Reproducibility, Replicability and Beyond: Assessing Production Readiness of Aspect Based Sentiment Analysis in the Wild

Rajdeep Mukherjee[1]([✉]), Shreyas Shetty[2], Subrata Chattopadhyay[1],
Subhadeep Maji[3], Samik Datta[3], and Pawan Goyal[1]

[1] Indian Institute of Technology Kharagpur, Kharagpur, India
rajdeep1989@iitkgp.ac.in, pawang@cse.iitkgp.ac.in
[2] Flipkart Internet Private Limited, Bengaluru, India
shreyas.shetty@flipkart.com
[3] Amazon India Private Limited, Bengaluru, India

Abstract. With the exponential growth of online marketplaces and user-generated content therein, aspect-based sentiment analysis has become more important than ever. In this work, we critically review a representative sample of the models published during the past six years through the lens of a practitioner, with an eye towards deployment in production. First, our rigorous empirical evaluation reveals poor reproducibility: an average 4–5% drop in test accuracy across the sample. Second, to further bolster our confidence in empirical evaluation, we report experiments on two challenging data slices, and observe a consistent 12–55% drop in accuracy. Third, we study the possibility of transfer across domains and observe that as little as 10–25% of the domain-specific training dataset, when used in conjunction with datasets from other domains within the same locale, largely closes the gap between complete cross-domain and complete in-domain predictive performance. Lastly, we open-source two large-scale annotated review corpora from a large e-commerce portal in India in order to aid the study of replicability and transfer, with the hope that it will fuel further growth of the field.

Keywords: Aspect based sentiment analysis · Aspect polarity detection · Reproducibility · Replicability · Transferability

1 Introduction

In recent times, online marketplaces of goods and services have witnessed an exponential growth in terms of consumers and producers, and have proliferated in a wide spectrum of market segments, such as e-commerce, food delivery,

healthcare, ride sharing, travel and hospitality, to name a few. The Indian e-commerce market segment alone is projected to grow to 300–350M consumers and \$100–120B revenue by 2025[1]. In the face of ever-expanding choices, purchase decision-making is guided by the reviews and ratings: Watson et al. [29] estimates that the average product rating is the most important factor in making purchase decisions for 60% of consumers. Similarly, the academic research on Aspect Based Sentiment Analysis (ABSA) has come a long way since its humble beginning in the SemEval-2014[2]. Over the past 6 years, the accuracy on a benchmark dataset for *aspect term polarity* has grown by at least 11.4%. We ask, is this progress enough to support the burgeoning online marketplaces?

We argue on the contrary. On one hand, industrial-strength systems need to demonstrate several traits for smooth operation and delightful consumer experience. Breck et al. [1] articulates several essential traits and presents a rubric of evaluation. Notable traits include: (a) "All hyperparameters have been tuned"; (b) "A simpler model is not better"; (c) "Training is reproducible"; and (d) "Model quality is sufficient on important data slices". On the other hand, recent academic research in several fields has faced criticisms from within the community on similar grounds: Dhillon et al. [6] points out the inadequacy of benchmark dataset and protocol for few-shot image classification; Dacrema et al. [4] criticises the recent trend in recommendation systems research on the ground of lack of reproducibility and violations of (a)–(c) above; Li et al. [14] criticises the recent trend in information retrieval research on similar grounds. A careful examination of the recent research we conduct in this work reveals that the field of ABSA is not free from these follies.

To this end, it is instructive to turn our attention to classic software engineering with the hope of borrowing from its proven safe development practises. Notably, Kang et al. [10] advocates the use of *model assertions* – an abstraction to monitor and improve model performance during the development phase. Along similar lines, Ribeiro et al. [20] presents a methodology of large-scale comprehensive testing for NLP, and notes its effectiveness in identifying bugs in several (commercial) NLP libraries, that would not have been discovered had we been relying solely on test set accuracy. In this work, in addition to the current practice of reporting test set accuracies, we report performance on two challenging data slices – e.g., *hard set* [31], and, *contrast set* [7] – to further bolster the comprehensiveness of empirical evaluation.

For widespread adoption, data efficiency is an important consideration in real-world deployment scenarios. As an example, a large e-commerce marketplace in India operates in tens of thousands of categories, and a typical annotation cost is 3¢ per review. In this work, we introduce and open-source two additional large-scale datasets curated from product reviews in lifestyle and appliance categories to aid replicability of research and study of transfer across domains and locales (text with similar social/linguistic characteristics). In particular, we note that

---

[1] *How India Shops Online* – Flipkart and Bain & Company.
[2] SemEval-2014 Task 4.

just a small fraction of the in-domain training dataset, mixed with existing in-locale cross-domain training datasets, guarantees comparable test set accuracies.

In summary, we make the following notable contributions:

– Perform a thorough reproducibility study of models sampled from a public leaderboard[3] that reveals a consistent 4–5% drop in reported test set accuracies, which is often larger than the gap in performance between the winner and the runner-up.
– Consistent with the practices developed in software engineering, we bolster the empirical evaluation rigour by introducing two challenging data slices that demonstrates an average 12–55% drop in test set accuracies.
– We study the models from the perspective of data efficiency and note that as little as 10–25% of the domain-specific training dataset, when used in conjunction with existing cross-domain datasets from within the same locale, largely closes the gap in terms of test set accuracies between complete cross-domain training and using 100% of the domain-specific training instances. This observation has immense implications towards reduction of annotation cost and widespread adoption of models.
– We curate two additional datasets from product reviews in lifestyle and appliances categories sampled from a large e-commerce marketplace in India, and make them publicly accessible to enable the study of replicability.

2   Desiderata and Evaluation Rubric

Reproducibility and replicability have been considered the gold-standard in academic research and has witnessed a recent resurgence in emphasis across scientific disciplines: see for e.g., McArthur et al. [18] in the context of biological sciences and Stevens et al. [23] in the context of psychology. We follow the nomenclature established in [23] and define *reproducibility* as the ability to obtain same experimental results when a different analyst uses an identical experimental setup. On the other hand, *replicability*, is achieved when the same experimental setup is used on a different dataset to similar effect. While necessary, these two traits are far from sufficient for widespread deployment in production.

Breck et al. [1] lists a total of 28 traits spanning the entire development and deployment life cycle. Since our goal is only to assess the production readiness of a class of models. We decide to forego all 14 data-, feature- and monitoring-related traits. We borrow 1 ("Training is reproducible") and 2 ("All hyperparameters have been tuned" and "Model quality is sufficient on important data slices") traits from the infrastructure- and modeling-related rubrics, respectively.

Further, we note that the ability to transfer across domains/locales is a desirable trait, given the variety of market segments and the geographic span of online marketplaces. In other words, this expresses data efficiency and has implications towards lowering the annotation cost and associated deployment hurdles. Given the desiderata, we articulate our production readiness rubric as follows:

---

[3] *Papers With Code*: ABSA on SemEval 2014 Task 4 Sub Task 2.

– *Reproducibility.* A sound experimental protocol that minimises variability across runs and avoids common pitfalls (e.g., hyperparameter-tuning on the test dataset itself) should reproduce the reported test set accuracy within a reasonable tolerance, not exceeding the reported performance gap between the winner and the runner-up in a leaderboard. Section 6 articulates the proposed experimental protocol and Sect. 7 summarises the ensuing observations.
– *Replicability.* The aforementioned experimental protocol, when applied to a different dataset, should not dramatically alter the conclusions drawn from the original experiment; specifically, it should not alter the relative positions within the leaderboard. Section 4 details two new datasets we contribute in order to aid the study of replicability, whereas Sect. 7 contains the ensuing observations.
– *Performance.* Besides overall test-set accuracy, an algorithm should excel at challenging data slices such as hard- [31] and contrast sets [7]. Section 7 summarises our findings when this checklist is adopted as a standard reporting practice.
– *Transferability.* An algorithm must transfer gracefully across domains within the same locale, i.e. textual data with similar social/linguistic characteristics. We measure it by varying the percentage of in-domain training instances from 0% to 100% and locating the inflection point in test set accuracies. See Sect. 7 for additional details.

Note that apart from the "The model is debuggable" and "A simpler model is not better" traits, the remaining traits as defined by Breck et al. [1] are independent of the choice of the algorithm and is solely a property of the underlying system that embodies it, which is beyond the scope of the present study. Unlike [1], we refrain from developing a numerical scoring system.

3   Related Work

First popularised in the SemEval-2014 Task 4 [19], ABSA has enjoyed immense attention from both academic and industrial research communities. Over the past 6 years, according to the cited literature on a public leaderboard[4], the performance for the subtask of *Aspect Term Polarity* has increased from 70.48% in Pontiki et al. [19], corresponding to the winning entry, to 82.29% in Yang et al. [32] on the laptop review corpus. The restaurant review corpus has witnessed a similar boost in performance: from 80.95% in [19] to 90.18% in [32].

Not surprisingly, the field has witnessed a phase change in terms of the methodology: custom feature engineering and ensembles that frequented earlier [19] gave way to neural networks of ever-increasing complexity. Apart from this macro-trend, we notice several micro-trends in the literature: the year 2015 witnessed a proliferation of LSTM and its variants [24]; years 2016 and 2017 respectively witnessed the introduction [25] and proliferation [2, 3, 16, 26] of memory networks and associated attention mechanisms; in 2018 research focused on

---

[4] *Papers With Code*: ABSA on SemEval 2014 Task 4 Sub Task 2.

CNN [31], transfer learning [13] and transformers [12], while memory networks and attention mechanisms remained in spotlight [9,11,15,27]; transformer and BERT-based models prevailed in 2019 [30,33], while attention mechanisms continued to remain mainstream [22].

While these developments appear to have pushed the envelope of performance, the field has been fraught with "winner's curse" [21]. In addition to the replicability and reproducibility crises [18,23], criticisms around inadequacy of baseline and unjustified complexity [4,6,14] applies to this field as well. The practice of reporting performance in challenging data slices [31] has not been adopted uniformly, despite its importance to production readiness assessment [1]. Similarly, the study of transferability and replicability has only been sporadically performed: e.g., Hu et al. [8] uses a dataset curated from Twitter along with the ones introduced in Pontiki et al. [19] for studying cross-domain transferability.

4   Dataset

For the *Reproducibility* rubric, we consider the datasets released as part of the SemEval 2014 Task 4 - Aspect Based Sentiment Analysis[5] for our experiments, specifically the Subtask 2 - Aspect term Polarity. The datasets come from two domains – Laptop and Restaurant. We use their versions made available in this Github[6] repository which forms the basis of our experimental setup.

The guidelines used for annotating the datasets were released as part of the challenge. For the *Replicability* rubric, we tagged two new datasets from the e-commerce domain viz., Men's T-shirt and Television, using similar guidelines.

The statistics for these four datasets are presented in Table 1. As we can observe, the sizes of the Men's T-shirt and Television datasets are comparable to the laptop and restaurant datasets, respectively.

Table 1. Statistics of the datasets showing the no. of sentences with corresponding sentiment polarities of constituent aspect terms.

Dataset	Train				Test			
	Positive	Negative	Neutral	Total	Positive	Negative	Neutral	Total
Laptop	994	870	464	2328	341	128	169	638
Restaurant	2164	807	637	3608	728	196	196	1120
Men's T-shirt	1122	699	50	1871	270	186	16	472
Television	2540	919	287	3746	618	257	67	942

For the *Performance* rubric, we evaluate and compare the models on two challenging subsets viz., *hard* as defined by Xue et al. [31] and *contrast* as defined by Gardner et al. [7]. We describe below the process to obtain these datasets:

---

[5] SemEval 2014: Task 4 http://alt.qcri.org/semeval2014/task4/.
[6] https://github.com/songyouwei/ABSA-PyTorch.

Table 2. Statistics of the Hard test sets

Dataset	Positive	Negative	Neutral	Total (% of Test Set)
Laptop	31	24	46	101 (15.8%)
Restaurants	81	60	83	224 (20.0%)
Men's T-shirt	23	24	1	48 (10.2%)
Television	43	40	19	102 (10.8%)

– **Hard data slice:** Hard examples have been defined in Xue et al. [31] as the subset of review sentences containing multiple aspects with different corresponding sentiment polarities. The number of such hard examples from each of the datasets are listed in Table 2.
– **Contrast data slice:** In order to create additional test examples, Gardner et al. [7] adds perturbations to the test set, by modifying only a couple of words to flip the sentiment corresponding to the aspect under consideration. For e.g., consider the review sentence: "I was happy with their service and food". If we change the word "happy" with "dissatisfied", the sentiment corresponding to the aspect "food" changes from positive to negative. We take a random sample of 30 examples from each of the datasets and add similar perturbations as above to create 30 additional examples. These 60 examples for each of the four datasets thus serve as our contrast test sets.

5   Models Compared

As part of our evaluation, we focus on two families of models which cover the major trends in the ABSA research community: (i) memory network based, and (ii) BERT based. Among the initial set of models for the SemEval 14 challenge, memory network based models had much fewer parameters compared to LSTM based approaches and performed comparatively better. With the introduction of BERT [5], work in NLP has focused on leveraging BERT based architectures for a wide spectrum of tasks. In the ABSA literature, the leaderboard[7] has been dominated by BERT based models, which have orders of magnitude more parameters than memory network based models. However, due to pre-training on large corpora, BERT models are still very data efficient in terms of number of labelled examples required. We chose three representative models from each family for our experiments and briefly describe them below:

– **ATAE-LSTM** [28] represents aspects using target embeddings and models the context words using an LSTM. The context word representations and target embeddings are concatenated and combined using an attention layer.

---

[7] https://paperswithcode.com/sota/aspect-based-sentiment-analysis-on-semeval.

– **Recurrent Attention on Memory (RAM)** [2] represents the input review sentence using a memory network, and the memory cells are weighted using the distance from the target word. The aspect representation is then used to compute attention scores on the input memory, and the attention weighted memory is refined iteratively using a GRU (recurrent) network.

– **Interactive Attention Networks (IAN)** [17] uses separate components for computing representations for both the target (aspect) and the context words. The representations are pooled and then used to compute an attention score on each other. Finally the individual attention weighted representations are concatenated to obtain the final representation for the 3-way classification task, with *positive*, *negative*, and *neutral* being the three classes.

– **BERT-SPC** [5] is a baseline BERT model that uses "[CLS] + context + [SEP] + target + [SEP]" as input for the sentence pair classification task, where '[CLS]' and '[SEP]' represent the tokens corresponding to *classification* and *separator* symbols respectively, as defined in Devlin et al. [5].

– **BERT-AEN** [22] uses an attentional encoder network to model the semantic interaction between the context and the target words. Its loss function uses a label smoothing regularization to avoid overfitting.

– **The Local Context Focus (LCF-BERT)** [33] is based on Multi-head Self-Attention (MHSA). It uses Context features Dynamic Mask (CDM) and Context features Dynamic Weighted (CDW) layers to focus more on the local context words. A BERT-shared layer is adopted to LCF design to capture internal long-term dependencies of local and global context.

6   Experimental Setup

We present an extensive evaluation of the aforementioned models across the four datasets: Laptops, Restaurants, Men's T-shirt and Television, as per the production readiness rubrics defined in Sect. 2. While trying to reproduce the reported results for the models, we faced two major issues; (i) the official implementations were not readily available, and (ii) the exact hyperparameter configurations were not always specified in the corresponding paper(s). In order to address the first, our experimental setup is based on a community designed implementation of recent papers available on GitHub[8]. Our choice for this public repository is guided by its thoroughness and ease of experimentation. As an additional social validation, the repository had 1.1k stars and 351 forks on GitHub at the time of writing. For addressing the second concern, we consider the following options; (a) use commonly accepted default parameters (for e.g., using a learning rate of $1e^{-4}$ for Adam optimizer). (b) use the public implementations to guide the choice of hyperparameters. The exact hyperparameter settings used in our experiments are documented and made available with our supporting code repository[9] for further reproducibility and replicability of results.

---

[8] https://github.com/songyouwei/ABSA-PyTorch.
[9] https://github.com/rajdeep345/ABSA-Reproducibility.

From the corresponding experimental protocols described in the original paper(s), we were not sure if the final numbers reported were based on the training epoch that gave the best performance on the test set, or whether the hyperparameters were tuned on a separate held-out set. Therefore, we use the following two configurations; (i) the test set is itself used as the held out set, and the model used for reporting the results is chosen corresponding to the training epoch with best performance on the test set; and (ii) 15% of the training data is set aside as a held out set for tuning the hyperparameters and the optimal training epoch is decided corresponding to the best performance on the held out set. Finally the model is re-trained, this time with all the training data (including 15% held out set), for the optimal no of epochs before evaluating the test set. For both the cases, we report mean scores over 5 runs of our experiments.

7   Results and Discussion: Production Readiness Rubrics

7.1   Reproducibility and Replicability

Tables 3(a) and 3(b) show our *reproducibility* study for the Laptop and Restaurant datasets, respectively. For both the datasets, we notice a consistent 1–2% drop in accuracy and macro-f1 scores when we try to reproduce the reported numbers in the corresponding papers. Only exceptions were LCF-BERT for Laptop and BERT-SPC for Restaurant dataset, where we got higher numbers than the reported ones. For ATAE-LSTM, the drop observed was much larger than other models. We notice an additional 1–2% drop in accuracy when we use 15% of the training set as a held-out set to pick the best model. These numbers indicate that the actual performance of the models is likely to be slightly worse than what is quoted in the papers, and the drop sometimes is larger than the difference between the performance of two consecutive methods on the leaderboard.

To study the *replicability*, Tables 3(c) and 3(d) summarise the performance of the individual models on the Men's T-shirt and Television datasets, respectively. We introduce these datasets for the first time and report the performance of all 6 models under the two defined configurations: test set as held out set, and 15% of train set used as held out set. We notice a similar drop in performance when we follow the correct experimental procedure (hyperparameter tuning on 15% train data as held-out set). Therefore, following a consistent and rigorous experimental protocol helps us to get a better sense of the true model performance.

7.2   Performance on the Hard and Contrast Data Slices

As per the *performance* rubric, we investigate the performance of all 6 models on both *hard* and *contrast* test sets, using the correct experimental setting (15% train data as held out set). The results are shown in brackets (in same order) in the last two columns of Tables 3(a), 3(b), 3(c), and 3(d) for the four datasets, respectively. We observe a large drop in performance on both these challenging data slices across models. LCF-BERT consistently performs very well on these test sets. Among memory network based models, RAM performs the best.

Table 3. Performance of the models on the four datasets. The first two dataset correspond to the reproducibility study, while the next two datasets correspond to the replicability study. Towards performance study, results on the hard and contrast data slices are respectively enclosed in brackets in the last two columns. All the reproduced and replicated results are averaged across 5 runs.

Model	Reported		Reproduced		Reproduced using 15% held out set	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
ATAE-LSTM	68.70	-	60.28	44.33	58.62 (33.47, 26.00)	43.27 (29.01, 22.00)
RAM	74.49	71.35	72.82	68.34	70.97 (56.04, 46.00)	65.31 (55.81, 43.16)
IAN	72.10	-	69.94	62.84	69.40 (48.91, 34.67)	61.98 (48.75, 33.40)
BERT-SPC	78.99	75.03	78.72	74.52	77.24 (59.21, 52.00)	72.80 (59.44, 48.67)
BERT-AEN	79.93	76.31	78.65	74.26	75.71 (46.53, 37.33)	70.02 (45.22, 36.20)
LCF-BERT	77.31	75.58	79.75	76.10	77.27 (62.57, 54.67)	72.86 (62.71, 49.56)

(a) Laptop

Model	Reported		Reproduced		Reproduced using 15% held out set	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
ATAE-LSTM	77.20	-	73.71	55.87	73.29 (52.41, 38.71)	54.59 (47.35, 33.13)
RAM	80.23	70.80	78.21	65.94	76.36 (59.29, 56.77)	63.15 (56.36, 56.12)
IAN	78.60	-	76.80	64.24	76.52 (57.05, 50.32)	63.84 (55.11, 48.19)
BERT-SPC	84.46	76.98	85.04	78.02	84.23 (68.84, 57.42)	76.28 (68.11, 57.23)
BERT-AEN	83.12	73.76	81.73	71.24	80.07 (51.70, 45.81)	69.80 (48.97, 46.88)
LCF-BERT	87.14	81.74	85.94	78.97	84.20 (69.38, 56.77)	76.28 (69.64, 57.81)

(b) Restaurant

Model	Replicated		Replicated using 15% held out set	
	Accuracy	Macro-F1	Accuracy	Macro-F1
ATAE-LSTM	83.13	55.98	81.65 (58.33, 40.67)	54.84 (39.25, 30.54)
RAM	90.51	61.93	88.26 (83.33, 46.00)	59.67 (56.01, 33.85)
IAN	87.58	59.16	87.41 (63.75, 42.67)	58.97 (42.85, 31.94)
BERT-SPC	93.13	73.86	92.42 (89.58, 66.00)	73.83 (60.62, 56.90)
BERT-AEN	88.69	72.25	87.54 (50.42, 58.67)	59.14 (32.96, 43.00)
LCF-BERT	93.35	72.19	91.99 (91.67, 71.33)	72.13 (62.30, 59.70)

(c) Men's T-shirt

Model	Replicated		Replicated using 15% held out set	
	Accuracy	Macro-F1	Accuracy	Macro-F1
ATAE-LSTM	81.10	53.71	79.68 (53.92, 25.33)	52.78 (39.13, 16.80)
RAM	84.29	58.68	83.02 (64.31, 53.33)	58.50 (50.07, 45.51)
IAN	82.42	57.15	80.49 (54.31, 32.00)	56.78 (41.67, 25.16)
BERT-SPC	89.96	74.68	88.56 (80.20, 62.67)	74.81 (74.32, 60.25)
BERT-AEN	87.09	67.92	85.94 (50.39, 50.66)	65.65 (38.08, 45.75)
LCF-BERT	90.36	76.01	90.00 (80.98, 66.67)	75.86 (73.72, 64.15)

(d) Television

7.3   Transferability Rubric: Cross Domain Experiments

In a production readiness setting, it is very likely that we will not have enough labelled data across individual categories and hence it is important to under-

stand how well the models are able to transfer across domains. To understand the transferability of models across datasets, we first experiment with cross domain combinations. For each experiment, we fix the test set (for e.g., Laptop) and train three separate models, each with one of the other three datasets as training sets (Restaurant, Men's T-shirt, and Television in this case). Consistent with our experimental settings, for each such combination, we use 15% of the cross-domain data as held-out set for hyperparameter tuning, re-train the corresponding models with all the cross-domain data and obtain the scores for the in-domain set (here Laptop) averaged across 5 different runs of the experiment.

Table 4. Transferability: Average drop between in-domain and cross-domain accuracies for each dataset pair for (a) BERT based and (b) Memory network based models. Rows correspond to the train set. Columns correspond to the test set.

	Laptop	Restaurant	Men's T-shirt	Television		Laptop	Restaurant	Men's T-shirt	Television
Laptop	0	4.50	3.84	3.53	Laptop	0	7.18	15.75	9.89
Restaurant	2.17	0	3.49	3.68	Restaurant	5.29	0	10.7	10.3
Men's T-shirt	9.59	7.57	0	2.00	Men's T-shirt	8.34	10.02	0	3.48
Television	3.85	5.65	2.08	0	Television	4.5	7.5	6.77	0

| (a) BERT based models | (b) Memory network based models |

Table 4 summarises the results averaged across the BERT-based models and Memory network based models, respectively on the four datasets. The rows and columns correspond to the train and test sets, respectively. The diagonals correspond to the in-domain experiments (denoted by 0) and each off-diagonal entry denotes the average drop in model performance for the cross-domain setting compared to the in-domain combination.

From Table 4 we observe that on an average the models are able to generalize well across the following combinations, which correspond to a lower drop in the cross domain experiments: (i) Laptops and Restaurants, and (ii) Men's T-shirt and Television. For instance, when testing on the Restaurant dataset, BERT based and memory network based models respectively show an average of ∼4 and ∼7 point absolute drops in % accuracies, when trained using the Laptop dataset. The drops are higher for the other two training sets. Interestingly, the generalization is more pronounced across locales rather than domains, contrary to what one would have expected. For e.g., we notice better transfer from Men's T-shirt → Television (similarity in locale) than in the expected Laptop → Television (similarity in domain). Given that our task is that of detecting sentiment polarities of aspect terms, this observation might be attributed to the similarity in social/linguistic characteristics of reviews from the same locale.

Further, in the spirit of *transferability*, we consider the closely related locales as identified above – {Laptop, Restaurant} and {Men's T-shirt, Television}, and conduct experiments to understand the incremental benefits of adding in-domain data on top of cross domain data, i.e., what fraction of the in-domain training instances can help to cover the gap between purely in-domain and purely

Table 5. Transferability: Results on including incremental in-domain training data. The rows correspond to cross-domain performance (0), adding 10%, 25% and 50% in-domain dataset to the cross-domain. To improve illustration, we repeat in-domain results. Inflection points for each dataset are boldfaced.

% in-domain	Laptop	Restaurant	Men's T-shirt	Television
0	74.6 (73.6, 74.6, 75.5)	78.3 (77.3, 77.8, 79.9)	88.6 (86.3, 89.6, 89.8)	86.1 (83.5, 87.5, 87.4)
10	**76.5** (73.9, 76.6, 78.9)	**81.6** (80.1, 81.5, 83.3)	88.9 (85.7, 90.6, 90.4)	83.8 (82.0, 86.1, 83.2)
25	76.3 (74.8, 77.0, 77.0)	82.1 (79.8, 82.8, 83.7)	**90.0** (87.2, 91.7, 91.0)	86.3 (83.8, 86.8, 88.2)
50	78.2 (76.4, 79.2, 78.9)	82.9 (80.8, 83.6, 84.4)	90.1 (86.8, 91.3, 92.3)	**87.2** (85.5, 88.2, 87.8)
In-domain	76.7 (75.7, 77.2, 77.3)	82.8 (80.1, 84.2, 84.2)	90.6 (87.5, 92.4, 92.0)	88.2 (85.9, 88.6, 90.0)

(a) Variance across BERT based models (BERT-AEN, BERT-SPC, LCF-BERT) is small.

% in-domain	Laptop	Restaurant	Men's T-shirt	Television
0	61.0 (58.6, 60.9, 63.6)	68.2 (68.3, 68.0, 68.3)	79.0 (76.6, 78.6, 81.9)	77.6 (75.4, 77.8, 79.6)
10	**65.1** (60.7, 65.6, 69.1)	**73.0** (70.1, 74.1, 74.9)	**83.8** (80.3, 84.1, 86.9)	**79.1** (77.1, 79.1, 81.2)
25	65.3 (59.9, 66.2, 69.8)	74.8 (72.2, 75.4, 76.6)	85.1 (82.9, 86.0, 86.4)	80.0 (78.7, 79.8, 81.5)
50	66.2 (60.5, 68.7, 69.5)	75.0 (72.9, 75.3, 76.8)	85.8 (82.7, 86.1, 88.4)	80.6 (78.8, 80.7, 82.4)
In-domain	66.3 (58.6, 69.4, 71.0)	75.4 (73.3, 76.5, 76.4)	85.8 (81.7, 87.4, 88.3)	81.1 (79.7, 80.5, 83.0)

(b) Variance across Memory network models (ATAE-LSTM, IAN, RAM) is significant.

cross-domain performance largely. For each test dataset, we take examples from the corresponding cross-domain dataset in the same locale as training set and incrementally add in-domain (10%, 25% and 50%) examples to evaluate the performance of the models. Table 5 summarises the results from these experiments for the BERT based models (a) and memory network based models (b). For instance, on the Restaurant dataset, the average cross-domain performance (i.e., trained on Laptop) across the three BERT-based models is 78.3 (first row), while the purely in-domain performance is 82.8 (last row). We observe that among all increments, adding 10% of the in-domain dataset (second row) gives the maximum improvement, and is accordingly defined as the inflection point, which is marked in bold. In Table 5(a), we report the accuracy scores (averaged over 5 runs) for the individual BERT based models (BERT-AEN, BERT-SPC, LCF-BERT) in brackets, in addition to the average numbers. As we can see, the variability in the numbers across models is low. For the memory network based models, on the other hand, the variability is not so low, and the corresponding scores have been shown in Table 5(b) in the order (ATAE-LSTM, IAN, RAM).

Interestingly, we notice that in most of the cases, the inflection point is obtained upon adding just 10% in-domain examples and the model performance reaches within 0.5–2% of purely in-domain performance, as shown in Table 6. While in a few cases, it happens by adding 25–50% in-domain samples. This is especially useful from the production readiness perspective since considerably good performance can be achieved by using limited in-domain labelled data on top of cross-domain annotated data from the same locale.

Table 6. Performance scorecard in accordance with the rubric: *reproducibility* – % drop in test set accuracy across Laptop and Restaurant, resp.; *replicability* – rank in leaderboard for Men's T-shirt and Television, resp. (rank obtained from avg. test set accuracy on Laptop and Restaurant); *performance* – % drop in test set accuracy (averaged across all four datasets) with hard and contrast-set data slices, resp.; *transferability* – % drop in test set accuracy in cross-domain setting, and upon adding in-domain training instances as per the inflection point, resp. (averaged over the four datasets)

Model	Reproducibility	Replicability	Performance	Transferability
ATAE-LSTM	(14.67, 5.06)	6, 6 (6)	(33.07, 55.31)	(4.60, 1.44)
RAM	(4.73, 4.82)	3, 4 (4)	(17.88, 36.12)	(8.06, 2.06)
IAN	(3.74, 2.64)	5, 5 (5)	(28.64, 48.93)	(9.22, 3.55)
BERT-SPC	(2.22, 0.27)	1, 2 (2)	(13.53, 30.58)	(3.83, 1.33)
BERT-AEN	(5.28, 3.67)	4, 3 (3)	(39.44, 41.88)	(2.61, 0.83)
LCF-BERT	(0.05, 3.37)	2, 1 (1)	(11.75, 27.55)	(3.14, 0.64)

7.4 Summary Comparison of the Different Models Under the Production Readiness Rubrics

We now make an overall comparison across different models considered in this study under our production readiness rubrics. Table 6 shows the various numbers across these rubrics. Under *reproducibility*, we observe a consistent drop in performance even for the BERT-based models, atleast for one of the two datasets, viz. Laptop and Restaurant. For Memory network based models, while there is a considerable drop across both the datasets, the drop for the Laptop dataset is quite noteworthy. Under *replicability*, we observe that the relative rankings of the considered models remain quite stable for the two new datasets, which is a good sign. Under *performance*, we note a large drop in test set accuracies for all the models across the two challenging data slices, with a minimum drop of 11–27% for LCF-BERT. Surprisingly, BERT-AEN suffered a huge drop in performance for both hard as well as contrast data slices. This is a serious concern and further investigation is needed to identify the issues responsible for this significant drop. Under *transferability*, while there is consistent drop in cross-domain scenario, the drop with the inflection point, corresponding to a meager addition of 10–25% of in-domain data samples, is much smaller.

7.5 Limitations of the Present Study

While representative of the modern trend in architecture research, memory network- and BERT-based models do not cover the entire spectrum of the ABSA literature. Important practical considerations, such as debuggability, simplicity and computational efficiency, have not been incorporated into the rubric. Lastly, a numeric scoring system based on the rubric would have made its interpretation objective. We leave them for a future work.

8   Conclusion

Despite the limitations, the present study takes an important stride towards closing the gap between empirical academic research and its widespread adoption and deployment in production. In addition to further strengthening the rubric and judging a broader cross-section of published ABSA models in its light, we envision to replicate such study in other important NLP tasks. We hope the two contributed datasets, along with the open-source evaluation framework, shall fuel further rigorous empirical research in ABSA. We make all the codes and datasets publicly available[10].

References

1. Breck, E., Cai, S., Nielsen, E., Salib, M., Sculley, D.: The ML test score: a rubric for ml production readiness and technical debt reduction. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 1123–1132 (2017). https://doi.org/10.1109/BigData.2017.8258038

2. Chen, P., Sun, Z., Bing, L., Yang, W.: Recurrent attention network on memory for aspect sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, pp. 452–461. Association for Computational Linguistics, September 2017. https://doi.org/10.18653/v1/D17-1047

3. Cheng, J., Zhao, S., Zhang, J., King, I., Zhang, X., Wang, H.: Aspect-level sentiment classification with heat (hierarchical attention) network. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, pp. 97–106. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3132847.3133037

4. Dacrema, M.F., Cremonesi, P., Jannach, D.: Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, pp. 101–109. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3298689.3347058

5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/N19-1423

6. Dhillon, G.S., Chaudhari, P., Ravichandran, A., Soatto, S.: A baseline for few-shot image classification. In: International Conference on Learning Representations (2020)

7. Gardner, M., et al.: Evaluating models' local decision boundaries via contrast sets. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1307–1323. Association for Computational Linguistics, November 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.117

---

[10] https://github.com/rajdeep345/ABSA-Reproducibility.

8. Hu, M., Wu, Y., Zhao, S., Guo, H., Cheng, R., Su, Z.: Domain-invariant feature distillation for cross-domain sentiment classification. 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), Hong Kong, China, pp. 5559–5568. Association for Computational Linguistics, November 2019. https://doi.org/10.18653/v1/D19-1558

9. Huang, B., Ou, Y., Carley, K.M.: Aspect level sentiment classification with attention-over-attention neural networks. In: Thomson, R., Dancy, C., Hyder, A., Bisgin, H. (eds.) SBP-BRiMS 2018. LNCS, vol. 10899, pp. 197–206. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93372-6_22

10. Kang, D., Raghavan, D., Bailis, P., Zaharia, M.: Model assertions for monitoring and improving ML models. In: Proceedings of the 3rd MLSys Conference, Austin, TX, USA (2020)

11. Li, L., Liu, Y., Zhou, A.: Hierarchical attention based position-aware network for aspect-level sentiment analysis. In: Proceedings of the 22nd Conference on Computational Natural Language Learning, Brussels, Belgium, pp. 181–189. Association for Computational Linguistics, October 2018

12. Li, X., Bing, L., Lam, W., Shi, B.: Transformation networks for target-oriented sentiment classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, pp. 946–956. Association for Computational Linguistics, July 2018

13. Li, Z., Wei, Y., Zhang, Y., Zhang, X., Li, X., Yang, Q.: Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. CoRR abs/1811.10999 (2018)

14. Lin, J.: The neural hype and comparisons against weak baselines. SIGIR Forum **52**(2), 40–51 (2019). https://doi.org/10.1145/3308774.3308781

15. Liu, Q., Zhang, H., Zeng, Y., Huang, Z., Wu, Z.: Content attention model for aspect based sentiment analysis. In: Proceedings of the 2018 World Wide Web Conference, WWW 2018. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 1023–1032 (2018). https://doi.org/10.1145/3178876.3186001

16. Ma, D., Li, S., Zhang, X., Wang, H.: Interactive attention networks for aspect-level sentiment classification. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 4068–4074 (2017). https://doi.org/10.24963/ijcai.2017/568

17. Ma, D., Li, S., Zhang, X., Wang, H.: Interactive attention networks for aspect-level sentiment classification. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 4068–4074. AAAI Press (2017)

18. McArthur, S.L.: Repeatability, reproducibility, and replicability: tackling the 3R challenge in biointerface science and engineering. Biointerphases **14**(2), 020201 (2019). https://doi.org/10.1116/1.5093621

19. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, pp. 27–35. Association for Computational Linguistics, August 2014. https://doi.org/10.3115/v1/S14-2004

20. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: behavioral testing of NLP models with CheckList. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4902–4912. Association for Computational Linguistics, July 2020. https://doi.org/10.18653/v1/2020.acl-main.442

21. Sculley, D., Snoek, J., Wiltschko, A.B., Rahimi, A.: Winner's curse? On pace, progress, and empirical rigor. In: ICLR (2018)

22. Song, Y., Wang, J., Jiang, T., Liu, Z., Rao, Y.: Targeted sentiment classification with attentional encoder network. In: Tetko, I.V., Kůrková, V., Karpov, P., Theis, F. (eds.) ICANN 2019. LNCS, vol. 11730, pp. 93–103. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30490-4_9

23. Stevens, J.R.: Replicability and reproducibility in comparative psychology. Front. Psychol. **8**, 862 (2017). https://doi.org/10.3389/fpsyg.2017.00862

24. Tang, D., Qin, B., Feng, X., Liu, T.: Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, pp. 3298–3307. The COLING 2016 Organizing Committee, December 2016

25. Tang, D., Qin, B., Liu, T.: Aspect level sentiment classification with deep memory network. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, pp. 214–224. Association for Computational Linguistics, November 2016. https://doi.org/10.18653/v1/D16-1021

26. Tay, Y., Tuan, L.A., Hui, S.C.: Dyadic memory networks for aspect-based sentiment analysis. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 107–116. ACM (2017)

27. Wang, B., Lu, W.: Learning latent opinions for aspect-level sentiment classification. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018), New Orleans, Louisiana, USA, 2–7 February 2018, pp. 5537–5544. AAAI Press (2018)

28. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, pp. 606–615. Association for Computational Linguistics, November 2016. https://doi.org/10.18653/v1/D16-1058

29. Watson, J., Ghosh, A.P., Trusov, M.: Swayed by the numbers: the consequences of displaying product review attributes. J. Mark. **82**(6), 109–131 (2018). https://doi.org/10.1177/0022242918805468

30. Xu, H., Liu, B., Shu, L., Yu, P.: BERT post-training for review reading comprehension and aspect-based sentiment analysis. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 2324–2335. Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/N19-1242

31. Xue, W., Li, T.: Aspect based sentiment analysis with gated convolutional networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, pp. 2514–2523. Association for Computational Linguistics, July 2018. https://doi.org/10.18653/v1/P18-1234

32. Yang, H., Zeng, B., Yang, J., Song, Y., Xu, R.: A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction. arXiv preprint arXiv:1912.07976 (2019)

33. Zeng, B., Yang, H., Xu, R., Zhou, W., Han, X.: LCF: a local context focus mechanism for aspect-based sentiment classification. Appl. Sci. **9**, 3389 (2019)