A Review of Datasets for Aspect-based Sentiment Analysis

Siva Uday Sampreeth Chebolu[1], Franck Dernoncourt[2], Nedim Lipka[3], Thamar Solorio[1,4]
[1]University of Houston, 4800 Calhoun Rd, Houston, TX, 77004, USA
[2]Adobe Research, 801 N 34th St, Seattle, WA 98103, USA
[3]Adobe Research, 321 Park Ave, San Jose, CA, 95110, USA
[4]MBZUAI, Masdar City, Abu Dhabi, UAE
[1]{sivauday.sampreeth8,thamar.solorio}@gmail.com
[2]franck.dernoncourt@gmail.com, [3]lipka@adobe.com

Abstract

Aspect-based sentiment analysis (ABSA) is a natural language processing problem that analyzes user-generated reviews to determine a) the target entity being reviewed, b) the high-level aspect to which it belongs, and c) the sentiment expressed toward the targets and the aspects. Numerous yet scattered corpora for ABSA make it difficult for researchers to identify corpora best suited for a specific ABSA subtask quickly. This study presents a database of corpora that can be used to train and evaluate autonomous ABSA systems. Additionally, we provide an overview of the major corpora for ABSA and its subtasks and highlight several features that researchers should consider when selecting a corpus. Finally, we discuss the advantages and disadvantages of existing dataset collection approaches and make recommendations for future corpora creation. This survey examines 98 publicly available ABSA datasets covering over 25 domains, including 77 English and 21 other languages datasets (https://github.com/RiTUAL-UH/ABSA-Datasets-Info).

1 Introduction

Consumers, product makers, and service providers benefit differently from user-generated reviews on e-commerce platforms. Reading about previous customer experiences can assist future customers in making informed decisions. At the same time, the characteristics that elicit user feedback may help manufacturers and merchants develop measures to enhance customer satisfaction. Furthermore, as the data grows daily at a rapid pace, there is a need to recognize and extract sentiment or opinion from text reviews automatically. Opinion mining or Sentiment Analysis (Pang et al., 2002; Turney, 2002) is a technology that combines computational linguistics and natural language processing to extract such opinions.

Aspects can be a feature, a trait, or a behavior of a product or an entity, like the atmosphere of a restaurant, the performance of a laptop, the display of a phone, and so on. The sentiment analysis focused on a finer degree, namely, aspect-based sentiment analysis (ABSA) (Hu and Liu, 2004a), determines the sentiment for each entity as well as its aspects (Poria et al., 2020). Many systems, metrics, and subtasks are created along with various corpora to solve the task. The goal of ABSA is to extract four elements: 1) the *target* entity of an opinion, 2) the high-level *aspect* the entity belongs to, 3) the actual *opinion* phrase, and 4) finally give a *sentiment* polarity to the specific target-aspect-opinion triple. For example, in the review "*The pizza is very tasty.*", *pizza* is the target entity, *FOOD* is the aspect category, *very tasty* is the opinion phrase, and the sentiment polarity for *pizza-FOOD-very tasty* is *positive*. The subtasks get their names from the subset of identified elements in that study.

Although there has been significant research on ABSA in the last two decades, it has become more popular after its formal introduction as a task in the SemEval-2014. SemEval-2015 consolidated its subtasks into a single framework in which all detected elements of expressed opinions (i.e., aspects, opinion and target expressions, and sentiment polarities) comply with a set of criteria and are related via sentence-level tuples. However, a user may be interested in the text's overall rating on a particular aspect. These ratings may be used to calculate the average sentiment for each aspect based on several sentences of a single review. Thus, in addition to sentence-level ABSA annotations, SemEval-2016 Task 5 included text-level ABSA , showing that ABSA can be performed at 1) Sentence-level and 2) Review-level (Chebolu et al., 2022).

Given the wide range of ABSA subtasks and techniques, researchers may find it challenging to establish which corpora are optimal for a specific research task. We want to solve this difficulty by providing an overview of available corpora and evaluating their applicability for fundamental

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 611–628
November 1–4, 2023.

ABSA tasks. Specifically, this research aims to review and summarize the literature on collecting text and categorical values for ABSA elements, explain what has been learned to date, and give recommendations for constructing future datasets.

1.1 What is different from Previous Surveys?

The primary difference between this survey and previous ones on ABSA (Laskari and Sanampudi, 2016; Schouten and Frasincar, 2016; Suresh and Raghavi, 2016; Sethi and Bhattacharyya, 2017; Sabeeh and Dewang, 2018; Do et al., 2019; Ahmet and Abdullah, 2020; Nazir et al., 2020; Brauwers and Frasincar, 2021) is that the previous work primarily focusses on the tasks, conduct a critical analysis of the techniques, and offer ideas and future directions for enhancing the performance of the tasks and addressing unresolved issues. In contrast, this research aims to review and summarize the literature on collecting text reviews and categorical values for ABSA elements, explain what has been learned to date, and give recommendations for constructing future datasets. Consequently, this survey complements the previous and more current ABSA surveys and critical retrospectives (Poria et al., 2020) that focus on definitions, methodology, and evaluations.

1.2 Contributions and Organization

We review 98 publicly available ABSA datasets in this survey that cover more than 25 domains, with 76 English and 22 other language datasets that help solve 12 different subtasks. The scope of this paper covers all the datasets specific to Aspect-based Sentiment Analysis rather than general sentiment analysis. We provide an overview of existing subtasks and current datasets, followed by a live version of the tables as a website allowing community additions. Following that, we look at what can be learned from current data collection approaches and provide a few suggestions for future ABSA datasets. We emphasize a few aspects including, the opinion phrase annotation and moving towards review-level ABSA from sentence-level, in the discussion section, that are particularly essential to the present ABSA research.

2 Tasks and Datasets Overview

This section will discuss the various tasks and subtasks associated with ABSA and the different datasets that help solve one or more of its subtasks independently or jointly.

2.1 Tasks Overview

ABSA comprises two sub-problems: 1) aspect extraction (for example, sushi, pasta, and well-behaved staff) and 2) identifying the polarity toward each aspect. Aspect extraction has two subtasks: a) extracting aspect terms/targets and b) categorizing/normalizing the extracted aspect terms into aspect categories. In turn, polarity detection has three subtasks: a) extract the opinion-oriented expression, b) determine the polarity of each category or each aspect word, and c) determine the joint polarity for aspect terms/targets and aspect categories. For example, we have a positive sentiment polarity for the aspect terms *value, dumplings, sushi*, and *service* and the respective aspect categories *Price, Food, Food*, and *Service* for the review: "Highly recommend this as great value for excellent dumplings, sushi, and service". The opinion phrases that are useful in determining the polarity are *great*, and *excellent*. Therefore, the four main elements that we can identify from a given data for different ABSA tasks and sub-tasks are 1) aspect terms/targets, 2) aspect categories, 3) opinion phrases, and 4) sentiment polarity.

Certain terms in Aspect-Based Sentiment Analysis (ABSA) research bear multiple labels, often used interchangeably across studies. Aspect terms are often referred to as targets and opinion target expressions. Alternative usages of aspect categories can be categories and entity-attribute pairs. Opinion expressions are a span of words/tokens in a text that provide a sentiment orientation towards an aspect term or target. These can be seen as opinion phrases, opinion terms, opinion words, sentiment phrases, sentiment words, and similar vocabulary. The term *aspect* can be ambiguous, as some researchers use it to denote aspect terms, while others use it for aspect categories. The exact meaning often hinges on the specific task being tackled and the elements emphasized within that task.

We present an overview of all the sub-tasks that stemmed from ABSA in Table 1. We provide the year and the paper in which the sub-task was first introduced along with the task name (as both acronym and full common name), the inputs of each task (as *Input*), and their identified subset of elements (as *Expected Output*) as the columns. For example, Aspect-Category Sentiment Analysis (ACSA) aims to identify the polarity of a

Restaurant Review (sentence): The ***pasta*** was ***very yummy*** but the ***place*** has some ***weird smell***.
List of Aspect Categories in the dataset: {*Food, Ambience, Service, Price, General*}

Year	Paper	Common Name	Acr.	Input	Expected Output
2004	Hu and Liu (2004a)	Aspect Term Extraction[1]	ATE	sentence	***pasta, place***
2004	Hu and Liu (2004a)	Aspect Term Sentiment Analysis[1]	ATSA	sentence, (***pasta, place***)	positive, negative
2014	Pontiki et al. (2014)	Aspect Category Detection[1]	ACD	sentence	*Food, Ambience*
2014	Pontiki et al. (2014)	Aspect Category Sentiment Analysis[1]	ACSA	sentence, (*Food, Ambience*)	positive, negative
2015	Pontiki et al. (2015)	Target Detection[1]	TD	sentence	***pasta, place***
2014	Pontiki et al. (2014)	Target Aspect Detection[2]	TAD	sentence	(***pasta***, *Food*), (***place***, *Ambience*)
2018	Schmitt et al. (2018)	Aspect Sentiment Joint Detection[2]	ASD	sentence	(*Food*, positive), (*Ambience*, negative)
2019	Li et al. (2018)	Target Sentiment Joint Detection[2]	TSD	sentence	(***pasta***, positive), (***place***, negative)
2019	Peng et al. (2020)	Aspect Sentiment Triplet Extraction[3]	ASTE	sentence	(***pasta***, ***very yummy***, positive), (***place***, ***weird smell***, negative)
2020	Wan et al. (2020)	Target Aspect Sentiment Detection[3]	TASD	sentence	(***pasta***, *Food*, positive), (***place***, *Ambience*, negative)
2020	Fan et al. (2019)	Target Opinion Word Extraction[2]	TOWE	sentence	(***pasta***, ***very yummy***), (***place***, ***weird smell***)
2021	Li et al. (2021)	Aspect-Sentiment-Opinion Triplet Extraction[3]	ASOTE	sentence	(***pasta***, positive, ***very yummy***), (***place***, negative, ***weird smell***)
2021	Cai et al. (2021)	Aspect-Category-Opinion-Sentiment[4±±]	ACOS	sentence	(***pasta***, *Food*, positive, ***very yummy***), (***place***, *Ambience*, negative, ***weird smell***)
2021	Zhang et al. (2021a)	Aspect Sentiment Quad Prediction[4±±]	ASQP	sentence	(***pasta***, *Food*, positive, ***very yummy***), (***place***, *Ambience*, negative, ***weird smell***)

Table 1: Common sub-tasks of ABSA and their relation with the identified elements Aspect Categories, Aspect Terms (a.k.a Targets), Opinion Phrases, and Sentiment Polarity. sentence: Review Sentence. Acr.: Acronym for the Common Name. [1]: single outcome sub-task, [2, 3, 4]: joint outcome sub-tasks. ±±: These tasks are similar but have been defined separately by different authors.

given aspect category. However, the Target-Aspect-Sentiment Detection (TASD) task jointly identifies the targets, aspect categories, and the polarity expressed towards the target-category pair. The last four rows in the Table are recently created tasks that include identifying opinion phrases in the given text that align with the sentiment polarity towards targets or aspects.

2.2 Task Challenges

There are a few major challenges for ABSA and its sub-tasks. Firstly, each of the elements described above is not independent but rather depends on other elements' detection. For example, aspect term extraction and aspect category detection tasks can be used in tandem to find terms and categories in a review (Wan et al., 2020; Xue et al., 2017). The aspect term extractor may extract related aspect terms and vice versa if it knows which aspect categories a review belongs to. In the review, "*However, it's the service that leaves a sour taste in my mouth*." The term *service* is explicitly used, indicating the aspect *service*. If the aspect term extractor is aware of the aspect category *Service*, it gives the word *service* in the review greater weight. Similarly, if the term *service* is given higher weight in the review, the aspect category detector can identify the *Service* category easily. Also, we need to detect the implicit phrase "sour taste in my mouth" as a sentiment indicator to know that the review conveys a negative sentiment polarity towards *service*. This phrase is an idiom with a negative connotation for service. The literal meaning should not be considered in this situation because the criticism is not directed toward any of the restaurant's food or beverages. There is some dedicated research on the implicit aspect and its sentiment detection (Cruz et al., 2014), which could be leveraged to improve the overall detection performance.

Another issue is ABSA's relevance and reliance on several other NLP tasks. It is worth noting that not every entity described in a text is an aspect. Entities that are the subject of an opinion are referred to as aspects. We require a sophisticated NER system to identify the names of foods, beverages, restaurants, computers, processors, hotels, and other items that may be possible targets/aspect-terms in the provided review sentence. To address the common opinion issue, we must find opinion phrases in the supplied review and link them to all of the proper entities. This issue is closely connected to the NLP Entity-Linking problem (Daiber et al., 2013; Kolitsas et al., 2018). *excellent* is used for both *food* and *margaritas* in the review "*the food was excellent, the margaritas too.*" On the other hand, there is no explicit reference to a target entity in this review "*creative and tasty but pricey*," thus it should be assumed that the opinion is conveyed on *FOOD*. However, we must conclude that the aspect is *RESTAURANT* rather than *FOOD* in the restaurant review "*Good and affordable.*" To address these issues, we must model the ABSA problem jointly with the related sub-problems such as aspect-term identification (NER), polarity and opinion detection (opinion phrases), syntactic simplification (Siddharthan, 2006; Scarton et al., 2017) to get separate sentences for each opinion-entity pair to solve the common-opinion problem.

2.3 Datasets Overview

All the publicly available datasets for ABSA are presented in Tables 7 and 8. We provide the paper introducing the dataset, with its citation, dataset's source and domain, number of reviews and the respective number of sentences, and other statistics such as the number of sentences that are annotated with positive, negative and neutral sentiment polarity for the aspect terms/targets or aspect categories. In Table 2, we show which ABSA sub tasks from Table 1 could be evaluated using the datasets from Table 7 and 8.

The SemEval challenge datasets and the recently published SentiHood and MAMS corpora are the most extensively used corpora for aspect-based sentiment analysis. The SemEval corpora were made public as part of a shared work held during the International Workshop on Semantic Evaluation, held annually from 2014 to 2016. The datasets are described in full in (Pontiki et al., 2014, 2015, 2016). Historically, ABSA was primarily concerned with aspect term extraction and sentiment analysis (Hu and Liu, 2004a). Before 2014, there has been very little research into aspect category detection and sentiment analysis. However, as ACD was formally presented at SemEval-2014, a slew of new challenges arose. ABSA has witnessed positive outcomes across all tasks thanks to the emergence of artificial neural networks.

Despite the popularity of SemEval datasets for this work, most sentences only include one or many aspects with the same sentiment polarity, reducing the ABSA task to sentence-level sentiment analysis. (Jiang et al., 2019) published a new large-scale Multi-Aspect Multi-Sentiment (MAMS) dataset, in which each phrase has at least two independent aspects with different sentiment polarity. Although Jiang et al. (2019) claimed that each sentence has more than one aspect-sentiment tuple, the approach they followed is not realistic. When there is only one opinion tuple in a sentence, they introduce either a "miscellaneous" category or another category with neutral sentiment as a second opinion tuple that does not have an opinion in the review. For instance, in the following review, *I like the smaller portion size for dinner.*, there is only one opinion, which is about the food's portion size. However, the actual annotation has two opinion tuples: one is for the food, and the other is a neutral opinion on the restaurant's *miscellaneous* aspect category. We do not dispute the legitimacy of this strategy, but we do not find it practical in the real world. Another drawback with the SemEval corpora is that they contain reviews about a single *target entity*, such as a laptop or restaurant. To overcome this, (Saeidi et al., 2016) created the SentiHood dataset to identify the sentiment towards each aspect of one or more entities.

As discussed previously, an opinion phrase is critical in determining the sentiment polarity towards an aspect or a target and, sometimes, determining to which target/aspect that opinion belongs. Fan et al. (2019); Peng et al. (2020) modified the SemEval datasets to account for the missing annotation of opinion phrases that lead to a specific sentiment polarity for the target, aspect term, or aspect category. However, this resulted in a few instances, coercing them to merge all the reviews from SemEval 2014 to 2016 into a single dataset.

2.4 Annotation Procedure and Dataset Source

Even though researchers use various annotation methods when building ABSA datasets, we explain the most frequent method here. One annotator (A) initially annotates a portion of the data, which is then checked by another annotator (B) for any corrections. The remainder of the sentences in the dataset will be annotated by annotator A, with additional instructions based on the nature of the earlier disagreements. When A lacked assurance, a decision was taken in collaboration with B. When A and B differed, they and a third expert annotator came to a judgment together. Another conflict resolution method was to take the vote of the majority and consider that as the correct annotation. Since most of the datasets follow this procedure where one annotator annotates and the expert annotator checks for the mistakes, many of the dataset papers lack the inter-annotator agreement scores.

The SemEval-2014 (SE-14) dataset was annotated in two stages. The first stage consisted of tagging and detecting the polarity of all single and multi-word words that designated certain aspects of the target item. The second step involves identifying the aspect categories and polarity of the sentences. Most datasets that include annotations simply for aspect terms/targets and their polarity, such as the Customer Review datasets (Hu and Liu, 2004b; Ding et al., 2008; Liu et al., 2015), TOWE (Fan et al., 2019), ASTE (Peng et al., 2020), follow the first stage of this process. The second stage is only implemented for datasets including aspect

Dataset Paper	ATE	ATSA	ACD	ACSA	TD	TSD	ASD	TAD	TASD	ASTE	TOWE	QUAD.Ex
Customer Reviews (Hu and Liu, 2004b)	✓	✓	-	-	-	-	-	-	-	-	-	-
Customer Reviews (Ding et al., 2008)	✓	✓	-	-	-	-	-	-	-	-	-	-
JDPA (Kessler et al., 2010)	-	-	-	-	✓	✓	-	-	-	-	-	-
Darmstadt Service (Toprak et al., 2010)	-	-	-	-	✓	✓	-	-	-	-	-	-
TripAdvisor Hotels (Wang et al., 2011)	-	-	✓	✓	-	-	✓	-	-	-	-	-
Czech Restaurants (Steinberger et al., 2014)	✓	✓	✓	✓	-	-	-	-	-	-	-	-
SE-14 Restaurants (Pontiki et al., 2014)	✓	✓	✓	✓	-	-	-	-	-	-	-	-
SE-14 Laptops (Pontiki et al., 2014)	✓	✓	-	-	-	-	-	-	-	-	-	-
Twitter Comments (Dong et al., 2014)	-	-	-	-	✓	✓	-	-	-	-	-	-
Customer Reviews (Liu et al., 2015)	✓	✓	-	-	-	-	-	-	-	-	-	-
HAAD (Al-Smadi et al., 2015)	-	-	✓	✓	✓	✓	✓	-	✓	-	-	-
SE-15 Restaurants (Pontiki et al., 2015)	-	-	✓	✓	✓	✓	✓	-	✓	-	-	-
SE-15 Laptops (Pontiki et al., 2015)	-	-	✓	✓	-	-	✓	-	-	-	-	-
SE-16 Rest & Hotels (Pontiki et al., 2016)	-	-	✓	✓	✓	✓	✓	-	✓	-	-	-
SE-16 Telecom (Pontiki et al., 2016)	-	-	✓	✓	✓	✓	✓	-	✓	-	-	-
SE-16 Laptops (Pontiki et al., 2016)	-	-	✓	✓	-	-	✓	-	-	-	-	-
SE-16 Mob.Phns.(Pontiki et al., 2016)	-	-	✓	✓	-	-	✓	-	-	-	-	-
SE-16 Dig.Cam. (Pontiki et al., 2016)	-	-	✓	✓	-	-	✓	-	-	-	-	-
SentiHood (Saeidi et al., 2016)	-	-	✓	✓	✓	✓	✓	-	✓	-	-	-
GermEval-2017 (Wojatzki et al., 2017)	-	-	✓	✓	✓	✓	✓	-	✓	-	-	-
BeerAdvocate, TripAdvisor (Yin et al., 2017)	-	-	✓	✓	-	-	✓	-	-	-	-	-
ABSITA-2018 (Basile et al., 2018)	-	-	✓	✓	-	-	✓	-	-	-	-	-
FiQA (de França Costa and da Silva, 2018)	-	-	✓	✓	✓	✓	✓	-	✓	-	-	-
Ba-Re-Cr (Rahman and Kumar Dey, 2018)	-	-	✓	✓	-	-	✓	-	-	-	-	-
MAMS (Jiang et al., 2019)	✓	✓	✓	✓	-	-	✓	-	-	-	-	-
TOWE (Fan et al., 2019)	-	-	-	-	✓	-	-	-	-	-	✓	-
Telugu Movies (Regatte et al., 2020)	✓	✓	✓	✓	-	-	✓	✓	-	-	-	-
ASTE (Peng et al., 2020)	-	-	-	-	✓	✓	-	-	-	✓	✓	-
ASOTE (Li et al., 2021)	-	-	-	-	✓	✓	-	-	-	✓	✓	-
ABSITA-2020 (De Mattei et al., 2020)	✓	✓	-	-	-	-	-	-	-	-	-	-
NewsMTSC (Hamborg and Donnay, 2021)	✓	✓	-	-	-	-	-	-	-	-	-	-
ASAP (Bu et al., 2021)	-	-	✓	✓	-	-	✓	-	-	-	-	-
ASQP (Zhang et al., 2021a)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ACOS (Cai et al., 2021)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DM-ASTE (Xu et al., 2023)	-	-	-	-	✓	✓	-	-	-	✓	✓	-
DE-ASTE (Chia et al., 2023)	-	-	-	-	✓	✓	-	-	-	✓	✓	-
MEMD-ABSA (Xu et al., 2023)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 2: Related subtasks of ABSA for each Dataset. QUAD.Ex: Quadruple Extraction. Ba-Re-Cr: Bangla Restaurants and Cricket Dataset. Note: The acronyms of the subtasks in the column names are according to the *Acr.* column in Table 1.

categories, such as the SemEval and FiQA datasets.

Few datasets, such as MAMS, give distinct annotations for ATSA and ACSA tasks where there is no one-to-one correspondence between aspect words, aspect categories, and their polarities. The restaurant datasets from SemEval (Table 7 and 8) are a subset of the datasets published by Ganu et al. (2009) that had only six aspect categories.

A typical approach followed by researchers is to take existing datasets and annotating them for missing items for an existing subtask or propose a new subtask for ABSA from the annotations. The TOWE and ASTE datasets (Table 7) are derived from SemEval restaurants and laptops. The authors included the opinion phrase information for the existing opinion tuples to propose a new subtask. The most common disagreements were noticed when annotating the multiword aspect term boundaries, aspect term vs. reference to target entity, neutral polarity ambiguity, and the problem of distinguishing aspect terms when they appear in conjunctions or disjunctions. The last one was resolved using the maximal phrase as the aspect term. Most of the English restaurant datasets, such as SemEval, MAMS, TOWE, and ASTE, are obtained from citysearch.com for New York restaurants. while the Laptop data were derived from laptop reviews on Amazon.com. Since aspect category detection task is formmaly introduced in SemEval-2014, all preceding datasets only include annotations for aspect words and their polarity.

3 Discussion and Future Directions

We explore several characteristics of the corpora in this section, including the formats, the need for joint datasets with opinion phrase annotations, and recommendations for future ABSA datasets.

3.1 Dataset Formats

The definition and format of ABSA components vary greatly depending on the dataset's source. SemEval-2014, for example, published a dataset with explicit and independent aspect categories, aspect terms, and corresponding sentiment polarity. Because there is no one-to-one correspondence between the terms and the categories, using aspect

terms and categories in a joint detection scenario is problematic. Therefore, one is forced to work on either the ATE and ASTE or the ACD and ACSA. However, in SemEval-2015 and SemEval-2016, the dataset structure is more unambiguous, establishing a link between the targets and the aspect categories. The sentiment polarity is linked to the target-aspect category combination. It allows the community to recognize a text's stated sentiment better using terms and categories. Again, in SemEval-2015 and SemEval-2016, the aspect category is divided into 1) Entity and 2) Attribute. Entities can be the reviewed entity itself, such as the *RESTAURANT*, a part/component of it, such as *AMBIENCE*, or another relevant entity, such as *DRINKS*. Attributes are facets of an entity such as *PRICE* or *QUALITY*.

Jiang et al. (2019), Regatte et al. (2020) and a few others followed the SemEval-2014's XML format and released new datasets in the recent past. However, (Fan et al., 2019; Peng et al., 2020) modified the datasets from all three SemEval shared tasks into another format to include the opinion phrase and released the datasets in an XML and NER task's BIO (beginning, inside, and outside) format.

On the other hand, the SentiHood dataset used a *JSON* format to provide the annotations for targets, aspects, and sentiments. But the definition of aspect category in the SentiHood dataset is the combination of the target and the aspect, leading to identifying ACD and ACSA tasks from Table 1. As mentioned in Section 2.3, aspect categories are formally introduced in SemeEval-2014. All the prior datasets only annotate the aspect terms and their sentiment polarity to solve the ATE and the ATSA tasks. Therefore, for more robust ABSA systems, we urge that the community use an already established structure and criteria for future datasets rather than introducing a new format or structure. The standardization of the annotation format would also benefit benchmarking and updating existing models without any adjustments to the architecture.

3.2 Opinion Phrases and Datasets Merging

As previously explained in Section 2.2, we must identify the opinion words in a given text to determine the sentiment polarity and the entities on which the opinion is conveyed, i.e., the aspect terms. It is evident from Table 2 that most recent tasks, such as the ASTE, TOWE, and ASQP, annotated the opinion words in the current SemEval shared task datasets to enhance the ABSA task.

In the original datasets of the SemEval challenge, the opinion targets (aspect terms) are annotated, but the opinion words and their correspondence with targets are not provided. In addition, most of the available benchmark corpora are small. This gives the opportunity to combine or merge datasets with similar characteristics. For instance, Fan et al. (2019) annotated the corresponding opinion words for the annotated targets. The sentences without targets or with implicit opinion expressions are not included. The original ASTE dataset does not contain cases where a single opinion span is associated with multiple targets. Consequently, Peng et al. (2020) refined the dataset with these additional missing triplets and expanded the corpora.

The community could focus on merging the existing datasets to obtain better quality corpora with increased sizes. Furthermore, researchers could annotate for the opinion words missing in most of the current datasets, which could greatly improve the overall performance of ABSA.

3.3 Need Large Datasets for Unified Models

Recent ABSA datasets are mostly drawn from SemEval shared challenges and include additional data processing and task-specific annotations. The small number of instances (for example, hundreds of phrases) in each dataset makes it challenging to compare models with reliability, particularly Transformer-based models with millions of parameters. Researchers currently evaluate a model's accuracy by averaging the results of numerous runs, but larger datasets would allow for more precise comparisons. However, more challenging datasets must still be provided to meet real-world scenarios that include reviews from many domains such as Xu et al. (2023); Chia et al. (2023) or languages, for example, can help evaluate multi-domain and multi-lingual ABSA systems.

Recently, the unified models built using the generative frameworks (Chebolu et al., 2021; Zhang et al., 2021b) yield SOTA performance on all the subtasks of ABSA by jointly solving for all the elements. The advantage of these unified models is that they could solve multiple subtasks without a change in the model architecture. Building more datasets similar to Zhang et al. (2021a); Cai et al. (2023), with annotations for all the elements, would be beneficial in developing and investigating these promising types of models.

Review: I picked the ***asparagus***, which was ***incredible***. It was ***steamed and tossed with garlic***. The ***steak*** was nice and juicy. It's ***served with either a peppercorn sauce or red wine reduction***. The ***service from the staff*** was ***extremely attentive and very friendly***. It was the highlight of our dinner.

Target	Aspect Cat.	Opinion Expr	Sent.
asparagus	FOOD#QLTY	incredible	Positive
NULL	FOOD#STY_OP	steamed and tossed with garlic	Positive
steak	FOOD#QLTY	nice and juicy	Positive
NULL	FOOD#STY_OP	served with either a peppercorn sauce or red wine reduction	Positive
service from the staff	SERVICE#GEN	extremely attentive and very friendly	Positive
NULL	FOOD#QLTY (OR) AMBIANCE#GEN	highlight	Positive

Table 3: Result of applying Sentence-Level methods to full-review to detect Target, Aspect, Opinion, and Sentiment.

3.4 Datasets in Low-Resource Languages

It is evident from the presented tables and preceding discussion that the majority of ABSA's available datasets are in English. Very few datasets are available for low-resource languages, where in most cases, just one dataset is accessible. Numerous studies employed cross-lingual approaches to automatically produce new datasets for low-resource languages, or to tackle the target language using data from the source language, or for other purposes. A few works first translated from the source to the target language with an off-the-shelf translation system and then aligned the labels using FastAlign or some other softwares (Barnes et al., 2016; Klinger and Cimiano, 2015). Others use the cross-lingual word embeddings pre-trained on large parallel bilingual corpus, or utilize the multilingual Pre-trained Language Models (mPLMs) such as multilingual BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), or the zero-shot learning (Jebbara and Cimiano, 2019). Despite these efforts, the quality of the datasets for low-resource languages are still under par. It could also be attributed to the fact that the XABSA problem is relatively under-explored compared to the monolingual ABSA. Although mPLMs are widely used for various cross-lingual NLP tasks nowadays, exploring their usage in the XABSA can be tricky since language-specific knowledge plays an essential role in any ABSA task (Zhang et al., 2022).

It explains the need to annotate datasets in different languages to push the boundaries of ABSA research ahead. Given the fact that sentiment analysis is a very subjective task, if researchers do not compromise on getting high inter-annotator agreement scores for annotating different elements of ABSA, the quality of the datasets could be ensured with high-confidence. As an initial step, taking the annotation costs into consideration, researchers can build small but high quality datasets in the low-resource languages.

3.5 Sentence-Level to Review-Level ABSA

In real-life scenarios, reviews often contain multiple sentences with overlapping contexts (Figure 1 in Appendix), making sentence-level ABSA methods less effective. These methods assign a NULL value for implicit targets and opinions within a sentence context and struggle to generalize to full reviews. In contrast, by considering the entire review, explicit targets in one sentence might be referred to implicitly in others, providing a richer and more nuanced analysis. The previous focus on sentence-level ABSA limits the applicability of these methods to real-world situations. Even tasks that included a Text-Level ABSA component, like SE16-ABSA (Pontiki et al., 2016; Chebolu et al., 2022), were aimed at summarizing opinions from individual sentences rather than capturing ABSA elements within the entire review context.

Moreover, applying sentence-level ABSA to full reviews often fails to accurately capture the intended targets of pronouns or implicit references (Table 3). For instance, the target of a pronoun like "it" will be marked as NULL (implicit) when sentences are considered in isolation. However, in the broader context of the review, "it" might refer to a specific entity mentioned in previous sentences. This lack of context makes it difficult for models to correctly assign opinions to targets.

Implicit targets could be effectively handled by considering the full context of a review. Likewise, the identification of correct categories also benefits from full context analysis (Table 6 in appendix). While previous benchmark datasets from SemEval competitions annotate opinions in a sentence based on review context, this approach is not guaranteed to generalize to all implicit cases, potentially leading to inconsistencies and reduced performance in system predictions. Hence, the

adoption of review-level ABSA and the creation of corresponding datasets can significantly improve the effectiveness of sentiment analysis.

3.6 Inter-Annotator Agreement

The inter-annotator agreement, a key metric for evaluating the quality of a dataset, can be quantified in various ways such as the renowned Cohen's Kappa and Fleiss's Kappa scores, which are used for classification tasks. However, according to a study (Hripcsak and Rothschild, 2005), Kappa may not be the best fit for span-extraction annotation in textual data. The limitation arises from the requirement of Kappa to compute the number of negative cases, which is unidentifiable for spans as they constitute sequences of words without a predetermined quantity of items for annotation in a text.

Due to these limitations of Kappa metric, the F-measure, which doesn't necessitate the calculation of negative cases, is often more suitable for gauging inter-annotator agreement in span extraction annotation tasks such as target and opinion phrase extraction (Deléger et al., 2012). In particular, for datasets that encompass both classification (category and polarity) and span extraction tasks, such as the ASQP dataset, the F-measure can effectively serve as the chief method for inter-annotator agreement computation.

3.7 Is ABSA only for reviews?

A significant limitation in the current landscape of ABSA datasets is the predominant focus on reviews and specific domains like restaurants and e-commerce platforms. This focus on customer reviews has been fostered by the abundance and accessibility of data in these areas. For instance, review websites and e-commerce platforms readily provide vast amounts of customer feedback data. While these applications have been successful, the over-reliance on specific data types curtails the broader applicability of ABSA and its potential to provide diverse insights across various sectors.

Diverse ABSA datasets are vital for advancing research and applications beyond the confines of reviews. By expanding datasets to encompass domains such as healthcare, education, finance, legal, and social issues, ABSA models can be trained to tackle real-world challenges and address broader problem areas. For instance, healthcare datasets could facilitate sentiment analysis of patient feedback, improving healthcare service quality. ABSA can help identify investor sentiment towards different aspects of a company or its financial performance, potentially predicting stock market trends (Sinha et al., 2022; de França Costa and da Silva, 2018; Ong et al., 2023; Hridoy et al., 2021). Education datasets could uncover student sentiments towards specific aspects of the learning environment, leading to targeted improvements (Alassaf and Qamar, 2020).

Expanding datasets to these areas would enable the development of more robust and generalizable models while enhancing decision-making processes, public opinion analysis, and customer experiences in diverse industries.

3.8 Why evaluate only on a few datasets?

While ABSA research has numerous datasets at its disposal, the focus often falls on a select few benchmark datasets. This practice, while providing consistency and quality control, might unintentionally narrow model adaptability and heighten biases. Benchmark datasets may not cover the breadth of linguistic diversity in real-world scenarios, potentially causing models to falter with different or new data. Over-reliance on these resources can instigate model bias as models may echo the limitations of their training data, compromising performance in varied contexts.

Addressing these issues necessitates the use of an array of datasets for evaluation and the creation of a central platform for evaluating all ABSA models across available datasets. By offering researchers a unified platform for accessing diverse datasets and generating standardized metrics, we can advance understanding of model performance. This crucial step in ABSA research can stimulate more adaptive and resilient model development. The proposed platform would not only assist in overcoming the limitations of benchmark datasets but also inspire the creation of new, superior ones. While similar initiatives have been seen in the broader NLP community (Aguilar et al., 2020), ABSA-specific platforms are still a necessity. This progressive move calls for the combined effort of researchers and professionals alike to pave the way for more effective and fair sentiment analysis tools.

3.9 Baseline Study: ABSA Task Variations Across Datasets

One of the main questions that comes to mind is why we need new datasets for the same task(s). Let us look at a sample of datasets from Table 2 and analyze the performance of six different ABSA

subtasks from Table 1 using the unified generative paraphrasing framework (Zhang et al., 2021a). The main objective of this experiment is to assess and contrast the intricacy of various datasets concerning multiple tasks. Additionally, we endeavor to illustrate the inter-dependencies among these tasks, emphasizing the imperative nature of addressing these dependencies to enhance task performance on a given dataset.

We chose three types of joint tasks for our experiments: (a) tuple extraction (ASD, TSD, AOPE), (b) triplet extraction (ASTE, TASD), and (c) quadruple extraction (ASQP). The emphasis was placed on tasks involving joint extraction rather than single-element extraction. The rationale behind this decision stems from findings in Chebolu et al. (2021), where the authors have shown that joint models, designed to handle multiple interrelated elements simultaneously, consistently outperformed models that were specifically fine-tuned for extracting a single element. The results of these experiments can be found in Table 4.

Dataset	ASD	TSD	AOPE	ASTE	TASD	ASQP
ACOS_Rest	77.03	75.01	71.25	68.77	71.71	65.71
ACOS_Lap	53.15	74.05	75.21	74.43	48.21	57.30
DM_ASTE_Beauty	-	65.05	42.94	45.26	-	-
DM_ASTE_Electronics	-	65.18	43.13	44.76	-	-
DM_ASTE_Home	-	67.85	43.93	44.44	-	-
DM_ASTE_Fashion	-	66.27	42.70	44.62	-	-
SE-15	71.10	69.31	57.38	62.56	63.06	46.93
SE-16	76.97	75.60	65.80	71.70	71.97	57.93

Table 4: Baseline experimental results for six tasks from Table 1 using Paraphrase-T5 method from Zhang et al. (2021a). - indicates the dataset doesn't have relevant data for that experiment.

Two combinations in these six tasks essentially affect ASQP positively or negatively: 1) ASD + TSD + AOPE and 2) ASTE + TASD. For the ACOS_Rest dataset, the performance of the ASTE and TASD task together helped the ASQP task to detect the quadruples decently. On the other hand, in the ACOS_Lap dataset, TSD, AOPE, and ASTE have significantly better performance when compared to ASD and TASD. This shows that identifying aspect categories in the laptop dataset is difficult compared to targets and opinion phrases, resulting in the low performance of the ASQP task. On a similar note, the DM_ASTE datasets have consistent performance across all the domains. However, the performance of the AOPE task is far from the TSD task, indicating that it is challenging to comprehend the complex target and opinion phrase relationship compared to the target sentiment relationship. It negatively affected the ASTE performance, a combination of TSD+AOPE tasks.

Datasets with higher performance in lower-level tasks don't always guarantee high performance in comprehensive tasks like ASQP. For instance, in SE-16, despite ASTE and TASD being reasonably high, ASQP was lower, pointing to complexities in merging the triplets. Given the presence of opinion phrases in AOPE, ASTE, and ASQP, the clarity and diversity of opinion expressions in a dataset can be pivotal. If AOPE scores are significantly lower than TASD or ASD, it can explain why ASQP might suffer. This is clearly evident in the se-15 and se-16 datasets. Furthermore, ASD and TSD scores in rest_acos show the dataset has clear sentiment expressions for categories and targets. This clarity in individual extraction didn't fully translate to ASTE and ASQP, suggesting that extracting opinion words/phrases concurrently with targets and aspects might be challenging.

In conclusion, it is evident that each of these datasets presents distinct challenges, ranging from identifying opinion terms in the SE-15 dataset to focusing on aspect categories in the ACOS_lap dataset. Further research is required to fully harness the inter-dependencies that exist among these tasks utilizing the datasets. Additionally, it is crucial to investigate whether the sub-optimal performance observed is attributed to the intrinsic complexity of the task, the inherent ambiguity of the elements, or the inadequacies in their representation.

4 Conclusion

In this survey, we highlighted the urgent need for standardization and diversity in ABSA datasets to ensure comparability and enhance model robustness. Emphasizing the value of opinion phrases and the potential benefits of merging similar datasets, we advocated for the creation of resources in low-resource languages and shifting focus to review-level ABSA. Highlighting the importance of robust inter-annotator agreement measures, we called for an expansion of ABSA beyond reviews and addressed the limitations of evaluating ABSA methods on only a few datasets. Consequently, we proposed a common platform for ABSA evaluation to foster comprehensive and fair assessments, promoting the development of more effective sentiment analysis tools. With these initiatives, we can advance ABSA research and ensure its applicability across diverse linguistic and domain contexts.

5 Limitations

While this survey endeavored to encompass a broad array of ABSA datasets, it is conceivable that some may have been unintentionally missed. Further, due to space constraints, we were unable to delve deeply into datasets from other crucial domains, including legal, healthcare, and education sectors. Even though we performed a few experiments to understand the interplay of datasets and tasks, we did not explore the option of combining the model predictions from ASTE + TASD and ASD + TSD + AOPE tasks, similar to an ensemble approach, to solve for ASQP. However, it is a non-trivial task to combine the predictions from the triplets, as target and sentiment polarity are the only common elements to merge, which may not be unique for an aspect category and opinion phrase pairs to form a quadruple. That is, the same target an sentiment polarity pair can exist for different aspect categories and different opinion phrases.

Although we have provided accessible links for public datasets, acquisition of other datasets necessitates direct requests to the original authors or proprietors. We encourage the community's active involvement in contributing to a live version of this review to address these identified gaps. Another potential limitation pertains to the absence of a discussion on data ownership and copyrights issues related to datasets obtained via web scraping. Web scraping often contravenes the terms of service of the scraped websites, and disseminating scraped content might infringe on copyright laws, particularly if the scraped data have been substantially altered. Hence, it is crucial to exercise due diligence in ensuring compliance with legal and ethical guidelines when using such datasets.

Acknowledgements

We thank the reviewers for their constructive feedback, which was pivotal to the addition of a few experimental results and analysis to the paper. This research was supported by Adobe Gift funding to the University of Houston collaboration and by the National Science Foundation (NSF) under grant #1910192. The authors acknowledge the use of the Sabine Cluster and the advanced support from the Research Community Data Core at the University of Houston to carry out the research presented here.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Ahmed Ahmet and Tariq Abdullah. 2020. Recent trends and advances in deep learning-based sentiment analysis.

Md Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Aspect based sentiment analysis: Category detection and sentiment classification for hindi. pages 246–257.

M. Al-Smadi, O. Qawasmeh, B. Talafha, and M. Quwaider. 2015. Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In *2015 3rd International Conference on Future Internet of Things and Cloud*, pages 726–730.

Manar Alassaf and Ali Mustafa Qamar. 2020. Aspect-based sentiment analysis of arabic tweets in the education sector using a hybrid feature selection method. In *2020 14th International Conference on Innovations in Information Technology (IIT)*, pages 178–185.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive Opinion Summarization in Quantized Transformer Spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Jeremy Barnes, Patrik Lambert, and Toni Badia. 2016. Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623, Osaka, Japan. The COLING 2016 Organizing Committee.

Pierpaolo Basile, Danilo Croce, Valerio Basile, and Marco Polignano. 2018. *Overview of the EVALITA 2018 Aspect-based Sentiment Analysis task (ABSITA)*, pages 10–16.

Gianni Brauwers and Flavius Frasincar. 2021. A survey on aspect-based sentiment classification. *ACM Comput. Surv.* Just Accepted.

Caroline Brun and Vassilina Nikoulina. 2018. Aspect based sentiment analysis into the wild. pages 116–122.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2069–2079, Online. Association for Computational Linguistics.

Hongjie Cai, Nan Song, Zengzhi Wang, Qiming Xie, Qiankun Zhao, Ke Li, Siwei Wu, Shijie Liu, Jianfei Yu, and Rui Xia. 2023. Memd-absa: A multi-element multi-domain dataset for aspect-based sentiment analysis. *ArXiv*, abs/2306.16956.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2021. Exploring conditional text generation for aspect-based sentiment analysis. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 119–129, Shanghai, China. Association for Computational Lingustics.

Siva Uday Sampreeth Chebolu, Paolo Rosso, Sudipta Kar, and Thamar Solorio. 2022. Survey on aspect category detection. *ACM Comput. Surv.* Just Accepted.

Yew Ken Chia, Hui Chen, Wei Han, Guizhen Chen, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2023. Domain-expanded aste: Rethinking generalization in aspect sentiment triplet extraction.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ivan Cruz, Alexander Gelbukh, and Grigori Sidorov. 2014. Implicit aspect indicator extraction for aspect based opinion mining. *Int. J. Comput. Linguistics Appl.*, 5:135–152.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, page 121–124, New York, NY, USA. Association for Computing Machinery.

Dayan de França Costa and Nadia Felix Felipe da Silva. 2018. Inf-ufg at fiqa 2018 task 1: Predicting sentiments and aspects on financial tweets and news headlines. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1967–1971, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Lorenzo De Mattei, Graziella De Martino, Andrea Iovine, Alessio Miaschi, Marco Polignano, and Giulia Rambelli. 2020. Ate absita@ evalita2020: Overview of the aspect term extraction and aspect-based sentiment analysis task. *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org*.

Louise Deléger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnár, Laura Stoutenborough, Michal Kouril, Keith A. Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:144–53.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 231–240, New York, NY, USA. Association for Computing Machinery.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications*, 118:272 – 299.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *The 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling.

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

Felix Hamborg and Karsten Donnay. 2021. NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1663–1675, Online. Association for Computational Linguistics.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *CoRR*, abs/1602.01585.

Md. Nasir Hossain Hridoy, Mohammad Mohitul Islam, and Ayesha Khatun. 2021. Aspect based sentiment analysis for bangla newspaper headlines. In *2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pages 1–4.

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.

Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Minqing Hu and Bing Liu. 2004b. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Soufian Jebbara and Philipp Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2486–2495, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. pages 6281–6286.

Jason S. Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. 2010. The 2010 icwsm jdpa sentment corpus for the automotive domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*.

Roman Klinger and Philipp Cimiano. 2015. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 153–163, Beijing, China. Association for Computational Linguistics.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Naveen Kumar Laskari and Suresh Kumar Sanampudi. 2016. Aspect based sentiment analysis survey.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2018. A unified model for opinion target extraction and target sentiment prediction. *CoRR*, abs/1811.05082.

Yuncong Li, Fang Wang, Wenjun Zhang, Sheng hua Zhong, Cunxiang Yin, and Yancheng He. 2021. A more fine-grained aspect-sentiment-opinion triplet extraction task.

Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 1291–1297. AAAI Press.

Natalia V. Loukachevitch, Pavel Blinov, Evgeniy V. Kotelnikov, Yuliya Rubtsova, Vladimir Ivanov, and E. Tutubalina. 2015. Sentirueval: testing object-oriented sentiment analysis systems in russian.

A. Nazir, Y. Rao, L. Wu, and L. Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, pages 1–1.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Conference on Empirical Methods in Natural Language Processing*.

Keane Ong, Wihan van der Heever, Ranjan Satapathy, Gianmarco Mengaldo, and Erik Cambria. 2023. Finxabsa: Explainable finance through aspect-based sentiment analysis.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

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.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research.

Md. Atikur Rahman and Emon Kumar Dey. 2018. Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data*, 3(2).

Yashwanth Reddy Regatte, Rama Rohit Reddy Gangula, and Radhika Mamidi. 2020. Dataset creation and evaluation of aspect based sentiment analysis in Telugu, a low resource language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5017–5024, Marseille, France. European Language Resources Association.

Ahmed A. Sabeeh and Rupesh Kumar Dewang. 2018. Comparison, classification and survey of aspect based sentiment analysis.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan. The COLING 2016 Organizing Committee.

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. 2017. MUSST: A multilingual syntactic simplification tool. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 25–28, Tapei, Taiwan. Association for Computational Linguistics.

Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. 2018. Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1114, Brussels, Belgium. Association for Computational Linguistics.

K. Schouten and F. Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

A. Sethi and P. Bhattacharyya. 2017. Aspect based sentiment analysis-a survey.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. Sentfin 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 73(9):1314–1335.

Josef Steinberger, Tomáš Brychcín, and Michal Konkol. 2014. Aspect-level sentiment analysis in Czech. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Baltimore, Maryland. Association for Computational Linguistics.

2 Suresh and 3 Raghavi. 2016. A comprehensive survey on aspect based sentiment analysis.

Kim Nguyen Thi Thanh, Sieu Huynh Khai, Phuc Pham Huynh, Luong Phan Luc, Duc-Vu Nguyen, and Kiet Nguyen Van. 2021. Span detection for aspect-based sentiment analysis in vietnamese. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 318–328, Shanghai, China. Association for Computational Lingustics.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Darmstadt service review corpus.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hai Wan, Y. Yang, Jianfeng Du, Y. Liu, Kunxun Qi, and J. Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *AAAI*.

Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 618–626, New York, NY, USA. Association for Computing Machinery.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12, Berlin, Germany.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Ting Xu, Huiyun Yang, Zhen Wu, Jiaze Chen, Fei Zhao, and Xinyu Dai. 2023. Measuring your ASTE models in the wild: A diversified multi-domain dataset for aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2837–2853, Toronto, Canada. Association for Computational Linguistics.

Wei Xue, Wubai Zhou, Tao Li, and Qing Wang. 2017. MTNA: A neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 151–156, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2054, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges.

A Appendix

Domain		Aspect Terms				
	#Sent	Pos	Neg	Neu	Con	Total
Laptops	348	185	33	169	1	388
Mobiles	1141	600	210	578	28	1416
Tablets	1244	418	157	479	2	1056
Cameras	150	107	11	64	1	183
Headphones	43	20	8	19	0	47
Home appliances	84	10	0	34	0	44
Speakers	47	20	3	25	0	48
Smartwatches	330	47	22	149	2	220
Televisions	135	41	3	99	1	144
Mobile apps	229	98	20	46	0	164
Travels	776	273	19	98	0	390
Movies	890	167	83	154	5	409
Overall	5417	1986	569	1914	40	4509

Table 5: Hindi Multi-Domain Dataset Statistics

A.1 Dataset Info Tables

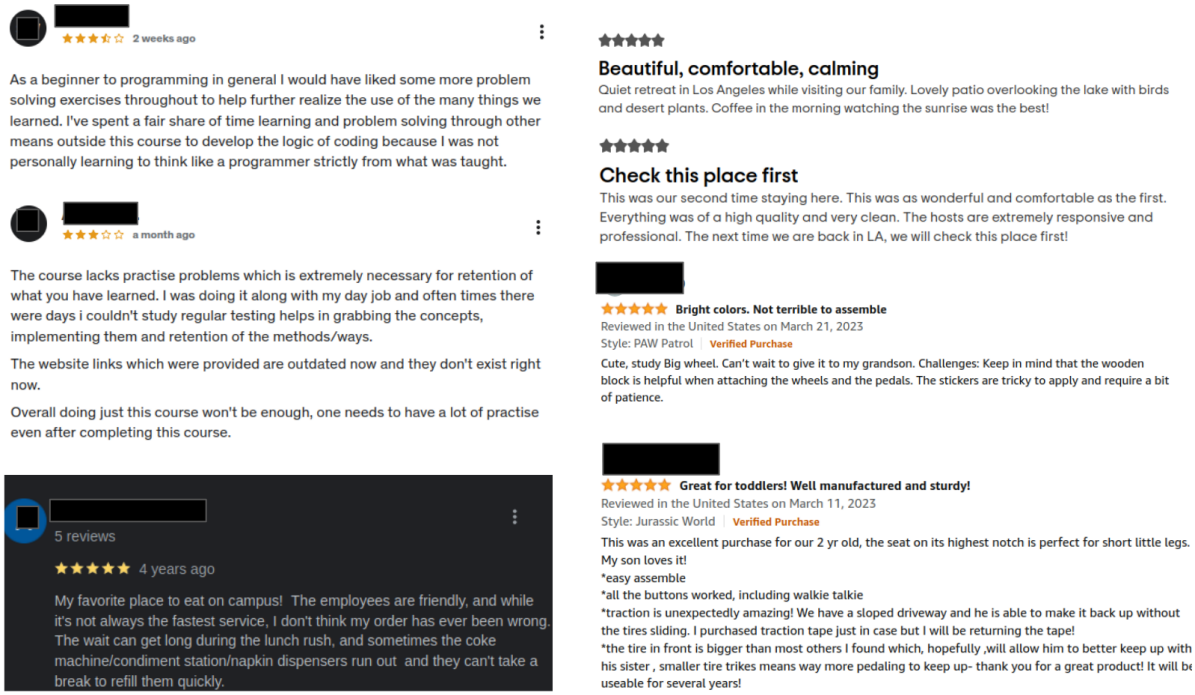

Figure 1: Multi sentence reviews from different platforms.

1 I picked 0 the asparagus , which turned out to be incredible and perfectly prepared . 0 It was steamed and tossed with garlic . 2 The steak was done to 1 my exact liking (medium rare) and was nice and juicy . 2 It 's served with 3 either a peppercorn sauce or red wine reduction , though 3 both were indistinguishable in taste , But 4 the service from the staff was extremely attentive and very friendly . 4 It was the highlight of our dinner .

Figure 2: Example of a full restaurant review with co-references resolved in each sentence using the context from previous sentences of the review.

Target	Aspect Cat.	Opinion Expr	Sent.
asparagus	FOOD#QLTY	incredible and perfectly prepared	Positive
	FOOD#STY_OP	NULL	Positive
steak	FOOD#QLTY	nice and juicy	Positive
	FOOD#STY_OP	served with either a peppercorn sauce or red wine reduction	Positive
	FOOD#STY_OP	indistinguishable	Negative
service from the staff	SERVICE#GEN	extremely attentive and very friendly	Positive
		highlight	Positive

Table 6: Transformation of Target, Aspect, Opinion, and Sentiment using the extended review level context, for Example in Figure 2.

Dataset Paper	Source	Domain	Lng	Stats				
				#Revs	#Sent	Target/Aspect Term		
						#pos	#neg	#neu
Customer Reviews (Hu and Liu, 2004b)	Amazon.com	Cameras	EN	34	346	172	31	-
		Phones	EN	41	546	252	86	-
		Electronics	EN	95	1716	514	331	-
		DVD-player	EN	99	739	195	235	-
		Cameras	EN	45	597	224	61	-
Customer Reviews (Ding et al., 2008)	Amazon.com	Cameras	EN	51	300	164	58	-
		Cameras	EN	1	229	121	27	-
		Routers	EN	31	312	186	79	-
		Phones	EN	49	594	310	148	-
		Routers	EN	48	577	154	64	-
		Ipod	EN	-	531	129	62	-
		Mp3 player	EN	50	1011	406	177	-
		Diaper Champ	EN	48	375	183	56	-
		Antivirus	EN	46	380	72	169	-
JDPA (Kessler et al., 2010)	J.D. Power and Associates	Cameras	EN	96	3527	-	-	-
		Cars	EN	111	4493	-	-	-
Darmstadt Service (Toprak et al., 2010)	tu-darmstadt.de	University reviews	EN	240	2786	-	-	-
	Darmstadt car service	Service Reviews	EN	234	6091	-	-	-
Twitter Comments (Dong et al., 2014)	www.twitter.com	Twitter comments	EN	-	6940	1734	1733	3371
SE-14 (Pontiki et al., 2014)	Amazon.com	Laptops	EN	-	3845	1331	994	629
Customer Reviews (Liu et al., 2015)	Amazon.com	Router	EN	-	879	185	122	-
		Computer	EN	-	581	270	84	-
		Speaker	EN	-	689	362	78	-
Hindi Multi-Domain (Akhtar et al., 2018)	News-, blog-, e-com. sites	Laptops & 11 others**	HN	-	5417	1986	569	1954
MAMS (Jiang et al., 2019)		Restaurants	EN	-	5297	4183	3418	6253
TOWE (Fan et al., 2019)	SE-14 Restaurants	Restaurants$^{\pm\pm}$	EN	-	2127	3508 targ.- op. pairs		
	SE-15 Restaurants	Restaurants$^{\pm\pm}$	EN	-	1079	1512 targ.- op. pairs		
	SE-16 Restaurants	Restaurants$^{\pm\pm}$	EN	-	1408	1969 targ.- op. pairs		
	SE-14 Laptops	Laptops$^{\pm\pm}$	EN	-	1501	2116 targ.- op. pairs		
ASTE-v1 (Peng et al., 2020)	SE-14 Restaurants	Restaurants$^{\pm\pm}$	EN	-	2119	3470	839	380
	SE-15 Restaurants	Restaurants$^{\pm\pm}$	EN	-	1059	1586	454	66
	SE-16 Restaurants	Restaurants$^{\pm\pm}$	EN	-	1372	2065	545	97
	SE-14 Laptops	Laptops$^{\pm\pm}$	EN	-	1487	1664	1012	365
ASTE-v2 (Xu et al., 2020)	SE-14 Restaurants	Restaurants$^{\pm\pm}$	EN	-	2119	2769	756	286
	SE-15 Restaurants	Restaurants$^{\pm\pm}$	EN	-	1057	1285	401	61
	SE-16 Restaurants	Restaurants$^{\pm\pm}$	EN	-	1372	1674	483	90
	SE-14 Laptops	Laptops$^{\pm\pm}$	EN	-	1487	1350	774	225
ABSITA-2020 (De Mattei et al., 2020)	e-Commerce platform	SD cards & 20 others$^{\mp\mp}$	IT	-	4363	7219	1577	-
ASOTE-v1 (Li et al., 2021)	SE-14 Restaurants	Restaurants$^{\pm\pm}$	EN	-	4828	2987	820	283
	SE-15 Restaurants	Restaurants$^{\pm\pm}$	EN	-	1741	1304	386	80
	SE-16 Restaurants	Restaurants$^{\pm\pm}$	EN	-	2355	1713	459	116
	SE-14 Laptops	Laptops$^{\pm\pm}$	EN	-	3021	1396	806	213
ASOTE-v2 (Li et al., 2021)	SE-14 Restaurants	Restaurants$^{\pm\pm}$	EN	-	6040	2987	820	283
	SE-15 Restaurants	Restaurants$^{\pm\pm}$	EN	-	2507	1304	386	80
	SE-16 Restaurants	Restaurants$^{\pm\pm}$	EN	-	3377	1713	459	116
	SE-14 Laptops	Laptops$^{\pm\pm}$	EN	-	4954	1392	806	213
N.MTSC (Hamborg and Donnay, 2021)	Financial News	News	EN	-	3021	1396	806	213
DM-ASTE (Xu et al., 2023)	Amazon.com (Ni et al., 2019)	Electronics$^{\pm\pm}$	EN	-	1994	5921	1316	306
		Fashion$^{\pm\pm}$	EN	-	1217	3578	918	215
		Beauty$^{\pm\pm}$	EN	-	766	2576	473	117
		Home$^{\pm\pm}$	EN	-	1503	4187	1091	180
		Book$^{\pm\pm}$	EN	-	484	1306	232	63
		Pet$^{\pm\pm}$	EN	-	507	1263	306	78
		Toy$^{\pm\pm}$	EN	-	527	1622	453	70
		Grocery$^{\pm\pm}$	EN	-	526	1597	285	80
Dom-Exp-ASTE (Chia et al., 2023)	SE-16 Restaurants	Restaurants$^{\pm\pm}$	EN	-	2942	-	-	-
	SE-16 Laptops	Laptops$^{\pm\pm}$	EN	-	1446	-	-	-
	Hotels (Angelidis et al., 2021)	Hotels$^{\pm\pm}$	EN	-	2136	-	-	-
	Cosmetics (He and McAuley, 2016)	Cosmetics$^{\pm\pm}$	EN	-	2468	-	-	-

Table 7: Publicly available ABSA datasets with no aspect category annotations. Lng: Language, #R: Number of Reviews, #S: Number of Sentences, #pos: Number of positive reviews, #neg: Number of negative reviews, #neu: Number of neutral reviews. EN: English, IT: Italian, HN: Hindi. N.MTSC: NewsMTSC Dataset. **: Full table in Appendix Table 5. $\pm\pm$: indicates that those datasets have the opinion phrase annotation along with other elements. $\mp\mp$: Irons, Water Bottles, Action Cameras, Razors, Phones, Printer Cartridges, Coffee Capsules, Backpacks, Hair Dryers, 2 different Movies, 2 different Books, Toy Phones, Car Light bulbs, Sweatshirts, Boots, Fans, Storage Chest, Shoe Cabinets, Personal Digital Assistants, TV streaming boxes/sticks. Note: The domain column has a downloadable link to each dataset.

Dataset Paper	Source	Domain	Lng	Stats							
				#Revs	#Sent	Target/Aspect Term			Aspect Category		
						#pos	#neg	#neu	#pos	#neg	#neu
TripAdvisor Hotels (Wang et al., 2011)	www.tripadvisor .com	Hotels**	EN	108K	1M	-	-	-	1.63M	153K	178K
SE-14 (Pontiki et al., 2014)	CSNY (Ganu et al., 2009)	Restaurants	EN	-	3841	2892	1001	829	2836	998	594
(Steinberger et al., 2014)	www.nejezto.cz	Restaurants	CZ	-	1244	679	725	403	521	569	246
HAAD (Al-Smadi et al., 2015)	LABR book reviews	Books	AR	-	2389	1376	1287	148	721	750	14
SE-15 (Pontiki et al., 2015)	CSNY (Ganu et al., 2009)	Restaurants	EN	350	2000	1326	496	73	1652	749	98
	Amazon.com	Laptops	EN	450	2500	-	-	-	1644	1094	185
SE-16 (Pontiki et al., 2016)	CSNY (Ganu et al., 2009)	Restaurants	EN	400	2286	1817	634	106	2268	953	145
	Amazon.com	Laptops	EN	530	3308	-	-	-	2118	1358	236
	-	Restaurants	ES	-	2691	1907	672	125	2675	948	168
	-	Restaurants	TR	339	1248	865	555	119	924	635	135
	-	Telecom	TR	-	3000	-	-	-	-	-	-
	-	Hotels	AR	2291	6029	7213	4003	824	7705	4556	852
	-	Restaurants	DU	400	2286	1016	546	145	1431	857	185
	-	Mobile Phones	DU	270	1697	-	-	-	1454	225	110
	(Loukachevitch et al., 2015)	Restaurants	RU	405	4699	3139	696	313	3973	1030	379
	-	Restaurants	FR	455	2429	1285	1061	289	1605	1646	233
	-	Mobile Phones	CH	200	9500	-	-	-	1168	794	-
	-	Digital Cameras	CH	200	8100	-	-	-	1153	587	-
SentiHood (Saeidi et al., 2016)	Yahoo Question Answering	Urban Neighborhoods	EN	-	5215	-	-	-	4305	1606	-
Customer Response (Yin et al., 2017)	www.beeradvocate .com	Beer Advocate**	EN	51K	552K	-	-	-	176K	8902	64K
	www.tripadvisor .com	Hotels**	EN	29K	375K	-	-	-	120K	66K	49.1K
GermEval-2017 (Wojatzki et al., 2017)	Internet crawling with search queries	Soc.Med., blogs, news	DE	-	27.8K	2802	12.5K	1459	2815	12.6K	13.9K
FiQA (de França Costa and da Silva, 2018)	Financial microblogs and headlines	Financial**	EN	1303	-	774	399	-	774	399	-
Bangla Rest., Cricket	FB, BBC, Daily Pronthom	Cricket	BG	-	2691	-	-	-	571	2157	266
(Rahman and Kumar Dey, 2018)	SE-14 Rest (Pontiki et al., 2014)	Restaurants	BG	-	1712	-	-	-	477	1226	371
ABSITA-2018 (Basile et al., 2018)	booking.com	Hotels	IT	-	9285	-	-	-	6893	5288	-
Foursquare (Brun and Nikoulina, 2018)	foursquare.com/	Restaurants	EN	-	1006	759	108	16	947	191	19
MAMS (Jiang et al., 2019)	CSNY (Ganu et al., 2009)	Restaurants	EN	-	3849	-	-	-	2415	2606	3858
Telugu Movies (Regatte et al., 2020)	123telugu, eenadu, samayam	Movies	TE	-	5027	2480	3251	1129	2480	3251	1129
Vietnam. Smartph. (Thanh et al., 2021)	e-commerce sites	Smartphones	VI	-	11122	-	-	-	21.7K	11.2K	2214
ASAP (Bu et al., 2021)	O2O e-commerce platforms	Restaurants	CH	46K	-	-	-	-	169K	35K	66K
ASQP (Zhang et al., 2021a)	SE-15 (Pontiki et al., 2015)	Restaurants	EN	-	1580	1407	489	68	1710	701	85
	SE-16 (Pontiki et al., 2016)	Restaurants	EN	-	2124	1811	613	110	2229	877	135
ACOS (Cai et al., 2021)	SE-16 (Pontiki et al., 2016)	Restaurants	EN	-	2287	2742	1518	259	3578	1879	316
	Amazon.com	Laptops	EN	-	4079	2004	663	114	2503	1007	151
MEMD-ABSA (Cai et al., 2023)	https://nijianmo.github.io/amazon/ (Ni et al., 2019)	Books	EN	986	2967	-	-	-	-	-	-
		Clothing	EN	928	2373	-	-	-	-	-	-
	https://www.yelp.com/dataset/download	Restaurant	EN	940	3526	-	-	-	-	-	-
	http://insideairbnb.com/get-the-data/	Hotel	EN	1029	5152	-	-	-	-	-	-
	Amazon.com	Laptops	EN	-	4076	-	-	-	-	-	-

Table 8: Publicly available ABSA datasets with aspect category annotations along with aspect terms/targets and polarity. Lng: Language, #Revs: Number of Reviews, #Sent: Number of Sentences, #pos: Number of positive reviews, #neg: Number of negative reviews, #neu: Number of neutral reviews. EN: English, AR: Arabic, IT: Italian, CZ: Czech, TU: Turkish, RU: Russian, FR: French, CH: Chinese, DE: German, BG: Bangla, TE: Telugu, VI: Vietnamese, ES: Spanish, DU: Dutch. **: indicates that the dataset has ratings converted to categorical sentiment polarities. Note: The domain column has a downloadable link to each dataset.