OPEDABSA: A Dataset for Open Domain Aspect-Based Sentiment Analysis from Public Reviews

Anonymous ACL submission

Abstract

Sentiment Analysis is core to customer management, product development and service delivery. In recent years, the need for Aspect-Based Sentiment Analysis (ABSA) has led to three shared tasks in SemEval (2014, 2015 and 2016), which attracted a large number of submissions from around the globe. Two challenges confronting ABSA are- **low amount of data** and **constrained domain coverage**. This work attempts to address these problems by presenting an open domain gold standard dataset (covering **111** fine-grained domains) curated from publicly available reviews. Along with the dataset, we also present strong baselines for four tasks- Aspect Term Extraction, Aspect Polarity Classification, Sentence Polarity Classification and End-to-End ABSA. We provide experimental results which show that our dataset helps models achieve a much better performance ($\sim$ **18.33**% absolute improvement, on average) in open domain ABSA tasks.

1 Introduction

Sentiment Analysis is one of the oldest application oriented domains of Natural Language Processing. The task has huge applications in the IT industry, a few include product improvement, customer segmentation, targeted marketing, etc. The primary premise is- given a sentence, the polarity/sentiment expressed is desired. However, with the increase in user reviews, it has been understood that more fine-grained sentiment analysis is necessary.

In order to aid the development of models for such applications, we introduce a gold standard dataset in this paper, that has been curated from Yelp reviews. Datasets (Pontiki et al., 2014, 2015, 2016; Pavlopoulos and Androutsopoulos, 2014; Jiang et al., 2019) previously posed for the task of Aspect-Based Sentiment Analysis (ABSA) have mostly been limited in size and the domains they cover. Our dataset introduces instances from a large set of domains. Yelp[1] offers open reviews in a variety of domains, which include (not exhaustive) hotels, restaurants, dentists, salons, dry cleaning, gyms, massage centres, *etc*. This supports the creation of an open domain ABSA dataset. Our dataset includes sentence level annotations of aspect boundaries, sentiments towards the aspects, sentiment of the overall sentence and domains.

The contributions of this work are- (*a*) A gold standard open domain dataset for ABSA and Sentiment Analysis, (*b*) Strong baselines for the possible tasks (Aspect Term Extraction, Aspect Polarity Classification, End-to-End ABSA and Sentence Polarity Classification), and (*c*) Demonstration of superiority of the dataset for open domain ABSA.

The rest of the paper is divided as- Section 2 highlights some of the previous works and draws motivation for this work, Section 3 provides details on the dataset and details on the annotation process, Section 4 provides strong baselines for the possible tasks, Section 5 highlights the superiority of the dataset for domain adaptation in ABSA, Section 6 presents a brief analysis of the results on domain adaptation and highlights the challenges that the dataset poses, and Section 7 provides a conclusion for the work.

2 Related Work

Ganu et al. (2009) produce one of the first works in ABSA. They provide a dataset annotated with the assumption that each sentence refers a single aspect. The dataset provides sentiment annotations for 6 classes/aspects- FOOD, SERVICE, PRICE, AMBIENCE, ANECDOTES and MISCELLANEOUS. Pontiki et al. (2014) extend on this by providing fine-grained aspect annotations along with their polarities. They annotate sentences from reviews in two domains- Laptops and Restaurants.

Pontiki et al. (2015) refine the task of ABSA

[1] yelp.com

by redefining the annotation guidelines to include implicit aspects too. Similar to Pontiki et al. (2014), they provide annotations for reviews in two domains- Laptops and Restaurants. Pontiki et al. (2016) internationalize the dataset, including annotations in 8 languages (English, Arabic, Chinese, Dutch, French, Russian, Spanish and Turkish), across 7 domains. However, the domain for English (in the training set) was still limited to Laptop and Restaurant.

Pavlopoulos and Androutsopoulos (2014) introduce a dataset specifically curated for Aspect Term Extraction, in the domains of Restaurant, Laptop and Hotel. Jiang et al. (2019) introduce a challenging dataset, curated to include sentences with multiple aspects and multiple polarities- each sentence contains at least two aspects with two different polarities. Although challenging, the dataset is synthetic and does not truly represent the real-world scenario for ABSA.

The datasets provided in the past have been mostly limited in the domains they cover. Motivated by this, we provide a dataset covering a large number of domains. Section 5 demonstrates the superiority of our dataset in domain adaptation tasks.

3 Dataset

Our dataset has been created from the publicly available Yelp reviews[2], in the format provided by Pontiki et al. (2014). It includes reviews from an array of domains (not exhaustive)- restaurants, salons/spas, hotels, clothing stores, clinics/hospitals/veterinary centres, clubs, vehicle repair shops, carwash, phone/laptop repair shops, supermarkets, tattoo shops, jewellery shops, concerts, bowling arenas, etc.

We provide four kinds of annotations with the dataset- **aspect boundaries**, **aspect polarities**, **sentence level sentiments** and **fine-grained domains**. We provide separate splits for training and testing of the models. Tables 1 and 2 present statistics of both the splits.

Each sentence in the dataset is annotated with the following details-

- **Review ID**: This is the unique identifier of the Yelp review from which the current sentence was chosen.

[2]Yelp reviews dataset- kaggle.com/yelp-dataset/

.	Split	
	Train	Test
# Sentences	8998	410
Average sentence length (characters)	70	53
# Positive sentiment	4429	116
# Negative sentiment	2391	126
# Neutral sentiment	2178	168

Table 1: Sentence level statistics of the dataset

.	Split	
	Train	Test
# Aspects	9799	584
# Positive sentiment	4565	172
# Negative sentiment	2230	153
# Neutral sentiment	2644	191
# Conflict sentiment	360	68

Table 2: Aspect level statistics of the dataset

- **Sentence Sentiment**: The sentiment label provided by the annotator for the current sentence.
- **Aspects**: The aspects (explicitly present in the sentence) that the user talks about. Each aspect, in turn, contains the following details-
  - **Term**: The sub-string of the sentence that denotes the aspect.
  - **From** and **To**: The start and end indices of the aspect term in the sentence string.
  - **Polarity**: The sentiment label provided by the annotator for the current aspect.

Along with that, we provide fine-grained domain annotations for each review in a separate file. The domains are annotated using the "*business categories*" details available from the Yelp dataset. We normalize the categories into several fine-grained domains, such as `restaurant`, `laundry_service`, `medical_service`, etc. Each review can potentially be linked to multiple fine-grained domains. The ideology follows from our observation that the businesses can provide an array of services. For example, some business may provide both laundry and sewing services, while some other may provide strictly only one of the two. Thus, we feel that categorizing businesses into single domains would

lead to noisy labels. Moreover, we feel that such a domain annotation scheme is much more applicable in real-world scenarios than coarse-grained domains. Motivated by this, we provide fine-grained annotations for domain, with **multiple domains for reviews** whenever applicable. The dataset (train + test) contains **111** fine-grained domains, with an average of **173.86** sentences per domain.

3.1 Annotation Guidelines

We follow the annotation guidelines provided by Pontiki et al. (2014) for aspect boundary and aspect polarity annotations, marking aspects that are explicitly present in sentences. We provide annotation guidelines for the sentiment annotations of the sentences, as follows-

1. A sentence should be annotated with *Positive* (*Negative*) sentiment if it expresses a positive (negative) view towards the business.

2. A sentence should be annotated with *Neutral* sentiment in the following cases-

    - The sentence expresses no explicit polarity towards the business- presents an opinion or stays neutral.
    - The sentence expresses both positive and negative view towards the business.

Additionally, we provide an explicit guideline while annotating aspects. Aspects follow an **abstract hierarchy**, *entity* $\rightarrow$ *aspect* $\rightarrow$ *aspects of aspect* and so on (example- *shopping center* $\rightarrow$ *shirt* $\rightarrow$ *color of the shirt*). Our guideline states to annotate only the first level aspect (*shirt* in the example).

The domain annotations also have been done manually. These annotations follow a normalization scheme, eliminating noise in the category annotations provided in the Yelp dataset. For example, the domain `restaurant` has been assigned to multiple possible categories such as `{Restaurant, Chinese}`, `{Restaurant, Indian Cuisine}`, `{Restaurant, Seafood}`, etc. Due to such variations, all representing the same feature, we normalize the categories into fine-grained domains, instead of using them directly. We provide a list of a few domains in Appendix B, Section 10.

3.2 Annotation Details

The annotation has been done by using an in-house annotation tool. Three annotators (A, B and C- all of them are post-graduate students with a background in Computer Science) have been employed to accomplish the annotation. Of the sentences, **1500** have been utilized to calculate the Inter-Annotator Agreement (IAA). Annotator C is an author of the project, who has formulated the additional guidelines of this annotation. Annotators A and B were provided with the annotation guidelines, and a calibration annotation of 100 representative sentences was done independently for A and B to test and calibrate their understanding.

Tables 3 and 4 report IAA scores for the annotations. We use Fleiss' Kappa (Fleiss, 1971) to report IAA for aspect polarity and sentence sentiment annotations. Aspect boundary annotation is similar to the annotation of Named Entities. Thus, we follow the argument put forward by Brandsen et al. (2020) and report pairwise F_1 scores as IAA for aspect boundary annotation. For two annotators, 1 and 2, we compute the F_1 score[3], as put forward by Hripcsak and Rothschild (2005), as-

$$F_1 = \frac{2 \times |A_1 \cap A_2|}{2 \times |A_1 \cap A_2| + |A_1 - A_2| + |A_2 - A_1|}$$

A_i denotes the set of aspects given by annotator i. $|A_i|$ denotes the cardinality of set A_i.

Annotation type	Fleiss' Kappa
Aspect polarity	0.78
Sentence sentiment	0.80

Table 3: Fleiss' Kappa for Aspect polarity and Sentence sentiment annotations

·	Annotator A	Annotator B
Annotator B	77.6%	-
Annotator C	84%	79%

Table 4: Inter Annotator Agreement for Aspect boundary annotation

In order to compute the IAA score for Aspect polarity specification, we take the aspects that are common to both the annotators in the concerned pair. The final annotation (for these 1500 sentences) has been taken by a voting methodology. For sentences where all the three annotators

[3]We urge the reader to view the reference for a detailed description of how this definition of F_1 score aligns with the Precision and Recall based definition.

Model	ATE			APC		
	Precision	Recall	F_1	Precision	Recall	F_1
BERT	**0.81**	**0.92**	**0.86**	**0.89**	**0.86**	**0.87**
Distill-BERT	0.77	0.89	0.82	0.86	0.81	0.82

Table 5: Macro-Average scores for Aspect Term Extraction (ATE) and Aspect Polarity Classification (APC)

Model	SPC			E2E ABSA		
	Precision	Recall	F_1	Precision	Recall	F_1
BERT	**0.88**	**0.89**	**0.88**	**0.76**	**0.75**	**0.73**
Distill-BERT	0.84	0.86	0.84	0.69	0.69	0.64

Table 6: Macro-Average scores for Sentence Polarity Classification (SPC) and End-to-End ABSA (E2E ABSA)

provided different annotations, sessions were conducted to review the annotation guideline, discuss the thought process of each annotator and arrive at a conclusion. Fortunately, there have been no cases where the annotators had any disagreements post session. The majority of the disagreements, which led to sessions, can be divided into two classes-

- **Nested aspect**- Nested aspects refer to those aspects below the first level in the abstract aspect hierarchy (*color of a shirt*). Initially, the external annotators had difficulty in deciding whether or not an aspect is nested. However, over the course of a few sessions, a systematic procedure was explained to the annotators, to identify the possible entity that the user is referring to in the sentence (explicitly mentioned or implicitly referred). This could then be used to identify whether the potential aspects are direct aspects of the entity or nested aspects of another. This led to annotations that are much more in agreement with the guidelines.

- **Entity/Aspect disambiguation**- Due to the open domain nature of our dataset, many entities can often turn into aspects, when the domain is changed. For instance, the word *restaurant* expresses a target entity for the domain `restaurant`, while the same word expresses an aspect for a domain such as `amusement_park` (restaurant inside Disney World). The source of this ambiguity is similar to the previous one, and applying the same solution alleviated it.

4 Experiments

In this section we detail the experiments conducted on the dataset. Specifically, we report figures for four tasks-

- **Aspect Term Extraction** (ATE)- This task attempts to extract all the aspects present in a given sentence. We formulate the task with a Sequence Labelling framework (similar to Named Entity Recognition), using BIO tagging scheme.

- **Aspect Polarity Classification** (APC)- This task attempts to classify the polarity towards a given aspect within the sentence. We formulate it as a Sequence Classification task.

- **Sentence Polarity Classification** (SPC)- This task attempts to classify the polarity of the entire sentence. Similar to the previous task, we formulate it as a Sequence Classification task.

- **End-to-End ABSA** (E2E ABSA)- This task attempts to extract aspects, along with a classification for the polarity of the extracted aspects. We frame this joint modelling task with a Sequence Labelling framework, using fine-grained BIO tagging scheme (*B-positive*, *B-conflict*, *I-neutral*, etc.).

4.1 Evaluation Measures

We follow the metric definitions provided by Pontiki et al. (2014) for ATE, APC and SPC. As E2E ABSA is formulated as a Sequence Labelling task,

Class	APC			E2E ABSA		
	Precision	Recall	F_1	Precision	Recall	F_1
Positive	0.87	0.99	0.93	0.59	0.91	0.71
Negative	0.84	0.92	0.88	0.69	0.86	0.77
Neutral	0.90	0.68	0.77	0.88	0.77	0.82
Conflict	0.96	0.85	0.91	0.86	0.47	0.61

Table 7: Class-wise scores for Aspect Polarity Classification (APC) and End-to-End ABSA (E2E ABSA)

Model	Accuracy	Macro-F1
AOA (Huang et al., 2018)	77.52	77.05
ATAE-LSTM (Wang et al., 2016)	76.32	75.83
Cabasc (Liu et al., 2018)	76.25	75.79
IAN (Ma et al., 2017)	77.83	77.42
MemNeT (Tang et al., 2016)	76.41	75.99
MGAN (Fan et al., 2018)	75.7	75.2
RAM (Chen et al., 2017)	75.97	75.48
TC-LSTM (Tang et al., 2015)	76.07	75.57
TD-LSTM (Tang et al., 2015)	76.2	75.7
TNet-LF (Li et al., 2018)	76.19	75.71

Table 8: Performance of baseline models for Aspect Polarity Classification (APC) task on our dataset

similar to ATE, we follow the same metric definitions. Following the definitions, we report macro-average values of Precision, Recall and F_1 measures[4] in this paper.

4.2 Baselines and Results

As baselines, we developed Transformer (Vaswani et al., 2017) based models using the HuggingFace transformers (Wolf et al., 2019) library. Specifically we fine-tune two pretrained models- BERT (Devlin et al., 2018) and Distil-BERT (Sanh et al., 2019). We report the results obtained on the test set (refer Tables 1 and 2 for stats). Tables 5 and 6 present the results for these two models. Appendix A, Section 9, provides details on training the models along with the compute requirements. We also report the class-wise performance results in Tables 7 and 9.

Additionally, the APC task (specifically) has garnered numerous baselines over the past years. We use the *PyABSA*[5] library to generate results for these baselines too. Table 8 reports these results.

[4]We use the Python library *seqeval* (Nakayama, 2018) to evaluate ATE and E2E ABSA

[5]github.com/PyABSA

Class	SPC		
	Precision	Recall	F_1
Positive	0.86	0.97	0.91
Negative	0.88	0.89	0.88
Neutral	0.91	0.82	0.86

Table 9: Class-wise scores for Sentence Polarity Classification (SPC)

5 Domain Adaptation

This section presents the open domain performance boosts that our dataset provides. We provide results on several splits to verify our claim.

In order to conduct these experiments, we design separate train-test splits from the combined dataset. We use the fine-grained domain annotations to embed the review sentences in 111 dimensions (one-hot for each fine-grained domain applicable for the sentence). We use these embeddings to accumulate sentences that have very few neighbours into the test set. We set 50 as the number of threshold neighbours within a ball of radius 2 as the criteria. Our experimentation revealed this to produce a reasonably sized test set. Remaining sentences are accumulated into the train split.

.	Split-I		Split-II	
	Train	Test	Train	Test
# Sentences	8958	450	8891	517
# Positive sentiment	4301	244	4202	343
# Negative sentiment	2402	115	2430	87
# Neutral sentiment	2255	91	2259	87

Table 10: Sentence level statistics of the dataset (domain-adaptation splits)

.	Split-I		Split-II	
	Train	Test	Train	Test
# Aspects	9911	472	9916	467
# Positive sentiment	4500	237	4445	292
# Negative sentiment	2272	111	2321	62
# Neutral sentiment	2728	107	2730	105
# Conflict sentiment	411	17	420	8

Table 11: Aspect level statistics of the dataset (domain-adaptation splits)

The analogy behind such a methodology is that sparsely spaced domains would be reasonably far from the domain distribution of the training data, thus providing a good testbed to judge performance on out-of-domain data. For further reference, we refer this split by **Split-I**.

Additionally, we also split the combined dataset by specifying the exact domains for the test set. The domains in Split-* are chosen based on themes. For example, in **Split-II**, the theme revolves around medical services. This ensures that there is no presence of related domains in the train and test split. We create 5 different test sets using this criterion-

1. **Split-II**- This test set contains fine-grained domains relevant to medicine and fitness only (`doctor`, `eyewear_shop`, `fitness_service`, `medical_service`, `sport_shop` and `surgery_service`).

2. **Split-III**- This test set contains fine-grained domains relevant to vehicles only (`car_dealer`, `automotive_service`, `automotive_parts`, `car_rental` and `automobile_repair`).

3. **Split-IV**- This test set contains fine-grained domains relevant to hair salons only (`salon` and `massage`).

4. **Split-V**- This test set contains fine-grained domains relevant to locations only (`amusement_park`, `museum`, `art_gallery`, `arcade`, `park`, `library`, `golf_club` and `casino`).

5. **Split-VI**- This test set contains fine-grained domains relevant to catering services only (`catering_service`).

We use **Split-I** and **Split-II** to demonstrate performance boost provided by our dataset in comparison to an allied dataset. Additionally, we report results, Table 13, using a strong baseline on the remaining splits (**Split-III**, **Split-IV**, **Split-V** and **Split-VI**).

We compare our dataset against that provided by Pontiki et al. (2014) (SE-14, for reference). The reason for choosing this dataset is that it follows a very close annotation scheme as our dataset. We demonstrate results on three tasks- ATE, APC and E2E ABSA for both the splits. We fine-tune identical BERT pre-trained models on both our train sets (**Split-I** and **Split-II**) and the training dataset available from SE-14 (we combine sentences from both `laptop` and `restaurant` domain in SE-14). Table 12 presents the results obtained for both the splits. It can be seen that our dataset leads to significantly better results under identical training conditions. This verifies our claim that our dataset provides much better training instances for models operating under open domain settings. For reference, we report the statistics of both the splits in

Split	Dataset	ATE			APC			E2E ABSA		
		P	R	F_1	P	R	F_1	P	R	F_1
Split-I	Ours	**0.50**	**0.82**	**0.62**	**0.85**	**0.70**	**0.68**	**0.57**	**0.60**	**0.56**
	SE-14	0.46	0.69	0.55	0.56	0.51	0.49	0.30	0.40	0.32
Split-II	Ours	**0.57**	**0.67**	**0.62**	**0.70**	**0.86**	**0.72**	**0.59**	**0.53**	**0.54**
	SE-14	0.46	0.50	0.48	0.51	0.54	0.52	0.27	0.32	0.28

Table 12: Comparison of SE-14 (Laptop + Restaurant, combined) and our dataset for tasks on domain-adaptation, using macro-average Precision (P), Recall (R) and F_1

Split	ATE			APC			E2E ABSA		
	P	R	F_1	P	R	F_1	P	R	F_1
Split-II	0.57	0.67	0.62	0.70	0.86	0.72	0.59	0.53	0.54
Split-III	0.41	0.72	0.53	0.87	0.81	0.83	0.57	0.66	0.59
Split-IV	0.63	0.85	0.73	0.74	0.82	0.77	0.49	0.51	0.45
Split-V	0.41	0.72	0.52	0.70	0.76	0.72	0.62	0.52	0.52
Split-VI	0.58	0.89	0.70	0.74	0.81	0.76	0.63	0.65	0.63

Table 13: Macro-Average Precision (P), Recall (R) and F_1 for tasks on theme oriented splits

Tables 10 and 11.

6 Analysis of Results

Table 13 presents the results obtained by fine-tuned BERT models for various domain adaptation splits. We see that the models perform better for APC, than other tasks. This section presents an analysis of such an observation and articulates the domain-related challenges that this dataset can help solve.

1. **Sentence**: *Rooms* kind of on the small side, but well taken care of and clean. MODEL-I: Positive. MODEL-II: Conflict **Ground Truth**: Conflict
2. **Sentence**: LOVE that there is actually a *parking lot* for the patrons, as it is way better than nothing at all, but it is pure chaos anytime I go. MODEL-I: Positive. MODEL-II: Conflict **Ground Truth**: Conflict
3. **Sentence**: They also have *perfume oils*. MODEL-I: Positive. MODEL-II: Neutral **Ground Truth**: Neutral

Table 14: Aspect Polarity predictions from models trained on SE-14 (MODEL-I) and our dataset (MODEL-II). The underlined and italicized phrase in a sentence signifies the aspect whose polarity is queried from the model.

1. **Sentence**: When they first opened, there seemed to be *attendants* directing traffic a bit, but now it is free for all. MODEL-I: No-aspects. MODEL-II: attendants
2. **Sentence**: OK, Here's the positives... Absolutely the most efficient *check-in* I've ever had! MODEL-I: No-aspects. MODEL-II: check-in
3. **Sentence**: My *tour guide* (Court) knew his stuff. MODEL-I: No-aspects. MODEL-II: tour guide

Table 15: Aspect Term predictions from models trained on SE-14 (MODEL-I) and our dataset (MODEL-II). The underlined and italicized phrases signify the aspects in the sentence.

We find that open domain ABSA is challenging due to two prime factors- *dependence of polarity on domain* and *dependence of aspects on domain*. We discuss these two dependencies in the rest of the section.

Dependence of polarity on domain: The polarity pertaining to an aspect depends on the domain of the review. Taking sentence 2 from Table 14 as an example, we can understand this dependence. A dataset constructed on `laptop` and `restaurant` domains would fail to understand the relation between *chaos* and *parking lot*. It is

thus expected that it would consider a *Positive* polarity as the most likely polarity by looking at the rest of the sentence. On the other hand, MODEL-II (trained on our dataset), understands such nuances much better and is able to predict the correct label.

Dependence of aspects on domain: Similarly, domain decides whether or not a phrase forms an aspect. For instance, *restaurant* can be an aspect when the domain is `amusement_park` but not when the domain is `restaurant`. We can see this in the examples provided in Table 15. In sentences 1, 2 and 3 (Table 15), we can see that MODEL-I fails to detect the presence of any aspect. It is quite unlikely for a dataset on `laptop` and `restaurant` domains to contain *attendants*, *tour guide* or *check-in* as aspects. Alternatively, although our train split for domain adaptation would not contain the same domains (as the test set), it provides enough variations for the model to decently capture domain rich distributional features of words (for instance, attendants might be an aspect in presence of parking lot/traffic), leading to better capture of out-of-domain aspects too.

Table 12 empirically establishes the superiority of our dataset. It establishes our dataset as a decent source to train models for open domain ABSA. Additionally, qualitative analysis also draws conclusion on why our dataset forms a better source-providing enough variations to tackle two key challenges of open domain ABSA.

7 Conclusion

In this paper, we propose an open domain gold standard dataset for Aspect-Based Sentiment Analysis. Our dataset differs from previous datasets by providing a larger training set and covering a wide range of domains. In addition to the dataset, we provide results obtained for a set of strong baselines. We also demonstrate the superiority of our dataset in achieving models that perform significantly well in open domain ABSA. Our results conclude that the dataset is well-suited for open domain ABSA modelling, covering two significant challenges appreciably. We strongly believe that the dataset would help researchers create competitive open domain ABSA models.

Although we serialize a large dataset, we realize that Yelp is an oceanic source of data for Sentiment Analysis, covering a large set of domains. As a future work, we would take up the task of enlarging the dataset. Along with that, we also realize the need of competent model to harness the knowledge within the dataset. To this end, we would also attempt to design model architectures that are specifically tailored to harness this knowledge.

8 Limitations and Ethical Statement

A big assumption of our dataset is that Yelp reviews cover all possible domains in ABSA. Although this is a bold assumption, it is trivial to see that Yelp covers a wide range of domains. Concluding from our Domain Adaptation experiments, we can thus posit that our dataset (consisting of multitude of domains) can be reliably used for ABSA in open domain setting.

Our annotations revealed that Yelp reviews can contain biased and hurtful reviews. We were careful in our annotations and refrained from adding reviews with gender or any other stereotypical biases into our dataset. Additionally, in order to preserve anonymity, we do not include the user data from Yelp in our dataset.

References

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Gayatree Ganu, Noémie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *WebDB*.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 197–206. Springer.

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, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. *arXiv preprint arXiv:1805.01086*.

Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. 2018. Content attention model for aspect based sentiment analysis. In *Proceedings of the 2018 World Wide Web Conference*, pages 1023–1032.

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

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

John Pavlopoulos and Ion Androutsopoulos. 2014. Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 44–52, Gothenburg, Sweden. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

9 Appendix A: Training Details and Compute Requirements

BERT-based models: All of our models use the pre-trained BERT model as the starting point. We fine-tune variants of the model (Token Classifier or Sequence Classifier, available from HuggingFace) using specifications in Table 16. We use DistilBERT to provide an additional baseline for our dataset. It uses the same hyper parameters specified in Table 16, except for the pre-trained model checkpoint. We use the "*distilbert-base-uncased*" (66M parameters) checkpoint to initialize the model. We use the respective tokenizers to tokenize input sentences. We urge the reader to use `huggingface.co/v3.0.2/model_doc/` as a reference to replicate the models we train.

We monitor the applicable metric on the validation set (15% of the training data) to judge convergence on training.

Aspect	Specification
Pre-trained model	bert-base-uncased (110M parameters)
Batch size	32
Optimizer	AdamW
Max number of epochs	50
Learning Rate	1×10^{-6}

Table 16: Hyper parameter specification

Baseline models from *PyABSA*: We use the default hyperparameter configurations (provided by the library) to train the model. We hold out 10% of the training data as the validation set and monitor the F_1 score to choose the best model.

Budget and Compute Requirements: All our experiments were run on the free tier of Google Colab[6] with Tesla T4 instances ($\sim$ 15GB RAM, single GPU). Each experimentation (on all tasks for Domain Adaptation and otherwise) lasted for a maximum of 4 hours. We derive our results from 10 runs for all the experiments.

10 Appendix B: Additional Details on Dataset

Our dataset includes reviews from 111 fine-grained domains, with 173.86 sentences per domain on an average. Yelp already provides fine-grained categories for each business. While using them directly is easier, we observed that the categories are quite sparse, with most of the variations implying the same domain. For example- `{Restaurant, Chinese}`, `{Restaurant, Indian Cuisine}`, `{Restaurant, Seafood}`, etc. all imply the same domain `restaurant`. We felt that merging such variations would lead to a much better domain representation within the dataset. However, we also observed variations in the services that businesses offer- some of them spanning a spectrum, others focusing on a single service. We felt that merging such spectrum into coarse-grained domains could lead to noisy, and potentially unusable, domain labels. Hence, we provide fine-grained annotations, which can always be converted coarse-grained labels as per requirements of dataset users.

Some of the domains included in our dataset are- `restaurant`, `event_planning_service`, `spa`, `hotel`, `massage`, `veterinary`, `fitness_service`, `catering_service`, `car_rental`, `tailoring_service`, `movie_hall`, `laundry_service`, `real_estate`, `wedding_chappel`, `wedding_planning_service`. Additionally, Figure 1 provides annotated examples to acclimatize the reader with our dataset.

[6] colab.research.google.com

11 Appendix C: Details on the Annotators

Three annotators (A, B and C- all of them are post-graduate students with a background in Computer Science) have been employed to accomplish the annotation. Annotator C, an author of the project, has drafted the additional annotation guidelines. In order to accustom annotators A and B with the scheme, a calibration annotation of 100 samples was done. Training sessions were held to resolve the doubts of the annotators during this calibration annotation and after annotators A and B gathered significance confidence, the annotation was started. The annotators were paid a reasonable compensation, decided mutually according to the mental effort and the time utilized for annotation. All the annotators are of Asian descent, with ages within the range 20 and 30, where one of them was a male and the other two were females.

```
<sentence review_id="lWC-xP3rd6obsecCYsGZRg"
 sent_id="lWC-xP3rd6obsecCYsGZRg_3" sentiment="Positive">
  <text>Waitstaff was warm but unobtrusive.</text>
  <aspectTerms>
    <aspectTerm term="Waitstaff" polarity="Positive"
     from="0" to="9"/>
  </aspectTerms>
</sentence>

<sentence review_id="c9I6y_xTGiLyAyZklv4WVw"
 sent_id="c9I6y_xTGiLyAyZklv4WVw_7" sentiment="Positive">
  <text>If you are an avid swimmer, you will also be glad to
   find an indoor pool here as well.</text>
  <aspectTerms>
    <aspectTerm term="indoor pool" polarity="Positive"
     from="61" to="72"/>
  </aspectTerms>
</sentence>

<sentence review_id="LMbMu_vmKY3jKD0sbovJHA"
 sent_id="LMbMu_vmKY3jKD0sbovJHA_1" sentiment="Positive">
  <text>Consistent performance - very trustworthy-
   good employees - typically very punctual.</text>
    <aspectTerms>
      <aspectTerm term="performance" polarity="Positive"
       from="11" to="22"/>
      <aspectTerm term="employees" polarity="Positive"
       from="48" to="57"/>
    </aspectTerms>
</sentence>
```

Figure 1: Examples from the annotated dataset