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 (∼ 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 Androuloupoulos, 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 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
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/

<table>
<thead>
<tr>
<th>.</th>
<th>Split</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
</tr>
<tr>
<td># Sentences</td>
<td>8998</td>
</tr>
<tr>
<td>Average sentence length (characters)</td>
<td>70</td>
</tr>
<tr>
<td># Positive sentiment</td>
<td>4429</td>
</tr>
<tr>
<td># Negative sentiment</td>
<td>2391</td>
</tr>
<tr>
<td># Neutral sentiment</td>
<td>2178</td>
</tr>
</tbody>
</table>

Table 1: Sentence level statistics of the dataset

<table>
<thead>
<tr>
<th>.</th>
<th>Split</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
</tr>
<tr>
<td># Aspects</td>
<td>9799</td>
</tr>
<tr>
<td># Positive sentiment</td>
<td>4565</td>
</tr>
<tr>
<td># Negative sentiment</td>
<td>2230</td>
</tr>
<tr>
<td># Neutral sentiment</td>
<td>2644</td>
</tr>
<tr>
<td># Conflict sentiment</td>
<td>360</td>
</tr>
</tbody>
</table>

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 do-
mains. 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 as-
pect polarity annotations, marking aspects that are
explicitly present in sentences. We provide annota-
tion 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 po-
  larity 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 ab-
tract hierarchy, entity → aspect → aspects of
aspect and so on (example- shopping center →
shirt → 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 normaliza-
tion 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 back-
ground 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 addi-
tional 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 annota-
tions. We use Fleiss’ Kappa (Fleiss, 1971) to report
IAA for aspect polarity and sentence sentiment an-
notations. Aspects 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 F1 scores as IAA for aspect
boundary annotation. For two annotators, 1 and
2, we compute the F1 score⁵, 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 \(A_i\).

<table>
<thead>
<tr>
<th>Annotation type</th>
<th>Fleiss’ Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect polarity</td>
<td>0.78</td>
</tr>
<tr>
<td>Sentence sentiment</td>
<td>0.80</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Annotator A</th>
<th>Annotator B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator B</td>
<td>77.6%</td>
<td>-</td>
</tr>
<tr>
<td>Annotator C</td>
<td>84%</td>
<td>79%</td>
</tr>
</tbody>
</table>

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 con-
cerned pair. The final annotation (for these 1500
sentences) has been taken by a voting methodol-
y. For sentences where all the three annotators

⁵We urge the reader to view the reference for a detailed
description of how this definition of F1 score aligns with the
Precision and Recall based definition.
Table 5: Macro-Average scores for Aspect Term Extraction (ATE) and Aspect Polarity Classification (APC)

<table>
<thead>
<tr>
<th>Model</th>
<th>ATE</th>
<th>APC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>BERT</td>
<td>0.81</td>
<td>0.92</td>
</tr>
<tr>
<td>Distill-BERT</td>
<td>0.77</td>
<td>0.89</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>SPC</th>
<th>E2E ABSA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>BERT</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>Distill-BERT</td>
<td>0.84</td>
<td>0.86</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;</th>
<th>Precision</th>
<th>Recall</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.87</td>
<td>0.99</td>
<td>0.93</td>
<td>0.59</td>
<td>0.91</td>
<td>0.71</td>
</tr>
<tr>
<td>Negative</td>
<td>0.84</td>
<td>0.92</td>
<td>0.88</td>
<td>0.69</td>
<td>0.86</td>
<td>0.77</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.90</td>
<td>0.68</td>
<td>0.77</td>
<td>0.88</td>
<td>0.77</td>
<td>0.82</td>
</tr>
<tr>
<td>Conflict</td>
<td>0.96</td>
<td>0.85</td>
<td>0.91</td>
<td>0.86</td>
<td>0.47</td>
<td>0.61</td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.86</td>
<td>0.97</td>
<td>0.91</td>
</tr>
<tr>
<td>Negative</td>
<td>0.88</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.91</td>
<td>0.82</td>
<td>0.86</td>
</tr>
</tbody>
</table>

### 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`<sup>5</sup> library to generate results for these baselines too. Table 8 reports these results.

---

<sup>4</sup>We use the Python library `seqeval` (Nakayama, 2018) to evaluate ATE and E2E ABSA

<sup>5</sup>github.com/PyABSA
Table 10: Sentence level statistics of the dataset (domain-adaptation splits)

<table>
<thead>
<tr>
<th></th>
<th>Split-I</th>
<th></th>
<th>Split-II</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Sentences</td>
<td>8958</td>
<td>450</td>
<td>8891</td>
<td>517</td>
</tr>
<tr>
<td># Positive sentiment</td>
<td>4301</td>
<td>244</td>
<td>4202</td>
<td>343</td>
</tr>
<tr>
<td># Negative sentiment</td>
<td>2402</td>
<td>115</td>
<td>2430</td>
<td>87</td>
</tr>
<tr>
<td># Neutral sentiment</td>
<td>2255</td>
<td>91</td>
<td>2259</td>
<td>87</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Split-I</th>
<th></th>
<th>Split-II</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Aspects</td>
<td>9911</td>
<td>472</td>
<td>9916</td>
<td>467</td>
</tr>
<tr>
<td># Positive sentiment</td>
<td>4500</td>
<td>237</td>
<td>4445</td>
<td>292</td>
</tr>
<tr>
<td># Negative sentiment</td>
<td>2272</td>
<td>111</td>
<td>2321</td>
<td>62</td>
</tr>
<tr>
<td># Neutral sentiment</td>
<td>2728</td>
<td>107</td>
<td>2730</td>
<td>105</td>
</tr>
<tr>
<td># Conflict sentiment</td>
<td>411</td>
<td>17</td>
<td>420</td>
<td>8</td>
</tr>
</tbody>
</table>

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...
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.

<table>
<thead>
<tr>
<th>Split</th>
<th>Dataset</th>
<th>ATE P</th>
<th>ATE R</th>
<th>ATE F1</th>
<th>APC P</th>
<th>APC R</th>
<th>APC F1</th>
<th>E2E ABSA P</th>
<th>E2E ABSA R</th>
<th>E2E ABSA F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split-I</td>
<td>Ours</td>
<td>0.50</td>
<td>0.82</td>
<td>0.62</td>
<td>0.85</td>
<td>0.70</td>
<td>0.68</td>
<td>0.57</td>
<td>0.60</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>SE-14</td>
<td>0.46</td>
<td>0.69</td>
<td>0.55</td>
<td>0.56</td>
<td>0.51</td>
<td>0.49</td>
<td>0.30</td>
<td>0.40</td>
<td>0.32</td>
</tr>
<tr>
<td>Split-II</td>
<td>Ours</td>
<td>0.57</td>
<td>0.67</td>
<td>0.62</td>
<td>0.70</td>
<td>0.86</td>
<td>0.72</td>
<td>0.59</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>SE-14</td>
<td>0.46</td>
<td>0.50</td>
<td>0.48</td>
<td>0.51</td>
<td>0.54</td>
<td>0.52</td>
<td>0.27</td>
<td>0.32</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Table 13: Macro-Average Precision (P), Recall (R) and F1 for tasks on theme oriented splits

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.

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.

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


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.
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 F1 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 (~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 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.

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.

\(^6\)colab.research.google.com
<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_xTGiLyAy2k1v4WVw"
    sent_id="c9I6y_xTGiLyAy2k1v4WVw_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