 (±1.18)	70.02 (±2.69)	70.96 (±2.11)	71.31 (±2.61)	69.91 (±4.00)	-
	LCF-ATEPC	70.71 (±1.68)	68.02 (±2.25)	69.94 (±0.60)	69.79 (±1.80)	67.96 (±1.59)	69.28 (±1.89)	-
ARS Accuracy	MGATN	11.68 (±0.83)	12.12 (±1.43)	11.14 (±1.78)	12.41 (±1.34)	13.87 (±0.93)	12.24 (±1.52)	-
	RGAT-BERT	34.31 (±6.26)	31.68 (±10.32)	34.84 (±3.83)	39.17 (±2.18)	34.01 (±6.34)	34.80 (±6.36)	-
	CapsNetBERT	35.52 (±10.83)	46.13 (±1.61)	41.75 (±3.66)	44.33 (±3.01)	42.34 (±2.90)	42.01 (±6.21)	25.86
	LCF-ATEPC	41.98 (±2.42)	37.77 (±4.95)	40.69 (±0.75)	40.94 (±4.09)	37.08 (±3.60)	39.69 (±3.73)	-

Table 4: Our performance results (mean ± standard deviation) for ATSC models. For SemEval-14 Restaurants and Laptops as well as for MAMS, no ARS Accuracy is measured.

Metric	Model	SemEval-14 Restaurant						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
F1 Micro	BERT+TFM	74.27 (±1.25)	74.90 (±0.84)	75.90 (±0.53)	74.55 (±0.54)	74.96 (±0.46)	74.91 (±0.91)	73.98
	GRACE	77.78 (±0.65)	77.40 (±0.54)	78.43 (±0.75)	77.90 (±0.95)	77.84 (±0.80)	77.87 (±0.76)	77.26
F1 Macro	BERT+TFM	66.71 (±1.52)	67.16 (±1.39)	69.37 (±0.73)	66.49 (±0.84)	67.63 (±1.20)	67.47 (±1.50)	-
	GRACE	72.05 (±0.88)	71.40 (±0.99)	72.41 (±1.22)	72.13 (±1.35)	71.36 (±1.49)	71.87 (±1.18)	-
Precision	BERT+TFM	74.25 (±1.46)	74.72 (±1.00)	76.04 (±0.86)	74.29 (±0.35)	75.46 (±0.85)	74.95 (±1.14)	-
	GRACE	76.25 (±0.79)	76.08 (±0.90)	77.17 (±0.82)	76.86 (±0.87)	76.35 (±0.83)	76.54 (±0.87)	-
Recall	BERT+TFM	74.30 (±1.30)	75.10 (±1.01)	75.78 (±0.57)	74.82 (±0.90)	74.48 (±1.07)	74.90 (±1.06)	-
	GRACE	79.37 (±0.75)	78.78 (±0.22)	79.75 (±0.87)	78.99 (±1.12)	79.41 (±0.83)	79.26 (±0.82)	-
ATE F1 Micro	GRACE	87.88 (±0.60)	88.29 (±0.30)	88.38 (±0.42)	88.64 (±0.41)	88.66 (±0.53)	88.37 (±0.51)	-
Metric	**Model**	**SemEval-14 Laptop**						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
F1 Micro	BERT+TFM	63.53 (±0.93)	63.92 (±0.81)	64.03 (±1.56)	64.16 (±0.99)	64.09 (±1.05)	63.95 (±1.03)	60.80
	GRACE	70.04 (±1.33)	68.84 (±0.27)	69.10 (±1.68)	69.10 (±1.17)	69.49 (±1.28)	69.31 (±1.21)	70.71
F1 Macro	BERT+TFM	56.92 (±2.33)	57.04 (±2.39)	57.92 (±2.66)	58.62 (±1.31)	58.09 (±1.49)	57.72 (±2.03)	-
	GRACE	65.29 (±1.90)	64.00 (±0.39)	64.95 (±2.42)	64.51 (±0.98)	65.06 (±1.57)	64.76 (±1.55)	-
Precision	BERT+TFM	65.57 (±1.16)	65.69 (±0.65)	65.19 (±1.61)	65.48 (±0.77)	65.35 (±1.02)	65.46 (±1.02)	63.23
	GRACE	69.77 (±1.47)	68.19 (±0.35)	68.18 (±1.78)	68.64 (±1.60)	68.63 (±1.31)	68.68 (±1.41)	72.38
Recall	BERT+TFM	61.65 (±1.38)	62.26 (±1.37)	62.94 (±1.79)	62.90 (±1.31)	62.90 (±1.33)	62.53 (±1.42)	58.64
	GRACE	70.32 (±1.27)	69.52 (±0.47)	70.06 (±1.69)	69.58 (±0.82)	70.38 (±1.38)	69.97 (±1.16)	69.12
ATE F1 Micro	GRACE	85.99 (±1.51)	85.18 (±0.60)	85.40 (±0.59)	85.98 (±0.72)	85.68 (±0.65)	85.64 (±0.87)	87.93
Metric	**Model**	**MAMS**						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
F1 Micro	BERT+TFM	-	-	-	-	-	64.94 (±1.47)	-
	GRACE	-	-	-	-	-	63.48 (±0.60)	-
F1 Macro	BERT+TFM	-	-	-	-	-	65.54 (±1.43)	-
	GRACE	-	-	-	-	-	64.59 (±0.61)	-
Precision	BERT+TFM	-	-	-	-	-	65.01 (±1.90)	-
	GRACE	-	-	-	-	-	62.63 (±0.98)	-
Recall	BERT+TFM	-	-	-	-	-	64.93 (±2.42)	-
	GRACE	-	-	-	-	-	64.37 (±0.86)	-
ATE F1 Micro	GRACE	-	-	-	-	-	75.96 (±0.42)	-
Metric	**Model**	**ARTS Restaurant**						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
F1 Micro	BERT+TFM	39.80 (±0.78)	39.34 (±0.44)	39.76 (±0.41)	39.29 (±0.56)	39.28 (±1.01)	39.50 (±0.66)	-
	GRACE	61.86 (±1.53)	63.22 (±1.04)	62.80 (±1.28)	62.44 (±1.71)	63.82 (±2.38)	62.83 (±1.66)	-
F1 Macro	BERT+TFM	36.83 (±0.90)	36.13 (±0.47)	36.80 (±0.50)	36.04 (±0.76)	36.19 (±1.27)	36.40 (±0.84)	-
	GRACE	55.91 (±2.11)	57.22 (±1.11)	56.89 (±1.80)	56.40 (±2.03)	57.18 (±3.46)	56.72 (±2.10)	-
Precision	BERT+TFM	28.21 (±0.62)	27.83 (±0.39)	28.22 (±0.28)	27.77 (±0.46)	27.97 (±0.56)	28.00 (±0.48)	-
	GRACE	60.76 (±1.67)	62.20 (±1.41)	61.63 (±1.62)	61.68 (±1.46)	62.56 (±2.38)	61.76 (±1.71)	-
Recall	BERT+TFM	67.55 (±1.17)	67.17 (±0.99)	67.33 (±0.85)	67.17 (±0.86)	66.01 (±2.72)	67.05 (±1.47)	-
	GRACE	63.02 (±1.65)	64.30 (±0.93)	64.02 (±1.00)	63.24 (±2.02)	65.14 (±2.38)	63.94 (±1.73)	-
ARS Accuracy	BERT+TFM	37.53 (±1.97)	35.60 (±2.25)	35.07 (±2.59)	35.83 (±2.43)	34.30 (±2.81)	35.67 (±2.94)	-
	GRACE	34.71 (±2.98)	38.39 (±3.00)	37.70 (±2.49)	36.78 (±3.81)	40.69 (±4.11)	37.66 (±3.64)	-
ATE F1 Micro	GRACE	50.53 (±0.32)	50.81 (±0.25)	50.78 (±0.26)	50.87 (±0.14)	51.02 (±0.33)	50.83 (±0.29)	-
Metric	**Model**	**ARTS Laptop**						
		Split 1	Split 2	Split 3	Split 4	Split 5	Overall	Reported
F1 Micro	BERT+TFM	34.56 (±1.88)	34.55 (±1.61)	35.06 (±1.64)	35.80 (±.075)	35.50 (±0.39)	35.09 (±1.36)	-
	GRACE	65.90 (±1.75)	64.63 (±3.57)	63.16 (±1.97)	64.36 (±2.47)	64.67 (±1.10)	64.54 (±2.30)	-
F1 Macro	BERT+TFM	31.70 (±2.60)	31.34 (±2.02)	32.44 (±2.22)	33.37 (±0.55)	33.12 (±0.64)	32.39 (±1.84)	-
	GRACE	63.98 (±1.92)	61.54 (±3.97)	60.24 (±2.27)	61.56 (±3.10)	61.90 (±1.85)	61.85 (±2.79)	-
Precision	BERT+TFM	25.91 (±1.29)	25.85 (±0.99)	26.06 (±1.00)	26.56 (±0.53)	26.41 (±0.15)	26.16 (±0.86)	-
	GRACE	66.81 (±2.20)	65.43 (±3.99)	63.83 (±2.04)	65.23 (±3.14)	65.41 (±2.23)	65.34 (±2.75)	-
Recall	BERT+TFM	51.91 (±3.33)	52.14 (±3.33)	53.62 (±3.45)	54.90 (±1.32)	54.15 (±1.42)	53.34 (±2.78)	-
	GRACE	65.03 (±1.48)	63.89 (±3.37)	62.51 (±1.96)	63.54 (±2.08)	64.00 (±1.34)	63.79 (±2.14)	-
ARS Accuracy	BERT+TFM	23.60 (±4.29)	23.26 (±4.83)	24.87 (±4.12)	26.91 (±2.10)	26.23 (±2.47)	24.97 (±3.70)	-
	GRACE	38.80 (±3.90)	36.40 (±3.85)	33.20 (±1.79)	32.80 (±3.03)	36.40 (±4.56)	35.52 (±3.97)	-
ATE F1 Micro	GRACE	52.97 (±0.53)	52.64 (±0.59)	52.62 (±0.36)	53.08 (±0.49)	52.82 (±0.37)	52.83 (±0.47)	-

Table 5: Our performance results (mean ± standard deviation) for ATE+ATSC models. For SemEval-14 Restaurants and Laptops as well as for MAMS, no ARS Accuracy is measured.