Artificial Intelligence Review (2026) 59:37
https://doi.org/10.1007/s10462-025-11437-x

Systematic review on aspect-based sentiment analysis in cross-domain

René Vieira Santin[1] · **Solange Oliveira Rezende**[1]

Received: 6 September 2024 / Accepted: 19 October 2025

Abstract
Aspect-level sentiment analysis is crucial for consumers and institutions, enabling them to monitor satisfaction regarding specific aspects of products and services through user reviews. Over time, various artificial intelligence techniques have been implemented with significant success. However, most of these techniques rely heavily on a substantial amount of labeled data. In this context, Cross-Domain Aspect-Based Sentiment Analysis emerges, leveraging data from source domains to enhance performance in the target domain. This systematic review contributes to this framework by outlining the primary solutions developed to tackle this challenge. It presents their data sources, compared methods, and the evolution of the main technologies adopted while identifying gaps that may inspire future research endeavors. A new classification of models is proposed here, considering the cross-domain approach. This fresh perspective aims to assist researchers in their quest for innovation, clarifying the context of their proposal and suggesting relevant comparisons with existing works.

Keywords Cross-domain · Sentiment analisys · ABSA · Aspect · Survey · Systematic review

1 Introduction

Users can express opinions on a wide range of topics in several ways on the Web. For instance, they may write a review of a hotel, or a restaurant, describe a financial asset, a medical treatment, share political opinions, or discuss a legal decision. These opinions can be expressed through different media of communication, including social networks, blogs, online platforms, and discussions in forums. Such interactions constitute an important

✉ René Vieira Santin
renevs@usp.br

Solange Oliveira Rezende
solange@icmc.usp.br

[1] ICMC, Universidade de São Paulo, Av. Trab. São Carlense, 400 - Centro, São Carlos, SP 13566–590, Brazil

Published online: 16 December 2025

source of information for other users, professionals, companies, and government (Marcacini et al. 2018; Zhang et al. 2022; Sinoara et al. 2021; Bashiri and Naderi 2024).

There are many applications for this data. Customer service teams can analyze user opinions as a primary source of positive and negative sentiment regarding aspects of products and services. This analysis can support strategies to track the sentiment over time, detect problems, make recommendations online, or even improve quality assurance. Financial investors can analyze the sentiment about stocks and their aspects, such as volatility and price, to make predictions. Regulators can observe an abnormal sentiment raised about a small stock that can be associated with illegal insider information. Politicians can monitor social media to gather information about the public opinion on positive and negative aspects about themselves and their opponents (Bashiri and Naderi 2024).

Due to the vast amount of information available, extracting knowledge from these diverse sources has become a challenging task when performed manually. Consequently, the employment of Artificial Intelligence (AI) opinion mining techniques has become a valuable solution (Marcacini et al. 2018; Zhang et al. 2022; Sinoara et al. 2021). Opinion mining or sentiment analysis seeks to extract the polarity of texts (e.g., positive, negative, or neutral) in an automated manner. It can be done at various levels: document, sentence, or aspect (Liu 2012; Machado and Pardo 2022). The document level is more general and does not always allow for the inference of exact sentiment. For example, a user may say that a restaurant has great food but terrible service. In this case, although the overall sentiment is neutral, they express a positive sentiment regarding the food aspect and a negative sentiment regarding the service aspect.

Aspect-Based Sentiment Analysis (ABSA) has received significant attention in recent years (Zhang et al. 2022). However, it has encountered several challenges. In particular, the models that have achieved state-of-the-art performance require a large amount of labeled data, which can be quite costly (Zhao et al. 2023; Zhang et al. 2022). An alternative approach is aspect extraction and classification in cross-domain settings. In this context, researchers leverage labeled data from one or more domains to apply in the desired domain, where labeled data are scarce. A key limitation of this approach is the assumption made by traditional sentiment classification methods that the training and testing data originate from the same independent and identically distributed (i.i.d.) distribution. In practice, however, domains often differ in their sentiment distributions and use distinct vocabularies to express similar opinions. As a result, models trained in one domain may lack prior knowledge of the terms, structures, and expressions commonly used in unseen domains (Zhao et al. 2021; Zhang et al. 2022).

Domain differences emerge in various ways, including variations in lexical distributions. For example, words such as *delicious*, *dessert*, and *waiter* are common in the restaurant domain, whereas words like *resolution* and *powerful* are more common in the laptop domain. However, these differences may involve more complex linguistic structures. In the financial domain, Du et al. (2024b) identify three particularly distinctive characteristics: (i) metaphorical expressions (e.g., *The market is riding a bull*, meaning that asset prices are rising); (ii) dependency on event direction (e.g., *profit raise* vs. *profit drop*, where the former conveys a positive and the latter a negative sentiment, despite both involving the positive term “profit”); and (iii) a mixture of qualitative and quantitative data. Beyond topical variation, a domain can also be defined by genre, style, level of formality, and other factors (Koehn and Knowles 2017).

Many works address the Cross-Domain ABSA; hence, a systematic review is critical so that new research can exploit existing information and compare their performance. Some secondary works synthesize publications on ABSA (Zhao et al. 2023; Zhang et al. 2022), partly focusing on cross-domains. However, to the best of our knowledge, no works have been found that systematically and specifically review this area. This systematic review was conducted specifically on the extraction and classification of aspect polarity in cross-domains.

Unlike previous ABSA reviews, the results of this study include general statistics and a list of the main journals and conferences that specifically address cross-domain ABSA. It also describes the main techniques employed in the literature and compares their usage over time. Furthermore, it introduces a mathematical intuition to guide the formulation of solutions, including an evolution of the formulation proposed by Gong et al. (2020) for cross-domain ABSA. Additionally, a new classification is proposed, which considers the adaptation focus and the primary techniques adopted in each method. The reviewed works are briefly presented, along with the metrics and data sources used for training and evaluating the proposed models. This study also provides a comprehensive overview of existing research on cross-domain ABSA, addressing subtasks such as aspect extraction, sentiment and category classification, levels of sentiment polarity granularity, requirements for labeled or unlabeled data in source and target domains, output formats generated by the models, and a list of available source codes for the reviewed methods.

This work is structured as follows: Sect. 2 presents the fundamental concepts for this review. Section 3 details the methodology and general results obtained for this review. Section 4 provides an overview of the reviewed works. Section 5 synthesizes the most common approaches, problems addressed, and models' performance. In Sect. 6, an overview of cross-domain ABSA is presented, analyzing the trends in the field and identifying some gaps. Finally, a conclusion about this work is presented in Sect. 7.

2 Fundamentals

This section presents the theoretical foundations of this review article. It begins by defining key concepts related to aspect extraction and classification in cross-domain scenarios. The standard categorization of Aspect-Based Sentiment Analysis (ABSA) works is introduced, based on the tasks they address, followed by an alternative categorization that considers the cross-domain perspective. Subsequently, the techniques frequently applied in the review studies are presented, offering a systematic understanding of the methods discussed in this work.

2.1 Definitions

ABSA encompasses several elements: aspect term, aspect category, opinion term, and sentiment polarity (Zhang et al. 2021). Consider the phrase "The lasagna is excellent." (Fig. 1). In this case, "lasagna" is the aspect term or simply aspect. Its category is "food". "Excellent" is the opinion term and has a positive sentiment polarity. All research involving any of these elements is considered ABSA. Thus, research that only performs sentiment analysis at the document or sentence level is not considered ABSA. Topic-based sentiment analysis is also

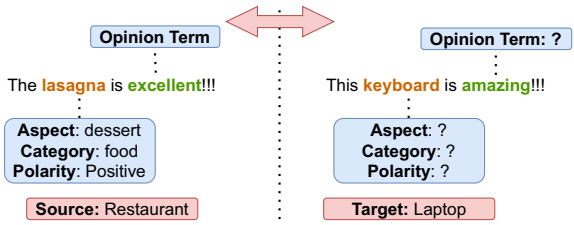

Fig. 1 Cross-domain aspect-based sentimen analysis overview

excluded from the definition. The topic model consists of groups of words built from documents that maintain a semantic relationship with each other (Li and Lei 2021). They do not have explicitly defined categories.

Cross-domain sentiment analysis requires the presence of at least two distinct domains. The term "domain" here will be used in a strict sense, excluding ABSA research in cross-lingual settings (Dang Van Thin and Nguyen 2023). Although it has some similarity to cross-domain, it deals with different situations. Similarly, cross-domain emotion analysis is excluded from the definition. Emotions have a more granular classification than polarity. One form of emotion classification categorizes sentiment into happy, excited, angry, sad, fearful, disgusted, and surprised (Peng et al. 2022; Zeng et al. 2023a). Emotions are short-term reactions to specific experiences, whereas sentiments represent more enduring evaluations that can influence future behavior and decision-making. Sentiment analysis, also referred to as opinion mining, seeks to identify the overall emotional tone in text to understand attitudes, opinions, or judgments expressed by individuals (Bashiri and Naderi 2024). Thus, reviews such as Wang et al. (2023) were not addressed in this paper.

Methods that merely apply general-purpose pre-trained models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019; Lopes et al. 2021; Pak and Gunal 2022), without any form of cross-domain evaluation or adaptation, are excluded from this review. These studies typically focus on general transfer learning or single-domain fine-tuning, involving only one domain. However, this study includes works that assess domain transferability—for example, those that train models on a source domain and test them on a different target domain—as such settings implicitly evaluate model robustness to domain shifts and inter-domain similarity (Zhao et al. 2022; van Berkum et al. 2022). Similarly, studies that rely solely on sentiment lexicons or ontologies are excluded (Liang et al. 2023; Guo et al. 2011), as they do not implement domain adaptation techniques per se, although these resources may serve as complementary components in cross-domain approaches.

Unsupervised models applied directly to the domain of interest are an alternative to sentiment analysis in unlabeled domains. These solutions only use the unlabeled desired domain, being able to infer aspects and sentiments using techniques such as syntactic relations and lexical dictionaries (Bagheri et al. 2013; Maharani et al. 2015). Since they do not use knowledge from a distinct domain to the one of interest, they diverge from the definition. Ideally, solutions using other domains can aggregate knowledge and bring better results. Publications using such models as benchmarks can be found (Li et al. 2012).

Thus, this review defines the scope of cross-domain ABSA as research and techniques involving the extraction of knowledge from one domain to apply it to aspect-based sentiment analysis tasks in another domain. The ABSA definition includes any works involving

one of the ABSA elements: category, aspect, opinion term, and polarity. In the considered definition, the term "cross-domain" includes works that involve at least two domains.

2.2 Cross-domain ABSA classification

We present two approaches for classifying ABSA-related studies. The first follows a standard perspective, organizing the works according to the specific tasks they address. The second is a novel classification scheme that considers how cross-domain adaptation is achieved.

2.2.1 Cross-domain ABSA classification considering the task

One way to classify cross-domain ABSA works is to examine the specific task they aim to address (Zhang et al. 2022). As described in Sect. 2.1, the following tasks are considered ABSA:

1. ACD (Aspect Category Detection)—detecting categories in a comment;
2. AE (Aspect Extraction)—extracting aspects;
3. ASC (Aspect Sentiment Classification)—classifying sentiment polarities of aspects or categories. It is also possible to find articles differentiating categories, Aspect-Category Sentiment Analysis (ACSA), and aspects, Aspect-Term Sentiment Analysis (ATSA); and
4. OTE (Opinion Term Extraction)—extracting opinion terms.

The following nomenclatures are used to combine these tasks:

5. AOPE (Aspect-Opinion Pair Extraction)—extracting aspects and opinion terms (AE + OTE);
6. E2E-ABSA (End-to-End ABSA)—extracting aspects and classifying their polarities (AE + ASC);
7. ACSA (Aspect Category Sentiment Analysis)—detecting categories and their sentiment polarities (ACD + ASC)[1];
8. ASTE (Aspect Sentiment Triplet Extraction)—triple extraction of aspect, opinion term, and sentiment polarity;
9. ACSD (Aspect-Category-Sentiment Detection)—triple extraction of aspect, category, and sentiment polarity; and
10. ASQP (Aspect Sentiment Quad Prediction)—quadruple extraction, including aspect, category, sentiment, and opinion term.

2.2.2 Cross-domain ABSA classification by cross-domain technique

Another way to categorize the works is related to cross-domain. The best solution to a cross-domain ABSA problem is the one that achieves the best score according to the desired

[1] Zhang et al. (2022) adopt *Aspect Category Sentiment Analysis* (ACSA) as an ACD+ASC task, different from what is mentioned in 2.2.1, where ACSA is exclusively the task of ASC for categories. Therefore, there is no unanimity in the use of this acronym.

metric regarding the target domain. The problem is mathematically formulated as follows, according to Gong et al. (2020):

$$f_t^* = \arg\min_{f \in H} \int_{(\mathrm{x},\mathrm{y})} \mathbb{P}_t(\mathrm{x},\mathrm{y}) L(\mathrm{x},\mathrm{y},f), \tag{1}$$

where:

- f is any function;
- H is the set of all possible functions;
- x is the feature vector and y is the corresponding label;
- t represents the target domain;
- $\mathbb{P}_t$ represents the probability of occurrence of the pair (x,y) in the target domain "t"; and
- L represents the loss function of the function f concerning the pairs (x,y).

The scientific literature (Sect. 3) presents two distinct approaches to using target domain data to achieve the objective defined in Eq. (1): *feature-based* and *instance-based*. The first one seeks to find a representation of the features invariant between the domains, called feature alignment. For example, syntactic relations of a sentence may indicate the presence of aspects and are invariant across domains. The second one seeks to reweight the distribution (x,y) of the target domain from the source domain. The term "instance-based" is not a consensus and can be referred to, for example, as "data-based" or "class-level". For instance, when synthetic reviews are generated for training a model, the weights of (x,y) are being redistributed in the loss function (Zhang et al. 2022; Ouyang and Shen 2023; Sun et al. 2023).

The way to solve Eq. (1) diverges in the studies found in this review. For example, Gong et al. (2020) mathematically simplify it by considering the use of data from the target domain to align the source domain:

$$\begin{aligned}
f_t^* &= \arg\min_{f \in H} \int_{(\mathrm{x},\mathrm{y})} \mathbb{P}_t(\mathrm{x},\mathrm{y}) L(\mathrm{x},\mathrm{y},f) \\
&= \arg\min_{f \in H} \int_{(\mathrm{x},\mathrm{y})} \frac{\mathbb{P}_t(\mathrm{x},\mathrm{y})}{\mathbb{P}_s(\mathrm{x},\mathrm{y})} \mathbb{P}_s(\mathrm{x},\mathrm{y}) L(\mathrm{x},\mathrm{y},f) \\
&\approx \arg\min_{f \in H} \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{\mathbb{P}_t(\mathrm{x}^s,\mathrm{y}^s)}{\mathbb{P}_s(\mathrm{x}^s,\mathrm{y}^s)} L(\mathrm{x}^s,\mathrm{y}^s,f) \\
&= \arg\min_{f \in H} \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{\mathbb{P}_t(\mathrm{y}^s \mid \mathrm{x}^s)}{\mathbb{P}_s(\mathrm{y}^s \mid \mathrm{x}^s)} \frac{\mathbb{P}_t(\mathrm{x}^s)}{\mathbb{P}_s(\mathrm{x}^s)} L(\mathrm{x}^s,\mathrm{y}^s,f).
\end{aligned} \tag{2}$$

In this case, note that a feature-based solution approximates the ratio $\frac{\mathbb{P}_t(\mathrm{y}^s|\mathrm{x}^s)}{\mathbb{P}_s(\mathrm{y}^s|\mathrm{x}^s)}$ to the value of one. On the other hand, an instance-based solution weighs the error function considering the proportion between the features in the source and target domains, i.e., $\frac{\mathbb{P}_t(\mathrm{x}^s)}{\mathbb{P}_s(\mathrm{x}^s)}$. For example, consider a model that establishes a syntactic relationship for aspect extraction very

common in the source domain that is also valid in the target domain. This is a feature-based solution. However, if this relationship is rare in the target domain, it will be useless.

Data from the source domain can be used to generate new examples in the target domain. Let D'_{syn} be the domain containing generated synthetic labeled examples. Gong et al. (2020)'s formula can be adapted as Eq. (3):

$$\begin{aligned} f_t^* &= \arg\min_{f \in H} \int_{(\mathrm{x},\mathrm{y})} \mathbb{P}_t(\mathrm{x},\mathrm{y}) L(\mathrm{x},\mathrm{y},f) \\ &= \arg\min_{f \in H} \int_{(\mathrm{x},\mathrm{y})} \frac{\mathbb{P}_t(\mathrm{x},\mathrm{y})}{\mathbb{P}_{\mathrm{syn}}(\mathrm{x},\mathrm{y})} \mathbb{P}_{\mathrm{syn}}(\mathrm{x},\mathrm{y}) L(\mathrm{x},\mathrm{y},f) \\ &\approx \arg\min_{f \in H} \frac{1}{N_{\mathrm{syn}}} \sum_{i=1}^{N_{\mathrm{syn}}} \frac{\mathbb{P}_t(\mathrm{x}^{\mathrm{syn}},\mathrm{y}^{\mathrm{syn}})}{\mathbb{P}_{\mathrm{syn}}(\mathrm{x}^{\mathrm{syn}},\mathrm{y}^{\mathrm{syn}})} L(\mathrm{x}^{\mathrm{syn}},\mathrm{y}^{\mathrm{syn}},f) \\ &= \arg\min_{f \in H} \frac{1}{N_{\mathrm{syn}}} \sum_{i=1}^{N_{\mathrm{syn}}} \frac{\mathbb{P}_t(\mathrm{y}^{\mathrm{syn}} \mid \mathrm{x}^{\mathrm{syn}})}{\mathbb{P}_{\mathrm{syn}}(\mathrm{y}^{\mathrm{syn}} \mid \mathrm{x}^{\mathrm{syn}})} \frac{\mathbb{P}_t(\mathrm{x}^{\mathrm{syn}})}{\mathbb{P}_{\mathrm{syn}}(\mathrm{x}^{\mathrm{syn}})} L(\mathrm{x}^{\mathrm{syn}},\mathrm{y}^{\mathrm{syn}},f). \end{aligned} \tag{3}$$

Models using synthetic data aim to approximate $\frac{\mathbb{P}_t(\mathrm{y}^{\mathrm{syn}}|\mathrm{x}^{\mathrm{syn}})}{\mathbb{P}_{\mathrm{syn}}(\mathrm{y}^{\mathrm{syn}}|\mathrm{x}^{\mathrm{syn}})}$ to one and ideally should consider the ratio $\frac{\mathbb{P}_t(\mathrm{x}^{\mathrm{syn}})}{\mathbb{P}_{\mathrm{syn}}(\mathrm{x}^{\mathrm{syn}})}$.

Cross-domain also includes cases where target domain data is labeled, and the goal is to generalize the characteristics of this data. Mathematically, consider the following formulation:

$$\begin{aligned} f_t^* &= \arg\min_{f \in H} \int_{(\mathrm{x},\mathrm{y})} \mathbb{P}_t(\mathrm{x},\mathrm{y}) L(\mathrm{x},\mathrm{y},f) \\ &= \arg\min_{f \in H} \int_{(\mathrm{x},\mathrm{y})} \mathbb{P}_t(\mathrm{y} \mid \mathrm{x}) \mathbb{P}_t(\mathrm{x}) L(\mathrm{x},\mathrm{y},f) \\ &\approx \arg\min_{f \in H} \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{P}_t(\mathrm{y}^t \mid \mathrm{x}^t) L(\mathrm{x}^t,\mathrm{y}^t,f). \end{aligned} \tag{4}$$

Some solutions aim to generalize target domain characteristics to increase the probability of $\mathbb{P}_t(\mathrm{y}^t \mid \mathrm{x}^t)$, minimizing the approximation error between steps 2 and 3. For example, models can extract the syntactic relationship from a training example, becoming more generic than models that consider individual words.

This generalization can be done implicitly by considering the source data. For example, consider a multitask model with two tasks, a main task, and an auxiliary task, mathematically represented in Eq. (5). If there is a strong relationship between the main and auxiliary tasks, the distributions become similar, $\mathbb{P}_t(\mathrm{x},\mathrm{y}) \sim \mathbb{P}_{\mathrm{aux}}(\mathrm{x},\mathrm{y}_{\mathrm{aux}})$, causing the search for the minimum value of the loss function of one task to contribute to the other.

$$
\begin{aligned}
f_t^* &= \underset{f, f_{\text{aux}} \in H}{\arg\min} \int_{(\mathrm{x},\mathrm{y})} \mathbb{P}_t(\mathrm{x},\mathrm{y}) L(\mathrm{x},\mathrm{y},f) + \int_{(\mathrm{x},\mathrm{y}_{\text{aux}})} \mathbb{P}_{\text{aux}}(\mathrm{x},\mathrm{y}_{\text{aux}}) L(\mathrm{x},\mathrm{y}_{\text{aux}},f_{\text{aux}}) \\
&\approx \underset{f, f_{\text{aux}} \in H}{\arg\min} \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{P}_t(\mathrm{y}^t \mid \mathrm{x}^t) L(\mathrm{x}^t,\mathrm{y}^t,f) + \sum_{i=1}^{N_{\text{aux}}} \mathbb{P}_{\text{aux}}(\mathrm{y}^t_{\text{aux}} \mid \mathrm{x}^t) L(\mathrm{x}^t,\mathrm{y}^t_{\text{aux}},f_{\text{aux}}) \\
&\approx \underset{f \in H}{\arg\min} \frac{1}{N_t + N_{\text{aux}}} \sum_{i=1}^{N_t+N_{\text{aux}}} \mathbb{P}_t(\mathrm{y}^t \mid \mathrm{x}^t) L(\mathrm{x}^t,\mathrm{y}^t,f).
\end{aligned}
\tag{5}
$$

Considering the different ways of conducting domain adaptation, a new hierarchical categorization of ABSA works is proposed here. The central idea is that this new categorization should assist researchers in understanding the techniques addressed and serve as a comparison for new models. At the first hierarchical level, the proposal divides ABSA models in domain adaptation into four groups (Fig. 2):

- **Independent**—models that use data from source domains but do not consider data from the target domain. For example, models that use syntactic relationships assuming they should work for any target domain. This type of solution is in accordance with Eq. (2).
- **Target → Source**—models that use target domain data to improve models using source

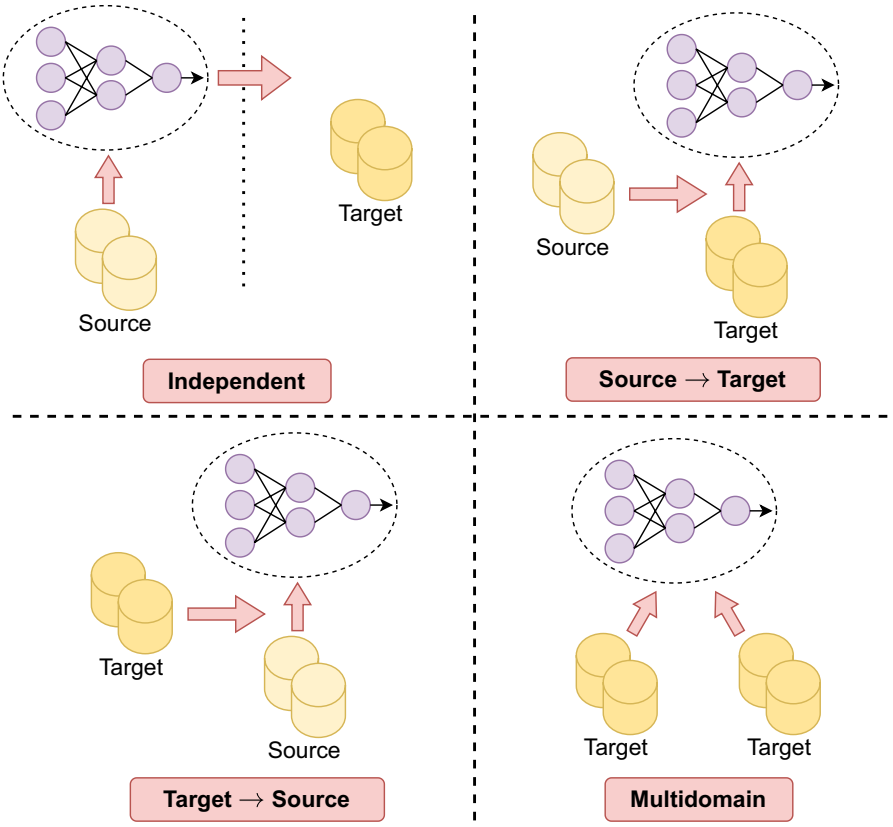

Fig. 2 Cross-domain ABSA classification by cross-domain technique

domain data, as in Eq. (2).
- **Source → Target**—models that use source data to improve models using target domain data. This can be done in accordance with Eqs. 3, 4, and 5.
- **Multidomain**—models where there is more than one target domain.

A second hierarchical level of the proposed classification, which accounts for the technique used for cross-domain adaptation, is introduced as part of the systematic literature review presented in Sect. 3.

2.3 Theoretical foundation: techniques

This subsection provides preliminary information on the techniques discussed in several works. These techniques include the use of: External Resources, Outputs from Generative, Classification, and Rule-Based Models, Conditional Random Fields (CRF), Latent Dirichlet Allocation (LDA), Gate Unit, Attention Mechanism, Transformer Architecture Models (BERT, BART, GPT, and T5), Graphs, Adversarial Networks, Measures of distance between probabilistic distributions, Pseudo-labeling, Teacher-Student Architecture, and Autoencoder.

2.3.1 External resources: lexical dictionaries, ontologies, semantic networks, and syntactic parses

Some articles used external resources to enhance their model. These include:

- *Lexical Dictionary*—a data structure containing information about words in a specific language. Specifically, sentiment lexicons were used, which are dictionaries describing the polarity of each word (Darwich et al. 2019).
- *Lexical Ontology*—an organized structure of knowledge about words. For example, WordNet is a database with nouns, adjectives, and adverbs grouped into sets of cognitive synonyms. These groups are interconnected by lexical and semantic-conceptual relations, representing relations such as synonyms, antonyms, hypernyms, meronyms, etc (Miller 1995).
- *Morphosyntactic Parser*—categorizes each word according to its morphological class: noun, verb, adverb, etc. It may also be referred to in the literature as *POS-Tagging* (Part-Of-Speech Tagging) (Marquez et al. 2000).
- *Syntactic Parser*—performs syntactic analysis of the sentence. The sentence can be structured into a tree or a dependency analysis can be made between words based on their syntactic role (Duran et al. 2023).
- *Semantic Ontology*—a knowledge graph that connects words and phrases by semantic relations. ConceptNet is an example of a semantic network. It contains relations such as "is a", "is used for", "made of", etc. (Speer et al. 2017).

2.3.2 Outputs from generative, classification, and rule-based models

Models that perform ABSA tasks in cross-domain settings can execute generative, classification, or rule-based tasks. A generative model produces tokens that represent the value for

the desired task. For example, a generative model that extracts aspects from the sentence "The pizza was great" will return the word "pizza". Conversely, a classification model that extracts aspects, assigns a label to each word identifying it as either an aspect or a non-aspect. Such classification models can label an aspect as "target", with the label "T", and a non-aspect as "outside", with the label "O". Models that classify each word or token into a label are specific to tasks that extract aspects or opinion terms from sentences. Finally, a rule-based model examines the presence of specific rules –morphological, syntactic, or semantic– within a sentence to identify the desired aspect or opinion term.

The classification of labels for each word can vary depending on the task to be performed. The task of extracting aspects or opinion terms may follow the Target–Outside (T–O) model. This classification approach has the limitation of being unable to distinguish between two consecutive aspects or opinion terms, so it is more common to adopt the pattern beginning of aspect (B), inside the aspect (I), and outside the aspect (O)—known as the BIO scheme (Begin–Inside–Outside). The BIO scheme can be extended depending on the task being performed. When the task involves extracting aspect and classifying polarity, models often classify labels as B-POS, B-NEG, B-NEU, I-POS, I-NEG, I-NEU, and O (Fig. 3a). For tasks that extract both aspects and opinion terms, the pattern commonly used is Begin-Aspect (BA), Begin-Opinion Term (BO), Inside-Aspect (IA), Inside-Opinion Term (IO), and None (N) (Fig. 3b). The None label is equivalent to the "Outside" (O) in the BIO scheme.

2.3.3 Conditional random fields (CRF)

Conditional Random Fields (CRF) is a conditional model that assigns a sequence of labels *Y* given a sequence of observations *X*. Each label depends on the observations and the other labels. However, the term CRF is commonly used in the literature as an abbreviation for *Linear Conditional Random Fields*, a simplification in which each label depends on the observations and the labels of its neighbors (Fig. 4) (Lafferty et al. 2001).

In a Linear CRF, the feature functions are defined according to the problem being modeled (Eq. 6). The weights (w_j) will be learned by adjusting the function in Eq. (7), which is done through maximum likelihood estimation. The function interacts over the N input vectors of X. The factor Z is a normalization factor.

$$F(X, y_{i-1}, y_i, i) = \sum_j w_j f_j(X, y_{i-1}, y_i, i) \tag{6}$$

$$\mathbb{P}(Y|X) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} F(X, y_{i-1}, y_i, i)\right) \tag{7}$$

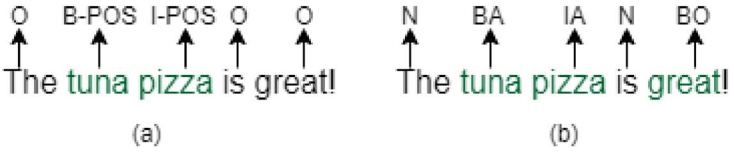

Fig. 3 BIO scheme. **a** represents the BIO + polarity scheme. **b** presents the BA-IA-BO-IO-N scheme for extracting aspects and opinion terms

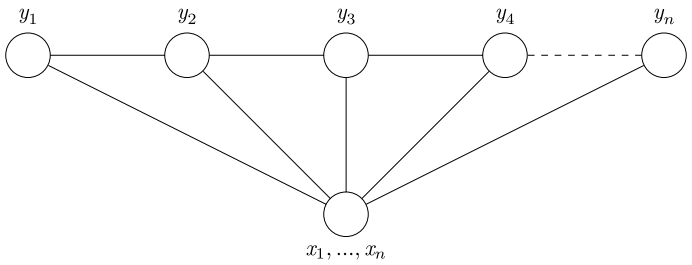

Fig. 4 Linear conditional random fields. Adapted from: Wallach (2004)

CRF is interesting for considering the prediction of its neighbors' labels. An aspect extraction scheme that labels using the BIO scheme (Begin–Inside–Outside), can penalize a sequence of labels O-I, for instance, since an aspect should begin with B.

Inference in this model is typically performed using the Viterbi algorithm. It processes the word labels one by one, calculating the probability of the label and the transition, and maintaining the path with the highest probability (Forney 1973).

2.3.4 Latent Dirichlet allocation (LDA)

LDA is an algorithm for generating documents topics (Blei et al. 2003). It assumes that documents are a random mixture of latent topics and that each topic is a distribution of words. Algorithm 1 illustrates document generation under this model. *Dir* represents the Dirichlet distribution. α is the parameter of the Dirichlet on the per-document topic distributions and β is the parameter of the Dirichlet on the per-topic word distribution.

1: Choose $N \sim \text{Poisson}(\xi)$
2: Choose $\theta \sim \text{Dir}(\alpha)$
3: **for** each of the N words w_n **do**
4: Choose a topic $z_n \sim \text{Multinomial}(\theta)$
5: Choose a word w_n from $\mathbb{P}(w_n|z_n, \beta)$, a multinomial probability conditioned on topic z_n,
6: **end for**

Algorithm 1 LDA

It is possible to perform the reverse process to discover topics, for example, using the Gibbs Sampling algorithm (Algorithm 2). To increase randomness, β is added to the value found in line 3, and α is added to the value in line 4.

1: Begin by assigning a topic to each word in each document
2: **for** each word in each document **do**
3: (i) Count how many times that word is assigned to each topic (excluding the current word).
4: (ii) Count how many words belong to each topic in the current document.
5: Multiply the two previous quantities (i and ii) and consider this as the probability for the new topic.
6: Randomly select a new topic considering the probabilities and assign it to the word.
7: **end for**

Algorithm 2 Gibbs Sampling

2.3.5 Gate unit

The *Gate Unit* originally emerged to enhance recurrent neural networks. Deep recurrent networks without this mechanism have difficulty capturing long-term dependencies due to vanishing and exploding gradient problems. By using information from both the previous and current steps, the mechanism informs how much emphasis to place on old information and how much to forget it (Cho et al. 2014; Hochreiter and Schmidhuber 1997).

The *minimal gate unit* is an example (Eq. 8). The vector f_t is the forget vector. Each element varies from 0 to 1, indicating how much information should be forgotten (Heck and Salem 2017).

$$\begin{aligned} \mathrm{f}_t &= \sigma(\mathrm{W}_f \mathrm{x}_t + \mathrm{U}_f \mathrm{h}_{t-1} + \mathrm{b}_f) \\ \hat{\mathrm{h}}_t &= \phi(\mathrm{W}_h \mathrm{x}_t + \mathrm{U}_h(\mathrm{f}_t \odot \mathrm{h}_{t-1}) + \mathrm{b}_h) \\ \mathrm{h}_t &= (1 - \mathrm{f}_t) \odot \mathrm{h}_{t-1} + \mathrm{f}_t \odot \hat{\mathrm{h}}_t, \end{aligned} \tag{8}$$

in which:

- $\odot$, σ, and ϕ: Hadamard product (element-wise multiplication), sigmoid function, and hyperbolic tangent function.
- x_t: input vector,
- h_t: output vector,
- $\hat{\mathrm{h}}_t$: candidate activation vector,
- f_t: forget vector, and
- W, U, and b: parameter matrices and vector.

2.3.6 Attention mechanism

Attention mechanisms in recurrent networks emerged in a translation model (Bahdanau et al. 2016). It models which words in the input data should receive more attention when translating each sentence segment, i.e., a context. The authors determine each word's context by running a bidirectional network centered on the desired word.

The attention mechanism created in the Transformer architecture, represented in Eq. (9), follows a similar principle. In this approach, the input vectors are of dimension d_k. These

vectors are projected into three subspaces: *Query*, *Key*, and *Value*. Then, the attention each word in the *Key* space receives for each Query vector is determined using an inner product. This value is rescaled by dividing by $\sqrt{d_k}$ to prevent a value explosion for huge vectors. The total attention is normalized by a Softmax function that weighs the projected vectors in the *Value* space (Vaswani et al. 2017).

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = Softmax\left(\frac{\text{QK}^T}{\sqrt{d_k}}\right)\text{V} \tag{9}$$

This mechanism weights the attention given to each word in a context. For example, if the model wishes to determine the sentiment of the aspect "service" in the sentence "The restaurant has great food, but terrible service", it should pay greater attention to the word "terrible" over the others.

2.3.7 BERT, GPT, BART, and T5

The models BERT (Devlin et al. 2019), GPT (Radford et al. 2019), BART (Lewis et al. 2020), and T5 (Raffel et al. 2020) are based on the *Transformers* architecture (Vaswani et al. 2017). It emerged as an improvement over recurrent networks and was initially applied for text translation. The *Transformers* architecture has two main components: an encoder and a decoder. The former transforms the data into an intermediate representation used by the latter to generate the translated text. This architecture employs the attention mechanism mentioned in Sect. 2.3.6.

Models based on this architecture utilize the encoder, the decoder, or both, and are typically pre-trained for one or more tasks using a substantial amount of text. This pretraining enables the fine-tuning of these models for specific tasks. BERT utilizes the encoder and is commonly used to create a contextualized representation of words for application in ABSA models. GPT utilizes the decoder and is occasionally used to generate synthetic sentences to train the cross-domain ABSA model. T5 and BART utilize both the encoder and the decoder and can also be used to generate synthetic examples.

In particular, BERT was originally trained for the Masked Language Model (MLM) and the Next Sentence Prediction (NSP) tasks. The input data consists of two sentences transformed into tokens as vectors (embeddings) and passed to the model in the format *<CLS>Sentence 1 <SEP>Sentence 2 <SEP>*, in which *<CLS>* and *<SEP>* are special tokens representing the start of the input and the separator of the sentences, respectively. Some input tokens are masked, and after processing the information, the model tries to predict these tokens in the output at their respective positions and determine whether one sentence follows the other using the output at the *<CLS>* token position.

2.3.8 Graphs and graph neural networks

A *graph* is formally defined as $G = (V, E, W)$, where V represents the set of nodes, E represents the set of edges connecting the nodes, and W represents the weights of these edges. Graphs can be heterogeneous networks, meaning they contain more than one node type. They can be intra-document or inter-document. An example of an intra-document is generating a syntactic dependency tree of words in the form of a graph. An inter-document

example is a graph that relates documents through common elements. For instance, Marcacini et al. (2018) linked each word of each document to its respective syntactic relation across different domains.

Related to graph theory, Graph Neural Networks (GNN) stand out. In particular, Graph Convolutional Networks (GCN) aggregate information from neighboring nodes to the current node. In the example presented in Fig. 5, node v_0 aggregates information from its neighbors, v_1, v_2, and v_3.

Equation 10 demonstrates how the information is encoded. The nodes are represented by information vectors h. The nearest neighbors (1-hop) of each node in the graph, represented by $\mathcal{N}(v_0)$, are projected by a matrix W and a bias $\mathbf{b}$ to be learned. Next, all the projections are summed. Finally, a non-linear operator, such as a ReLU function,[2] is applied. This process can be repeated k times to aggregate information from their more distant neighbors.

$$\mathrm{h}^{(k)} = \mathrm{ReLU}\left(\sum_{v \in \mathcal{N}(v_0)} \mathrm{W}^{(k)} \mathrm{h}_v^{(k-1)} + \mathrm{b}^{(k)}\right). \tag{10}$$

Figure 5 contains a self-loop at node v_0. Typically, it needs to be artificially constructed so that the node aggregates information from itself during the process.

Equation 10 can be rewritten using an adjacency matrix $\mathbf{A}$, which indicates "1" for the connection between nodes and "0" otherwise. First, we add the self-loop:

$$\hat{\mathrm{A}} = \mathrm{A} + \mathrm{I}, \tag{11}$$

where $\mathbf{I}$ is the identity matrix. Then, this matrix must be normalized based on the degree of each node, i.e., the number of neighbors. This is important because nodes with a higher degree will have contributions from a larger number of other nodes:

$$\hat{\mathrm{D}}^{-\frac{1}{2}} \hat{\mathrm{A}} \hat{\mathrm{D}}^{-\frac{1}{2}}, \tag{12}$$

where $\mathbf{D}$ is a diagonal matrix with the degree of each node from the matrix $\hat{\mathrm{A}}$. Finally, Eq. (10) becomes:

$$\mathrm{H}^{(k)} = \mathrm{ReLU}(\hat{\mathrm{D}}^{-\frac{1}{2}} \hat{\mathrm{A}} \hat{\mathrm{D}}^{-\frac{1}{2}} \mathrm{H}^{(k-1)} \mathrm{W}^{(k)} + \mathrm{b}^{(k)}) \tag{13}$$

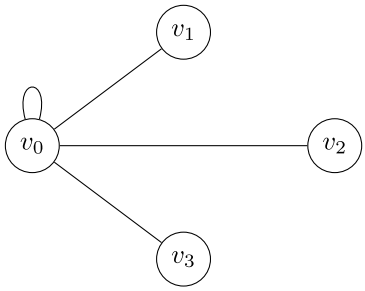

Fig. 5 Graph convolutional networks—GCN. Adapted from: Karagiannakos (2021)

[2] $\mathrm{ReLU}(x) = \max(0, x)$

A variation of GCN is the Relational Graph Convolutional Network (R-GCN). These networks use the representations from GCN to attempt predictions of the types of edges connecting the nodes. In doing so, the nodes are enriched with information from these connections.

Another approach used to obtain representations of vectors enriched by neighbors is to use the Node2Vec algorithm (Grover and Leskovec 2016). It starts from each node and performs random walks of a certain length. The nodes in this path are fed into a Skipgram network where, for each node, it tries to predict the presence of the others. By doing this, the algorithm generates a node representation with context.

2.3.9 Adversarial networks

Adversarial networks emerged as a generative model that trains two different models with distinct objectives. A generative model (G) attempts to learn the distribution of the training data, while another model, the discriminator (D), tries to discriminate whether the examples come from the training data or the generative model. In this model, the generative network aims to maximize the error of D, while D aims to minimize the error of its discriminator (Goodfellow et al. 2020).

This technique was adapted for domain adaptation. The features extracted to classify a model in a labeled domain should be invariant with respect to the target domain. Therefore, a commonly used approach in various papers is to use a domain classifier with an adversarial network. The features are extracted by maximizing the error of the domain discriminator. At the same time, the domain classifier must minimize its error to predict correctly whenever possible. The model uses these same features to perform the ABSA task in the labeled domain (Ganin et al. 2017).

The most common way to apply this technique in models is to attach a gradient reversal layer after generating the features in the neural network for the domain classification submodel. During the forward pass, this layer is null, meaning it does nothing. However, during the neural network's backpropagation phase in training, the gradient is multiplied by -1, causing the error to be maximized rather than minimized. Figure 6 illustrates this process. In this figure, the features extracted by G_f with parameters θ_f to be trained can be observed. These features are simultaneously passed to both the labeled task classifier G_y, with parameters θ_y, and to the domain classifier G_d. The gradient reversal layer is placed before the domain classifier. The loss functions of both submodels are trained simultaneously.

2.3.10 Distance measures between distributions

The distances between two distributions P and Q can be measured in various ways. The Maximum Mean Discrepancy (MMD) is one such measure. Let x and y be samples from P and Q respectively, the MMD can be described as:

$$MMD(P, Q, \mathcal{F}) = \sup_{f \in \mathcal{F}} |\mathbb{E}[f(\mathrm{x})] - \mathbb{E}[f(\mathrm{y})]|, \tag{14}$$

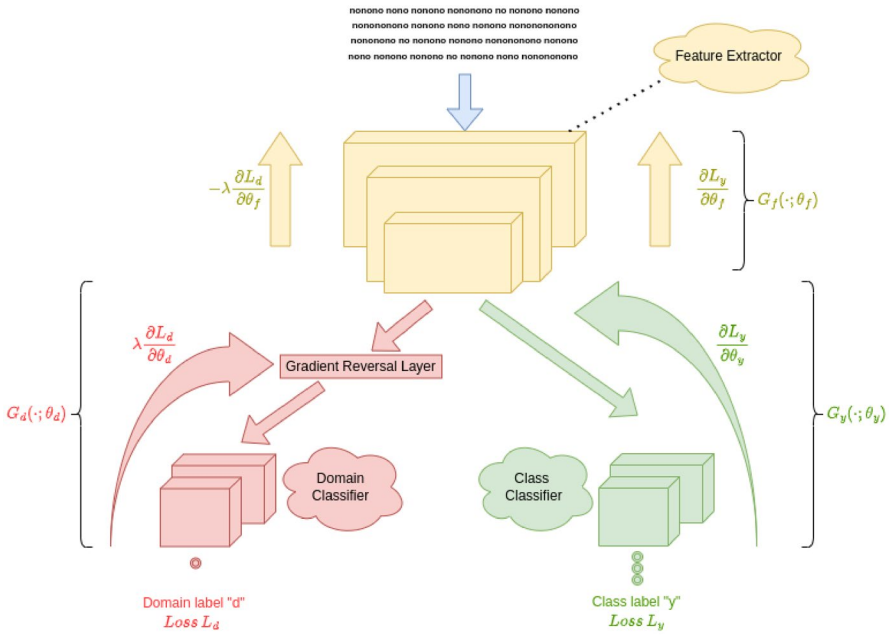

Fig. 6 Adversarial domain network. Adapted from Ganin et al. (2017)

where $\mathcal{F}$ is the space containing all continuous functions. Gretton et al. (2012) propose limiting $\mathcal{F}$ to the unit ball in the Reproducing Kernel Hilbert Space (RKHS), denoted by $\mathcal{H}$. The RKHS possesses a series of properties, such as a well-defined inner product.

This restricted MMD is called kernel-based MMD and is described as follows:

$$\begin{aligned} MMD^2(P,Q,\mathcal{F}) &= \sup_{||f||_{\mathcal{H}}\leq 1} |\mathbb{E}[f(x)] - \mathbb{E}[f(y)]|^2 \\ &= ||\mu_p - \mu_q||^2_{\mathcal{H}}, \end{aligned} \tag{15}$$

where μ_p and μ_q are the expected values of the kernel functions applied to the distributions, i.e., $\mathbb{E}_x[k(\cdot,\mathrm{x})]$ and $\mathbb{E}_y[k(\cdot,\mathrm{y})]$, respectively. They are called mean embeddings of the distributions. MMD is zero if and only if two distributions are identical when the kernel is characteristic[3] and bounded: $|k(\cdot,\cdot)| < +\infty$ (Gretton et al. 2012; Gao et al. 2021). Equation 15 can be rewritten as follows:

$$\begin{aligned} MMD^2(P,Q,\mathcal{F}) &= ||\mu_p - \mu_q||^2_{\mathcal{H}} \\ &= \mathbb{E}_{x,x'}[k(\mathrm{x},\mathrm{x}')] - 2\mathbb{E}_{x,y}[k(\mathrm{x},\mathrm{y})] + \mathbb{E}_{y,y'}[k(\mathrm{y},\mathrm{y}')] \end{aligned} \tag{16}$$

An estimate of Eq. (16) from a sample is obtained by:

[3] A *kernel* is characteristic if it satisfies certain conditions. For more details, see Gretton et al. (2012).

$$MMD(P,Q,\mathcal{F})^2 \approx \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(\mathrm{x}_i,\mathrm{x}_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(\mathrm{y}_i,\mathrm{y}_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(\mathrm{x}_i,\mathrm{y}_j) \tag{17}$$

When m and n are equal, a simpler version can be used (Gretton et al. 2012). Let $Z = (\mathrm{z}_1, ..., \mathrm{z}_m)$ be a sample of size m i.i.d. Where $z = (\mathrm{x}, \mathrm{y}) \sim p \times q$:

$$MMD(P,Q,\mathcal{F})^2 \approx \frac{1}{m(m-1)}\sum_{i\neq j}^{m} h(\mathrm{z}_i,\mathrm{z}_j), \quad h(\mathrm{z}_i,\mathrm{z}_j) = k(\mathrm{x}_i,\mathrm{x}_j) + k(\mathrm{y}_i,\mathrm{y}_j) - k(\mathrm{x}_i,\mathrm{y}_j) - k(\mathrm{x}_j,\mathrm{y}_i). \tag{18}$$

A commonly used kernel that satisfies the restrictions is the Gaussian kernel, also called the Radius Basis Function Kernel (RBF) (Zhang et al. 2023b). It maps an n-dimensional vector into an infinite-dimensional vector:

$$k(\mathrm{x},\mathrm{y}) = \exp\left(-\frac{||\mathrm{x}-\mathrm{y}||^2}{2\sigma^2}\right) \tag{19}$$

Another distance measure between distributions is the Kullback–Leibler (KL) divergence (Cao et al. 2021). It represents the distance between distributions based on entropy (H), where rare events bring more information. The entropy of a discrete distribution is defined as:

$$H(P) = -\sum_{i=1}^{n} P[x_i]\log_b x_i. \tag{20}$$

The KL divergence is:

$$KL(P||Q) = \sum_{i=1}^{n}(P[x_i]\log_b x_i - Q[x_i]\log_b x_i) \tag{21}$$

Equation 21 is not symmetric. Jensen–Shannon created a symmetric approximation adapted from KL.

2.3.11 Pseudo-labeling/teacher-student model

Pseudo-labeling is the process of iteratively adding unlabeled samples to the training data by labeling them with a weak model. The final model is trained using a combination of labeled and pseudo-labeled samples. The most common form of *pseudo-labeling* is a process that starts by training a regular classifier with labeled data. In the next step, predictions of unlabeled data with high confidence levels are added to the training set for a new itera-

tion. This process repeats several times until the classifier no longer finds confident predictions in the unlabeled set (Cascante-Bonilla et al. 2021).

Formally, the classifier model is adjusted for its labels. In this case, an example x will have a certain probability of having a label $\hat{\mathrm{y}}$, given by the highest probability. If $\hat{\mathrm{y}}$ has a high probability, it means the label would occur with that value for a large number of times considering the distribution $\mathbb{P}(\hat{\mathrm{y}})$. Since the data from $\mathbb{P}(\hat{\mathrm{y}})$ has been adjusted by the training data, there is a high chance that this predicted label will coincide with the actual value for a large number of times (Zheng et al. 2020).

When generating pseudo-labeled data, the model assumes the role of a teacher. When it trains with these data, it assumes the role of a student, taking the form of a teacher-student model. As the model trains with its own data, it is subject to a problem called confirmation bias, as the pseudo-labels may be entirely incorrect. Models must be adjusted to minimize this problem (Tarvainen and Valpola 2017).

2.3.12 Autoencoder

An autoencoder, or auto-associative, is a neural network that learns a compact representation of unlabeled data. It consists of an encoder, which generates an intermediate representation of the data, and a decoder, which regenerates the original data. This type of network is commonly used for data compression (Kramer 1991). Mathematically, let $\mathbf{x}$ be the input data, an encoder $\mathrm{h} = f(\mathrm{x})$ is applied, and a decoder $\tilde{x} = g(\mathrm{h})$ is applied, where h is the intermediate representation.

This type of network aims to learn the identity function. However, since the intermediate representation acts as a bottleneck, the model learns to compress relationships between the data. If the input data is completely random, this representation is infeasible. At the same time, if the intermediate layer has dimensions larger than or equal to the input data, the solution becomes trivial.

2.4 Final remarks on the theoretical foundations

This section presented definitions of cross-domain ABSA and introduced two approaches for classifying existing works. The first follows a traditional perspective, based on the specific ABSA task addressed. The second is a proposed classification approach that focuses on how cross-domain adaptation is performed. The main techniques applied in cross-domain ABSA studies were summarized.

The following section (Sect. 3) outlines the methodology and general results of the systematic review on cross-domain ABSA. It includes general statistics on the reviewed studies, data sources, and metrics.

3 Systematic review methodology and general results

This section describes the systematic review approach and presents the general results. The general results are composed of statistics, data sources, and metrics found.

3.1 Methodology

The methodology employed in the systematic review follows that of Scannavino et al. (2017). First, a protocol is defined to be followed, containing:

- General information—includes the title, description, and objectives;
- Research questions—objectives described in the form of questions;
- Study identification—describes the search criteria, list of sources, and strategy;
- Selection and evaluation of studies—includes items for the inclusion and exclusion of studies, strategies for selection and evaluation of quality;
- Presentation of results—data is synthesized and presented. It includes data extraction and summarization strategies, as well as their publications.

The following sections and subsections highlight the entire process and the results obtained, mapping the works in the area.

3.2 General information, objectives, and research questions

The title of the systematic review is "Aspect-Based Sentiment Analisys in Cross-Domain". The main objective is to list and map the works on aspect extraction and classification in cross-domain settings. The secondary objectives are (i) to identify how the respective models' tests are conducted, (ii) to map where related studies are being published, and (iii) to explore gaps in this type of solution. The research questions are enumerated as follows:

1. When and where have the studies been published?
2. What cross-domain ABSA techniques are currently being addressed?
3. What are the data sources, and which domains have been used to test the solution?
4. What are the main metrics used to evaluate the results in tests?
5. What are the main gaps and opportunities for future work?

3.3 Study identification

The following mechanisms and bibliographic databases were considered in the search: ACM digital library (https://dl.acm.org), ACL Anthology (https://aclanthology.org/), IEEE Xplore digital library (https://ieeexplore.ieee.org), Scopus (https://www.scopus.com), Scien ceDirect (https://www.sciencedirect.com), and Web of Science (https://www.webofscience.com/). Only works in English were searched, when the search mechanism or bibliographic database allowed it.

After a preliminary analysis, three main cores were considered to assemble the search string:

- Sentiment analysis;
- Aspect; and
- Cross-domain.

Sentiment analysis or opinion mining determines the type of work sought. The term aspect, also referred to as target, attribute, or feature (Zhang et al. 2022; Pak and Gunal 2022; Guo et al. 2013), describes at what level sentiment analysis should be considered. It is worth noting that "target" and "aspect" refer to the same concept in this work, unlike the work of some authors who differentiate between them (Pak and Gunal 2022).

A preliminary research was essential to analyze the terms "feature" and "attribute" in the search string. They are commonly used in the context of any machine learning model, and their inclusion could considerably increase the number of works to be analyzed. The exclusion of such terms does not affect the objective of this review, as they were less common and predominantly found in earlier studies, while currently the term "aspect" is almost consolidated (Zhang et al. 2022; Pak and Gunal 2022; Liu 2012). Nevertheless, in order to minimize the number of omitted works, it was decided to replace them with the compound term "product feature".

The *cross-domain* core has more undefined boundaries. In principle, any model that uses data from a domain different from the target domain employs cross-domain techniques. This includes any model that executes transfer learning, meaning any model trained using data from a distinct domain. For this core, terms such as *cross-domain*, *domain adaptation*, *BERT*, *fine-tuning*, *pre-trained*, *transfer learning*, and *domain-oriented* are searched. The search follows the definitions and restrictions established in Sect. 2.1.

The search string was created and adapted according to the search mechanisms and bibliographic databases. The following is the one used in Scopus. Others can be found in Appendix A. The string was validated to ensure it captures at least the articles from the "cross-domain" section of the review made by Zhang et al. (2022).

```
TITLE-ABS-KEY ( ( "aspect" OR "aspected" OR "product feature" OR "target" ) AND
( "sentiment analysis" OR "sentiment classification" OR "opinion" ) AND ( "domain
adaptation" OR "cross domain" OR "domain-invariant" OR "different domains" OR
"across domains" OR "transferable" OR "multiple domains" OR "other domains"
OR "target domain" OR "transfer learning" OR "fine tuning" OR "pre-trained" OR
"pre-training" OR "domain-oriented" ) ) AND ( LIMIT-TO ( DOCTYPE , "cp" ) OR
LIMIT-TO ( DOCTYPE , "ar" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) )
```

Despite the search string being quite comprehensive, it does not include articles that do not contain the mentioned words. Therefore, to mitigate the risk of disregarding any known work, it was decided to apply the backward snowballing technique in this review. The strategy consists of identifying relevant studies based on the bibliographic references of other articles. Specifically, the articles cited by at least two others during the comparison of model performance and that meet the selection criteria described in the following section were considered.

3.4 Selection and evaluation of studies

The sought studies necessarily need to be about aspect extraction and/or classification in cross-domains. Thus, they need to be tested in ABSA and use a domain different from the objective. After the searches, the following exclusion criteria were applied:

- The study is an older version of another study already considered;
- It is not primary research;
- The study lacks an abstract;
- It is a description of a course, editorial, lecture abstract, workshop, or tutorial;
- The study is a dissertation or thesis;
- The study is not accessible; and
- The study is not in English.

In particular, a study is accessible if there is a free means of obtaining it.

For the systematic review, the searches were conducted in three stages:

(a) Studies up to July 2023. A total of 2219 studies were found using the search strings. After removing duplicates and analyzing the title and abstract, 80 articles remained. After an in-depth reading, 59 articles remained.
(b) Studies up to December 2023. A total of 472 studies were found. Where possible, the search string was altered to exclude studies prior to July, already covered in the previous search phase. After removing duplicates and analyzing the title and abstract, 11 articles remained. After an in-depth reading, six articles remained.
(c) Studies up to early April 2024. A total of 644 studies were found. Where possible, the search string was altered to exclude studies prior to December 2023, already covered in the previous search phase. After removing duplicates and analyzing the title and abstract, eight articles remained. After an in-depth reading, four articles remained.

After conducting the initial searches, the backward snowballing technique was applied, resulting in the identification of five additional articles: Chen and Wan (2022), Chen and Qian (2021), Chen and Qian (2019), He et al. (2018), and Kiritchenko et al. (2014). As a result, the total number of selected articles increased to 74 (Fig. 7).

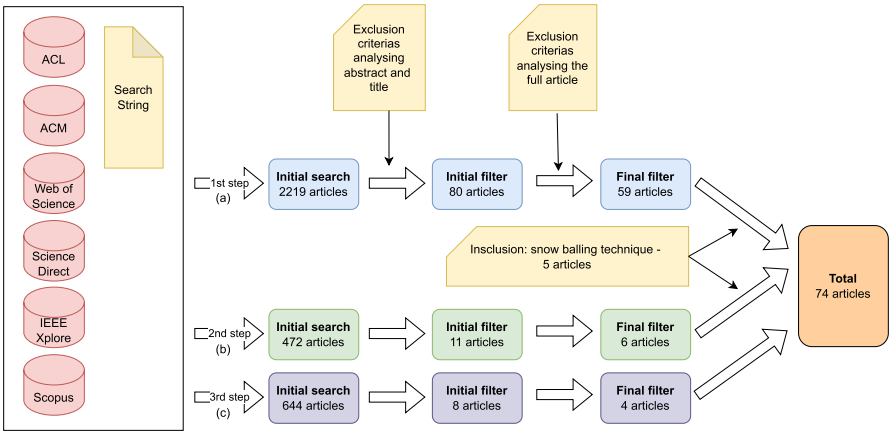

Fig. 7 Overview of the systematic review conducted

3.5 Some general statistics of the review results

This subsection presents general statistics to answer the first question of the systematic review: "When and where have the studies been published?". Figure 8 presents the number of articles published per year. The first one found in this review is from 2010. Since then, the production of publications has been growing. However, it is important to note that these numbers may underestimate reality. As described in Sect. 2.1, the term "aspect" has recently consolidated, which implies that the count may not reflect all existing articles. Still, the number of articles has been increasing, especially in recent years, when the term became more homogeneous. In 2023, 16 articles were published, indicating that the topic has received considerable attention recently.

A total of 30 articles were found published in 23 journals. Table 1 presents the journals where the articles were found, the respective quantities, and their classifications. The Journal Impact Factor (JIF) indicates the level of the journal and is calculated based on the Web of Science compilation. The Journal Citation Indicator (JCI) indicates the relative number of citations, with a value of "1" representing the average impact. JIF and JCI are indices from Clarivate (https://jcr.clarivate.com/). The Qualis CAPES index is an indirect qualification of intellectual production maintained by the Coordination for the Improvement of Higher Education Personnel (CAPES), a foundation linked to the Ministry of Education of Brazil (https://sucupira.capes.gov.br/sucupira/). The concepts range from A1 to A4, B1 to B4, and C, with A1 being the maximum and B4 the minimum. The C concept comprises journals that do not have any of the indicators used in categorization or do not meet good editorial practices. The journals are ordered in descending order according to the JIF. Only two journals were included with more than two publications: (i) Transactions on Affective

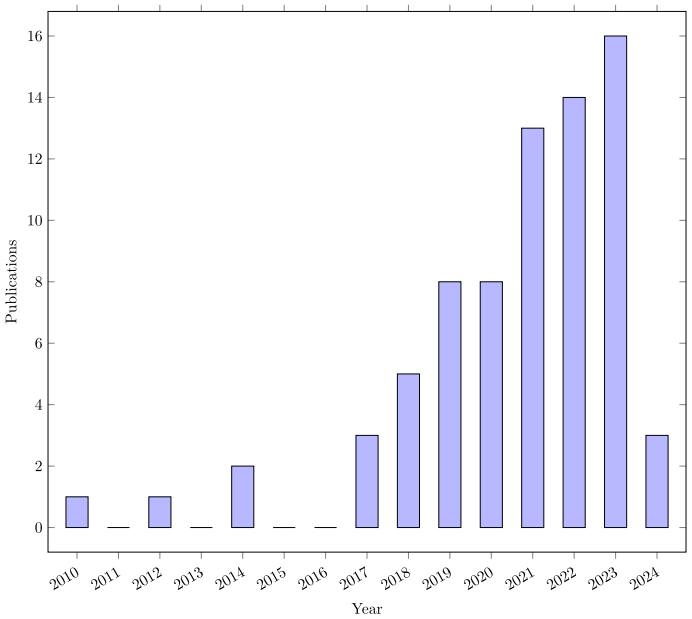

Fig. 8 Number of cross-domain ABSA articles published per year

Table 1 Qualitative classifications of journals containing cross-domain ABSA articles

Qty.	Journal	JIF	JCI	Qualis CAPES
4	IEEE—Transactions on Affective Computing	11.2	1.88	A1
1	IEEE—Transactions on Neural Networks and Learning Systems	10.4	2.63	A1
1	MIT—ACL-Computational Linguistics	9.3	3.52	A1
1	IEEE—Transactions on Knowledge and Data Engineering	8.9	1.83	A1
1	Elsevier—Knowledge-Based Systems	8.8	1.5	A1
1	Elsevier—Applied Soft Computing Journal	8.7	1.57	A1
1	Elsevier—Information Sciences	8.1	2.21	A1
1	Elsevier—Decision Support Systems	7.5	1.36	A1
1	IEEE—Transactions on Big Data	7.2	2.45	A3
1	Elsevier—Neurocomputing	6	1.02	A1
2	Springer—Neural Computing & Applications	6	0.94	A2
3	IEEE/ACM—Transactions on Audio, Speech, and Language Processing	5.4	1.53	A1
1	Taylor & Francis—Connection Science	5.3	0.94	–
2	IEEE—Access	3.9	0.89	A3
1	Springer—World Wide Web	3.7	0.9	A2
1	Springer—Journal of Supercomputing	3.3	0.72	A2
1	Hindawi—Computational Intelligence and Neuroscience	3.12	0.73	A1
1	MDPI—Electronics (Switzerland)	2.9	0.64	A4
1	Springer—Wireless Personal Communications	2.2	0.49	A4
1	Tech Science Press—Computer Systems Science and Engineering	2.2	0.58	–
1	ACM—ACM Transactions on Asian and Low-Resource Language Information Processing	2.0	0.33	–
1	Elsevier—Intelligent Systems with Applications	–	–	–
1	PIAP—Journal of Automation, Mobile Robotics and Intelligent Systems	–	–	–

"Qty." refers to the quantity of articles found. JIF and JCI are indices from Clarivate, with the former being the Journal Impact Factor and the latter a relative citation index. Qualis CAPES (CAPES) is an indirect qualification of intellectual production, ranging from A1 to A4, B1 to B4, and C

Computing, with four publications, and (ii) Transactions on Audio, Speech, and Language Processing, with three publications. This highlights the diversity of publications in this area.

A total of 44 articles were found, published in 21 conferences and workshops. Table 2 contains a list of the conferences and workshops, along with the respective quantity of articles. As there are no Clarivate indexes for conferences, the ordering was done by the Qualis CAPES index, from best to worst classification, followed by the respective quantity of articles in descending order. Unlike journals, a higher concentration of articles was observed in conferences: Annual Meeting of the Association for Computational Linguistics (ACL), Conference on Empirical Methods in Natural Language Processing (EMNLP),

Table 2 Qualitative classifications of conferences and workshops containing cross-domain ABSA articles

Qty.	Conferences/workshops	Qualis CAPES
8	ACL—Annual Meeting of the Association for Computational Linguistics (ACL)	A1
7	ACL—Conference on Empirical Methods in Natural Language Processing (EMNLP)	A1
4	AAAI—Conference on Artificial Intelligence (AAAI)	A1
3	ACL—North American Chapter of the Association for Computational Linguistics Annual Meeting (NAACL)	A1
2	ACL—International Conference on Computational Linguistics (COLING)	A1
2	ACM—International World Wide Web Conferences (WWW)	A1
2	ACM—International Conference on Information and Knowledge Management (CIKM)	A1
1	ACM—Conference Series on Recommender Systems (RecSys)	A1
1	ACL—International Workshop on Semantic Evaluation (SemEval)	A1
1	ACL—International Joint Conference on Natural Language Processing (IJCNLP)	A2
1	ACM—Symposium on Applied Computing (SAC)	A2
2	ACL—Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)	A3
2	ACM—IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)	A3
1	Hal Science—International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)	A4
1	IEEE—International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)	–
1	IEEE—International Conference on Computer Communication and Informatics (ICCCI)	–
1	IEEE—International Conference on Advanced Computer Science and information Systems (ICACSIS)	–
1	IEEE—Euro-Asia Conference on Frontiers of Computer Science and Information Technology (FCSIT)	–
1	ACL—International Conference on Natural Language and Speech Processing (ICNLSP)	–
1	IEEE—International Conference on Computing and Communication Technologies (RIVF)	–
1	Arxiv	–

"Qty." refers to the quantity of articles found. Qualis CAPES (CAPES) is an indirect qualification of intellectual production, ranging from A1 to A4, B1 to B4, and C

AAAI—Conference on Artificial Intelligence (AAAI), and North American Chapter of the Association for Computational Linguistics Annual Meeting (NAACL).

Journals, conferences, and workshops received good ratings in the Qualis CAPES index, with a few exceptions. This can be interpreted as a subject of interest for the top journals. Regarding the classifications, there are some caveats: (i) the absence of a JCR index or Qualis CAPES should not be interpreted solely as a means of low-quality publication; these evaluations occur periodically, and the publication may be too new to be included in the indexes, and (ii) a low classification by an index may be outdated and does not necessarily indicate that the article is of low quality.

3.6 Data sources

This subsection answers the research question "What are the data sources, and which domains have been used to test the solution?". The articles listed in this review reference several data sources from various domains for training and testing, labeled for one or more ABSA tasks or for document-level sentiment analysis. They are listed in tables with their respective statistics and aim to assist future research. Data sources that are subsets or aggregations of others listed here are not included. Also, sources that are not accessible are excluded.

Table 3 contains the found data sources, their URLs, and for which tasks the data is labeled, following the nomenclature described in Sect. 2.2.1: Document sentiment classification—Doc., aspect or category sentiment classification—ASC, aspect extraction—AE, opinion term extraction—OTE, and category detection—ACD. The SemEval 2014 and SemEval 2016 sources were the most cited data sources in model evaluations. Therefore, it is suggested that new research considers them for comparative purposes.

These datasets are described as follows, and all of them are in English, unless explicitly stated otherwise:

1. **SemEval 2014 – Task 4** (Pontiki et al. 2014) – The Aspect-Based Sentiment Analysis (ABSA) task was initially introduced in SemEval-2014 (Task 4), featuring sentence-level datasets in English. These datasets, derived from the restaurant review corpus by Ganu et al. (2009), were enriched with additional restaurant and laptop reviews manually annotated for aspect terms (e.g., "mouse", "pizza") and their associated sentiment polarities. For the restaurant domain, broader aspect categories (e.g., "food") were also provided. Although primarily composed of tweet-like sentences, the dataset includes instances from other sources as well.
2. **SemEval 2015 – Task 12** (Pontiki et al. 2015) – This task extended the 2014 formulation into a unified framework, where aspects, opinion targets, and sentiment polarities were jointly annotated as structured tuples at the sentence level. The annotations identify the exact textual span expressing the sentiment. In addition to the laptop and restaurant domains, the hotel domain was introduced, though only with test data.
3. **SemEval 2016 – Task 5** (Pontiki et al. 2016) – The 2016 edition broadened the task by incorporating text-level ABSA annotations, enabling sentiment analysis toward aspect categories over entire reviews. In this edition and in SemEval 2017, sentiment polarity was labeled on a five-point ordinal scale: `HIGHLYPOSITIVE`, `POSITIVE`, `NEUTRAL`, `NEGATIVE`, and `HIGHLYNEGATIVE`.

Table 3 List of databases used in cross-domain ABSA articles

Id.	Description	URL	Task
1	SemEval 2014—Task 4 (Pontiki et al. 2014)	https://alt.qcri.org/semeval2014/task4/	ABSA: AE / ASC / ACD
2	SemEval 2015—Task 12 (Pontiki et al. 2015)	https://alt.qcri.org/semeval2015/task12/	ABSA: AE / ASC / ACD
3	SemEval 2016—Task 5 (Pontiki et al. 2016)	http://alt.qcri.org/semeval2016/task5/	ABSA: AE / ASC / ACD
4	Twitter (Dong et al. 2014)	http://goo.gl/5Enpu7	ABSA: AE / ASC
5	Amazon *Review* (Ni et al. 2019)	https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/	Doc. / ABSA: ACD
6	Yelp Open Dataset (Yelp Inc 2024)	https://www.yelp.com/dataset	Doc. / ABSA: ACD
7	Reviews: camera, cellphone, mp3, and dvd (Hu and Liu 2004)	https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html	ABSA: AE / ASC
8	SemEval 2017—Task 4 (Rosenthal et al. 2017)	https://alt.qcri.org/semeval2017/task4	Doc.—Entidade
9	Reviews: computer, speaker, and router (Liu et al. 2015)	https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html	ABSA: AE / ASC
10	Web Services (Toprak et al. 2010)	https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2448	ABSA: AE / OTE / ASC
11	Reviews: camera, diaper, MP3, router, and antivirus (Ding et al. 2008)	https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html	ABSA: AE / ASC
12	Automobiles and cameras (Kessler et al. 2010)	https://verbs.colorado.edu/jdpacorpus/	ABSA: AE / ASC / OTE
13	Clothing, bags, and shoes (includes buyers, sellers, categories)—THGRL (Jiang et al. 2019b)	https://github.com/lzswangjian/THGRL	ABSA: ACD
14	MGAN YelpAspect (Li et al. 2019b)	https://github.com/hsqmlzno1/MGAN	ABSA: ACD / ASC (Cat.)
15	TSA-MD (various domains) (Toledo-Ronen et al. 2022)	https://github.com/IBM/yaso-tsa/tree/master/TSA-MD	ABSA: AE / ASC
16	MAM-for-ABSA—Restaurants (Jiang et al. 2019a)	https://github.com/siat-nlp/MAMS-for-ABSA	ABSA: ACD / AE / ASC
17	Twitter Sentiment Dataset (Hussein 2021)	https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset	Doc
18	OTE Annotations for SemEval 2014 (Wang et al. 2016)	https://github.com/happywwy/Recursive-Neural-Conditional-Random-Field	ABSA: OTE + polarity
19	USAGE (Klinger and Cimiano 2014)	http://dx.doi.org/10.4119/unibi/citec.2014.14	ABSA: AE / OTE / ASC
20	Reviews: books (Álvarez-López et al. 2018)	https://www.gti.uvigo.es/index.php/en/book-reviews-annotated-dataset-for-aspect-based-sentiment-analysis	ABSA: AE / ASC / ACD

4. **Twitter** (Dong et al. 2014) – This dataset consists of tweets related to celebrities, products, and companies. Tweets were retrieved using keywords such as "bill gates", "taylor swift", "xbox", "windows 7", and "google". Two annotators labeled randomly sampled tweets, and the dataset was balanced to include 25% negative, 50% neutral, and 25% positive instances.

5. **Amazon Reviews** (Ni et al. 2019) – This dataset comprises product reviews collected from the Amazon website,[4] including metadata such as rating, category, and technical specifications. However, the collection methodology is not clearly documented.
6. **Yelp Open Dataset** (Yelp Inc 2024) – Provided by Yelp for academic use, this dataset includes reviews across 29 product and service categories, such as automotive, books, movies, and musical instruments.
7. **Reviews: Camera, Cellphone, MP3, and DVD** (Hu and Liu 2004) – User reviews were collected from Amazon.com and C|net.com, with up to 100 reviews per product. Manual annotation was performed to identify opinionated sentences and extract corresponding aspect terms and sentiment polarities. Non-opinionated sentences were excluded.
8. **SemEval 2017 – Task 4** (Rosenthal et al. 2017) – This dataset was built upon the SemEval 2016 dataset by incorporating additional tweet-based instances to better capture the informal and noisy nature of social media language.
9. **Reviews: Computer, Speaker, and Router** (Liu et al. 2015) – An extension of the dataset by Hu and Liu (2004), this collection includes additional product categories and was annotated independently by two annotators. Agreement analysis was conducted to ensure annotation quality.
10. **Web Services** (Toprak et al. 2010) – Consumer reviews were collected from RateItAll and Epinions. A two-stage annotation scheme was applied: (i) sentence relevance to a predefined topic and whether the sentence expresses an evaluative judgment concerning that topic, and (ii) expression-level annotation for semantic orientation, intensity, and evaluative structure (e.g., opinion terms, targets, holders).
11. **Reviews: Camera, Diaper, MP3, Router, and Antivirus** (Ding et al. 2008) – Also an extension of Hu and Liu (2004), this dataset includes product reviews from amazon.com in additional categories.
12. **Automobiles and Cameras (JDPA Corpus)** (Kessler et al. 2010) – Composed of blog posts retrieved through web search queries, this corpus includes annotations for opinion expressions, coreference, meronymy, sentiment polarity, and modifiers (e.g., negators, intensifiers). Compared to formal news texts, the language is more expressive but remains grammatically structured.
13. **Clothing, Bags, and Shoes (THGRL)** (Jiang et al. 2019b) – This is a three domain-specific e-commerce dataset derived from Taobao, a large-scale consumer-to-consumer marketplace managed by Alibaba. Each dataset includes detailed records of user interactions, such as purchases, and user-generated product reviews. The aspect categories mentioned in the reviews were manually annotated.
14. **MGAN YelpAspect** (Li et al. 2019b) – Built on the Yelp recommendation dataset with three domains: Restaurant, Beauty Spa, and Hotel. Aspect categories and sentiment labels were first identified by an automated parser and then manually double-checked to correct incorrect annotations. Finally, the authors selected additional instances containing negation, contrastive structures, and questions to make the dataset more challenging.
15. **TSA-MD (various domains)** (Toledo-Ronen et al. 2022)—Domain-specific reviews were generated by crowd workers, who were allowed to select the topic of their review. Subsequently, the reviews were annotated for ABSA by instructing annotators to identify all sentiment-bearing targets within each sentence. The annotation was made by

[4] https://www.amazon.com/

crowd reviewers and not reviewed; thus, the resulting data contains a higher degree of noise and should be used only for model training, not as a benchmark.

16. **MAM-for-ABSA—Restaurants** (Jiang et al. 2019a)—This dataset is composed of sentences with at least two aspects with different sentiment polarities. It is based on the corpus by Ganu et al. (2009).
17. **Twitter Sentiment Dataset** (Hussein 2021)—Built from Twitter data, but limited information is available regarding its construction process.
18. **OTE Annotations for SemEval 2014** (Wang et al. 2016)—The authors manually annotated the SemEval 2014 dataset to identify the opinion target for the OTE task.
19. **USAGE** (Klinger and Cimiano 2014)—The Bielefeld University Sentiment Analysis Corpus for German and English (USAGE) is a corpus annotated from Amazon product reviews and has labels for aspect terms and their associated subjective expressions. The corpus employed an annotation methodology with details on inter-annotator agreement.
20. **Reviews: books** (Álvarez-López et al. 2018)—A total of 300 reviews were randomly selected from the INEX Amazon/LibraryThing Book Corpus (Koolen et al. 2016). Three expert annotators performed sentence-level ABSA annotations.

Table 4 summarizes the data sources listed in Table 3 and is independent of labels. Statistical data may not always be available, so they are omitted accordingly.

The following tables are specific according to their labels. Table 5 contains statistics of aspects and polarities for the ASC and AE tasks. Many of these databases are divided into training and testing sets, which facilitates algorithm comparison. Some also have a second division for validation data, which is not included in the table. In this latter case, the statistics of the validation data were added to the training data. The columns present the sentiment polarity, with emphasis on the "+−" column, which is used to show conflicting polarity in the aspect. Conflicting polarity differs from neutral polarity, as more than one sentiment is referenced to the aspect with different polarities.

Table 6 contains statistics of data sources labeled at the category level. The "#Cat" column indicates the number of categories, as a review can be associated with several categories. In some cases, they may have more than one hierarchical level. Similar to the aspects table, statistics related to polarities are listed.

A specific task in aspect extraction and classification is the extraction and classification of entities: personalities, locations, products, etc. Table 7 shows the SemEval 2017 data source and the statistics of labeled data for this task. In practical terms, this dataset can be considered as an aspect-based dataset. Table 8 presents the labeled datasets at the document level. As mentioned earlier, many of the datasets did not provide the statistics listed in this review and, therefore, were omitted.

Finally, Table 9 summarizes the accessible datasets listed in Table 3 used by each article. The listing of datasets is not exhaustive, indicating that the articles may use other datasets not listed. For instance, the article by Jakob and Gurevych (2010) does not utilize any of the listed datasets.

3.7 Metrics for evaluating model performance

This subsection addresses the research question "What are the main metrics used to evaluate the results in tests?". It describes quantitative and qualitative evaluations and lists the most

Table 4 General statistics of ABSA databases: total reviews and sentences

Id	Description	Reviews			Sentences		
		Train	Test	Total	Train	Test	Total
1	SemEval 2014				6086	1600	7686
	(a) Restaurant Base				3041	800	3841
	(b) Laptop Base				3045	800	3845
2	SemEval 2015	531	299	830	3054	1712	4766
	(a) Restaurant Base	254	96	350	1315	685	2000
	(b) Laptop Base	277	173	450	1739	761	2500
	(c) Hotel Base		30	30		266	266
3	SemEval 2016	800	170	970	4500	1484	5984
	(a) Restaurant Base	350	90	440	2000	676	2676
	(b) Laptop Base	450	80	530	2500	808	3308
4	Twitter	6248	692	6940			
5	Amazon			233 M			
6	Yelp			7 M			
7	Reviews						
	(a) Camera 1						597
	(b) Camera 2						346
	(c) Cellphone						546
	(d) MP3						1716
	(e) DVD						740
8	SemEval 2017						
	(a) Subtask A	50,333	1284	51,617	50,333	1284	51,617
	(b) Subtask B	20,508	6185	26,693			
	(c) Subtask C	30,623	6100	36,723			
9	Reviews						
	(a) Computer						531
	(b) Speaker						879
	(c) Router						689
10	Web Services				1492	747	2239
11	Reviews						
	(a) Camera 1						229
	(b) Camera 2						300
	(c) Diaper						375
	(d) Router 1						312
	(e) Ipod						531
	(f) Router 2						577
	(g) MP3						1011
	(h) Cellphone						554
	(i) Antivirus						380
12	Automobiles and Cameras			335			13,126
15	TSA-MD				952		952
17	Twitter Sentiment Dataset						162,980
18	OTE Annotations						7692
	(a) Restaurant Base				3044	800	3844
	(b) Laptop Base				3048	800	3848
19	USAGE			622			
20	Reviews: books			300	2219	758	2977

Table 5 Statistics of aspect databases with their respective polarities

Id	Description	Training					Testing					Total
		+	0	−	+−	Tot.	+	0	−	+−	Tot.	
1	SemEval 2014											
	(a) Restaurant	2164	633	805	91	3693	728	196	196	14	1134	4827
	(b) Laptop	987	460	866	45	2358	341	169	128	16	654	3012
2	SemEval 2015(1)											
	(a) Restaurant	1198	53	403		1654	454	45	346		845	2499
	(c) Hotel						243	12	84		339	339
3	SemEval 2016(1)											
	(a) Restaurant					2507					859	3366
	(b) Laptop					2909					808	3717
4	Twitter	1561	3127	1560		6248	173	346	173		692	6940
7	*Reviews:*											
	(a) Camera 1											237
	(b) Camera 2											174
	(c) Cellphone											302
	(d) MP3											674
	(e) DVD											296
9	*Reviews:*											
	(a) Computers											354
	(b) Speakers											307
	(c) Routers											440
16	MAM-for-ABSA											
	(a) Aspects	3783	5646	3089		12,518	400	607	329		1336	13,854
19	USAGE	3426(2)		1799(2)		8545						8545
20	Reviews: book	726	296	1663		2685	230	88	501		819	3504

(1) Approximate values. (2) Polarity is on the opinion term, meaning the total aspects will differ

commonly used models for baseline. The identification of quantitative evaluations encompasses the main metrics used in ABSA tasks that will be described below.

The *precision* metric measures the frequency with which the model classifies instances as true when they are indeed true. It is calculated as the ratio of True Positives (TP) to the sum of True Positives (TP) and False Positives (FP) (Eq. 22). The *recall* metric measures the model's coverage, i.e., the proportion of true instances identified by the model out of the total existing instances. Recall is calculated as the ratio of TP to the sum of TP and False Negatives (FN) (Eq. 23). These metrics have been addressed in different ways in cross domain ABSA models. For example, when using this metric in an Aspect Term Extraction model (ATE), it must be determined whether a match occurs only when all words belonging to an aspect are identified. For instance, in the sentence "Tuna pizza is excellent" "Tuna pizza" is the aspect. Some models consider a match only when all words are identified (*exact match*), while others consider a "partial match" if only *pizza* is identified as an aspect. Conversely, BIO models may evaluate precision and recall considering each word (*tag evaluation*) as an independent prediction.

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

Models that consider a match only as *exact match* should analyze the *golden labels* of the examples and consider as TP only those aspects that are identical to them. Conversely, false positives are all the aspects identified in the examples that do not belong to this set. False negatives are all the aspects from the *golden labels* that were not identified. For instance, a generator model with exact match that extracts the aspects *pizza* and *salad* from the example "I love the salad and the tuna pizza" will have the following sets:

- Golden labels = 'salad', 'tuna pizza',
- TP = 'salad',
- FP = 'pizza', and
- FN = 'tuna pizza'.

Thus, with this single example, the precision and recall are 0.5.

To assess the overall "correctness" of the model, accuracy (Eq. 24) can be used. This metric has been widely used to evaluate sentiment polarity classification models but is not used for aspect extraction and opinion term models. In sentences where there are multiple aspects, each consisting of more than one word, determining the True Negative (TN) can be challenging.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{24}$$

The main mean of evaluation in the articles is a quantitative comparison among the models using the F1 metric. F1 is a harmonic mean of precision and recall (Eq. 25). When the

Table 6 Statistics of category databases with their respective polarities

Id	Description	#Cat	Train					Test					Total
			+	0	−	+−	Tot.	+	0	−	+−	Tot.	
1	SemEval 2014	5	2179	500	839	163	3713	657	94	159	52	1025	
2	SemEval 2015[2]												
	(a) Restaurant Base	6 / 5[1]	1198	53	403		1654	454	45	346		845	2499
	(b) Laptop Base	22 / 9[1]	1103	106	765		1974	541	79	329		949	2923
	(c) Hotel Base	7 / 8[1]						243	12	84		339	339
3	SemEval 2016												
	(a) Restaurant Base	6 / 5[1]					1435					404	1839
	(b) *Laptop* Base	22 / 9[1]					2082					545	2627
14	MGAN *YelpAspect*												
	(a) Restaurant	68	46,315	45,815	16,020		108,150	5207	4944	1743		11,894	120,044
	(b) Beautyspa	45	45,770	42,580	16,023		104,643	5056	4793	1823		11,672	116,315
	(c) Hotel	44	40,775	36,901	20,864		98,540	4418	4048	2450		10,916	109,366
16	MAM-for-ABSA												
	(b) Categories	8	2170	3465	2343		7978	245	393	263		901	8879

[1]Includes categories and subcategories

[2] Approximate values

data can be classified into multiple classes (e.g., positive, negative, and neutral), there are two types of F1 scores: micro and macro. Micro-F1 considers the classes collectively. For example, precision is calculated based on the correct predictions regardless of the class. In these cases, micro-F1 and accuracy result in the same value. Macro-F1 is obtained by calculating the F1 for each class independently and then averaging these values. This is useful in situations where the classes are imbalanced. In this review, the most commonly used metric for sentiment polarity classification tasks of aspects and categories was Macro-F1. The most used metric for aspect and opinion term extraction models was Micro-F1. When authors apply exact match in aspect extraction, they consider two classes: *aspect* and *non-aspect*. However, since aspects can consist of more than one word, it is not possible to calculate the F1 for *non-aspects*, thus making it impossible to calculate Macro-F1. An exception for calculating Macro-F1 in aspect extraction arises where authors calculate the F1 for each class in the BIO scheme by considering each classified word individually. In this situation, the exact match is not calculated.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall} \tag{25}$$

Another metric used is the *Hamming Score*, which is comparable to accuracy but applicable to multilabel situations. This metric calculates the percentage of correct labels for each example and averages them. For example, consider a multilabel classifier that can classify an example into A, B, and C. If it classifies an example into A when it should be labeled as A and B, it will have correctly identified 2/3 of the labels. The advantage of this metric is that it considers "partial correctness," unlike accuracy, which only considers correct or incorrect. De Clercq et al. (2017) used the *Hamming Score* for category classification where each example can belong to more than one category. The calculation is thus based on the presence or absence of categories for each example. However, like accuracy, this metric has a bias. The number of categories per example is usually small, and a model that always denies the existence of categories tends to have a high value for this metric.

A metric used to evaluate model performance on datasets with imbalanced labels is the Kappa metric (Zorn et al. 2020), which assesses the model's reliability compared to random chance. Equations 26 and 27 describe the Kappa calculation for the confusion matrix described in Table 10. In this indicator, a value of 1 indicates high reliability, while a zero or negative value indicates that the model may be guessing randomly. Ramos and Fuentes (2023) and Zhang et al. (2023c) used this metric.

$$\text{ExpAccuracy} = \frac{P_1 \times R_1 + P_2 \times R_2 + P_3 \times R_3}{(P_1 + P_2 + P_3)^2} \tag{26}$$

$$Kappa = \frac{Accuracy - ExpAccuracy}{1 - ExpAccuracy} \tag{27}$$

The last quantitative metric considered in this review is Rouge-N, where N describes the number of N-grams. Rouge-1 and Rouge-2 refer to unigram and bigram comparisons, respectively. This metric evaluates the similarity between two texts and is used to asses text generation models. Both the generated text and the original text are divided into n-grams.

Table 7 Statistics of entity databases with their respective polarities

Id	Description	Train					Test					Total
		#Ent.	+	0	-	Tot.	#Ent.	+	0	-	Tot.	
8	SemEval 2017–											
	(b) Subtask B	373	14,951	1544	4013	20,508	125	2463		3722	6185	26,693
	(c) Subtask C	200	13,942	12,993	3697	30,632	125	2453	6194	3722	12,379	43,011

Precision and recall are then calculated, followed by the F1 score. This metric was used by the model of Yang et al. (2020), which generates text for the Cross Domain ABSA solution.

Table 11 presents the metrics used by the articles. It was not possible to determine for all articles whether the F1 score was macro or micro, and these cases are classified as "not specified". Models that evaluate the metric for each word or token are identified by the column "Tags evaluation". However, in most cases, the models were not explicit about whether the evaluation was performed for each predicted token/word, and this was inferred to the best of our efforts. Another observation is that the exact match inference for aspects and opinion terms requires the model to use F1-micro, as explained earlier. Some models opted to illustrate accuracy and F1 values per class. Finally, to mitigate the risk of model performance results being conditioned on chance, many authors considered initializing model variables and randomly splitting the data several times, obtaining an average of the results (Chen and Qian 2019, 2021; Chen and Wan 2022). This last point is not reflected in the table.

Certain models were more commonly used than others in performance comparison, and some factors could be identified as root causes. The first is that newer models will naturally be less cited. A second factor may be linked to simplicity, efficiency, and the pre-existence of source code. Some models are complex and take longer to execute, making their comparison difficult. It is also natural for models using a specific technique to compare with other models using similar ones. For example, it is natural for a text generation model to compare with others that also perform text generation. Figure 9 illustrates the relationships between model tests. For better visualization, models not used in the tests were excluded.[5] The arrowhead indicates which model was tested in a particular article.

The top three most cited models, according to Fig. 9, were the models by Wang and Pan (2018), with 12 citations; Jakob and Gurevych (2010), with 10 citations; and the model by Li et al. (2012), with eight citations. However, as explained, articles that were not referenced were removed. Upon including them, the top three most cited models become the models by Wang and Pan (2018), with 19 citations; Jakob and Gurevych (2010), with 14 citations; and the model by Li et al. (2019a), with 12 citations. On the other hand, articles that have compared the proposed model with a greater number of models tend to provide a more robust comparison. The top three models that were compared with a greater number of other models, according to Fig. 9, were the models by Chen and Qian (2021), with nine comparisons; Li et al. (2022), with eight comparisons; and Chen and Wan (2022), with seven comparisons. However, upon including the articles that were not referenced, this order changes to Shi et al. (2023), with 11 comparisons; Chen and Qian (2021), with nine comparisons; and Li et al. (2022), with eight comparisons. The citations and comparisons exclude any models not mentioned in this review.

A widespread test found in the articles was the ablation test. This test aims to demonstrate that a certain mechanism is contributing to the model's performance. The authors remove such a component or nullify its effect, showing a drop in the model's performance. It was also possible to find authors who conducted qualitative tests. For example, Yang et al. (2020) evaluated manually generated summaries and using metrics such as ROUGE-1. Yang et al. (2021) demonstrated the attention given to opinion words in the classification of an aspect.

An interesting analysis of the models is to demonstrate that the distributions of the source and target domains are close after the model processing to verify if the feature alignment

[5] By removing models that were not cited, the edges of the models they compared were also removed, leaving some orphan nodes in the graph, although they were cited.

Table 8 Statistics of review databases with their respective polarities

Id	Description	Train				Test				Total
		+	0	-	Tot.	+	0	-	Tot.	
8	SemEval 2017—(a) Subtask A[(3)]	19,902	22,591	7840	50,333	2375	5937	3972	12,284	62,617
17	*Twitter Sentiment Dataset*	72,249	55,212	35,509	162,980					162,980

mechanism is working. Wang and Pan (2019a) calculated the Maximum Mean Discrepancy (MMD) of the embeddings before and after the model processing, showing a much smaller distance in the second case. Another example was from the model of Yu et al. (2021), which generates sentences for training in the target domain. Using the same metric, the authors calculated the distance between the distribution of the vector representation of the generated sentences and the sentences of the target domain, showing them to be very close. The vector representation of the sentence was obtained using BERT-PT (Xu et al. 2019). There are also cases of authors who preferred to do this visually, such as Sun et al. (2023), who depicted the characteristics of these source and target domains in a t-SNE plot. This illustration demonstrates that there is no well-defined boundary after processing in their model.

4 Reviewed works

This section presents a summary of the works found in this review. Before presenting the summary, the studies are organized using the Cross-Domain ABSA Classification by the Cross-Domain Technique (Sect. 2.2.2). Moreover, for this classification, we present a second category hierarchy level. After that, all the works are described, grouping them into the four proposed categories (Independent, Target → Source, Source → Target, and Multidomain) and the appropriate subcategory. In this section and the next, we provide an answer to the second research question of the systematic review: "What cross-domain ABSA techniques are currently being addressed?"

4.1 Cross-domain ABSA classification results by cross-domain technique

The studies were categorized based on the cross-domain technique (Sect. 2.2.2). Figure 10 presents the percentage distribution of articles by group identified in this systematic review.

These groups are subdivided according to the techniques used for knowledge transfer between domains, generating the second hierarchical level (Fig. 11). Many works used more than one technique and could be classified into more than one category. However, for didactic purposes, the category in which the article placed more emphasis was sought. The next subsections detail each of these groups and their subgroups, describing the articles found.

4.2 ABSA: independent

This subsection describes the models that use source domain data to perform ABSA tasks, disregarding the target domain data. The articles can be grouped into three categories: *casual, using data*, or by *independent rules*, as described in the following sections. Figure 12 illustrates the percentage of articles in each subgroup.

Table 9 Accessible datasets used by the articles

Article	1.SemEval 2014	2.SemEval 2015	3.SemEval 2016	4. Twitter (Dong et al. 2014)	5.Amazon	6.Yelp	7.Reviews: (Hu and Liu 2004)	8.SemEval 2017	9.Reviews: (Liu et al. 2015)	10.Web Services	11. Reviews: (Ding et al. 2008)	12. Automobiles and cameras	13. Clothing, bags, and shoes	14.MGAN YelpAspect	15.TSA-MD	16.MAM-for-ABSA	17. Twitter (Hussein 2021)	18.OTE Annotations	19. USAGE	20. Reviews: books
Anand and Mampilli (2021)	x	x					x		x	x	x	x								
van Berkum et al. (2022)	x	x					x													x
Bhattacharjee et al. (2021)	x		x	x																
Cao et al. (2021)	x			x										x						
Chauhan et al. (2020)			x																	
Chen et al. (2024a)	x						x		x		x									
Chen et al. (2024b)	x																			
Chen and Qian (2019)	x				x	x														
Chen and Qian (2021)	x	x					x													
Chen and Qian (2022)	x	x	x				x			x										

Table 9 (continued)

Article	1.SemEval 2014	2.SemEval 2015	3.SemEval 2016	4. Twitter (Dong et al. 2014)	5.Amazon	6.Yelp	7.Reviews: (Hu and Liu 2004)	8.SemEval 2017	9.Reviews: (Liu et al. 2015)	10.Web Services	11. Reviews: (Ding et al. 2008)	12. Automobiles and cameras	13. Clothing, bags, and shoes	14.MGAN YelpAspect	15.TSA-MD	16.MAM-for-ABSA	17. Twitter (Hussein 2021)	18.OTE Annotations	19. USAGE	20. Reviews: books
Chen and Wan (2022)	x	x	x				x			x										
Chernyshevich (2014)	x																			
De Clercq et al. (2017)																				
Deng et al. (2023)	x	x	x				x			x										
Ding et al. (2017)	x	x								x		x								
Gong et al. (2020)	x	x	x				x			x										
He et al. (2018)	x	x	x		x	x														
Howard et al. (2022)	x	x					x											?		
Hu et al. (2019)	x			x																
Hu et al. (2022)	x	x												x						
Huang et al. (2023)	x					x														

Table 9 (continued)

Article	1.SemEval 2014	2.SemEval 2015	3.SemEval 2016	4. Twitter (Dong et al. 2014)	5.Amazon	6.Yelp	7.Reviews: (Hu and Liu 2004)	8.SemEval 2017	9.Reviews: (Liu et al. 2015)	10.Web Services	11. Reviews: (Ding et al. 2008)	12. Automobiles and cameras	13. Clothing, bags, and shoes	14.MGAN YelpAspect	15.TSA-MD	16.MAM-for-ABSA	17. Twitter (Hussein 2021)	18.OTE Annotations	19. USAGE	20. Reviews: books
Jakob and Gurevych (2010)																				
Jiang et al. (2019b)													x							
Jiang et al. (2024)	x						x													
Kan and Chang (2022)	x																			
Kannan et al. (2023)	x	x	x																	
Ke et al. (2021)	x						x		x		x									
Kiritchenko et al. (2014)	x				x	x														
Klein et al. (2022)	x	x																		
Knoester et al. (2023)			x		x															
Lark et al. (2018)		x	x																	

Table 9 (continued)

Article	1.SemEval 2014	2.SemEval 2015	3.SemEval 2016	4. Twitter (Dong et al. 2014)	5.Amazon	6.Yelp	7.Reviews: (Hu and Liu 2004)	8.SemEval 2017	9.Reviews: (Liu et al. 2015)	10.Web Services	11. Reviews: (Ding et al. 2008)	12. Automobiles and cameras	13. Clothing, bags, and shoes	14.MGAN YelpAspect	15.TSA-MD	16.MAM-for-ABSA	17. Twitter (Hussein 2021)	18.OTE Annotations	19. USAGE	20. Reviews: books
Lee et al. (2023)	x	x					x													
Li et al. (2012)																				
Li et al. (2019a)	x	x	x				x			x										
Li et al. (2019b)	x			x										x						
Li et al. (2022)	x	x					x													
Liang et al. (2022)	x	x					x			x										
Liu et al. (2021)	x															x				
Liu et al. (2023)	x	x	x	x												x				
Liu and Zhao (2022)																				
Majumder et al. (2022)	x																			
Marcacini et al. (2018)	x						x													
Ouyang and Shen (2023)	x	x	x				x			x										

Table 9 (continued)

Article	1.SemEval 2014	2.SemEval 2015	3.SemEval 2016	4. Twitter (Dong et al. 2014)	5.Amazon	6.Yelp	7.Reviews: (Hu and Liu 2004)	8.SemEval 2017	9.Reviews: (Liu et al. 2015)	10.Web Services	11. Reviews: (Ding et al. 2008)	12. Automobiles and cameras	13. Clothing, bags, and shoes	14.MGAN YelpAspect	15.TSA-MD	16.MAM-for-ABSA	17. Twitter (Hussein 2021)	18.OTE Annotations	19. USAGE	20. Reviews: books
Pak and Gunal (2022)	x						x													
Pastore et al. (2021)	x	x	x						x											
Pereg et al. (2020)	x	x					x											x		
Ramos and Fuentes (2023)	x																			
Ruskanda et al. (2019)			x				x													
Santos et al. (2021)	x	?					?		?		x									
Shi et al. (2023)	x	x	x				x			x	x									
Sun et al. (2023)	x	x	x				x			x										
Toledo-Ronen et al. (2022)	x														x	x				
Tran et al. (2021)			x																	

Table 9 (continued)

Article	1.SemEval 2014	2.SemEval 2015	3.SemEval 2016	4. Twitter (Dong et al. 2014)	5.Amazon	6.Yelp	7.Reviews: (Hu and Liu 2004)	8.SemEval 2017	9.Reviews: (Liu et al. 2015)	10.Web Services	11. Reviews: (Ding et al. 2008)	12. Automobiles and cameras	13. Clothing, bags, and shoes	14.MGAN YelpAspect	15.TSA-MD	16.MAM-for-ABSA	17. Twitter (Hussein 2021)	18.OTE Annotations	19. USAGE	20. Reviews: books
Wang and Pan (2018)	x	x					x													
Wang and Pan (2019a)	x	x					x													
Wang and Pan (2019b)	x	x					x													
Wang et al. (2018)																				
Xu et al. (2018)	x		x		x	x														
Xu et al. (2019)			x		x	x														
Xu et al. (2020)	x		x		x	x														
Xue et al. (2023)	x			x	x	x		x												
Yang et al. (2020)					x															
Yang et al. (2021)	x	x	x																	
Yang et al. (2023)	x				x	x											x			
Yu et al. (2021)	x	x	x				x			x										

Table 9 (continued)

Article	1.SemEval 2014	2.SemEval 2015	3.SemEval 2016	4. Twitter (Dong et al. 2014)	5.Amazon	6.Yelp	7.Reviews: (Hu and Liu 2004)	8.SemEval 2017	9.Reviews: (Liu et al. 2015)	10.Web Services	11. Reviews: (Ding et al. 2008)	12. Automobiles and cameras	13. Clothing, bags, and shoes	14.MGAN YelpAspect	15.TSA-MD	16.MAM-for-ABSA	17. Twitter (Hussein 2021)	18.OTE Annotations	19. USAGE	20. Reviews: books
Yu et al. (2023)	x	x	x				x			x										
Zeng et al. (2023b)	x	x	x				x			x										
Zhang et al. (2023a)	x			x																
Zhang et al. (2023b)	x			x										x						
Zhang et al. (2023c)			x																	
Zhao et al. (2024)	x	x	x																	
Zhao et al. (2022)	x																			
Zheng et al. (2020)	x		x		x	x														
Zhou et al. (2020)	x	x	x		x	x														
Zhou et al. (2021b)	x	x	x				x			x										

Table 9 (continued)

Article	1.SemEval 2014	2.SemEval 2015	3.SemEval 2016	4. Twitter (Dong et al. 2014)	5.Amazon	6.Yelp	7.Reviews: (Hu and Liu 2004)	8.SemEval 2017	9.Reviews: (Liu et al. 2015)	10.Web Services	11. Reviews: (Ding et al. 2008)	12. Automobiles and cameras	13. Clothing, bags, and shoes	14.MGAN YelpAspect	15.TSA-MD	16.MAM-for-ABSA	17. Twitter (Hussein 2021)	18.OTE Annotations	19. USAGE	20. Reviews: books
Zou and Wang (2025)	x	x																		

Table 10 Confusion Matrix for 3 Classes

True class	Predicted class			Total
	Class 1	Class 2	Class 3	
Class 1	A_{11}	A_{12}	A_{13}	R_1
Class 2	A_{21}	A_{22}	A_{23}	R_2
Class 3	A_{31}	A_{32}	A_{33}	R_3
Total	P_1	P_2	P_3	

4.2.1 Independent: casual

In this section, we describe articles whose models have been validated for cross-domain, even though no specific technique for this purpose has been implemented, meaning they are inherently robust models. Pastore et al. (2021) conduct aspect extraction training considering that some neural network models are robust enough to operate in domains different from the one they were trained on. They use the BIO scheme with BERT and variants of the Bi-LSTM network, employing more than one form of embedding and confirming that such models are relatively well-behaved in cross-domain. Following the same line, Zhao et al. (2022) demonstrate that BERT improves the ASC task (see Sect. 2.2.1) after being trained in a domain different from the target domain.

Kannan et al. (2023) propose the Aspect Based Sentiment Aware method (ABWE) for opinion term extraction from a sentence and an aspect. They aim to pass the aspect context so that the model pays greater attention to related words. First, a sentence with the desired aspect position is passed. Two LSTM networks process from the sentence extremes towards the aspect (inward). Additionally, two more networks start from the aspect towards the extremes (outward). The inward and outward embeddings are concatenated. In parallel, the sentence is passed through a traditional Bi-LSTM network, generating global embeddings. The global embeddings and those generated by in/outward are passed to a Bi-GRU network that labels the opinion terms in the BIO scheme.

Liu et al. (2021) propose independent models for category extraction and sentiment classification within the category. The authors solve the problem using text generation. They argue that traditional methods of pre-trained models (BERT/BART) that add an extra layer on top for classification are indirect, resulting in performance loss. First, because this extra layer does not exist in the original model and has not undergone pre-training. Second, because an extra word is passed in the input that is not in the form the model was originally trained on. Instead, the authors consider using these models for text generation. *Templates* are defined. For sentiment classification, they use: *The sentiment polarity of* $< a_i >$ *is* $< p_k >$. During inference, they input the review to detect sentiment. The output sentence with the highest probability is chosen by substituting $< a_i >$ for the category and $< p_k >$ for each of the three possibilities of sentiment polarity: positive, negative, or neutral. Two templates are created to detect categories: *The* $< a_i >$ *category is discussed / The* $< a_i >$ *category is not discussed*. They compare the output with the phrases "... *is discussed*" and "... *is not discussed*" filled in with each category. The phrase with the highest probability indicates the presence or absence of the category.

Table 11 Metrics used by the articles

Article	Task	Tags evaluation	Precision	Recall	Accuracy	Accuracy-by class	F1-macro	F1-micro	F1-unespecified	F1-by class	Exact match	Kappa	Hamming	Rouge
Anand and Mampilli (2021)	AE	x	x	x				x						
van Berkum et al. (2022)	ASC				x									
Bhattacharjee et al. (2021)	ASC						x			x				
Cao et al. (2021)	ASC				x		x							
Chauhan et al. (2020)	AE	x	x	x										
Chen et al. (2024a)	ASC				x		x							
Chen and Qian (2019)	ASC				x		x							
Chen and Qian (2021)	AE	?							x					
Chen and Qian (2022)	E2E-ABSA							x			x			
Chen and Wan (2022)	ASCAE							x			x			
Chernyshevich (2014)	AE	x						x						
De Clercq et al. (2017)	ACSD	x[1]	x[1]	x[1]									x[2]	
Deng et al. (2023)	ASTE							x			x			
Ding et al. (2017)	AE							x			x			
Gong et al. (2020)	E2E-ABSA							x			x			
He et al. (2018)	ASC				x		x							

Table 11 (continued)

Article	Task	Tags evaluation	Precision	Recall	Accuracy	Accuracy-by class	F1-macro	F1-micro	F1-unespecified	F1-by class	Exact match	Kappa	Hamming	Rouge
Howard et al. (2022)	AE								x					
Hu et al. (2019)	ASC				x		x							
Hu et al. (2022)	ASC				x		x							
Huang et al. (2023)	ASC				x				x					
Jakob and Gurevych (2010)	AE		x	x	x			x			x			
Jiang et al. (2019b)	ACD						x	x						
Jiang et al. (2024)	ASC								x					
Kan and Chang (2022)	ASC						x	x						
Kannan et al. (2023)	OTE		x	x					x					
Ke et al. (2021)	ASC				x		x							
Kiritchenko et al. (2014)	ACD/ASC/AE		x[3]	x[3]	x[4]				x[3]					
Klein et al. (2022)	AE							x			x			
Knoester et al. (2023)	ASC				x	x								
Lark et al. (2018)	ACD/AE		x	x					x					
Lee et al. (2023)	ASC				x									

Table 11 (continued)

Article	Task	Tags evaluation	Precision	Recall	Accuracy	Accuracy-by class	F1-macro	F1-micro	F1-unespecified	F1-by class	Exact match	Kappa	Hamming	Rouge
Li et al. (2012)	AOPE		x	x					x					
Li et al. (2019a)	E2E-ABSA							x		x				
Li et al. (2019b)	ASC				x		x							
Li et al. (2022)	AOPE							x						
Liang et al. (2022)	AE							x			x			
Liu et al. (2021)	ACSA		x[3]	x[3]	x[4]			x[3]						
Liu et al. (2023)	ASC				x		x							
Liu and Zhao (2022)	ASC				x			x						
Majumder et al. (2022)	ASC						x							
Marcacini et al. (2018)	AE	x	x	x				x						
Ouyang and Shen (2023)	AE/ E2E-ABSA							x			x			
Pak and Gunal (2022)	AE								x					
Pastore et al. (2021)	AE								x					
Pereg et al. (2020)	AOPE							x			x			
Ramos and Fuentes (2023)	ASC				x		x					x		

Table 11 (continued)

Article	Task	Tags evaluation	Precision	Recall	Accuracy	Accuracy-by class	F1-macro	F1-micro	F1-unespecified	F1-by class	Exact match	Kappa	Hamming	Rouge
Ruskanda et al. (2019)	AOPE		x	x				x			x			
Santos et al. (2021)	AE	x			x		x							
Shi et al. (2023)	AE							x						
Sun et al. (2023)	AE							x			x			
Toledo-Ronen et al. (2022)	E2E-ABSA		x	x				x			x			
Tran et al. (2021)	E2E-ABSA	x	x	x	x				x					
Wang and Pan (2018)	AOPE							x			x			
Wang and Pan (2019a)	AOPE							x			x			
Wang and Pan (2019b)	AOPE							x			x			
Wang et al. (2018)	ASTE	x	x	x					x					
Xu et al. (2018)	AE								x					
Xu et al. (2019)	ASC/AE						x[4]	x[1]						
Xu et al. (2020)	E2E-ABSA				x[4]		x[4]	x[1]						
Xue et al. (2023)	ASC				x[4]		x[4]							
Yang et al. (2020)	E2E-ABSA													x

Table 11 (continued)

Article	Task	Tags evaluation	Precision	Recall	Accuracy	Accuracy-by class	F1-macro	F1-micro	F1-unespecified	F1-by class	Exact match	Kappa	Hamming	Rouge
Yang et al. (2021)	ASC				x				x					
Yang et al. (2023)	ASC/AE				x[4]		x[4]							
Yu et al. (2021)	E2E-ABSA							x			x			
Yu et al. (2023)	E2E-ABSA							x			x			
Zeng et al. (2023b)	E2E-ABSA							x						
Zhang et al. (2023a)	ASC								x					
Zhang et al. (2023b)	ASC				x									
Zhang et al. (2023c)	ASC			x	x				x			x		
Zhao et al. (2024)	ASC							x						
Zhao et al. (2022)	ASC			x										
Zheng et al. (2020)	ASC				x		x							
Zhou et al. (2020)	ASC				x		x							
Zhou et al. (2021b)	E2E-ABSA							x			x			

[1]Used for ATE

[2]Used for ACD/ASC

[3]Used for ACD or AE

[4]Used for ASC

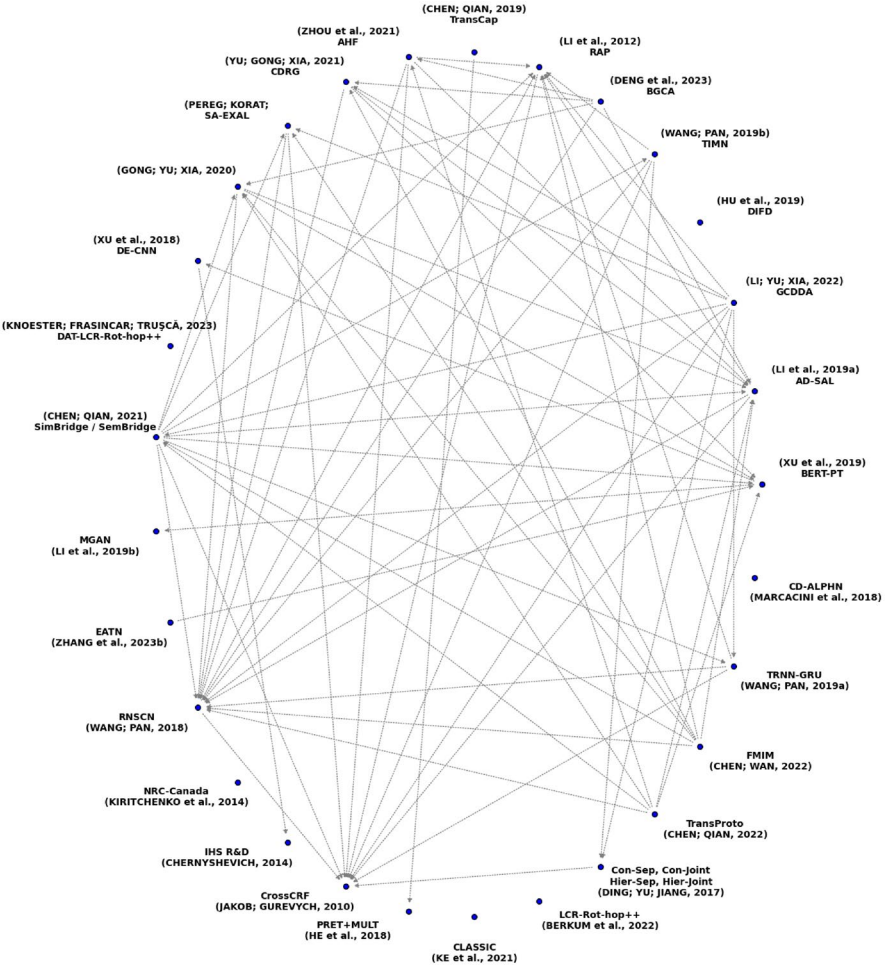

Fig. 9 Algorithms compared for performance analysis. Models that were not mentioned in tests of other models were omitted

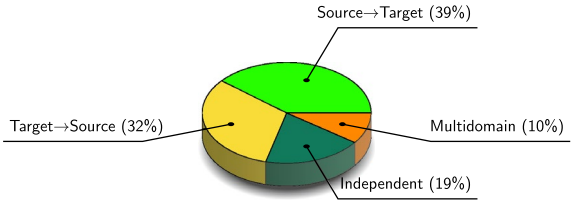

Fig. 10 Percentage of articles per group

4.2.2 Independent: data

One approach some authors use to make models robust for cross-domain is to train them with different data sources. Santos et al. (2021) fine-tune BERT on various different domains for the aspect extraction task. They test their Multidomain Aspect Extraction using Bidi-

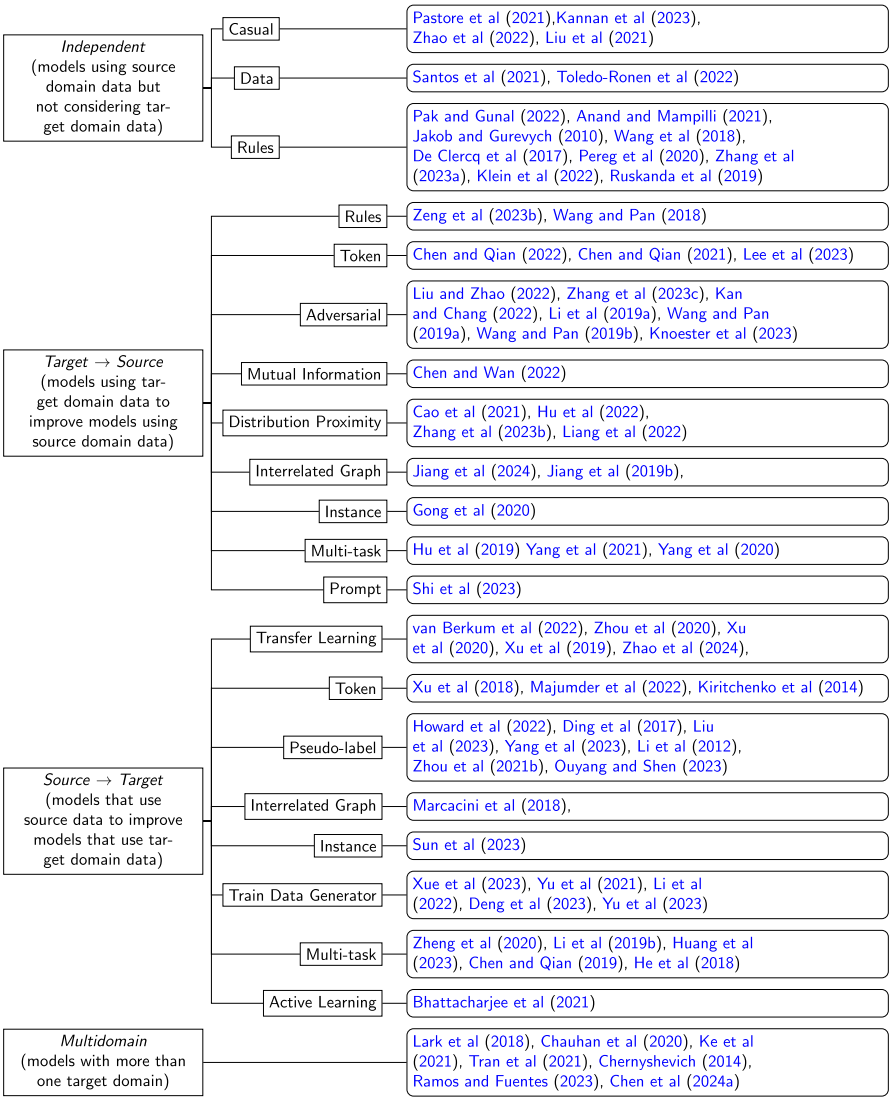

Fig. 11 Proposed classification considering cross-domain

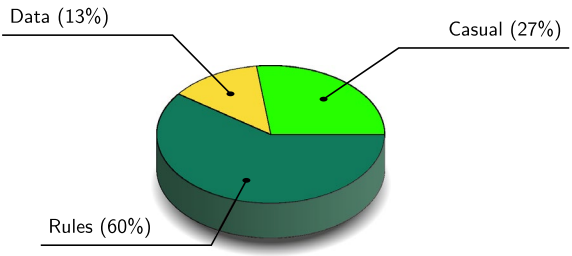

Fig. 12 *Independent* group distribution

rectional Encoder Representations from Transformers model (MDAE-BERT) in a target domain different from those used in training. The model shows an improvement compared to the model trained on a single data source.

Toledo-Ronen et al. (2022) propose something similar but using a single labeled data source and other unlabeled sources. First, they train the model on the labeled database. Using the labeled model, they generate pseudo-labels on the other data sources and retrain the model. They repeat this process for several steps. The final model generalizes cross-domain operation.

4.2.3 Independent: rules

This group includes models that extract characteristics independent of the domain. Pak and Gunal (2022) use multiple annotated source domains to create a set of rules. The generated rules consider only words, POS-tagging, and a lexicon of opinion words. They are applied to a target domain that must achieve a specific score for aspects to be extracted. They differentiate between targets and opinions of aspects, where the latter is an aspect for which an opinion has been expressed.

Ruskanda et al. (2019) developed an aspect and opinion term extraction algorithm that dynamically extracts syntactic relations from labeled source data. The algorithm starts by extracting syntactic dependency rules with the fewest possible steps between the aspect and the opinion term and expands them to cover the labeled source data. During the validation of the effectiveness of the obtained rules, the extracted words are compared for similarity with some pre-determined aspect and opinion words and must meet a minimum threshold. Ultimately, the obtained rules are applied to the target domain to extract aspects and opinion terms.

Anand and Mampilli (2021) use a genetic algorithm applied to labeled source data to generate aspect extraction rules. Some initial rules are constructed using snippets of syntactic dependency trees expressing relations between nodes containing words, POS-tagging, and sentiment polarity extracted from a lexicon, among other features. These rules must contain the labeled aspect. Then, the generated rules undergo crossover and mutation and are partially re-evaluated on the source data, where they must achieve a minimum score. Crossover involves randomly choosing two rules and exchanging the graph segment from a common node, generating two new offspring. Mutation can occur in two ways. One way is to completely regenerate new rules, replacing the initial ones and generating new offspring. Another way is to perform a specific operation, such as adding a new edge, removing one, changing a label, etc. The final rules are applied directly to the target domain, or a classifier is trained where the input features are vectors with the application of the rules to each word in the sentence.

Jakob and Gurevych (2010) trained a CRF model for aspect extraction on labeled source domain data in the BIO scheme. They used structural features such as the words themselves, POS-tagging, whether they have a dependency relation with sentiment expression, and proximity to sentiment expression, among others. Because these are domain-independent relations, the same model performed well when tested in another domain. An interesting test the authors conducted was removing words as features. This improved the result in cross-domain scenarios, as words are domain-dependent. Similarly, Wang et al. (2018) generated a CRF considering POS-tagging and syntactic dependency relations. Along the same lines,

De Clercq et al. (2017) used CRF to train a model for aspect extraction employing POS-tagging and syntactic relations, among other features.

Pereg et al. (2020) developed the Syntactically-Aware EXtended Attention Layer model (SA-EXAL) for aspect and opinion term extraction in the BIO scheme. They adapted the encoding layer (encoder) of the BERT model (Fig. 13) by modifying the existing attention mechanism (see Sect. 2.3.6) to incorporate syntactic information. A parallel mechanism, the Syntax-Aware Self-Attention (SA_{parse}), is created to avoid retraining the entire BERT model. The output of the attention mechanism is the same as the original BERT, Multi-Head Self Attention (SA), but now with the addition of attention given to syntactic data. The calculation of SA_{parse} is described in Eq. (28). P_{parse} is an N by N square matrix, where N is the number of tokens. Each position represents the probability of a token being the head of another in a syntactic dependency relation, obtained from a pre-existing library. The "$\odot$" operator is the Hadamard product, i.e., element-wise multiplication.

$$\begin{aligned} \mathrm{A}_{parse} &= Softmax\left(\frac{\mathrm{Q}_{parse}\mathrm{K}_{parse}^{T}}{\sqrt{\mathrm{d}_k}} \odot \mathrm{P}_{parse}\right), \\ SA_{parse} &= FF(\mathrm{A}_{parse}\mathrm{V}_{parse}), \\ \mathrm{h}^{(l+1)} &= LN(\mathrm{h}^{(l)} + SA(\mathrm{h}^{(l)}) + SA_{parse}(\mathrm{h}^{l})), \end{aligned} \tag{28}$$

in which LN and FF are the normalization and feed-forward layers of the BERT model.

Zhang et al. (2023a) developed the Transformer-based Semantic-Primary Knowledge Transferring network model (TSPKT) considering syntactic relationships and generalizing semantic relations. They argue that high-level abstractions of aspects and opinion terms help in cross-domain scenarios. For example, the generalization of "pizza" is "food" and the generalization of "delicious" is "joy". Considering this, a graph is generated with all the words from the Senticnet ontology (Cambria et al. 2014), connecting words that have a semantic relationship. All words from the source domain are mapped, and a new graph is constructed

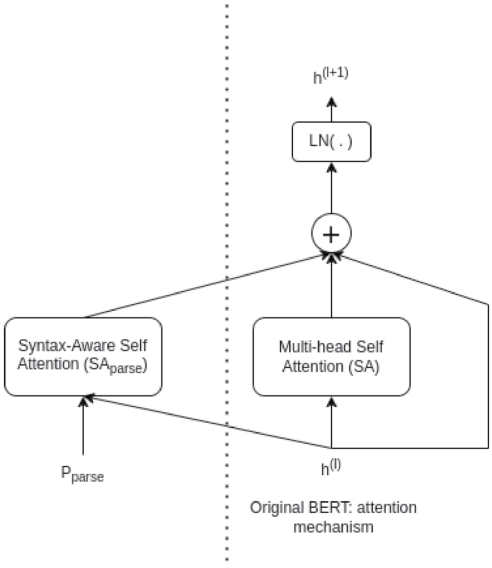

Fig. 13 SA-EXAL Model. The mechanism represented on the left was added to BERT to incorporate syntactic attention. Adapted from Pereg et al. (2020)

considering k-hops from them. Then, emotions are assigned to each word using the sentiment lexicon "EmoLex" (Mohammad and Turney 2013). A breadth-first search is performed from each word to obtain more general words with the same emotions, generating words with primary meanings. These words are incorporated into the model through a modified Bi-LSTM network. The outputs of this network pass through an attention mechanism that incorporates information from the syntactic dependency tree. An attention mechanism is also created for the aspect for which sentiment classification is desired. A softmax output is created to label the sentiment polarity. The authors considered Glove and BERT vectors as input to the model (Pennington et al. 2014).

Klein et al. (2022) used an auxiliary task to incorporate syntactic or semantic information. They based their work on the BERT and SA-EXAL models for aspect extraction in the BIO scheme as the main task. The auxiliary task exists only during training and considers the opinion term vector to predict the aspect (ASP) or the dependency relation between the opinion term and the aspect (PATT). A significant improvement was observed in the SA-EXAL model with semantic relations, as it already incorporates syntactic relationships. The authors analyzed the difference between source and target distributions using Jensen-Shannon. They observed that more similar source data distributions have a greater chance of transferring syntactic and semantic knowledge.

4.3 ABSA: target to source

This subsection contains models whose cross-domain adaptation considers data from the target domain to approximate the model using data from the source domain. It is subdivided into *Rules, Token, Adversarial, Mutual Information, Distribution Proximity, Interrelated Graph, Instance, Multi-task*, and *Prompt*. Figure 14 illustrates the percentage of articles in each subgroup.

4.3.1 Target to source: rules

The rule-based works in this subsection bear similarity to those in Sect. 4.2.3. However, target domain data is employed to enhance performance. The model by Wang and Pan (2018), Recursive Neural Structural Correspondence Network (RNSCN), obtains aspects and opinion terms from a review. It recursively generates new vectors based on the syntactic dependency relation of the others. The model attempts to predict these relations for both source and target domains as an auxiliary task and, in doing so, improves cross-domain performance. For the source domain, the projected vectors of each word in the review are passed

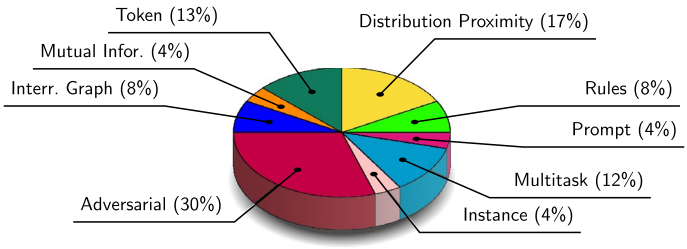

Fig. 14 *Target→Source* Group Distribution

to a recurrent network that will predict the labels. They also enhance this model. The parsers that extract syntactic dependency relations may have errors, so they seek to group them. For this, a predetermined number of vectors to be learned during training is considered. A new vector is created by applying an attention mechanism that relates the original to each of the vectors in the group. It is used for the auxiliary task of syntactic dependency classification. To ensure representativeness, an autoencoder is used to regenerate the original vector. Additionally, the model aims to make the vectors as orthogonal as possible by penalizing non-orthogonal vectors.

Zeng et al. (2023b) argue that merely learning the syntactic relation for term extraction is not sufficient. For instance, considering the English sentences "The best bagel in New York" and "Windows being the main issue", the syntactic relation "opinion term → adjectival modifier (*amod*) → aspect" is valid for the first sentence ("best bagel") but not for the second ("main issue"). The authors then utilize Conceptnet, a concept structure (Sect.2.3.1). However, to use it, they need to separate concepts by domain. For example, "windows" could refer to a part of a car or an operating system. They developed their Knowledge-Enhanced and Topic-Guided Cross-Domain ABSA model (KETGM). To do this, the authors first use LDA to extract associated words. Then, for each review, they start from nouns, adjectives, and adverbs as seeds and, along with the extracted words, filter Conceptnet considering up to 2 hops. For example: pizza $\xrightarrow{\text{IsA}}$ food $\xrightarrow{\text{AtLocation}}$ restaurant. By combining the trees, they obtain Conceptnet filtered by domain. Next, the authors assemble an R-GCN. An encoder is used to represent the characteristics of each node, and a decoder is used to predict the existence of each relation between two nodes (isA, AtLocation, etc.). The model is trained, and vectors that abstract the conceptual relations are found. These vectors undergo an autoencoder to compress them. A weighted attention mechanism with the topic words obtains a second vector for each word in the review. A third vector abstracting the syntactic relations is created using a BERT model that attempts to predict POS-tagging and dependency relations. These three vectors are concatenated, and a Softmax is used to classify the tags into BIO+polarity.

4.3.2 Target to source: token

The models described in this topic focus on the words (tokens) of the reviews. Chen and Qian (2022) created the method *TransProto* for aspect extraction and classification. The central idea of the authors is to append representations of target domain words that have similar semantics and syntax to the source domain representations, making the vectors more independent. The model considers POS-tagging, dependency relations, semantic vector representations such as Word2Vec, and polarity obtained from a sentiment lexicon. The model discovers the most similar target words for each word in the review and concatenates an average of Word2Vec vector representations or contextualized BERT vectors of these words. The model is trained using labeled source domain data. Additionally, an adversarial network is employed with data from both domains to approximate the features.

The models *SynBridge* and *SemBridge* follow the same idea (Chen and Qian 2021). For the *SynBridge* model, POS-tagging and dependency relations are added to the embeddings. For *SemBridge*, target domain words are associated with the source domain using a procedure similar to the *TransProto* method, where POS-tagging, dependency relations, and semantic vector representations are considered. Using these enriched representations, each

model process the reviews through a convolutional network until to obtain the BIO classification for aspect extraction.

The model DIWS-LCR-Rot-hop++ (DIWS, Domain-Independent Word Selector) classifies aspects (Lee et al. 2023). The data passes through a domain classifier that uses an attention mechanism in the last layer. Those vectors that obtained a higher weight to classify the domain are discarded in the next step. This way, the model does not depend on the domain. Next, the model generates vector representations for the aspect, the data to the left and right of the aspect, and associates attention mechanisms until classification.

4.3.3 Target to source: adversarial

The models listed here focus on adversarial networks to address the cross-domain problem. Knoester et al. (2023) created the DAT-LCR-Rot-hop++ model (DAT—Domain Adversarial Training), similar to the DIWS-LCR-Rot-hop++ model (see Sect. 4.3.2). Instead of explicitly discarding specific domain tokens like the DIWS-LCR-Rot-hop++ model, a domain adversarial network is learned and attached to the last layer. This way, the model learns to classify sentiments with source domain data using invariant features.

Liu and Zhao (2022) created an aspect classification model. The data passes through BERT. Then, the output of the main vector goes through convolution filters and Gate Units. In the end, the model contains the aspect classifier and an adversarial network for the domain.

Zhang et al. (2023c) created the SKEP_Gram-CDNN model for sentiment classification. Using a combination of techniques, they replicate the generator-discriminator model of adversarial networks. The generator model is applied to the target domain data. During training, the model tries to predict if the embeddings came from the source or target domain. At the same time, annotated source domain data is used to train the sentiment classifier.

Kan and Chang (2022) created a model to classify the sentiment of a category. They noticed that training a model with all categories together led to higher errors in categories with fewer training examples. Training separately for each category worsened the model even more. They devised the Adversarial Reptile algorithm. They use various labeled source domains, which can be one domain per category. During training, k steps are executed at a time for each source domain, but the model parameters related to features and the classifier are updated only after all source domains are processed. This process is repeated several times until the final model.

The Aspect Detection—Selective Adversarial Learning model (AD-SAL) is used for aspect extraction and classification (Li et al. 2019a). The authors employ both local and global memories to infer the latent relationship between aspect and opinion words. First, the review is passed through a Bi-LSTM network, which utilizes local memory. Next, there are two global memory vectors, one for the aspect and one for the opinion term, which are learned during training. Associations are made between these global vectors and the output vectors of the Bi-LSTM network using matrices that are also learned during model training. There are three types of association matrices: aspect-aspect and opinion-opinion (intra-relationship) and aspect-opinion (inter-relationship). Two relation vectors are extracted with the associations for each output. The first one, with aspect-aspect and aspect-opinion associations, represents aspect relations. A relation where the output vector is highly intra-related to the aspect memory and inter-related to the opinion memory should indicate an aspect.

The second one represents opinion term relations, relating the outputs and memories using opinion-opinion and aspect-opinion associations.

The global memories are refined using an attention mechanism applied to the output vectors based on the obtained relations. The memories are used to generate new relations, and the process repeats several times. The final relations are used as attention mechanisms to classify aspects and opinion terms as auxiliary tasks. At the end of the process, a recurrent network is applied to these relations, generating labels in the BIO + polarity scheme of the main task. In this paper, the authors consider that only aspects need to be aligned across domains. They apply an adversarial network word by word, emphasizing those with greater weight in the aspect attention mechanism.

Similar to the AD-SAL model, Wang and Pan (2019b) created the Transferable Interactive Memory Network (TIMN) model to extract aspects and opinions. The model follows the same principles of local and global memory. Vectors incorporating local memory are obtained by a GRU recurrent network applied to words. Relation vectors are obtained by associating the global memories of aspect and opinion. However, unlike AD-SAL, the authors believe aligning aspects across domains is unfeasible. The authors propose aligning opinion terms and the existing relation between aspects and opinion terms. Thus, an adversarial network is applied to the global memory of opinion term, and another one is applied to the concatenation of the global memories of aspect and opinion. Additionally, they believe that the associations of aspect-aspect, opinion-opinion, and aspect-opinion are invariant across domains. Using the associations between global memories in different layers, the model attempts to predict the respective type of association.

Wang and Pan (2019a) enhanced the RNSCN algorithm (Sect.4.3.1) to include the adversarial network, creating the Transferable Recursive Neural Network model (TRNN-GRU) for aspect and opinion term extraction. The vector of each word is associated with a domain classifier with a reverse gradient. However, the authors consider that there is more than one type of distribution of words regarding their syntactic function. Instead of passing only the word through the adversarial network, the authors choose also to pass the type of syntactic dependency relation between the head and the word. This is done as a one-hot vector, meaning a vector with a value of "0" for all possible syntactic relations and a value of "1" for the existing one. For the alternative version of the model that groups syntactic relations, the value of "1" is passed to the most likely group, i.e., the one with the highest weight in the attention mechanism.

4.3.4 Target to source: mutual information

Chen and Wan (2022) proposed the *Finegrained Mutual Information Maximization* (FMIM) technique to enhance the results obtained by classification models. They observed two problems in adapting ABSA models in cross-domains. First, adapted models tend to classify all words into a single class. For example, in a BIO scheme, words tend to be classified as "O". Second, even if the model distinguishes the labels, it is not confident enough, indicating a low probability. Therefore, the authors suggest maximizing the mutual information between the distribution of predicted labels (Y) and the extracted features (X):

$$\begin{aligned} I(x;y) &= H(y) - H(y|x) \\ &= \mathbb{E}_y[\log p(y)] - \mathbb{E}_{(x,y)}[\log p(y \mid x)] \end{aligned} \tag{29}$$

By maximizing Eq. (29), the entropy of Y ($H(Y)$) is increased, meaning there is a higher possibility of the label predictions being something other than "O". Maximum entropy occurs when the distribution is uniform. At the same time, it attempts to minimize the entropy of Y when X is known ($H(Y|X)$), meaning a uniform probability distribution is undesirable when X is known.

This equation becomes a component of the cost function. Since the real distribution is unknown, each batch containing examples from the source and target domains is considered an approximation of the real distribution during training. The authors tested it on a BERT model with BIO labeling, with the cost function being a composition of the respective label's error function and the proposed mutual information component, achieving good results.

4.3.5 Target to source: distribution proximity

In this group are the models whose goal was to decrease the distance between the source and target distributions using some distribution distance metric. Zhang et al. (2023b) created the *An Efficient Adaptive Transfer Network for Aspect-Level Sentiment Analysis* (EATN) model for sentiment classification. First, reviews from the source and target domains are passed through BERT, generating contextualized vectors. The process continues with the contextualized vectors undergoing some operations and an attention mechanism relating the review representations to the aspect's representation. Then, the vectors go through several dense layers until sentiment classification and domain classification by an adversarial network. In this architecture, the authors also seek to minimize the distance between the vectors obtained from the source and target domains in the last layers. They consider the last layers to be more domain-specific and, by doing so, improve applicability in the target domain. They apply MMD with RBF kernel (Eq. 19) for each training batch, aiming to minimize the error (distance) by Eq. (18).

Hu et al. (2022) aim to classify the sentiment of categories in cross-domain. They use various techniques, including adversarial network. The authors first manually separate the source and target domains into subdomains according to the category at a higher level, generating the multi-source multi-target transfer network (MMTN) model. The model input consists of the review, the category at a lower level, and the subdomain. The review goes through a Bi-LSTM network to generate contextualized vectors. Then, an attention mechanism is applied to these contextualized vectors to generate the feature vector. Attention is applied to the concatenation of two vectors, which are to be learned and represent the category and the subdomain. The feature vectors of examples from source and target domains go through an adversarial network. During training, labeled examples from the source domain go to the sentiment classifier of their respective subdomain. In parallel, during training, the model minimizes the distance between the feature vector and the domain vector, propagating the error only to the latter. Similarly, it does so between the aspect vector and the domain vector. Finally, during inference, there are no sentiment classifiers for the target subdomains, necessitating a composition of the existing ones. To ensure this works, during training, two similarities are considered for examples from the source domain: (i) between the subdomain vector of the example and the other subdomain vectors and (ii) between the feature vector with all other subdomain vectors. A Gate Unit is applied to split the weights between these two types, and a combination that considers the similarities found is calculated. This combi-

nation serves to weigh the weights of the average of the classifiers' estimates, which should approximate the original label.

The *Deep Transfer Learning Mechanism* (DTLM) model minimizes the difference between the source and target domains using the KL divergence (Cao et al. 2021). The labeled source data and unlabeled target data pass through a BERT. This BERT is used to classify the sentiment of an aspect. During training, the loss function considers: (1) the prediction of the label of the source data; (2) the KL divergence between the source and target data; (3) minimum entropy in the predicted labels by the target data, where lower entropy ensures that class predictions are less uniform; and (4) data augmentation by translating a sentence to a different language from the original and translating it back.

The *Multi-level Sentence-Word Interaction Transfer* (MSWIT) model is used for aspect extraction (Liang et al. 2022). It assumes a source with aspect and category labels and a target with category labels only. The central idea is that the relationship between category and aspects is domain-invariant. It applies a Bi-LSTM network and an attention mechanism to the review, generating contextualized vectors. Then, the neural network is divided into two modules. The first module goes through several dense layers until ending in a CRF and classifies the aspect labels. The second module undergoes an attention mechanism, condensing the vectors into a sentence and ending by indicating the source and target data categories. Finally, it seeks to approximate the vectors used to classify the aspects and the vector used to classify the sentence. The vectors from each layer of the first module are summed, projected into another space, and the Euclidean distance with the vector used to classify the sentence is minimized.

4.3.6 Target to source: interrelated graph

The *Interrelated Graph* section contains models that approximate the features of the source data to the target data by connecting the reviews through a graph. Jiang et al. (2024) constructed two models using tripartite graphs[6] for aspect classification. The first model builds the *word-topic-instance* graph, connecting the instances, the instance words, and the topics. To construct this graph, the authors rely on the idea that instances are compositions of topics and that each word belongs to a topic, following the concept of LDA. To obtain the topic distribution of each document, the authors use a neural topic model: from the bag of words of a document, a topic distribution is inferred using a variational autoencoder.[7] This topic distribution is applied to a word matrix for each topic, regenerating the document. This word matrix should be learned during training. The topics are generated to be common to the source and target domains, allowing connection between the instances. Three graphs are built: one for each domain and one with shared source and target domains, connecting words and topics. Each instance node is composed of a contextualized vector after BERT processing. These graphs are processed by GCNs, merged, and aspect classification is performed for each instance.

The second model is based on a *word-pivot-instance* graph, where words, instances, and the pivot are connected. The pivot is a text that summarizes the instance's characteristics and is domain-invariant. For example, a review warning about a product's problems would

[6] Tripartite graphs contain nodes of three distinct classes.

[7] The difference between a common autoencoder and a variational autoencoder is that the latter learns the parameters of a distribution instead of a compact vector.

generate the pivot: "The quality is not good". It is generated for each review in two steps. In the first step, a BERT is used to extract the review section containing the aspect. For the second step, a BERT is used to perform two tasks, using the masking of input elements accordingly. The first task obtains the pivot using the extracted section in the first step and latent vectors. The second task consists of learning these latent vectors, which are obtained from the pivot and the extracted section. Finally, using the graphs of the source, target, and both domains, GCNs are applied and merged for a classifier application.

The *Traceable Heterogeneous Graph Representation Learning* (THGRL) model is used to identify the categories addressed in a review (Jiang et al. 2019b). A heterogeneous graph is constructed connecting the reviews, words, products they refer to, categories, the people who wrote them, and the stores of the products. Among the edges, word co-occurrences are included. Then, several random walks are performed to be used by the *node2vec* method (see Sect.2.3.8). This form of random walk can leave out aspects and add irrelevant contexts causing noise. Thus, the authors sought a more global representation. They apply the idea of LDA, considering that the nodes obtained in each walk form the documents and thus obtain the vertex-tracers, equivalent to topics. Skip-gram is used to obtain representations of the vertices obtained by random walks and their associated topics. In a subsequent step, the words of each review are transformed by concatenated vectors of the vertex representation and the vertex-tracer, and an average is taken, which is used for a category classifier.

4.3.7 Target to source: instance

The main difference of the model by Gong et al. (2020) is to align the proportion between the examples found in the source and target during domain adaptation (Eq. 2). After obtaining contextualized words with a BERT, the Unified Domain Adaptation (UDA) model is divided into feature-based and instance-based components. The first module performs auxiliary tasks of predicting POS-tagging and dependency of syntactic relations, making the features more invariant. The second module classifies the domain of each word. The probability of belonging to each domain will give the ratio $\frac{\mathbb{P}_t}{\mathbb{P}_s}$ used to align the instances. Thus, these values are used to weigh the loss function during training.

4.3.8 Target to source: multi-task

Models can benefit from auxiliary tasks. Hu et al. (2019) created the *Domain-Invariant Feature Distillation* (DIFD) model for sentiment classification using aspect extraction as an auxiliary task. The authors consider sentiment classification and aspect extraction to be orthogonal tasks, with the former being more domain-independent than the latter. The model works as follows: first, contextualized representations of each word are obtained using a BiLSTM network. Then, the words are separated between sentiment classification and aspect extraction tasks. To achieve this, an attention mechanism with two positions is applied for each word, summing to "1" for each of them. Each of the positions is applied to the contextualized vector, generating two representations, one to be used by the auxiliary task of extracting aspects and another used to classify sentiment. The sentiment classification task still applies attention regarding the aspect before classifying the sentiments of the source data. An adversarial network achieves the invariance of feature vectors. Instead of using a reverse gradient, the authors perform specific training. Each step consists of two

stages. In the first stage, the loss function is a composition of the sentiment classifier for the source domain data, the domain classifier with inverted labels, plus the aspect extractors from the source and target domains. This is done by freezing the domain classifier. In the second stage, only the domain classifier is trained; in this case, the labels are not inverted, as the idea is to train the classifier.

Yang et al. (2021) created the *Neural Attentive model for cross-domain Aspect-level sentiment CLassification* (NAACL). The model classifies sentiments of categories considering an altered LDA, called wsLDA (weak supervised LDA). In this altered LDA, each document is considered a composition of words originating from distributions obtained by a hierarchical category. The first level separates the word distribution of sentiment, aspect, and others. These groups are subdivided into topics. The Gibbs Sampling algorithm (see Sect. 2.3.4) is adapted to consider these levels. The model user must provide some words that will be used as seeds affecting the probability of belonging to each group at the first level. The wsLDA is applied to determine the topics for each group. A vector with the probability distribution of all words for each category is generated. These distribution vectors are applied to an attention mechanism in a network that classifies the sentiment of the category. The model has an auxiliary task to classify the domain. The choice to classify this model in this subsection is because, in this case, the authors do not use an adversarial network.

Similar to NAACL, Yang et al. (2020) created the Neural Attentive model for Cross-domain Aspect/Sentiment-aware Abstractive review Summarization (CASAS). This model uses the wsLDA mechanism and attention. However, this model generates text instead of just classifying the category. When manipulating text, the authors use some additional techniques. They use a mechanism that reuses words from the source data with a certain probability, as the reviews may contain words the model has never seen. During training, the model is penalized if the generated sentence is close to a random one.

4.3.9 Target to source: prompt

Language models need to receive extra information as input data so they can perform a specific task, which is called a *prompt*. For example, consider the sentence "Translate to Russian: How are you?" as input to a language model. In this sentence, there is a command (*prompt*), "Translate to Russian:", and the data that should undergo the action of the command, "How are you?". The way this command is passed as input to the model can affect the response.

Shi et al. (2023) propose to learn the prompt dynamically. Using a pre-trained T5 model, the vector representations of the words from the review and their POS-tagging are passed through the encoder and decoder to obtain the prompt. This prompt is fed back to the T5 model along with the representations of the words from the review and their POS-tagging. The encoder output is used to extract aspects in the BIO scheme. Ideally, the prompt generates vectors that represent the domain of the reviews. For training the prompt, the words that most represent each domain are identified considering mutual information. Each review will use a subset of the vector representations closest to the words of its respective domain.

4.4 ABSA: source to target

In this group are the models that used the source data to enrich the target data. These works are grouped according to Fig. 11 into: *Transfer Learning, Token, Pseudo-label, Interrelated Graph, Instance, Train Data Generator, Multi-task* and *Active learning*. Figure 15 illustrates the percentage of articles in each subgroup.

4.4.1 Source to target: transfer learning

This subsection deals with models that were pre-trained with source data. van Berkum et al. (2022) uses the LCR-Rot-hop++ model for aspect polarity classification. The model generates contextualized vectors for the aspect and its left and right contents using Bi-LSTM networks. An attention mechanism is applied between the aspect and the vectors of its left and right contents, generating two vectors: r_l and r_r. These vectors are used in new attention mechanisms applied to the left and right contents, generating two new vectors, r_{lt} and r_{rt}. This process repeats successively. Finally, the vectors r_l, r_r, r_{lt}, and r_{rt} are concatenated, and classification is performed. To perform transfer learning, the authors train the model on the source data, freeze the first layers, and train on the few labeled target data. The lower layers encode generic language features, and the higher layers encode domain-specific language features with their respective sentiments.

Zhao et al. (2024) used *transfer learning* in their aspect classification model. They start from the problem where there are labeled data for the source domain but few data for the target domain. The model begins by transforming the review into Glove embeddings. Then, the aspect and the review are passed through BiLSTM networks, generating contextualized representations. A graph is constructed for the review, considering the syntactic dependency relations. A GCN is applied to this graph. These three outputs, aspect embeddings, GCN, and review embeddings, go through successive attention mechanisms until their classification. The chosen *transfer learning* strategy was "unfreezing the first n layers". It works as follows: first, pre-training is done with the source domain data. A copy of the model structure is made. The first n layers are initialized with the parameters trained in the original model. The remaining layers are initialized with a uniform distribution. All variables are updated during training. Data from both domains, source and target, are used in the fine-tuning, but with different weights in the loss function.

Another form of *transfer learning* performed by some authors was pre-training models using auxiliary tasks. They suggest that auxiliary tasks can improve the performance of the main task. Xu et al. (2020) noticed that pre-training BERT on domains similar to the target domain could positively influence the model, so they created the *Domain-oriented Language Model for Aspect-based Sentiment Analysis* (DomBERT). They retrain BERT on the masked language modeling (MLM) task but emphasize examples close to the target domain.

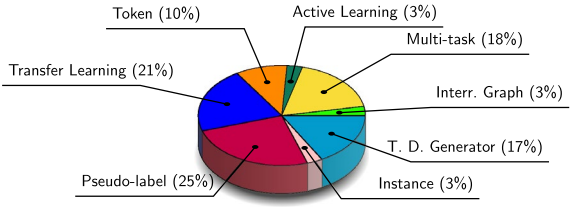

Fig. 15 *Source→Target* Group Distribution

To do this, they append a domain classifier to BERT, and at each step, examples from each source domain are chosen based on their similarity to the target domain.

Xu et al. (2019) followed the same approach and created the *BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis* (BERT-PT). Several tasks are performed on a pre-trained BERT. They train it on MLM and next sentence prediction (NSP) in a domain similar to the desired one. At the same time, they train it on SQUAD, a question-answering task with a large amount of data from diverse domains (Rajpurkar et al. 2016). The use of question-answering helps in aspect extraction and classification tasks by covering all types of content. After this, the model is fine-tuned for the specific task of aspect extraction or classification for the target domain.

Zhou et al. (2020) created the *position-aware hierarchical transfer* (PAHT) model for sentiment classification. Each review is broken down into sentences using Rhetorical Structure Theory (RST). The words of each segment are passed through a Bi-LSTM network to generate contextualized vectors, as well as the aspect. Additionally, the position of each word relative to the aspect is added. The words undergo an attention mechanism concerning the aspect and are merged into a sentence for each segment. Similarly, the sentences are passed through another Bi-LSTM network, adding the relative position of the sentence in relation to the one containing the aspect. An attention mechanism is applied, the sentences are concatenated, and the aspect is classified. The authors suggest performing transfer learning to enhance the model in the target domain. They perform sentence classification using a labeled source domain. They choose 3 strategies to train the model: (1) Random sampling; (2) Sampling of sentences containing the aspects present in the target domain; and (3) sampling considering sentences from the source domain similar to those from the target. In this case, the target sentences serve as queries. This is done using the BM25 algorithm (Robertson and Zaragoza 2009). They also consider transferring the parameters of the trained model to the final model at various depths.

4.4.2 Source to target: token

Methods in this category extract knowledge from words or their vector representations from the source domain to enrich the tokens in the target domain. Xu et al. (2018) proposed using two types of embeddings in their models, one trained on a large corpus of data like GloVe (Pennington et al. 2014), and the other trained on domains similar to those used in ABSA. The former has broader coverage, and the latter is more contextualized. They then used labeled data from the target domain and created the *Dual Embeddings CNN* (DE-CNN) model using both as input.

An algorithm can leverage the embeddings from the aspect extraction task for the sentiment classification task. Majumder et al. (2022) trained an initial model on labeled source domain data by passing a sentence through a Bi-GRU network that extracts aspects using the BIO scheme (Cho et al. 2014). Then, this model is used in the target domain. The contextualized embeddings output from this Bi-GRU are concatenated with the input embeddings of the sentence. They are then passed to a new model to classify polarity. According to the authors, aspect extraction is a syntactic activity that can be used in cross-domain tasks. By performing aspect extraction, the model identifies opinion terms, assisting in the classification task.

The *NRC-Canada* model, also known as SVM-*feature*, uses sentiment-annotated corpora at the document level as the source domain (Kiritchenko et al. 2014). The Pointwise Mutual Information (PMI) metric is applied to assess the correlation between the frequency of each word in relation to the sentiments and categories of the documents, thereby generating sentiment lexicons and a score for each category. A third category is generated for each word using the *Brown* algorithm, which classifies the words into 1000 categories (Brown et al. 1992). These measures are associated with the words in the target domain reviews to generate input features for the model to be trained. These features are then used in SVM models to classify categories or determine the sentiment of the aspect and the category.

4.4.3 Source to target: pseudo-label

In this subsection, we discuss models in which pseudo-label generation is one of the main techniques used for cross-domain adaptation. The *Adaptive Hybrid Framework* (AHF) is an algorithm for aspect extraction and classification that follows the student-teacher technique (Zhou et al. 2021b). The student is trained using labeled data from the source domain and pseudo-labeled data from the target domain, generated by the teacher. Additionally, an adversarial network with reverse gradient is applied to this student to classify the domain. The utilized pseudo-labels are only those predicted by the teacher submodel that exceed a certain threshold. This threshold is dynamic and depends on the similarity with the source data, which is obtained through the confidence level of the domain classifier. The teacher model has the same structure as the student classifier model, but its labels are updated more slowly with a moving average.

Howard et al. (2022) devised a technique for generating pseudo-labels in their aspect extraction model. First, they retrieve the top-k nouns (noun phrases) from the target domain using word frequency. Words are expanded using the ConceptNet ontology (Sect. 2.3.1) and the COMET algorithm (Bosselut et al. 2019), which predicts related words given the original words and semantic relationships, generating a knowledge graph. Next, using POS-tagging and syntactic dependencies, they extract the most likely words to be aspects. These words are filtered using the knowledge base. Pseudo-label generation is performed and tested in two ways. In the first approach, the reviews from the target data are modified. Tokens are placed before the extracted words as cues for the aspect extractor. Tokens are also added for the source data, but to make the model robust, tokens are inserted with some randomness according to a precision and recall rule. Using DeBERTa and BERT for BIO classification, the DeBERTa-PT and BERT-PT models were created (Chen and Qian 2022).

The second approach alters DeBERTa, generating the DeBERTa-MA model. DeBERTa features a disentangled attention mechanism:

$$\begin{aligned} \mathrm{A}_{i,j}^{c2c} &= \mathrm{Q}_i^c \mathrm{K}_j^{cT}, \mathrm{A}_{i,j}^{c2p} = \mathrm{Q}_i^c \mathrm{K}_{\delta(i,j)}^{pT}, \mathrm{A}_{i,j}^{p2c} = \mathrm{K}_j^c \mathrm{Q}_{\delta(j,i)}^{pT}, \\ \mathrm{A}_{i,j} &= \mathrm{A}_{i,j}^{c2c} + \mathrm{A}_{i,j}^{c2p} + \mathrm{A}_{i,j}^{p2c}, \\ \mathrm{H} &= \left(\frac{\mathrm{A}}{\sqrt{3\mathrm{d}}} \right) \mathrm{V}^c, \end{aligned} \tag{30}$$

in which $\mathrm{Q}^c, \mathrm{K}^c, \mathrm{V}^c \in \mathbb{R}^{N \times d}$ represent N content embeddings (*Query, Key,* and *Value*) of dimension *d*, Q^p and $\mathrm{K}^p \in \mathbb{R}^{N \times d}$ are embeddings representing the relative position $\delta(i, j)$.

The model alters this mechanism by introducing new embeddings, m^+ and $\mathrm{m}^- \in \mathbb{R}^d$, representing hints of what can or cannot be an aspect. Content vectors relate to these embeddings as follows:

$$\begin{aligned}
\mathrm{A}_{i,j}^{c2c} &= \mathrm{Q}_i^c \mathrm{K}_j^{c^T}, \mathrm{A}_{i,j}^{c2p} = \mathrm{Q}_i^c \mathrm{K}_{\delta(i,j)}^{p^T}, \mathrm{A}_{i,j}^{p2c} = \mathrm{K}_j^c \mathrm{Q}_{\delta(j,i)}^{p^T}, \\
\mathrm{A}_{i,j}^{c2m} &= \mathrm{Q}_i^c \mathrm{K}_j^{m^T}, \mathrm{A}_{i,j}^{m2c} = \mathrm{Q}_i^m \mathrm{K}_j^{c^T}, \\
\mathrm{A}_{i,j} &= \mathrm{A}_{i,j}^{c2c} + \mathrm{A}_{i,j}^{c2p} + \mathrm{A}_{i,j}^{p2c} + \mathrm{A}_{i,j}^{c2m} + \mathrm{A}_{i,j}^{m2c}, \\
\mathrm{H} &= \left(\frac{\mathrm{A}}{\sqrt{5\mathrm{d}}} \right) V^c,
\end{aligned} \tag{31}$$

where $\mathrm{Q}^m, \mathrm{K}^m \in \mathbb{R}^{N \times d}$ are learned projections representing *Query* and *Key* for m^+ and m^-. This model learns to pay attention to the hints passed as parameters. Although both approaches showed promising results in the authors' tests, the first one exhibited better performance.

Ding et al. (2017) created models for aspect extraction, considering labeled source data and unlabeled target data. First, they generate labels based on syntactic rules for both source and target data. The source and target data pass through a recurrent network that classifies the generated pseudo-labels in the BIO scheme. Then, two models are created. The first one passes the source data through a second recurrent network. The contextualized vectors from the first and second networks are concatenated (Con model) to predict the actual labels. The second model is a hierarchical model (Hier) that passes the contextualized vectors from the first recurrent network to a second one and attempts to predict the actual labels. Training was done in two ways, either training with pseudo-labels first and then the actual labels (Sep) or training everything together (Joint). Combining the two models with the two training methods resulted in 4 models: Con-Sep, Con-Joint, Hier-Sep, and Hier-Joint, with the latter showing the best average performance.

Liu et al. (2023) created the Unified Instance and Knowledge Alignment Pretraining (UIKA) model for sentiment aspect classification. It is used when the target domain has labeled data but aims to enrich it with labeled source domain data at the sentence level. Firstly, sentences from the source domain similar to the target domain are separated using a fast method. Then, a refinement of this separation is made considering a text embedding. Pseudo-aspects are created for these sentences based on POS-tagging and frequency. A first model is trained using these pseudo-aspects and the sentence's sentiment. In the next step, a student-teacher model is created, starting with the same weights as the model trained in the previous step and using the target data for training. During training, the teacher model is penalized when it deviates too much from the student model. The goal is not to forget what was learned in the previous step. This penalty decreases towards the end of training. Meanwhile, the student model has its weights updated slowly using a moving average. In the last step, the teacher model is discarded, and the student model is trained and fine-tuned for the target data. The first step aims to align the data at the instance level. The second step aligns the characteristics.

The S^3Map was created for aspect extraction and classification with labeled source data and unlabeled target data (Yang et al. 2023). First, the aspects of the sentences from the source data are masked along with an equal number of random tokens. The sentences go

through a BERT model that attempts to predict whether each masked token is an aspect. This model is used to identify what may or may not be aspects in the target data. The sentences from the target domain pass through the model, and the top-k most likely words to be aspects are selected. The sentence is masked with these words and passed through the model again, which discards those that did not reach a minimum probability of being an aspect. The word embeddings and selected aspects from the source and target data go through one of the four types of encoders, which return the contextualized sentence and aspect: (1) attention regarding the aspect, (2) self-attention of the sentence and the aspect, (3) a GCN graph on top of the syntactic dependency tree to obtain contextualized vectors, and (4) a Dual-GCN, which incorporates attention between sentences into the GCN adjacency matrix. After passing through one of these mechanisms, the sentence and aspect vectors are concatenated, and sentiment prediction is made. In the case of target data, there is no sentiment label. A new pseudo-label is used during the training. This pseudo-label is the sharpened label itself by the formula:

$$\tilde{y_j} = \frac{(p_j)^{\frac{1}{\tau}}}{\sum_{j'=1}^{\mathrm{M}} (p_{j'})^{\frac{1}{\tau}}}, \tag{32}$$

where τ is a hyperparameter, M is the total number of sentiment polarities, and p_j is the probability of each sentiment. The higher the value of the hyperparameter, the more certainty is attributed to the predicted label. Although the model can extract aspects, the authors focused the tests on sentiment classification.

The Relational Adaptive bootstraPping (RAP) algorithm is used for aspect and opinion term extraction in cross-domain scenarios (Li et al. 2012). It starts by extracting aspects and opinion terms from the target domain using syntactic rules learned from the source and employing similarity metrics between the source and target domains. These terms are pseudo-labels to train Support Vector Machines (SVMs) classifier models for aspect and opinion term extraction. With these models, scores are obtained for each word to determine its likelihood of being an aspect and opinion term still unlabeled. A graph is constructed connecting the aspects and sentiments by their respective syntactic rules. Different weights are assigned to newly and previously classified words, and then the graph is regularized. Those words that reach a certain threshold are incorporated, and the process is repeated.

The model by Ouyang and Shen (2023) is used for aspect extraction and sentiment classification in the BIO + polarity scheme. The source and target domain data pass through a BERT model, referred to as "A". In parallel, data augmentation is performed by randomly replacing tokens with synonyms, generating new data, which will be the input for a BERT "B", similar to model "A". The source domain data are used to predict their respective labels. There is a professor submodule for each of the A and B models. These professor submodules are updated with a temporal average of their respective students. The professor submodules generate cross-labeled data for the target domain, i.e., the A professor is used by B and vice versa. Typically, student-teacher models suffer from a problem called error amplification, meaning if the teacher mislabels, the student learns incorrectly, and the error propagates. By performing this cross-labeling, this error is mitigated. For better performance of this model, the authors aim to maximize the mutual information between the predicted labels (Y) distribution and the extracted features (X). InfoNCE (NCE—noise contrastive estima-

tion) is used: $I(X;Y) = H(Y) - H(Y \mid X)$ (Oord et al. 2019). As presented in Sect. 4.3.4, by increasing the mutual information, we increase the entropy of Y (H(Y)), i.e., increase the possibility of the labels predicting something different from the "O" tag. At the same time, it is desirable to minimize the entropy of Y when X is known ($H(Z \mid X)$). Training is done by contrast (contrastive learning). It is considered that the vector X of an example and the predicted distribution of its labels, Y, are true, and the other N-1 label distributions from the training batch are false.

4.4.4 Source to target: interrelated graph

The Cross-Domain Aspect Label Propagation through Heterogeneous Networks (CD-ALPHN) model by Marcacini et al. (2018) was singled out in a separate topic for exclusively employing graph techniques to propagate labels in a transductive model. A heterogeneous graph is constructed, linking words from the source and target domains to their syntactic dependency relations. Labels are propagated, classifying them as "1" or "0", i.e., whether they are aspects or not. This model is conditioned on the prior existence of all data to be classified.

4.4.5 Source to target: instance

Sun et al. (2023) highlight that an adversarial network aligns features but ignores labels, leading to suboptimal results (Eq. 2). They observed that feature alignment causes some labels to change in the classifier applied to the target domain (Fig. 16). Therefore, they developed the Self-training through Classifier Disagreement (SCD) student-teacher model for aspect extraction. Firstly, the model is trained with source domain data. In the second stage, the teacher is kept fixed while the student learns an adversarial network and pseudo-labels of target domain data, generated by the student itself. However, the weights assigned

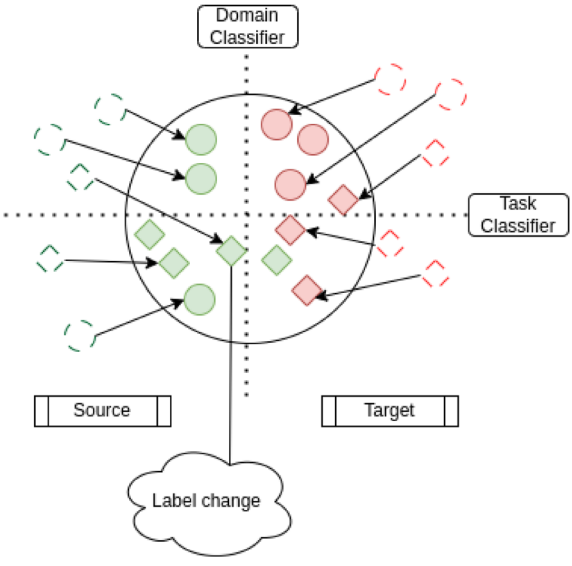

Fig. 16 Label change after feature alignment. Adapted from Sun et al. (2023)

to the error related to pseudo-labels are different if there is a label swap between the student being trained and the fixed teacher.

4.4.6 Source to target: train data generator

Some models have chosen to generate new labeled texts for training, such as the Cross-Domain Generative Data Augmentation framework (CDGDA). Xue et al. (2023) built a model that generates extra sentences for ABSA to classify polarity. A pre-trained T5 model is fine-tuned using unlabeled sentences from the source data, the aspect, and occasionally, the labeled polarity of the target data to generate the original sentence of the target data. Then, the parameters of the generator model are frozen. Four sentences are generated by passing a sentence from the source domain, some aspect from the target, and optionally, a polarity. The sentences are filtered using an entropy filter, sorted from highest to lowest entropy, and the first one(s) are selected. In the final step, the generated and annotated sentences are combined to train the BERT classification model.

Yu et al. (2021) proposed the Cross-Domain Review Generation (CDRG) to generate training examples for aspect extraction and sentiment classification. The algorithm starts by identifying aspect and opinion terms in the domains. To accomplish this in the unlabeled target domain, the syntactic rule-based algorithm Double Propagation is used (Qiu et al. 2011). Then, unique terms from the sentences of each domain are isolated and masked by a [MASK]. In parallel, pre-training is done on a BERT model in the masked language model (MLM) task for the target domain. This pre-trained BERT is applied to each masked sentence to predict aspects and sentiments, thus constructing sentences for the target domain. Finally, all source domain sentences containing aspects and sentiments common to both domains, along with the generated sentences, are used for training.

Li et al. (2022) proposed the Generative CrossDomain Data Augmentation Framework (GCDDA) to expand target domain data using the source. The authors considered labels important in the process of generating examples. First, a BERT with CRF is trained on the source domain to predict aspects and opinion terms in the BIO scheme with polarity. Then, *pseudo-labels* are created on the target using this model. Meanwhile, domain-specific terms are discovered in each domain using a frequency metric. To train the model, annotated source and pseudo-annotated target data are used as input to a BART model's encoder. In this model, specific words are masked, and labels are passed as embedding. The BART decoder is used to predict the masked tokens (words) and labels. In addition to the data processed by the encoder, a value discriminating between source and target domains is passed, as appropriate. During inference, examples from the source with masked specific words are passed to the encoder, and the target domain value is fixed in the decoder. The generated sentences are filtered by a consistency series to finally be used to train a classifier.

Deng et al. (2023) created the *Bidirectional Generative Cross-domain ABSA* (BGCA) model. First, a model like T5 or BART is trained to generate texts explicitly stating labeled source data's aspects, opinion terms, and polarity. The output format includes tags indicating the element type (aspect, opinion term, or polarity) and the respective words. Once trained, during inference, possible predictions considering words not in the text or not indicative tags of the elements are discarded. Then, the model is trained in the opposite direction: generating texts of reviews from the found aspects, opinion terms, and polarity. Finally, the model generates examples for the target domain and thus obtains data augmentation. The

ABSA elements are extracted from the target domain. It passes the extracted terms to generate phrases. It performs some consistency validations with the generated phrases, analyzing whether it generated phrases without the passed terms, invalid phrases, whether all the passed terms were used, and only the passed terms. In the latter case, it passes the generated sentence back to identify the ABSA elements and verify consistency. The model is retrained using the generated phrases and labeled data from the source domain.

The Domain-Adaptive Language Modeling model (DA^{2}LM) is used for aspect extraction and classification (Yu et al. 2023). It consists of three stages. The first stage's goal is to generate pseudo-labels for the target domain data. Aspects of the target domain are annotated using the syntactic rule-based algorithm Double Propagation (Qiu et al. 2011). Then, reviews from the source and target domains undergo BERT to generate contextualized vectors. Maximum Mean Discrepancy (MMD) is applied between the source and target domain aspects to make them invariant. CRF is used to train the classification model. This model is used to generate pseudo-labels in the target. In the second stage, an LSTM or GPT-2 model is trained to generate annotated sentences for the target domain. The training of each example involves passing information on whether the sentence is from the source or target domain and taking the current word and the sentiment of the previous word as input to generate the next word and the sentiment of the current word. With the model trained, we move on to the final stage. Sentences are generated for the target domain. At each step, a word is chosen among the top-k as the next word, and the predicted label by the model is considered. Consistency is applied to remove generated sentences with nonsensical label sequences. Additionally, the sentence is passed through the initially trained classifier, and sentences with conflicting labels are removed. Finally, a BERT with CRF is trained.

4.4.7 Source to target: multi-task

Zheng et al. (2020) created two models that utilize the auxiliary task "Z" of Document Sentiment Classification (DSC) for the task "Y" of classifying the sentiment of aspects in another domain (ASC). The first model, Anchored Model Transfer (AMT), has a common submodel for both tasks called submodel "A." The output of "A" is then fed into two separate submodels: submodel "B," used for the ASC task, and submodel "C," used for the DSC task. The hypothesis is that "A", common to both tasks, should be similar, serving to map the characteristics of the input data. To allow for a slight variation, the parameters of model A will be given by $\theta + \Delta Y$ and $\theta + \Delta Z$ for ASC (Y) and DSC (Z) models, respectively. The parameters ΔY and ΔZ undergo L2 regularization in the cost function, meaning they cannot deviate too much from θ. Since they are similar tasks, the generated characteristics should be close.

The second model, Soft Instance Transfer (SIT), is a pseudo-labeling model. The error of the cost function of labeled data for "Y" is minimized. In parallel, the data from the DSC task is applied to the ASC classifier. The obtained labels are used to minimize the error of the ASC classifier on this data, meaning the cost function is between $P(\hat{y})$ and $\hat{y}$, proportional to the confidence level in the predicted label. For example, if the model believes the prediction to be 95% correct, the cost in the model's error function will be higher than if the model has 55% confidence. By not completely discarding examples with low confidence levels, it tends to perform better when the domains are very distinct, as tested by the authors.

Finally, the authors considered training both models simultaneously, with pseudo-labels generated for both tasks in the second model.

Huang et al. (2023) proposed the Transfer Learning with Document-level Data Augmentation (TL-DDA) model. This model is trained for the task of document-level sentiment classification in one domain to enhance the sentiment classification of aspects in another domain. For the document classification task training, small reviews are selected. In addition to them, new documents are generated by concatenating one with positive sentiment and another with negative sentiment. During training, words are converted into GloVe or BERT vectors and passed through a Bi-LSTM network for contextual representations. Then, using an attention mechanism, the model learns to classify the original documents and individually the two concatenated documents. According to the authors, this attention mechanism emphasizes sentiment words. To classify aspects, the reviews are passed to this model and a second model. This second model first learns contextual word vectors. These vectors then go through an attention mechanism regarding the aspect along with the learned mechanism to emphasize sentiment words in document(s) classification. This mechanism that emphasizes sentiment words in aspect classification considers its distance to each word using a syntactic dependency tree. Finally, the aspect is classified.

Li et al. (2019b) created the Multi-Granularity Alignment Network (MGAN) model. The model uses data with labeled sentiment at the category level from the source domain to enhance the sentiment classification model at the aspect level in the target domain. For this, separate submodels were created for the source and target domain data. The submodels consist of Bi-LSTM networks and attention mechanisms to obtain contextual representations of reviews, categories, and aspects. In the next step, only the source domain submodel uses an attention mechanism to generate a vector for the fictitious task of classifying the category itself. The authors argue that the attention mechanism of this task points to the likely aspect words, achieving alignment with the target domain. As the aspect is not always explicit in the review in the source domain, the generated vector goes through a Gate Unit, which may give more weight to the contextualized category vector before the attention mechanism is applied. In the last step of the submodels, a final attention mechanism is applied that considers the distance of each word to the aspect. Closer words are usually more important for the respective aspect. In the case of the source domain, which does not have the exact distance, the submodel considers the weight given to each word in the attention mechanism used in category classification. Finally, sentiment classifications are made by each submodel. The union of the submodels is done by Contrastive Feature Alignment (CFA). The goal is to approximate the representations of the final vectors of the source and target domains if they have the same label or to distance them if they have different labels.

The PRE+MULT (He et al. 2018) and Transcap (Chen and Qian 2019) models follow the same principle: the auxiliary task of sentiment classification in documents helps aspect classification. PRE+MULT has two phases. In the first phase, an LSTM network is pre-trained to classify documents. In the second phase, this network is used to classify aspects and documents. The TransCap uses capsule networks to perform both tasks simultaneously (Hinton et al. 2011). Capsule networks resemble CNNs, but instead of the convolutional operator returning a scalar value, it returns vectors representing various features and their intensity.

4.4.8 Source to target: active learning

The method by Bhattacharjee et al. (2021) utilizes active learning to train a cross-domain aspect sentiment classification model. First, the model is trained with labeled data from the source domain. Then, documents from the target domain are selected based on two criteria:

1. Documents that are difficult to classify. The choice is based on examples with the highest class entropy (higher entropy means less informative distribution, i.e., more uniform), and
2. Adjectives that are frequent in the target but not in the source domain.

4.5 ABSA: multidomain

In this group are the models where there is more than one target domain, and each domain acts as the source for the other domain. This group was not subdivided because the number of works found in the search was small.

The first one to be described is the model by Lark et al. (2018), used to extract aspects. They propose that besides indicating which words are aspects in the labeled training data of various domains, the category to which they belong should also be annotated. The theory is that it is not the domains of the data that are different but the categories. For example, reviews of hotels, museums, and restaurants may all discuss the "location" category. Using a CRF model, the two approaches are compared: (i) predicting what an aspect is only and (ii) predicting what an aspect is and to which category it belongs. The authors found that in the second approach, the gain after training the model is more significant when using various domains in the training. In their tests, the authors compared category by category and found that the most significant gains were in the categories common to the domains.

Chernyshevich (2014) trained a CRF model. Various word features (the word itself, most common sentiment of the word in a set of documents, POS-tagging, etc.), relationships with nouns, and semantics were included. The authors demonstrated that the model trained on a mixture of data from the source and target domains did not lose power. The good performance may be related to the considered invariant characteristics, such as syntactic relationships.

The model by Chauhan et al. (2020) can be considered an unsupervised aspect extraction model since no labeled data exists. However, this model was included because there is knowledge sharing among unlabeled domains, making it a cross-domain model. The algorithm generates pseudo-labels for the involved domains considering syntactic rules and then, using Bi-LSTM networks and an attention mechanism that considers the domain, extracts aspects in the BIO scheme.

The CLASSIC model is used for multidomain sentiment classification (Ke et al. 2021). The authors built a model that leverages learning from various tasks while avoiding catastrophic forgetting. Catastrophic forgetting occurs when training the model for a new task negatively interferes with the learning of another task learned so far. The model consists of a pre-trained BERT with frozen weights. The layers of this BERT encoder are modified by adding dense networks that will be learned. Each neuron in this network has a Gate Unit. Each trained task, or each trained domain, will associate "0" and "1" values to the Gate Unit. Neurons with "1" values are used in the task in question and cannot be forgotten; thus, these

neurons are frozen at each training of a new task. The various models are combined in the final model using contrastive learning. The cost function of an example in a training batch should have a closer classification between the current model and the previous ones than the other training examples in the same batch. Contrastive learning is also performed between the output of BERT for the current task and the other tasks for which a weighted average is considered, taking into account the similarity between them. The outputs of examples with the same labels between the current task and this composition of the other tasks should be more similar than those with different labels. A final contrastive learning is performed considering the labels of the training examples, in addition to cross-entropy for the current task.

Another model in this topic is by Tran et al. (2021). In this model, each word in a review is transformed into a vector composed of GloVe representation and the POS-tagging of the vector. These vectors go through a composition of CNN with Bi-LSTM networks to obtain the domain, aspect, and sentiment. The use of syntactic information favors multidomain capabilities.

The multidomain model BSp_{LLA} by Ramos and Fuentes (2023) considers possible aspects using POS-tagging and classifies them. Using BERT, the model is trained across multiple domains continuously (continual learning). A regularization of the parameters between the current model and the previous model is considered to avoid catastrophic forgetting.

Finally, the Continual Adapter Tuning for aspect sentiment classification (CAT) model is a multidomain model also based on continual learning (Chen et al. 2024a). An *Adapter*-BERT layer (Houlsby et al. 2019) is created for each new domain. The *Adapter*-BERT is a frozen BERT with attached parallel layers to be learned. In addition to the sentence and the aspect, all possible labels (e.g., positive and negative) are passed as tokens in the input. The authors consider that the semantics of the labels contribute to the classifier. The classifier considers the output of the CLS token and the outputs related to the tokens of the labels passed as input. Using contrastive learning, the model training aims to (i) maximize the inner product between the CLS token and the correct label token, and (ii) minimize the inner product with the other labels. It aims to (iii) maximize the relationship between the CLS token of an example and the respective label token of other examples if they have the same classification, and (iv) minimize it otherwise. Finally, it aims to (v) maximize the correct label token relative to the others for each example. To use the knowledge from other models learned so far, the authors chose three strategies to initialize the parameters for each new trained domain: use the last model, use a random one, or use the one with the best performance in validation on a dataset. Finally, during testing, it is unknown to which domain the example belongs. Thus, all models are applied, and a vote is made to decide the correct label.

5 Summary of cross-domain ABSA models

This section provides a summary of some points from the articles identified during this systematic review. It presents overviews of the works classification by the task, techniques, sentiment polarity granularity, model outputs, labeling of data from the source and target domains, a source code list for the methods, and a preliminary comparison among a few methods.

5.1 Cross-domain ABSA classification results considering the task

In Sect. 2.2, we introduced two approaches to classify the reviewed documents. The results for the proposed method, which focuses on how each work performs cross-domain adaptation, were presented in Sect. 4. In this subsection, we present the results for the other approach, the more standard approach, which classifies cross-domain ABSA works based on the specific task they address (Sect. 2.2.1). The results are summarized in Table 12. No studies performing ASQP for cross-domain were found.

5.2 Summary of techniques used for cross-domain ABSA

The described techniques are used in the articles' models as shown in Table 13. This is only a subset of them, which are listed here because they are used in more than one article. Although the articles using BERT/BART/GPT/T5 indirectly use the attention mechanism, the "attention mechanism" column will only be filled if it is mentioned as a separate mechanism in the model. The same applies to the Gate Unit in relation to recurrent networks.

This review provides an overview of the evolution of the use of these techniques over the years in the application of ABSA models in cross-domain, allowing us to infer which ones are used by state-of-the-art models and which are being abandoned or underexplored. Figure 17 illustrates the evolution of each technique individually. Each bar shows the percentage of articles that used a particular technique over a period.

The first observation is that language models (LMs) have increased participation in recent years. After 2022, approximately 85% of the listed articles used a LM in some form in their models, making them a fundamental component in many state-of-the-art models.

A second observation is that the use of external resources, such as lexicons and grammatical and semantic parsers, has been and continues to be widespread, with half of the authors applying them in some way in their models over the years. However, the paradigm has shifted. When crossed with the graph showing the use of LMs, it can be observed that these resources have transitioned from being a primary tool to enriching the models. Another commonly employed technique is attention mechanisms, which have followed the growth of LMs. In recent years, approximately 42% of the articles used some form of attention mechanism in their models. It is important to note that even models that did not directly use attention mechanisms may have indirectly utilized them, as attention mechanisms are a crucial component of LM models.

The use of recurrent networks, LSTM, Bi-LSTM, GRU, etc., has remained constant. The same is true for the Gate Unit, an essential component of these networks. Many articles describe models composed of combinations of LMs and recurrent networks. Regarding graphs, the "U" shape of the graph may indicate a paradigm shift where models have transitioned from using simple graphs for label propagation to incorporating them into neural networks. Techniques such as adversarial networks and pseudo-labeling were applied in the construction of approximately a quarter of the models, becoming essential mechanisms in this activity. Distance measures are an alternative to approximate invariant vector representations between domains, complementing techniques such as adversarial networks.

CRF was widely used in the early solutions of cross-domain ABSA, but its usage has declined in the past five years. However, the graph in Fig. 17 illustrates that it is still in use. Again, a supposition is a paradigm shift. The early models used pure CRF, while the newer

Table 12 Classification of the articles according to the task performed

Article	ACD	ASC	AE	OTE	AOPE	E2E-ABSA	ACSA	ASTE	ACSD
Anand and Mampilli (2021)			x						
van Berkum et al. (2022)		x							
Bhattacharjee et al. (2021)		x							
Cao et al. (2021)		x							
Chauhan et al. (2020)			x						
Chen and Qian (2019)		x							
Chen and Qian (2021)			x						
Chen and Qian (2022)						x			
Chen and Wan (2022)		x	x						
Chen et al. (2024a)		x							
Chernyshevich (2014)			x						
De Clercq et al. (2017)									x
Deng et al. (2023)								x	
Ding et al. (2017)			x						
Gong et al. (2020)						x			
He et al. (2018)		x							
Howard et al. (2022)			x						
Hu et al. (2019)		x							
Hu et al. (2022)		x							
Huang et al. (2023)		x							
Jakob and Gurevych (2010)			x						
Jiang et al. (2019b)	x								
Jiang et al. (2024)		x							
Kan and Chang (2022)		x							
Kannan et al. (2023)				x					
Ke et al. (2021)		x							
Kiritchenko et al. (2014)	x	x	x						
Klein et al. (2022)			x						
Knoester et al. (2023)		x							
Lark et al. (2018)	x[1]		x						
Lee et al. (2023)		x							
Li et al. (2012)					x				
Li et al. (2019a)						x			
Li et al. (2019b)		x							
Li et al. (2022)					x				
Liang et al. (2022)			x						
Liu and Zhao (2022)		x							
Liu et al. (2021)							x		
Liu et al. (2023)		x							
Majumder et al. (2022)		x							
Marcacini et al. (2018)			x						
Ouyang and Shen (2023)			x			x			
Pak and Gunal (2022)			x						
Pastore et al. (2021)			x						
Pereg et al. (2020)					x				
Ramos and Fuentes (2023)		x							
Ruskanda et al. (2019)					x				
Santos et al. (2021)			x						

Table 12 (continued)

Article	ACD	ASC	AE	OTE	AOPE	E2E-ABSA	ACSA	ASTE	ACSD
Shi et al. (2023)			x						
Sun et al. (2023)			x						
Toledo-Ronen et al. (2022)						x			
Tran et al. (2021)						x			
Wang and Pan (2018)					x				
Wang and Pan (2019a)					x				
Wang and Pan (2019b)					x				
Wang et al. (2018)								x	
Xu et al. (2018)			x						
Xu et al. (2019)		x	x						
Xu et al. (2020)						x			
Xue et al. (2023)		x							
Yang et al. (2020)						x[2]			
Yang et al. (2021)		x							
Yang et al. (2023)		x	x[3]						
Yu et al. (2021)						x			
Yu et al. (2023)						x			
Zeng et al. (2023b)						x			
Zhang et al. (2023a)		x							
Zhang et al. (2023b)		x							
Zhang et al. (2023c)		x							
Zhao et al. (2022)		x							
Zhao et al. (2024)		x							
Zheng et al. (2020)		x							
Zhou et al. (2020)		x							
Zhou et al. (2021b)						x			

[1]Indirectly classifies the category

[2]Generates a summary considering both the aspect and the sentiment

[3]Extracts aspects, although this is not its primary focus

models use it to supplement neural network models. LDA and autoencoder are used, but are not among the most popular techniques.

5.3 Summary of sentiment polarity granularity

Sentiment polarity can be positive, negative, or neutral. While most authors consider this classification, some others have chosen to classify the sentiment of aspects or categories as either positive or negative only. Additionally, a few authors also considered the polarity of opinion terms. For example, consider the sentence: "I would say that the new keyboard is both good and bad at the same time." The term *keyboard* is associated with two opinion terms, *good* and *bad*, which have positive and negative polarities, respectively. These few authors approached such terms in two ways: (i) some considered a fourth sentiment possibility: *conflicting*, and (ii) others have chosen to label the polarity of the opinion term rather than the sentiment. The various ways of categorizing sentiment polarity are illustrated in Table 14.

Table 13 Technologies Present in Each Article Found in the Systematic Review

Article	External Resources	Gate Unit	Recurrent Networks	Attention Mechanism	CRF	BERT/BART/GPT/T5	LDA	Graph	Adversarial Network	Distance Measures	Pseudo-labeling	Autoencoder
Anand and Mampilli (2021)	x											
van Berkum et al. (2022)			x	x		x						
Bhattacharjee et al. (2021)	x		x									x
Cao et al. (2021)						x				x	x	
Chauhan et al. (2020)	x		x	x					x		x	
Chen and Qian (2019)		x										
Chen and Qian (2021)	x	x							x			
Chen and Qian (2022)	x	x		x		x			x			
Chen and Wan (2022)						x						
Chen et al. (2024a)						x						
Chernyshevich (2014)	x				x							
De Clercq et al. (2017)	x				x							
Deng et al. (2023)						x						
Ding et al. (2017)	x		x								x	

Table 13 (continued)

Article	External Resources	Gate Unit	Recurrent Networks	Attention Mechanism	CRF	BERT/BART/GPT/T5	LDA	Graph	Adversarial Network	Distance Measures	Pseudo-labeling	Autoencoder
Gong et al. (2020)	x					x						
He et al. (2018)			x	x								
Howard et al. (2022)	x			x		x		x			x	
Hu et al. (2019)			x	x					x			
Hu et al. (2022)		x	x	x						x		
Huang et al. (2023)	x		x	x		x						
Jakob and Gurevych (2010)	x				x							
Jiang et al. (2019b)							x	x				x
Jiang et al. (2024)						x	x	x				x
Kan and Chang (2022)						x			x			
Kannan et al. (2023)			x	x								
Ke et al. (2021)		x		x		x			x			
Kiritchenko et al. (2014)	x											
Klein et al. (2022)	x					x					x	
Knoester et al. (2023)			x	x		x			x			
Lark et al. (2018)	x				x						x	

Table 13 (continued)

Article	External Resources	Gate Unit	Recurrent Networks	Attention Mechanism	CRF	BERT/BART/GPT/T5	LDA	Graph	Adversarial Network	Distance Measures	Pseudo-labeling	Autoencoder
Lee et al. (2023)			x	x		x						
Li et al. (2012)	x				x			x			x	
Li et al. (2019a)	x		x	x					x			
Li et al. (2019b)		x	x	x						x		
Li et al. (2022)					x	x					x	
Liang et al. (2022)			x	x	x					x		
Liu and Zhao (2022)		x				x			x			
Liu et al. (2021)						x						
Liu et al. (2023)	x					x		x			x	
Majumder et al. (2022)			x		x							
Marcacini et al. (2018)	x							x				
Ouyang and Shen (2023)	x					x					x	
Pak and Gunal (2022)	x											
Pastore et al. (2021)			x			x						
Pereg et al. (2020)	x					x						
Ramos and Fuentes (2023)	x					x						

Table 13 (continued)

Article	External Resources	Gate Unit	Recurrent Networks	Attention Mechanism	CRF	BERT/BART/GPT/T5	LDA	Graph	Adversarial Network	Distance Measures	Pseudo-labeling	Autoencoder
Ruskanda et al. (2019)	x											
Santos et al. (2021)						x						
Shi et al. (2023)	x					x						
Sun et al. (2023)			x			x			x		x	
Toledo-Ronen et al. (2022)	x					x					x	
Tran et al. (2021)	x		x									
Wang and Pan (2018)												
Wang and Pan (2019a)	x		x						x			x
Wang and Pan (2019b)			x	x					x			
Wang et al. (2018)	x				x							
Xu et al. (2018)												
Xu et al. (2019)						x						
Xu et al. (2020)						x						
Xue et al. (2023)						x						
Yang et al. (2020)	x		x	x			x					
Yang et al. (2021)	x		x	x			x		x			

Table 13 (continued)

Article	External Resources	Gate Unit	Recurrent Networks	Attention Mechanism	CRF	BERT/BART/GPT/T5	LDA	Graph	Adversarial Network	Distance Measures	Pseudo-labeling	Autoencoder
Yang et al. (2023)	x			x		x		x			x	
Yu et al. (2021)	x					x						
Yu et al. (2023)	x		x		x	x				x		
Zeng et al. (2023b)	x			x		x	x	x				x
Zhang et al. (2023a)	x		x	x		x		x				
Zhang et al. (2023b)				x		x			x	x		
Zhang et al. (2023c)				x		x			x			
Zhao et al. (2022)						x						
Zhao et al. (2024)	x		x	x		x						
Zheng et al. (2020)			x								x	
Zhou et al. (2020)			x	x								
Zhou et al. (2021b)			x						x		x	

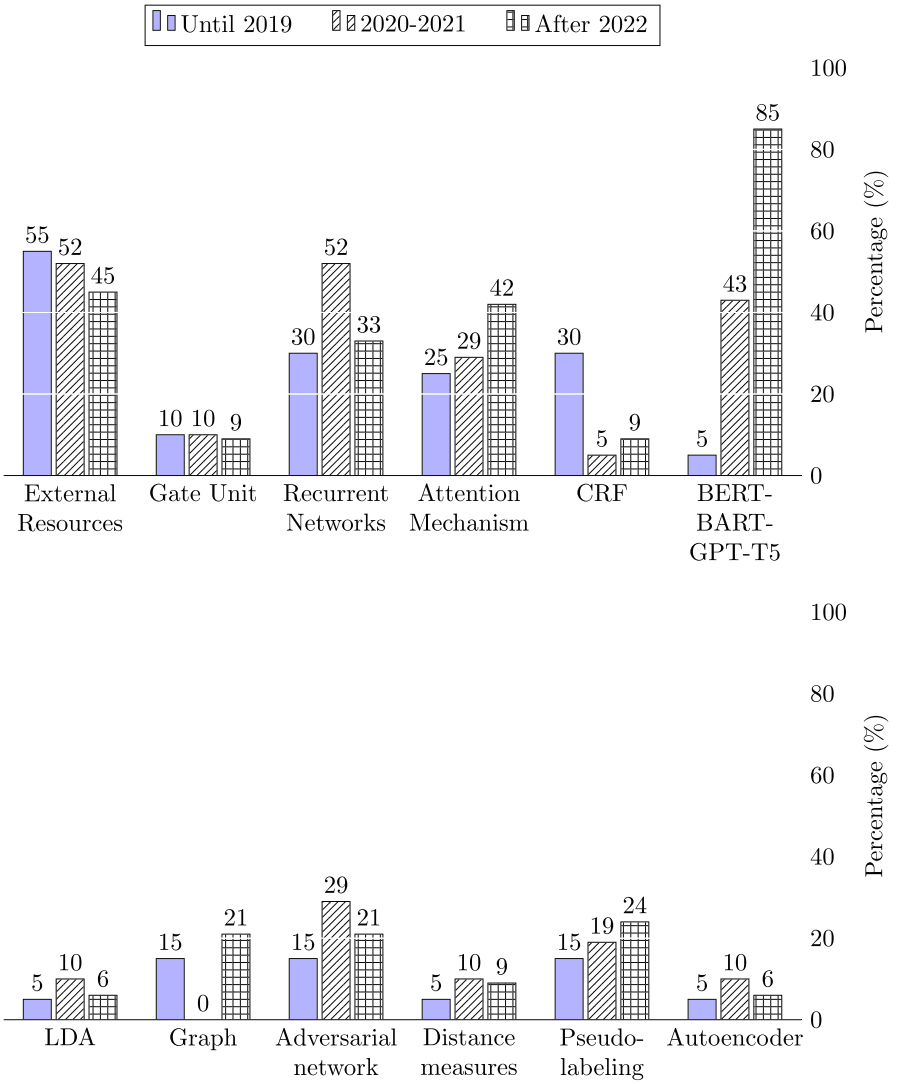

Fig. 17 Percentage of articles that used a particular technique in ABSA models in cross-domain over the years

5.4 Summary of model outputs

A model that exclusively performs sentiment classification of an aspect or category can either generate a word representing the sentiment (positive, negative, or neutral) or classify it. Conversely, a model that performs aspect or opinion term extraction can do so in various ways, as illustrated in Table 15. Generally, models follow the approaches outlined in Sect. 2.3.2: (i) generating words for aspects or opinion terms, (ii) applying morphological, syntactic, or semantic rules, or (iii) executing a BIO classification. It is important to note that Table 15 depicts the schema rather than the exact output. For instance, Marcacini et al. (2018)

Table 14 Summary of Sentiment Polarity Granularity

Article	Task	Granularity
van Berkum et al. (2022)	ASC	POS/NEG/NEU
Bhattacharjee et al. (2021)	ASC	POS/NEG/NEU
Cao et al. (2021)	ASC	POS/NEG/NEU
Chen and Qian (2019)	ASC	POS/NEG/NEU
Chen and Qian (2022)	E2E-ABSA	POS/NEG/NEU
Chen and Wan (2022)	ASCAE	POS/NEG/NEU
Chen et al. (2024a)	ASC	POS/NEG/NEU
Deng et al. (2023)	ASTE	POS/NEG/NEU
Gong et al. (2020)	E2E-ABSA	POS/NEG/NEU
He et al. (2018)	ASC	POS/NEG/NEU
Hu et al. (2019)	ASC	POS/NEG/NEU
Hu et al. (2022)	ASC	POS/NEG/NEU
Huang et al. (2023)	ASC	POS/NEG/NEU
Jiang et al. (2024)	ASC	?[1]
Kan and Chang (2022)	ASC	POS/NEG/NEU
Ke et al. (2021)	ASC	POS/NEG/NEU
Kiritchenko et al. (2014)	ACD/ASC/AE	POS/NEG/NEU/CONFLICT
Knoester et al. (2023)	ASC	POS/NEG/NEU
Lee et al. (2023)	ASC	POS/NEG/NEU
Li et al. (2019a)	E2E-ABSA	POS/NEG/NEU
Li et al. (2019b)	ASC	POS/NEG/NEU
Liu and Zhao (2022)	ASC	POS/NEG
Liu et al. (2021)	ACSA	POS/NEG/NEU
Liu et al. (2023)	ASC	POS/NEG/NEU
Majumder et al. (2022)	ASC	POS/NEG/NEU
Ouyang and Shen (2023)	AE/E2E-ABSA	POS/NEG/NEU
Ramos and Fuentes (2023)	ASC	POS/NEG/NEU
Toledo-Ronen et al. (2022)	E2E-ABSA	POS/NEG
Tran et al. (2021)	E2E-ABSA	POS/NEG/NEU
Wang et al. (2018)	ASTE	OTE $\rightarrow$ POS/NEG[2]
Xu et al. (2019)	ASC/AE	POS/NEG/NEU
Xu et al. (2020)	E2E-ABSA	POS/NEG/NEU
Xue et al. (2023)	ASC	POS/NEG/NEU
Yang et al. (2021)	ASC	POS/NEG/NEU
Yang et al. (2023)	ASC/AE	POS/NEG/NEU
Yu et al. (2021)	E2E-ABSA	POS/NEG/NEU
Yu et al. (2023)	E2E-ABSA	POS/NEG/NEU
Zeng et al. (2023b)	E2E-ABSA	POS/NEG/NEU
Zhang et al. (2023a)	ASC	POS/NEG/NEU
Zhang et al. (2023b)	ASC	POS/NEG/NEU
Zhang et al. (2023c)	ASC	POS/NEG/NEU
Zhao et al. (2024)	ASC	POS/NEG/NEU
Zhao et al. (2022)	ASC	POS/NEG/NEU
Zheng et al. (2020)	ASC	POS/NEG/NEU
Zhou et al. (2020)	ASC	POS/NEG/NEU
Zhou et al. (2021b)	E2E-ABSA	POS/NEG/NEU

[1]The granularity level is not precise

[2]The sentiment polarity is annotated in opinion terms

Table 15 Summary of model outputs

Article	Task	Output
Anand and Mampilli (2021)	AE	Rules (Tagging)
Chauhan et al. (2020)	AE	BIO
Chen and Qian (2021)	AE	BIO
Chen and Qian (2022)	E2E-ABSA	BIO + Polarity
Chen and Wan (2022)	ASCAE	BIO
Chernyshevich (2014)	AE	FA, FH, FPA, O
De Clercq et al. (2017)	ACSD	AE = BIO
Deng et al. (2023)	ASTE	Generative: <aspect-pos/neg/neu>... <opinion>..
Ding et al. (2017)	AE	BIO
Gong et al. (2020)	E2E-ABSA	BIO+Polarity
Howard et al. (2022)	AE	BIO
Jakob and Gurevych (2010)	AE	BIO
Kannan et al. (2023)	OTE	BIO
Kiritchenko et al. (2014)	ACD/ASC/AE	AE = Target-Outside (T-O)
Klein et al. (2022)	AE	BIO
Lark et al. (2018)	ACD/AE	<Category>-Outside
Li et al. (2012)	AOPE	Target (0=aspect, 1=opinion term, 2=none)
Li et al. (2019a)	E2E-ABSA	B, I, E, S, O + Polarity
Li et al. (2022)	AOPE	BA, IA, BO, IO, N
Liang et al. (2022)	AE	BIO
Marcacini et al. (2018)	AE	Target-Outside (T-O)
Ouyang and Shen (2023)	AE/E2E-ABSA	BIO+Polarity
Pak and Gunal (2022)	AE	Rules (Tagging)
Pastore et al. (2021)	E2E-ABSA	BIO
Pereg et al. (2020)	AOPE	BA, IA, BO, IO, N
Ruskanda et al. (2019)	AOPE	Rules (Tagging)
Santos et al. (2021)	AE	BIO
Shi et al. (2023)	AE	BIO
Sun et al. (2023)	AE	BIO
Toledo-Ronen et al. (2022)	E2E-ABSA	O-P-N
Tran et al. (2021)	E2E-ABSA	BIO+Polarity
Wang and Pan (2018)	AOPE	BA, IA, BO, IO, N
Wang and Pan (2019a)	AOPE	BA, IA, BO, IO, N
Wang and Pan (2019b)	AOPE	BA, IA, BO, IO, N
Wang et al. (2018)	ASTE	FB, FI, PB, PI, CB, CI, N, O
Xu et al. (2018)	AE	BIO
Xu et al. (2019)	ASC/AE	BIO
Xu et al. (2020)	E2E-ABSA	BIO / BIO+Polarity
Yang et al. (2023)	ASC/AE	AE = Target-Outside (T-O)
Yu et al. (2021)	E2E-ABSA	BIO+Polarity
Yu et al. (2023)	E2E-ABSA	BIO+Polarity
Zeng et al. (2023b)	E2E-ABSA	BIO+Polarity
Zhou et al. (2021b)	E2E-ABSA	BIO+Polarity

label words with "0" and "1", which is similar to the *Target*—O/T schema represented in the table. In particular, the generative schema can occur in various forms. For example, the output of Deng et al. (2023) consists of words preceded by tags indicating the aspect with the respective sentiment or opinion terms, as applicable.

There are a few variations of the described schemas, such as those occurring in the BIO schema. For example, Li et al. (2019a) created a more complex version of the BIO schema incorporating polarity. They classify each word with the tags B-I-E-S-O, representing: beginning of, inside of, end of, single-word, and no aspect term, along with sentiment tags POS, NEG, and NEU. Another variant was introduced by Chernyshevich (2014), who created the FA-FH-FPA-O schema for extracting aspects. In this schema, FH denotes the head of a group of words forming an aspect; FA and FPA represent the preceding and following words of this group, respectively; and O for words that do not form an aspect. This approach is suggested to enable the model to generate tags consistently, unlike the BIO schema. They present the example: *camera* vs *compact camera*. In the suggested example, *camera* will receive the FH tag in both cases, unlike the BIO schema, which would assign B and I tags, respectively.

Wang et al. (2018) classify the sentiment of opinion words instead of aspects. Thus, they adapted the BIO+polarity schema, creating the schema: FB-FI-PB-PI-CB-CI-N-O, described as follows:

- FB: The beginning of topic words,
- FI: The midst of topic words,
- PB: The beginning of positive sentiment words,
- PI: The midst of positive sentiment words,
- CB: The beginning of negative sentiment words,
- CI: The midst of negative sentiment words,
- N: Negative adverbs, and
- O: Other words

Variants of the Outside-Target (O-T) schema, which is a simplified version of the BIO schema, have also been identified. Li et al. (2012) adapted the Outside-Target schema to include the extraction of opinion terms. Toledo-Ronen et al. (2022) included positive and negative sentiment, resulting in the O–P–N schema, i.e., outside, positive, and negative.

The BIO schema can lead to inconsistencies. For example, the model might assign the tag "I" (inside) to a word that follows another word tagged as "O". Models can address such situations by changing these occurrences to "B" (begin). Similarly, a model that classifies words in BIO+polarity might assign conflicting sentiments to words. Therefore, algorithms performing sentiment extraction and classification with the BIO+polarity schema need to define a protocol. Chen and Qian (2022) decided that when a selected aspect consists of more than one word, the sentiment of the first word prevails. In the case of the O–P–N tags, Toledo-Ronen et al. (2022) consider that tags with different sentiments represent different aspects.

5.5 Labeling of source and target domain data for models

Each solution presented in this review considers the data from the source and target domains labeled differently. Table 16 illustrates the labeling situation for each domain according to the presented paper. Models classified as *independent* (Sect. 2.2.2) should not have labeled data in the target domain. Models classified as *multidomain* generally have labeled data for both the source and target domains. An exception is the model by Chauhan et al. (2020), which does not have labeled data in either domain. *Multitask* models, from the Target → Source group, have labeled data from the source domain for tasks different from those in the target domain.

5.6 Source code repositories

Several of the methods included in this review provide access to their source code, which is essential for ensuring reproducibility, enabling comparative experiments, and promoting further research. To identify the available implementations, we systematically examined the articles for explicit links to code repositories, such as GitHub, and also consulted supplementary materials, including appendices, author webpages, and academic or technical blogs. Table 17 summarizes the methods for which source code was found, the task objective, along with the corresponding repository links.

5.7 Brief summary for cross-domain ABSA performance

To indicate the average performance of these methods in ABSA, we conducted preliminary experiments in a cross-domain setting, focusing on aspect extraction and classification (E2E-ABSA). The selection of methods was guided by criteria such as the availability of source code and reported performance in the literature. Although this is a preliminary study, a more comprehensive evaluation remains a subject for future research. While existing works often include comparisons with other methods, such comparisons may be subject to biases depending on the experimental setup or evaluation criteria.

The selected datasets–restaurants (R), *laptops* (L), devices (D), and services (S) (see Sect. 3.6)–were used in a cross-domain setting for aspect extraction and classification (E2E-ABSA). Cross-combinations between *laptops* and devices were excluded due to the similarity in their data distributions.

The models selected for the preliminary experiments were chosen based on the availability of source code and reported performance. The evaluated models are:

- $AD-SAL$ (Li et al. 2019a)—Sect. 4.3.3.
- $AD-AL$ (Li et al. 2019a)—similar to AD-SAL, but without the adversarial network used to align aspects across domains—Sect. 4.3.3.
- $CDRG-BERT_B$(Yu et al. 2021)—a model that generates sentences using the masked language modeling (MLM) task, based on BERT (Devlin et al. 2019)—Sect. 4.4.6.
- $CDRG-BERT_E$ (Yu et al. 2021)—similar to $CDRG-BERT_B$, but using BERT-PT (Xu et al. 2019) as the base model—Sect. 4.4.6.
- *BGCA* (Deng et al. 2023)—a model that leverages a language model for data augmentation—Sect. 4.4.6.

Table 16 Labeling of source and target domain data for models

Article	Task	Source	Target
Anand and Mampilli (2021)	AE	Labeled	Does not exist
van Berkum et al. (2022)	ASC	Labeled	Sparsely labeled
Bhattacharjee et al. (2021)	ASC	Labeled	Unlabeled
Cao et al. (2021)	ASC	Labeled	Unlabeled
Chauhan et al. (2020)	AE	Unlabeled	Unlabeled
Chen and Qian (2019)	ASC	Labeled for Document	Labeled
Chen and Qian (2021)	AE	Labeled	Unlabeled
Chen and Qian (2022)	E2E-ABSA	Labeled	Unlabeled
Chen and Wan (2022)	ASCAE	Labeled	Unlabeled
Chen et al. (2024a)	ASC	Labeled	Labeled
Chernyshevich (2014)	AE	Labeled	Labeled
De Clercq et al. (2017)	ACSD	Labeled	Does not exist
Deng et al. (2023)	ASTE	Labeled	Unlabeled
Ding et al. (2017)	AE	Labeled	Unlabeled
Gong et al. (2020)	E2E-ABSA	Labeled	Unlabeled
He et al. (2018)	ASC	Labeled for document	Labeled
Howard et al. (2022)	AE	Labeled	Unlabeled
Hu et al. (2019)	ASC	Labeled	Unlabeled
Hu et al. (2022)	ASC	Labeled	Unlabeled
Huang et al. (2023)	ASC	Labeled for document	Labeled
Jakob and Gurevych (2010)	AE	Labeled	Does not exist
Jiang et al. (2019b)	ACD	Labeled	Sparsely labeled
Jiang et al. (2024)	ASC	Labeled	Unlabeled
Kan and Chang (2022)	ASC	Labeled	Unlabeled
Kannan et al. (2023)	OTE	Labeled	Does not exist
Ke et al. (2021)	ASC	Labeled	Labeled
Kiritchenko et al. (2014)	ACD/ASC/AE	Labeled for document	Labeled
Klein et al. (2022)	AE	Labeled with aspects + Opinion terms	Does not exist
Knoester et al. (2023)	ASC	Labeled	Unlabeled
Lark et al. (2018)	ACD/AE	Labeled	Labeled
Lee et al. (2023)	ASC	Labeled	Unlabeled
Li et al. (2012)	AOPE	Labeled	Unlabeled
Li et al. (2019a)	E2E-ABSA	Labeled	Unlabeled
Li et al. (2019b)	ASC	Labeled with category sentiment	labeled with aspect sentiment
Li et al. (2022)	AOPE	Labeled	Unlabeled
Liang et al. (2022)	AE	labeled	Labeled for category
Liu and Zhao (2022)	ASC	labeled	Unlabeled
Liu et al. (2021)	ACSA	Labeled	Does not exist
Liu et al. (2023)	ASC	Labeled for document	Sparsely labeled
Majumder et al. (2022)	ASC	Labeled with aspects	Labeled
Marcacini et al. (2018)	AE	Labeled	Unlabeled
Ouyang and Shen (2023)	AE	Labeled	Unlabeled
Pak and Gunal (2022)	AE	Labeled	Does not exist
Pastore et al. (2021)	AE	Labeled	Does not exist
Pereg et al. (2020)	AOPE	Labeled	Does not exist

Table 16 (continued)

Article	Task	Source	Target
Ramos and Fuentes (2023)	ASC	Labeled	Labeled
Ruskanda et al. (2019)	AOPE	Labeled	Sparsely labeled
Santos et al. (2021)	AE	Labeled	Does not exist
Shi et al. (2023)	AE	Labeled	Unlabeled
Sun et al. (2023)	AE	Labeled	Unlabeled
Toledo-Ronen et al. (2022)	E2E-ABSA	One labeled source and other unlabeled sources	does not exist
Tran et al. (2021)	E2E-ABSA	Labeled	Labeled
Wang and Pan (2018)	AOPE	Labeled	Unlabeled
Wang and Pan (2019a)	AOPE	labeled	unlabeled
Wang and Pan (2019b)	AOPE	Labeled	Unlabeled
Wang et al. (2018)	ASTE	Labeled	Does not exist
Xu et al. (2018)	AE	Labeled	Unlabeled
Xu et al. (2019)	ASC/AE	Labeled	Sparsely labeled
Xu et al. (2020)	E2E-ABSA	Unlabeled	Labeled
Xue et al. (2023)	ASC	Labeled	Unlabeled
Yang et al. (2020)	E2E-ABSA	Labeled	Unlabeled
Yang et al. (2021)	ASC	Labeled	Sparsely labeled
Yang et al. (2023)	ASC/AE	Labeled	Unlabeled
Yu et al. (2021)	E2E-ABSA	Labeled	Unlabeled
Yu et al. (2023)	E2E-ABSA	Labeled	Unlabeled
Zeng et al. (2023b)	E2E-ABSA	Labeled	Unlabeled
Zhang et al. (2023a)	ASC	Labeled	Does not exist
Zhang et al. (2023b)	ASC	Labeled	Unlabeled
Zhang et al. (2023c)	ASC	Labeled	Unlabeled
Zhao et al. (2024)	ASC	Labeled	Sparsely labeled
Zhao et al. (2022)	ASC	Labeled	Does not exist
Zheng et al. (2020)	ASC	Labeled for document	labeled
Zhou et al. (2020)	ASC	Labeled for document	Sparsely labeled
Zhou et al. (2021b)	E2E-ABSA	Labeled	Unlabeled

Although source code was provided for most models, several adaptations were necessary. In many cases, the code depended on undocumented or outdated third-party libraries that were no longer available. The evaluation metric used in these non-financial domain experiments was the micro-F1 score. Each model was run with three different random seeds, and the final results are the simple average across these runs. Table 18 summarizes the results.

Among the evaluated models, BGCA achieved the best overall performance, except for the domain pairs $L \rightarrow S$ and $S \rightarrow D$, where CDRG variants outperformed it. These results suggest that generative models like BGCA have significant potential for future cross-domain ABSA approaches. Comparative evaluations against other approaches are commonly reported in the studies reviewed, including, for instance, the work of Deng et al. (2023).

The next topic presents the trend analysis and future work for Cross-Domain ABSA.

Table 17 Source code repositories

Model	Article	Task	Source code repository
LCR-Rot-hop++	van Berkum et al. (2022)	ASC	https://github.com/stefanvanberkum/CD-ABSC
TransCap	Chen and Qian (2019)	ASC	https://github.com/NLPWM-WHU/TransCap
SimBridge / SemBridge	Chen and Qian (2021)	AE	https://github.com/NLPWM-WHU/BRIDGE
FMIM	Chen and Wan (2022)	ASCAE	https://github.com/CasparSwift/DA_MIM
BGCA	Deng et al. (2023)	ASTE	https://github.com/DAMO-NLP-SG/BGCA
	Gong et al. (2020)	E2E-ABSA	https://github.com/NUSTM/BERT-UDA
PRET+MULT	He et al. (2018)	ASC	https://github.com/ruidan/Aspect-level-sentiment
CLASSIC	Ke et al. (2021)	ASC	https://github.com/ZixuanKe/PyContinual
ASP / PATT	Klein et al. (2022)	AE	https://github.com/IntelLabs/nlp-architect/tree/libert-path-amtl/nlp_architect/models/libert
DAT-LCR-Rot-hop++	Knoester et al. (2023)	ASC	https://github.com/jorisknoester/DAT-LCR-Rot-hop-PLUS-PLUS
DIWS-LCR-Rot-hop++	Lee et al. (2023)	ASC	https://github.com//ejoone/DIWS-ABSC
AD-SAL	Li et al. (2019a)	E2E-ABSA	https://github.com/hsqmlzno1/Transferable-E2E-ABSA
GCDDA	Li et al. (2022)	AOPE	https://github.com/nustm/gcdda
	Liu et al. (2021)	ACSA	https://github.com/lgw863/ACSA-generation
UIKA	Liu et al. (2023)	ASC	https://github.com/WHU-ZQH/UIKA
	Ouyang and Shen (2023)	AE	https://github.com/PhSe-coder/MMT-ABSA
BSpLLA	Ramos and Fuentes (2023)	ASC	https://github.com/dionis/ABSA-DeepMultidomain/
DE-CNN	Xu et al. (2018)	AE	https://github.com/howardhsu/DE-CNN
BERT-PT	Xu et al. (2019)	ASC/AE	https://github.com/howardhsu/BERT-for-RRC-ABSA
CDRG	Yu et al. (2021)	E2E-ABSA	https://github.com/NUSTM/CDRG
DA2LM	Yu et al. (2023)	E2E-ABSA	https://github.com/NUSTM/DALM

6 Overview of cross-domain ABSA

This section addresses the question: "What are the main gaps and opportunities for future work?". The following are descriptions of some analyses of the found works. Next, the potential future work is presented.

Table 18 Performance of the models for aspect extraction and classification (micro-F1 average over 3 seeds)

	R→L	R→D	R→S	L→R	L→S	D→R	D→S	S→R	S→L	S→D
A	0.259	0.234	0.221	0.336	0.311	0.248	0.273	0.360	0.299	0.333
B	0.258	0.264	0.245	0.386	0.320	0.329	0.279	0.382	0.300	0.338
C	0.381	0.267	0.354	0.510	0.414	0.486	0.363	0.476	0.340	0.334
D	0.424	0.303	0.418	0.586	**0.440**	0.560	0.410	0.558	0.347	**0.380**
E	**0.457**	**0.343**	**0.424**	**0.611**	0.410	**0.597**	**0.486**	**0.560**	**0.368**	0.368

A: AD-AL (Li et al. 2019a), **B**: AD-SAL (Li et al. 2019a), **C**: CDRG-BERT$_B$ (Yu et al. 2021), **D**: CDRG-BERT$_E$ (Yu et al. 2021), **E**: BGCA (Deng et al. 2023). The bold values indicate the best performance for each column.

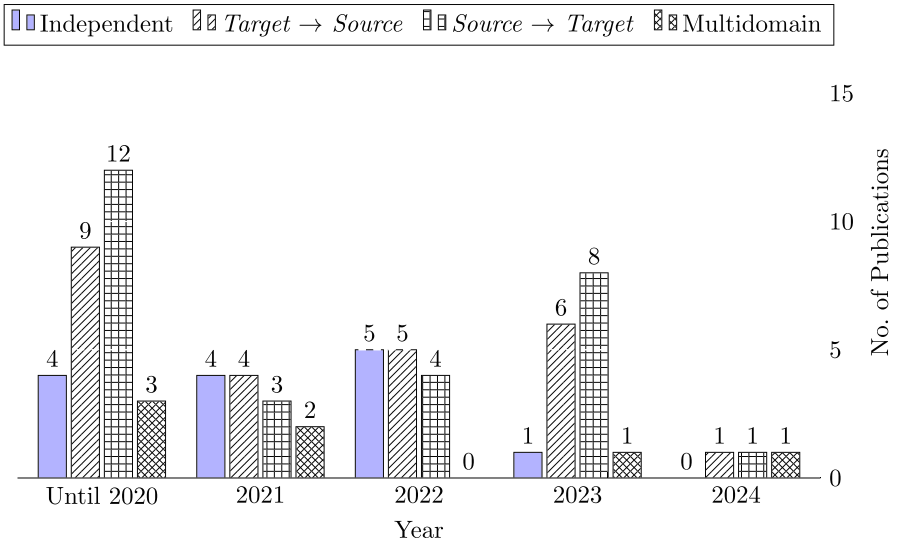

Fig. 18 Number of cross-domain ABSA articles by Group and Year

6.1 Trend analysis

Figure 18 illustrates the distribution of cross-domain ABSA (Aspect-Based Sentiment Analysis) articles published per year, categorized into four methodological groups: *Independent*, *Target → Source*, *Source → Target*, and *Multidomain*. The temporal axis includes cumulative publications until 2020, and yearly data from 2021 to 2024.

Between 2021 and 2023, we observed a clear upward trend in publication count, peaking in 2023 with a total of 16 articles. The most prominent group in this peak is *Source → Target*, with 8 publications, followed by *Target → Source* with 6. This indicates a growing emphasis on transfer strategies that adapt models trained in one domain using information from the other domain.

The *Independent* group has remained relatively stable over the years, with contributions ranging from 1 to 5 publications annually. The *Multidomain* approach, in contrast, shows limited usage throughout the entire period, with only 3 publications until 2020 and sparse presence thereafter (never exceeding 1 per year). This suggests that, while multidomain

strategies were explored early on, recent efforts have shifted toward more explicit and directional adaptation methods.

It is also important to note that the bar labeled "2024" includes only partial data, as not all publications from 2024 had been collected at the time of writing. Thus, the apparent decline in that year (a total of 3 articles) should not be interpreted as a definitive downward trend.

A classification of cross-domain ABSA models can be made by considering the use of language models. Models created before 2019 belong to the first generation, characterized by the absence of pre-trained Language Models (LMs). Although they might use recurrent neural networks, these were not trained on a large scale for a wide variety of domains. The second generation emerged in 2019, with the introduction of pre-trained LMs such as BERT and GPT, which are incorporated into ABSA models. As discussed in Sect. 5.2, this technique has been predominant in recently developed models. The third and newest generation begins in 2024, with the adoption of large language models (LLMs). LLMs are larger models trained with a vast amount of data, which should have a positive impact if used as sub-models.

Another point indicated in this review is the performance of the models over time. Figure 19 illustrates the F1 scores of the models compared by Shi et al. (2023). These values are the averages obtained for models trained on the SemEval datasets: Laptop → Restaurant and Restaurant → Laptop (Pontiki et al. 2014, 2015, 2016). It is noteworthy that the 2023 value is from the author's own model, which may contain bias. The growth of the graph over time shows the advancement of the academic community. The recent achievement of increasingly better results indicates room for new work in this research area.

Despite the large number of studies, some gaps remain, which are described in the following subsection.

6.2 Future work

This section explores promising directions for future research in cross-domain aspect-based sentiment analysis (ABSA), emphasizing the key open challenges identified throughout this review. Key issues include inconsistencies and limitations in existing datasets, the lack of standardized and comprehensive model comparisons, and the need for studies in specific domains such as finance. Additionally, this section addresses the underexplored potential of large language models (LLMs), the challenges of integrating multimodal data (e.g., text with images), and the importance of going beyond accuracy by incorporating considerations of efficiency and interpretability. Each of these aspects presents concrete research oppor-

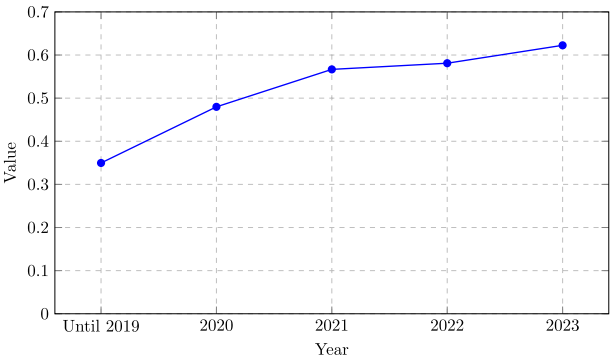

Fig. 19 Average Micro F1 of Cross-Domain ABSA Models Over the Years. The graph was constructed based on the article by Shi et al. (2023). An average of the values per year and between the cross-domains Laptop → Restaurant and Restaurant → Laptop from the SemEval dataset Pontiki et al. (2014, 2015, 2016) was made

tunities to advance the development of more robust and applicable cross-domain ABSA systems.

6.2.1 Datasets

The comparison between models depends on a uniform dataset. The most commonly used dataset for aspect extraction and classification in cross-domain scenarios is a composition of *SemEval* 2014, 2015, and 2016 Pontiki et al. (2014, 2015, 2016). This makes it the natural choice as a *benchmark* for most cross-domain ABSA models. This composition includes annotated data for aspect extraction and classification for restaurants in 2014, 2015, and 2016, and *laptops* in 2014, along with their respective categories. In addition to the SemEval datasets, two other frequently used datasets are the *devices* dataset (Hu and Liu 2004) and the *services* dataset (Toprak et al. 2010). Over time, these datasets have been adapted depending on the model, resulting in various variants.

The original SemEval datasets consist of collections of XML files with elements listing the aspects of sentences. Yu et al. (2021) and Deng et al. (2023) used an adaptation of this dataset where each word receives a *tag* in the T-O scheme, with T representing an aspect. This type of representation generates problems, as it cannot differentiate consecutive aspects. Additionally, there are discrepancies with the original aspect annotations. For example, consider the sentence "*Strengths:Well-shaped Weaknesses:A bad videocard!*" from the *laptops* dataset. The original annotation indicates the words "*shaped*" and "*videocard*" as aspects, while the adapted variant considers only "*videocard*" as an aspect. Out of 3045 sentences, 76 (2.5%) had different annotations in the *laptops* dataset. Yu et al. (2021) and Deng et al. (2023) combined the restaurant datasets from 2014, 2015, and 2016, creating a single dataset with 3877 sentences. During this merging process, 7 sentences were removed, and there are 124 sentences (3.2%) with labels different from the original merge.

This inconsistency can affect both the training and evaluation stages. During training, label errors in the original dataset may degrade the model's learning process, resulting in an unfair scenario where different methods are trained using slightly different versions of the same dataset. During evaluation, even a small proportion of mislabeled data can introduce uncertainty in the reported performance.

To illustrate this effect, consider a test scenario using the SemEval 2014 restaurant dataset, which contains 800 sentences and 1134 annotated aspects (see Sect. 3.6). Suppose a model yields 700 true positives (TP), 434 false positives (FP), and 434 false negatives (FN). The resulting precision and recall are both approximately 0.618, producing an F1-score of 0.618 (see Sect. 3.7). Now, assume that 3.2% of the data (about 36 sentences) contain incorrect labels that cause 36 of the TPs to be misclassified as FNs. In this case, the updated counts become TP = 664 and FN = 470. The resulting precision is 0.605 and recall is 0.586, yielding a new F1-score of approximately 0.595. Although a decrease of 2.3 percentage points in the F1-score may appear small, it is substantial when evaluating state-of-the-art models whose performance differences often fall within this margin (see Sect. 6.1).

These datasets present other issues. The *devices* dataset is labeled only as negative and positive, not including a neutral polarity. This creates a labeling discrepancy compared to the other listed datasets, which poses a problem in cross-domain scenarios. Additionally, it is common for current models to use pre-trained language models. These models may have used data from these datasets during their pre-training. For example, BERT (Devlin et al.

2019) is a 2019 model and the listed datasets emerged before 2016. The problem is further exacerbated with LLMs, which often do not detail what was used in their training.

All these problems can be summarized into the following future activities:

- standardize datasets labeled to follow the same rules;
- map which variant of each dataset is being used in published works;
- create new datasets to revalidate the models from the published articles.

6.2.2 Models comparison

We presented a preliminary and independent analysis of various methods for cross-domain aspect extraction and classification in Sect. 5.7. Although this comparison offers initial insights into the behavior of these approaches, it remains limited in scope, both in terms of the number of models evaluated and the variety of metrics considered.

A careful review of the existing literature reveals an absence of a comprehensive and standardized evaluation framework encompassing all major methods in this field. To the best of our knowledge, no prior study has conducted a broad and fair comparison involving multiple models applied to a uniformly annotated dataset, evaluated across a consistent set of performance metrics. This absence makes it difficult to draw robust conclusions about the relative strengths and weaknesses of each method and limits reproducibility and progress in the area.

To address these limitations, one promising direction for future work is:

- To carry out a thorough and unbiased comparative study of state-of-the-art cross-domain aspect-based sentiment analysis models. Such a study should evaluate a broad set of representative methods using a standard, uniformly annotated benchmark dataset and employ a diverse range of evaluation metrics (e.g., precision, recall, F1-score, macro/micro averages) to provide a better understanding of each method's effectiveness.

6.2.3 Studies in specific domains

The models studied in this article aim to demonstrate their performance in a generic cross-domain setting. However, since they always use the same datasets to demonstrate their performance, questions arise about their behavior in specific domains. Even if a cross-domain model can perform well in a source/target scenario, it can fail in another, with a different domain.

Saunders (2022) defines domain as a composition of topic and genre. Texts often encompass content from multiple subject areas–commonly referred to as topics–such as healthcare, finance, or digital entertainment. Each topic tends to exhibit its characteristic distribution of words (Bashiri and Naderi 2024; Saunders 2022). Genre has attributes as function, register, syntax, and style (Santini 2004; Saunders 2022). E-mails, blogs, and meeting minutes all have different genres for the same topics, even if they may share the same vocabulary and entity names. Thus, for example, two domains can be very similar genres, which makes the model perform well in a cross-domain scenario; however, it may not work very well with a completely different genre.

Assume a financial domain context. The financial domain significantly differs from traditional domains (restaurant, *laptop*, devices, or services) in aspect extraction and classification. Du et al. (2024a) points out three major challenges that differ from other domains:

1. Metaphorical expressions. Some expressions are particular to the financial domain, such as "The market is riding a bull", which means that the market is on the rise.
2. Financial sentiment analysis (FSA) often relies on the direction of events or changes, emphasizing the importance of contextual interpretation. For example, an increase in profit is generally perceived as positive, while a decrease is typically interpreted as negative.
3. Financial texts frequently combine qualitative discourse with quantitative data. For example: "In the four weeks following its release, the standard iPhone 15 sold 130.6% more units than the standard iPhone 14 did during the same period in the previous year."

For the financial context, we present a preliminary study with some presented methods in this review. Table 19 contains the F1-micro scores obtained from them trained on traditional datasets (Sect. 6.2.1): Restaurants (R), Laptops (L), Devices (D), and Services (S), and applied to a Financial domain (F). The Financial dataset used in the tests is *SEntFiN* (Sinha et al. 2022), which is a dataset for financial news microblogs, with aspects and polarities annotated. The tests were executed three times with different seeds, and the average F1-micro score was calculated. The last column of the table presents the performance of models trained and tested in the same financial domain. As illustrated, the performance of these models was low.

The experiments indicated that the tested models fail when applied to the financial domain using traditional datasets as the source domain. Thus, this indicates the need for studies in specific domains, such as economic, financial, medical, and other domains. In other words, future work should focus on:

- ABSA studies in cross-domain applications for specific domains.

6.2.4 Use of LLMs for cross-domain ABSA

Another important aspect is the use of LLMs, which remains underexplored in this subject. During the period covered by this systematic review, we could not find any relevant work. However, some new works have been developed. Given the current importance of the LLMs, we added a section in the appendix (Appendix B) supplementing

Table 19 Performance of models for aspect extraction and classification using financial data as the target domain

	R	L	D	S	F (in-domain)
$AD-AL$	0.026	0.020	0.020	0.022	0.688
$AD-SAL$	0.055	0.030	0.032	0.027	0.695
$CDRG-BERT_B$	0.066	0.076	0.045	0.073	0.832
$CDRG-BERT_E$	0.071	0.080	0.045	0.090	0.833
BGCA	0.103	0.106	0.095	0.121	0.849

Table 20 Prompts used in aspect extraction tests using LLM

Prompt	Text
A	`Extract the aspects terms (Aspect-based sentiment analisys) in the following sentence. Dont write anything except the terms separated by pipe("|").`

B	`Extract the aspects terms (Aspect-based sentiment analisys) only to following sentence.`
	`Dont write anything but the terms separated by pipe.`
	`Example: I went to the Abdalas, in the center. The garlic bread was delicious, their fantastic pasta was terrible, and the wine was ok, with expensive prices.`
	`Answer: garlic bread|pasta|wine|prices`
	`Example: If you' re in New York, you do not want to miss this place.`
	`Answer:`
	`Example: The hotel is amazing, but the rooms are very small.`
	`Answer: rooms`

Table 21 Performance of LLAMA without fine-tuning for aspect extraction. The bold values indicate the best performance for each column

	R	L	D	S
Prompt A	0.220	0.156	0.093	0.188
Prompt B	**0.501**	**0.342**	**0.252**	**0.409**

this topic. This is a recent technology, and new works employing this technique will likely emerge soon.

This paper presents some initial tests with LLMs utilizing LLAMA (Touvron et al. 2023). LLAMA has several versions, with version 3 being used for this paper. Version 3 is subdivided into 8B and 70B versions, with 8 billion and 70 billion parameters, respectively. Each of these versions has variants, notably pre-trained and instruction versions. The pre-trained version is the base version, trained to generate text, while the instruction version (*instruct*) is fine-tuned to follow instructions in text generation (AI@Meta 2024). After some exploratory tests, the *instruct* version was selected for this study because it provides better results without any further fine-tuning. The 8B version was chosen because it can be run on accessible graphics cards for this research, specifically an NVIDIA GeForce RTX 4090 with 24GB. The 70B version requires additional computational resources.

A preliminary test using LLMs was the task of aspect extraction (AE) on the datasets from Sect. 6.2.1, without cross-domain. Two prompts were considered for the model to determine the effect. The first prompt (*Prompt* A) is a pure instruction to the LLM model, while the second (*Prompt* B) also includes some examples. Table 20 illustrates the two prompts used in the tests. Table 21 presents the F1-micro values obtained for the two prompts applied to the datasets from Sect. 6.2.1. *Prompt* B performed better across all datasets, demonstrating that the prompt can be decisive for this type of activity. However, the computational cost increases significantly with a larger prompt.

Table 22 Performance of LLAMA with *Prompt* tuning applied in-domain. The bold values indicate the best performance for each column

	R	L	D	S
0 epochs	0.220	0.156	0.093	0.188
1 epoch	0.695	0.506	0.156	0.647
3 epochs	0.761	0.806	0.469	0.715
10 epochs	**0.810**	**0.831**	**0.594**	**0.732**

Table 23 Performance of LLAMA with *Prompt* tuning applied in cross-domain. The bold values indicate the best performance for each column

Model	R → L	R → D	R → S	L → R	L → S	D → R	D → S	S → R	S → L	S → D
$LLAMA$ – Prompt Tuning —0 epochs	0.156	0.093	0.188	0.220	0.188	0.220	0.188	0.220	0.156	0.093
$LLAMA$ – Prompt Tuning —1 epoch	0.429	0.362	**0.276**	0.259	0.182	0.047	0.040	0.379	0.309	0.277
$LLAMA$ – Prompt Tuning —3 epochs	**0.456**	**0.416**	0.259	0.344	**0.280**	0.214	0.125	0.517	0.396	0.340
$LLAMA$ – Prompt Tuning —10 epochs	0.311	0.312	0.195	**0.345**	0.277	**0.397**	**0.241**	**0.542**	**0.415**	**0.357**

Given the importance of the initial prompt and the computational cost, the use of *Soft Prompt* was considered. Lester et al. (2021) developed the "Prompt Tuning" system. They observed that instructions and examples provided to the model affect the final performance and that these can be learned. During the model execution, these instructions are transformed into vectors. Considering this, they proposed fixing the model weights and learning these vectors, resulting in *soft prompts*, a more cost-effective option. In contrast to *hard prompts*, *soft prompts* do not represent actual words and symbols, making them even more effective. *Prompt* A underwent a fine-tuning process (*Prompt Tuning*) using a quantized[8] version of LLAMA. The model was trained for 1, 3, and 10 epochs using three different seeds for each dataset. The F1-micro values obtained are shown in Table 22. The model fine-tuned for *Prompt* A for a single epoch performed better on 3 out of 4 datasets compared to *Prompt* B without fine-tuning. From three epochs onward, the model showed improved performance across all datasets, achieving the best performance when trained for 10 epochs. Based on these results, the next tests were conducted in a cross-domain setting.

The cross-domain test in this paper using LLMs followed a strategy from the Casual models group, which is Independent. For this, the datasets from Sect. 6.2.1 were used. All cross-domain combinations were tested except between *device* and *laptop*, as they are similar. Tests were performed with the *Soft Prompt* trained for 0, 1, 3, and 10 epochs, each with three seeds. The results are shown in Table 23. The performance in cross-domain is considerably lower than when trained with in-domain data indicating that the model is learning a *prompt* somewhat tied to the domain. Despite this, the model performed better than the version without fine-tuning. Another important point is that the model is experiencing overfitting in some cases. For instance, the model trained for three epochs on the restaurant dataset performed better in cross-domain tests with other datasets compared to the model trained for 10 epochs. This difference relative to other

[8] Quantization allows the model parameters to be represented using fewer memory bits with slightly reduced performance.

datasets is likely due to the fact that the restaurant training dataset has a larger number of examples.

This simple strategy of training the prompt in one domain and applying it to another was not sufficient for an LLM model to outperform state-of-the-art models. Therefore, a point for future work is:

- Explore the use of LLMs in cross-domain models.

6.2.5 Multimodal ABSA: challenges and opportunities

In the early days, product or service reviews were expressed in text form. Sentiment analysis was initially performed on the entire document and was gradually refined to the aspect level (ABSA). However, with the evolution of social media, users can now express their opinions not only through text but also through other modalities, such as images or videos. A common form is the combination of text with these other modalities, i.e., multimodality.

Aspect-level sentiment analysis based solely on text is infeasible in some contexts, necessitating the use of multimodality. For example, the phrase "I went to the new restaurant and tried the dessert" (Fig. 20). Without an image, the aspect *dessert* has a neutral polarity. However, Fig. 20a expresses a negative polarity and Fig. 20b, a positive one. De Bruyne et al. (2022) constructed a dataset using the Adidas Instagram page by collecting 4900 comments on 175 Instagram images, and annotating them with aspect categories and emotional information. They performed an analysis of this dataset to determine the importance of images. Of these, 2285 comments did not express any sentiment. For the remaining comments, it was impossible to determine sentiment in 87% of them without the aid of the image. Similarly, 60% contained an implicit aspect or used terms such as this, that, these, those, and it.

There are some works on multimodal ABSA. Hu and Yamamura (2023) developed a multimodal ABSA sentiment classifier, while Wang and Guo (2024) created a multimodal extractor and classifier (E2E-ABSA). However, no multimodality works in cross-domain ABSA were found. Cross-domain ABSA can be done in various ways. An image can assist in sentiment classification and be domain-agnostic. For example, just as Fig. 20, a and b determine the sentiment regarding the dessert at a restaurant,

Image:		
	(a)	(b)
Comment:	I went to the new restaurant and tried the dessert.	I went to the new restaurant and tried the dessert.
Aspect:	dessert	dessert
Polarity:	Negative	Positive

Fig. 20 Comments with the same text but different images. **a** determines the negative polarity of the first comment and **b** determines the positive polarity

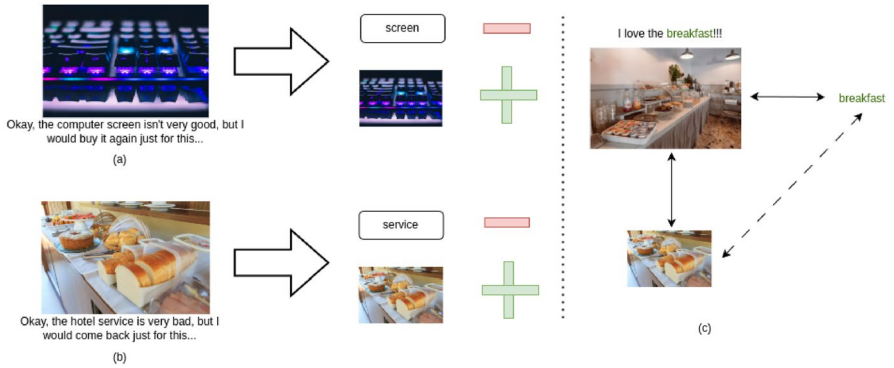

Fig. 21 Multimodal aspect extraction. In (**a**) and (**b**), similar phrases show that the image represents an aspect. In (**c**), although the image in review (**b**) is not associated with an aspect text, it is possible to infer the aspect text by considering a similar review

they could be used to express negative or positive sentiments about the speed of a new computer.

In terms of aspect extraction, the image itself can express an aspect. Therefore, these aspects should not be classified as *hidden*, since they are explicitly presented, albeit in a modality other than text. We propose a cross-domain model that could learn to return the *image*, or part of it, as an aspect instead of text. Identifying that an image is an aspect is a pattern that can be learned and applied in a cross-domain context. For example, Fig. 21a and b illustrate similar comments in different domains, notebook and hotel, respectively. In both cases, the comments are indicating that an aspect will be shown in the image. In Fig. 21a, the sentence "Okay, the *computer screen* isn't very good, but I would buy it again just for this..." indicates a negative sentiment towards the computer screen and a positive sentiment towards the keyboard, which is presented as an image. In Fig. 21b, the sentence "Okay, the hotel *service* is very bad, but I would come back just for this..." indicates a negative sentiment towards the hotel service but a positive sentiment towards the breakfast, which is shown in the image. Thus, a model could use an annotated notebook domain to learn that this type of comment indicates that the presented image is the associated aspect and infer in a hotel domain.

Regarding the aspect, a model can learn to associate the image with the aspect text. Consider Fig. 21b, c. From the comment in Fig. 21c, "I love the *breakfast*", it is possible to associate the image with the aspect *breakfast*. Thus, even though the comment in Fig. 21b does not mention the text *breakfast*, it is possible to infer it by observing that the images 21b and c are similar.

One of the factors that could be hindering the evolution of these models is that until recently, there was a lack of labeled data at the aspect level. For instance, the work of Anschütz et al. (2023) searches for images and, using these images, obtains the texts that will be used for ABSA classification. However, the model does not utilize the

images in the proposed sentiment prediction model, and its training uses the SemEval 2017 dataset (Rosenthal et al. 2017).

There are a few datasets that approach aspect-level annotation for ABSA. De Bruyne et al. (2022) developed an aspect-level dataset but removed sentences containing more than one aspect. The main characteristic of aspect-level ABSA is the ability to determine more than one aspect in a sentence with different polarities, which limits this dataset. Yu and Jiang (2019) annotated the TWITTER-15 and TWITTER-17 datasets for sentiments. These are multimodal datasets for named entities, with 5,338 and 5,972 entities, respectively. While named entities share a significant similarity with aspect extraction, they are different. An aspect can be a characteristic of an entity, such as the keyboard of a computer or the breakfast at a hotel. Zhou et al. (2021a) created a multimodal dataset annotated with sentiment at the category level, consisting of 38,000 examples. Notably none of these datasets annotate the relationship between the image, or part of it, and the aspect. Furthermore, multimodality can include audio and video, which have not been addressed in any work.

The lack of annotated datasets for Multimodal ABSA highlights the gaps for future work in ABSA, which can be summarized as follows:

- Lack of annotated data for Multimodal ABSA. This includes datasets annotated at the aspect level and datasets that associate images with aspects.
- Multimodal ABSA models. This includes works that return images, audio, videos, or text as aspects.
- Works that learn to associate multimodality with the aspect text.

6.2.6 Beyond accuracy: efficiency and interpretability considerations

Most ABSA studies in cross-domain settings rely on evaluation metrics that focus exclusively on the model's ability to correctly extract and classify aspects. These include not only accuracy and F1-score, but also other similar metrics (e.g., precision, recall) which are designed to quantify how often the model is right or wrong. However, such metrics fail to address crucial practical dimensions, such as computational efficiency and interpretability, which are essential in real-world deployments.

From an efficiency perspective, many recent methods are based on large language models that demand significant computational resources, including GPUs, to be executed. This imposes limitations in time-sensitive or resource-constrained environments. For example, running a LLaMA 2 70B model on an *ml.g5.12xlarge* AWS SageMaker instance (equipped with four NVIDIA A10G GPUs, totalizing 96GB of memory) yields a throughput of just 33.3 tokens per second (Face 2023; Amazon Web Services 2024). Give that each polarity classification requires approximately five tokens, processing 100,000 instances would take over 4 h – an unacceptable delay in time-critical scenarios.

Additionally, these models often need to be executed on different machines from where the data is stored, requiring secure and high-performance data transfer infrastructures. This requirement adds to the deployment complexity and operational cost.

Interpretability is another essential but frequently overlooked dimension. In some domains, such as finance or healthcare, models must provide transparent and explainable reasoning. While complex neural architectures offer strong predictive performance, they usually function as black boxes. On the other hand, models in the "Rules" subgroup of the "Independent" and "Target→Source" categories are typically more interpretable and thus more suitable for such use cases.

Recommendations for future work:

- Extend evaluation practices to include computational cost (e.g., inference time, GPU requirements, memory usage);
- Analyze the model's feasibility for deployment in scenarios with time or budget constraints;
- Include assessments of interpretability, especially in regulated or sensitive domains;
- Consider trade-offs between predictive performance, efficiency, and explainability.

6.3 Final considerations of the overview

This section presented an overview of cross-domain ABSA, analyzing the evolution of models from various perspectives. It also identified several gaps for future work, though this list is not exhaustive. For example, one area that could be further explored, not discussed as a future gap, is the potential for solution improvement using approximations of the *Source* → *Target* category, particularly concerning instance approximation. For example, not all models that generate text for training in the target domain consider the relationship between $\frac{\mathbb{P}_t(\mathbf{x}^{sin})}{\mathbb{P}_{sin}(\mathbf{x}^{sin})}$ (Eq. 3), and mechanisms that help balance this ratio may bring improvement to the models. The conclusion of this review is presented next.

7 Conclusion

Throughout the text, answers to questions related to this work's objective have been presented. The review demonstrated that this is a highly explored topic, with new models emerging annually in a growing trend. The journals, conferences, and workshops that have published the most on the subject were also presented (Sect. 3.5).

To the best of our efforts, works addressing the topic were sought and summarized (Sect. 4). Standard techniques were listed and summarized (Sect. 2.3). For these techniques, a detailed analysis of their evolutions and applications in models was conducted, highlighting those that are emerging and declining and the relationship among them (Sect. 5.2). A listing of the works showing the tasks performed is presented in Sect. 5.1. An analysis was conducted on the sentiment polarity granularity of the models (Sect. 5.3), the outputs of the models (Sect. 5.4), and the labeling (Sect. 5.5). These analyses should assist researchers and institutions in selecting solutions.

The problem of ABSA in cross-domain can be addressed by various strategies, aiming to align the domains to mitigate the lack of labeled data in the target domain. A new classification of ABSA models in cross-domain contexts has been proposed and presented in this review, emphasizing the cross-domain solution approach (Sects. 2.2.2 and 4.1). Traditionally, classifications have focused on the specific task of the model. With the new classification, researchers can more intuitively understand how the cross-domain problem has been addressed in ABSA and identify those who have used similar techniques to those proposed in their work, regardless of the task type. This framework also facilitates considering complementary cross-domain solutions based on existing implementations, thereby improving model performance. Additionally, since the first hierarchical level of this classification followed a mathematical intuition adapted and expanded from the work of Gong et al. (2020), researchers can classify their work and observe gaps. For example, a model that generates examples to train the model in the target domain using data from the source domain, classified as *Source* $\rightarrow$ *Target*, should consider the data distribution in the target domain to achieve better results.

This review listed the data sources used, presented statistics, and analyzed the feasibility of obtaining them according to the articles, updating this information if necessary (Sect. 3.6). In addition, the main metrics and both quantitative and qualitative evaluations performed were highlighted. A graph with the mentioned models was included, indicating which models are most frequently used as references for others. The metrics of each work were listed (Sect. 3.7).

After all, an overview of the Cross-Domain ABSA context was presented (Sect. 6). Analyses were conducted considering the historical evolution of the models (Sect. 6.1). The trend analysis highlighted which group, according to the taxonomy of cross-domain models, received the most emphasis over the years. The increasing dependence on language models and a perspective on the use of LLMs were also presented. Subsequently, opportunities for future work were outlined, considering datasets, models comparison, studies in specific domains, multimodality, and new performance metrics (Sect. 6.2).

Finally, this systematic review may also have gaps and, like any work, is subject to bias. However, in the best of our efforts, this work aimed to present the main points independently to contribute to related areas.

Appendix A: Search strings

Table 24 lists all search strings used in all sources of the systematic review, except for the *ACL Anthology* source. It does not have a search engine, only being possible to obtain a file with all published articles. In this case, a Python code was written to filter these articles, as listed in source code 1.

Listing 1 is the source code used to filter articles from the ACL Anthology.

```
import bibtexparser
from bibtexparser.bwriter import BibTexWriter
from bibtexparser.bibdatabase import BibDatabase
import unicodedata

def remover_acentos(texto):
    # Remove acentos utilizando a biblioteca unicodedata
    texto_sem_acentos = ''.join(c for c in \
                                unicodedata.normalize('NFD', texto)
                                if unicodedata.category(c) != 'Mn')
    return texto_sem_acentos

def limpar_string(texto):
    # Remove enter e substitui por espaco em branco
    texto = texto.replace('\n', '')
    texto = texto.replace('-', '' )

    # Remove espacos em branco do inicio e final da string
    # texto = texto.strip()
    # Remove acentos da string
    texto = remover_acentos(texto)

    return texto


nome_arquivo_entrada = 'anthology+abstracts.bib'
nome_arquivo_saida = "arquivo_pesquisado.bib"

print( 'Lendo', nome_arquivo_entrada )
with open( nome_arquivo_entrada, "r") as arquivo_entrada:
    parser = bibtexparser.bparser.BibTexParser( \
             interpolate_strings=False)
    bib_database = bibtexparser.load(arquivo_entrada, \
                   parser=parser)

buscas = [
["aspect", 'product-feature', 'target'],
["sentiment-analysis", "sentiment-classification", "opinion"],
["domain-adaptation","cross-domain","domain-invariant", \
"domain-invariant", "different-domains", "across-domains", \
"transferable", "multiple-domains", "other-domains", \
"target-domain", "transfer-learning", "fine-tuning", \
"pre-trained", "pre-trained", "pre-training", \
"pre-training", "domain-oriented", "domain-oriented"]
]

print('Exemplo', bib_database.entries[0] )
print('Total de registros para busca:', \
      len( bib_database.entries ) )
entradas = []
for entrada in bib_database.entries:
    achou = True
    i = 0
```

```
while achou and i < len( buscas ):
    achou_interno = False
    j = 0
    while not achou_interno and j < len( buscas[i] ):
        if buscas[i][j] in limpar_string( \
            entrada.get("title",'').lower() + '-' + \
            entrada.get("author",'').lower() + '-' + \
            entrada.get("abstract",'').lower() ):
             achou_interno = True
        j+=1
    achou = achou_interno

        i+=1
    if achou:
        entradas.append( entrada )
bib_database.entries = entradas
print('Total de registros encontrados:', \
      len( bib_database.entries ) )


writer = BibTexWriter()
with open( nome_arquivo_saida, 'w') as bibfile:
   bibfile.write(writer.write(bib_database))


print("Busca completa.")
```

Appendix B: Works with large language models

Given the growing relevance of *Large Language Models* (LLMs), we extended our review to include recent studies related to this topic. To achieve this, we performed a complementary search using the Scopus database covering studies until June of 2025, and using the following simplified query:

(LLM OR "large language model") AND "cross-domain" AND "aspect"

Although the search string is simpler than those used in the main protocol and was applied only to Scopus (unlike the original multi-database approach), we note that Scopus indexes a broad range of relevant literature, including journals from IEEE, ACM, ScienceDirect, and others. Furthermore, the keywords used are now widely recognized in the field, which we believe helps mitigate potential omissions.

A total of 12 studies were found using the search string. After removing non English studies, nine articles remained. Following an in-depth reading, two articles remained: Chen et al. (2024b) and Zou and Wang (2025).

Chen et al. (2024b) address the challenge of constructing prompts for sentiment classification within the ABSA framework and conclude that there is no single prompt capable of handling all cases effectively. To overcome this limitation, the authors propose a multi-step method. First, the GPT 3.5 Turbo (Mann et al. 2020), a very large language model, is used

Table 24 Search strings used in the systematic review

Source	Search String
Scopus	TITLE-ABS-KEY (("aspect" OR "aspected" OR "product feature" OR "target") AND ("sentiment analysis" OR "sentiment classification" OR "opinion") AND ("domain adaptation" OR "cross domain" OR "domain-invariant" OR "different domains" OR "across domains" OR "transferable" OR "multiple domains" OR "other domains" OR "target domain" OR "transfer learning" OR "fine tuning" OR "pre-trained" OR "pre-training" OR "domain-oriented")) AND (LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English"))
ACM Digital Library	(Title:("aspect*" OR "product feature*" OR "target*") OR Abstract:("aspect*" OR "product feature*" OR "target*") OR Keyword:("aspect*" OR "product feature*" OR "target*")) AND (Title:("sentiment analysis" OR "sentiment classification" OR "opinion") OR Abstract:("sentiment analysis" OR "sentiment classification" OR "opinion") OR Keyword:("sentiment analysis" OR "sentiment classification" OR "opinion")) AND (Title:("domain adaptation" OR "cross domain" OR "domain-invariant" OR "different domains" OR "across domains" OR "transferable" OR "multiple domains" OR "other domains" OR "target domain" OR "transfer learning" OR "fine tuning" OR "pre-trained" OR "pre-training" OR "domain-oriented") Abstract:("domain adaptation" OR "cross domain" OR "domain-invariant" OR "different domains" OR "across domains" OR "transferable" OR "multiple domains" OR "other domains" OR "target domain" OR "transfer learning" OR "fine tuning" OR "pre-trained" OR "pre-training" OR "domain-oriented") Keyword:("domain adaptation" OR "cross domain" OR "domain-invariant" OR "different domains" OR "across domains" OR "transferable" OR "multiple domains" OR "other domains" OR "target domain" OR "transfer learning" OR "fine tuning" OR "pre-trained" OR "pre-training" OR "domain-oriented"))
IEE Digital Library	("All Metadata":"aspect" OR "All Metadata":"aspected" OR "All Metadata":"product feature" OR "All Metadata":"target") AND ("All Metadata":"sentiment analysis" OR "All Metadata":"sentiment classification" OR "All Metadata":"opinion") AND ("All Metadata":"domain adaptation" OR "All Metadata":"cross domain" OR "All Metadata":"domain-invariant" OR "All Metadata":"different domains" OR "All Metadata":"across domains" OR "All Metadata":"transferable" OR "All Metadata":"multiple domains" OR "All Metadata":"other domains" OR "All Metadata":"target domain" OR "All Metadata":"transfer learning" OR "All Metadata":"fine tuning" OR "All Metadata":"pre-trained" OR "All Metadata":"pre-training" OR "All Metadata":"domain-oriented")
Science Direct	("aspect" OR "aspected" OR "product feature" OR "target") AND ("sentiment analysis" OR "sentiment classification" OR "opinion") AND ("domain adaptation" OR "cross domain" OR "domain-invariant" OR "different domains" OR "across domains" OR "transferable" OR "multiple domains" OR "other domains" OR "target domain" OR "transfer learning" OR "fine tuning" OR "pre-trained" OR "pre-training" OR "domain-oriented")

Table 24 (continued)

Source	Search String
Web of Science	`(TS="aspect*" OR TS="product feature*" OR TS="target*") AND (TS="sentiment analysis" OR TS="sentiment classification" OR TS="opinion") AND (TS="domain adaptation" OR TS="cross domain" OR TS="domain-invariant" OR TS="different domains" OR TS="across domains" OR TS="transferable" OR TS="multiple domains" OR TS="other domains" OR TS="target domain" OR TS="transfer learning" OR TS="fine tuning" OR TS="pre-trained" OR TS="pre-training" OR TS="domain-oriented")`

to generate synthetic sentences for a given aspect and sentiment. Next, prompt candidates are created using a pre-trained T5 model. These prompts are then applied to real sentences through a selection process that combines two criteria: (i) the semantic similarity between the target sentence and the synthetic sentences, computed using Sentence-BERT with cosine similarity; and (ii) the success rate of each prompt across similar examples. This model is tested in a cross-domain scenario. Table 25 presents some model characteristics.

Zou and Wang (2025) propose a method that combines dependency syntax, large language models, and soft prompting. The model first processes each input sentence using the LLaMA-8B model (AI@Meta 2024) to obtain contextualized embeddings. Based on these embeddings, two classifiers are trained: one to identify aspect attributes (i.e., whether a word belongs to an aspect term) and another to extract sentiment-bearing expressions. Subsequently, a transformer is constructed with an attention mechanism constrained by syntactic dependencies: each word is allowed to attend only to syntactically related words in the dependency tree, and aspect terms are restricted to attend only to other words within the same aspect term. The contextual embeddings are then concatenated with syntactic embed-

Table 25 Article Summary: Aspect target sentiment classification with instance retrieval-based prompt template selection—Chen et al. (2024b)

Conference	IEEE—International Conference on Electrical Engineering Big Data and Algorithms (EEBDA)—not evaluated in JIF, JCI or Qualis
Metrics	Accuracy and Macro-F1
Tasks	ASC
Cross domain	Independent—Casual
Polarity	POS/NEG/NEU
Labeling	Source: labeled / Target: does not exist
Model output	Generative: <sentiment> >Z>

Table 26 Article Summary: Large language model augmented syntax-aware domain adaptation method for aspect-based sentiment analysis—Zou and Wang (2025)

Journal	Elsevier—Neurocomputing—JIF: 6 / JCI: 1.02 / Qualis: A1
Metrics	Macro-F1, Precision and Recall
Tasks	E2E-ABSA
Cross domain	Target → Source—Prompt
Polarity	POS/NEG/NEU
Labeling	Source: labeled / Target: not labeled
Model output	BIO+Polarity

dings of the aspect terms and fed into an adversarial network to predict the domain. Next, automatic soft prompt learning is performed using domain-topic and task-relevant features to help the model capture domain-specific semantics in a cross-domain setting. Finally, all intermediate representations are passed through a gated unit to predict BIO tags along with their corresponding sentiment polarities. Table 26 presents some model characteristics.

Author contributions All authors contributed equally to the work.

Funding This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. Lastly, this work is supported by CNPq (grant #309575/2021-4 and #307184/2025-0).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

AI@Meta (2024) Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Accessed 08 Sept. 2024

Álvarez-López T, Fernández-Gavilanes M, Costa-Montenegro E et al (2018) A Proposal for Book Oriented Aspect Based Sentiment Analysis: Comparison over Domains. In: Proceedings of the Natural Language Processing and Information Systems, Paris, France, pp 3–14, https://hal.science/hal-01958697

Amazon Web Services (2024) Amazon ec2 instance types. https://aws.amazon.com/pt/ec2/instance-types/. Accessed 06 Sept 2024

Anand D, Mampilli BS (2021) A novel evolutionary approach for learning syntactic features for cross domain opinion target extraction. Appl Soft Comput 102:107086. https://doi.org/10.1016/j.asoc.2021.107086

Anschütz M, Eder T, Groh G (2023) Retrieving users' opinions on social media with multimodal aspect-based sentiment analysis. In: Proceedings of the 2023 IEEE 17th International Conference on Semantic Computing. IEEE, Laguna Hills, CA, USA, ICSC'23, pp 1–8, https://doi.org/10.1109/ICSC56153.2023.00008

Bagheri A, Saraee M, de Jong F (2013) Care more about customers: unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. Knowl-Based Syst 52:201–213. https://doi.org/10.1016/j.knosys.2013.08.011

Bahdanau D, Cho K, Bengio Y (2016) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. arXiv:1409.0473

Bashiri H, Naderi H (2024) Comprehensive review and comparative analysis of transformer models in sentiment analysis. Knowl Inf Syst 66(12):7305–7361. https://doi.org/10.1007/s10115-024-02214-3

Bhattacharjee K, Gangadharaiah R, Muresan S (2021) Domain and task-informed sample selection for cross-domain target-based sentiment analysis. In: Proceedings of the 4th International Conference on Natural Language and Speech Processing. Association for Computational Linguistics, Online, ICNLSP'21, pp 252–256, https://aclanthology.org/2021.icnlsp-1.29

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Bosselut A, Rashkin H, Sap M et al (2019) COMET: Commonsense transformers for automatic knowledge graph construction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, ACL'19, pp 4762–4779, https://doi.org/10.18653/v1/P19-1470

Brown PF, Della Pietra VJ, deSouza PV et al (1992) Class-based n-gram models of natural language. Comput Linguist 18(4):467–480

Cambria E, Olsher DJ, Rajagopal D (2014) Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In: Proceedings of the 28th AAAI conference on artificial intelligence. AAAI Press, Québec City, Québec, Canada, AAAI'14, https://api.semanticscholar.org/CorpusID:14206596

Cao Z, Zhou Y, Yang A et al (2021) Deep transfer learning mechanism for fine-grained cross-domain sentiment classification. Connect Sci 33(4):911–928. https://doi.org/10.1080/09540091.2021.1912711

Cascante-Bonilla P, Tan F, Qi Y et al (2021) Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI'21, vol 35. AAAI Press, Online, pp 6912–6920, https://doi.org/10.1609/aaai.v35i8.16852

Chauhan GS, Kumar Meena Y, Gopalani D et al (2020) An unsupervised multiple word-embedding method with attention model for cross domain aspect term extraction. In: Proceedings of the 2020 3rd international conference on emerging technologies in computer engineering: machine learning and Internet of Things. IEEE, Jaipur, India, ICETCE'20, pp 110–116. https://doi.org/10.1109/ICETCE48199.2020.9091738

Chen Z, Qian T (2019) Transfer capsule network for aspect level sentiment classification. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, Italy, ACL'19, pp 547–556, https://doi.org/10.18653/v1/P19-1052

Chen Z, Qian T (2021) Bridge-based active domain adaptation for aspect term extraction. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, ACL-IJCNLP'21, vol 1. Association for Computational Linguistics, Online, pp 317–327. https://doi.org/10.18653/v1/2021.acl-long.27

Chen Z, Qian T (2022) Retrieve-and-edit domain adaptation for end2end aspect based sentiment analysis. IEEE/ACM Trans Audio Speech Language Process 30:659–672. https://doi.org/10.1109/TASLP.2022.3146052

Chen X, Wan X (2022) A simple information-based approach to unsupervised domain-adaptive aspect-based sentiment analysis. arXiv preprint arXiv:2201.12549

Chen Q, Huang J, Wen W et al (2024a) Cat: continual adapter tuning for aspect sentiment classification. Neurocomputing 580:127423. https://doi.org/10.1016/j.neucom.2024.127423

Chen Z, Zhang J, Yuan H (2024b) Aspect target sentiment classification with instance retrieval-based prompt template selection. In: 2024 IEEE 3rd international conference on electrical engineering, big data and algorithms (EEBDA), IEEE, pp 371–375, https://doi.org/10.1109/EEBDA60612.2024.10486026

Chernyshevich M (2014) IHS R &D Belarus: Cross-domain extraction of product features using CRF. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). Association for Computational Linguistics, Dublin, Ireland, SemEval'14, pp 309–313, https://doi.org/10.3115/v1/S14-2051

Cho K, van Merriënboer B, Gulcehre C et al (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Doha, Qatar, EMNLP'14, pp 1724–1734. https://doi.org/10.3115/v1/D14-1179

Darwich M, Mohd SA, Omar N et al (2019) Corpus-based techniques for sentiment lexicon generation: a review. J Digit Inf Manag 17(5):296. https://doi.org/10.6025/jdim/2019/17/5/296-305

De Bruyne L, Karimi A, De Clercq O, et al (2022) Aspect-based emotion analysis and multimodal coreference: a case study of customer comments on adidas Instagram posts. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 574–580. https://aclanthology.org/2022.lrec-1.61

De Clercq O, Lefever E, Jacobs G et al (2017) Towards an integrated pipeline for aspect-based sentiment analysis in various domains. In: Proceedings of the 8th workshop on computational approaches to subjectivity, sentiment and social media analysis. Association for Computational Linguistics, Copenhagen, Denmark, WASSA'17, pp 136–142, https://doi.org/10.18653/v1/W17-5218

Deng Y, Zhang W, Pan SJ et al (2023) Bidirectional generative framework for cross-domain aspect-based sentiment analysis. In: Proceedings of the 61st annual meeting of the association for computational linguistics, ACL'23, vol 1. Association for Computational Linguistics, Toronto, Canada, pp 12272–12285. https://doi.org/10.18653/v1/2023.acl-long.686

Devlin J, Chang MW, Lee K et al (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL'19, vol 1. Association for Computational Linguistics, Minneapolis, MN, USA, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423

Ding X, Liu B, Yu PS (2008) A holistic lexicon-based approach to opinion mining. In: Proceedings of the 2008 international conference on web search and data mining. association for computing machinery, Palo Alto, CA, WSDM'08, pp 231–240. https://doi.org/10.1145/1341531.1341561

Ding Y, Yu J, Jiang J (2017) Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In: Proceedings of the 31st AAAI conference on artificial intelligence, AAAI'17, vol 31. AAAI Press, San Francisco, CA, USA. https://doi.org/10.1609/aaai.v31i1.11014

Dong L, Wei F, Tan C et al (2014) Adaptive recursive neural network for target-dependent Twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, ACL'14, vol 2. Association for Computational Linguistics, Baltimore, Maryland, pp 49–54. https://doi.org/10.3115/v1/P14-2009

Du K, Xing F, Mao R et al (2024a) An evaluation of reasoning capabilities of large language models in financial sentiment analysis. In: 2024 IEEE conference on artificial intelligence (CAI), IEEE, pp 189–194. https://doi.org/10.1109/CAI59869.2024.00042

Du K, Xing F, Mao R et al (2024b) Financial sentiment analysis: techniques and applications. ACM Comput Surv 56(9):1–42. https://doi.org/10.1145/3649451

Duran MS, Nunes MdGV, Pardo TAS (2023) Construções sintaticas do português que desafiam a tarefa de parsing: uma analise qualitativa. In: Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival. Association for Computational Linguistics, Belo Horizonte, MG, Brazil, UDFest-BR'23, pp 432–441, https://aclanthology.org/2023.udfestbr-1.2

Face H (2023) Llama on sagemaker benchmark. https://huggingface.co/blog/llama-sagemaker-benchmark. Accessed: 13 Aug. 2024

Forney G (1973) The viterbi algorithm. Proc IEEE 61(3):268–278. https://doi.org/10.1109/PROC.1973.9030

Ganin Y, Ustinova E, Ajakan H et al (2017) Domain-adversarial training of neural networks. Springer, Cham, pp 189–209. https://doi.org/10.1007/978-3-319-58347-1_10

Ganu G, Elhadad N, Marian A (2009) Beyond the stars: improving rating predictions using review text content. In: WebDB, pp 1–6. https://www.cs.rutgers.edu/~amelie/papers/2009/WebDB2009.pdf

Gao R, Liu F, Zhang J et al (2021) Maximum mean discrepancy test is aware of adversarial attacks. In: Proceedings of the 38th international conference on machine learning, ICML'21, vol 139. Morgan Kaufmann, Vienna, Austria (Online), pp 3564–3575, https://proceedings.mlr.press/v139/gao21b.html

Gong C, Yu J, Xia R (2020) Unified feature and instance based domain adaptation for aspect-based sentiment analysis. In: Proceedings of the 2020 conference on empirical methods in natural language processing. Association for Computational Linguistics, Online, EMNLP'20, pp 7035–7045,https://doi.org/10.18653/v1/2020.emnlp-main.572

Goodfellow I, Pouget-Abadie J, Mirza M et al (2020) Generative adversarial networks. Commun ACM 63(11):139–144. https://doi.org/10.1145/3422622

Gretton A, Borgwardt KM, Rasch MJ et al (2012) A kernel two-sample test. J Mach Learn Res 13(25):723–773

Grover A, Leskovec J (2016) node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, San Francisco, CA, USA, SIGKDD'16, p 855–864. https://doi.org/10.1145/2939672.2939754

Guo JL, Peng JE, Wang HC (2013) An opinion feature extraction approach based on a multidimensional sentence analysis model. Cybern Syst 44:379–401. https://doi.org/10.1080/01969722.2013.789649

Guo H, Zhu H, Guo Z et al (2011) Domain customization for aspect-oriented opinion analysis with multi-level latent sentiment clues. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, Glasgow, Scotland, UK, CIKM'11, pp 2493–2496. https://doi.org/10.1145/2063576.2064000

He R, Lee WS, Ng HT et al (2018) Exploiting document knowledge for aspect-level sentiment classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL'18, vol 2. Association for Computational Linguistics, Melbourne, Australia, pp 579–585. https://doi.org/10.18653/v1/P18-2092

Heck JC, Salem FM (2017) Simplified minimal gated unit variations for recurrent neural networks. In: Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems. IEEE, Boston, MA, USA, MWSCAS'17, pp 1593–1596, https://doi.org/10.1109/MWSCAS.2017.8053242

Hinton GE, Krizhevsky A, Wang SD (2011) Transforming auto-encoders. In: Proceedings of the Artificial Neural Networks and Machine Learning. Springer, Espoo, Finland, ICANN'11, pp 44–51, https://doi.org/10.1007/978-3-642-21735-7_6

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Houlsby N, Giurgiu A, Jastrzebski S et al (2019) Parameter-efficient transfer learning for nlp. In: Proceedings of the 36th International Conference on Machine Learning. Morgan Kaufmann, Long Beach, CA, USA, ICML'19, pp 2790–2799, https://proceedings.mlr.press/v97/houlsby19a.html

Howard P, Ma A, Lal V et al (2022) Cross-domain aspect extraction using transformers augmented with knowledge graphs. In: Proceedings of the 31st ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, Atlanta, GA, USA, CIKM'22, pp 780–790. https://doi.org/10.1145/3511808.3557275

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, Seattle, WA, USA, SIGKDD'04, pp 168–177. https://doi.org/10.1145/1014052.1014073

Hu X, Yamamura M (2023) Hierarchical fusion network with enhanced knowledge and contrastive learning for multimodal aspect-based sentiment analysis on social media. Sensors 23(17):7330. https://doi.org/10.3390/s23177330

Hu M, Wu Y, Zhao S et al (2019) Domain-invariant feature distillation for cross-domain sentiment classification. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, Hong Kong, China, EMNLP-IJCNLP'19, pp 5559–5568. https://doi.org/10.18653/v1/D19-1558

Hu M, Gao H, Wu Y et al (2022) Fine-grained domain adaptation for aspect category level sentiment analysis. IEEE Trans Affect Comput, pp 1–12. https://doi.org/10.1109/TAFFC.2022.3228695

Huang X, Li J, Wu J et al (2023) Transfer learning with document-level data augmentation for aspect-level sentiment classification. IEEE Trans Big Data 9(6):1643–1657. https://doi.org/10.1109/TBDATA.2023.3310267

Hussein S (2021) Twitter sentiments dataset. https://doi.org/10.17632/Z9ZW7NT5H2.1

Jakob N, Gurevych I (2010) Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Cambridge, Massachusetts, USA, EMNLP'10, p 1035–1045, https://aclanthology.org/D10-1101/

Jiang Q, Chen L, Xu R et al (2019a) A challenge dataset and effective models for aspect-based sentiment analysis. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, Hong Kong, China, EMNLP-IJCNLP'19, pp 6280–6285

Jiang Z, Wang J, Zhao L et al (2019b) Cross-domain aspect category transfer and detection via traceable heterogeneous graph representation learning. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, Beijing, China, CIKM'19, p 289–298, https://doi.org/10.1145/3357384.3357989

Jiang X, Bai R, Wang Z et al (2024) Cross-domain aspect-based sentiment classification with tripartite graph modeling. IEEE/ACM Trans Audio Speech Language Process 32:1623–1635. https://doi.org/10.1109/TASLP.2024.3365975

Kan TJ, Chang CH (2022) Home appliance review analysis via adversarial reptile. In: Proceedings of the IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. Association for Computing Machinery, Melbourne, VIC, Australia, WI-IAT'21, pp 64–70, https://doi.org/10.1145/3486622.3493958

Kannan GT, Gunasekar M, Ponnazhagan N et al (2023) Aspect based sentiment aware word embedding for cross domain sentiment analysis. In: Proceedings of the 2023 international conference on computer communication and informatics. IEEE, Coimbatore, India, ICCCI'23, pp 1–5. https://doi.org/10.1109/ICCCI56745.2023.10128251

Karagiannakos S (2021) Best graph neural networks architectures: Gcn, gat, mpnn and more. The AI Summer https://theaisummer.com/gnn-architectures/

Ke Z, Liu B, Xu H et al (2021) CLASSIC: Continual and contrastive learning of aspect sentiment classification tasks. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, EMNLP'21, pp 6871–6883, https://doi.org/10.18653/v1/2021.emnlp-main.550

Kessler JS, Eckert M, Clark L et al (2010) The 2010 icwsm jdpa sentment corpus for the automotive domain. In: Proceedings of the 4th international AAAI conference on weblogs and social media data workshop challenge. AAAI Press, Washington, DC, USA, ICWSM-DWC'10. http://www.cs.indiana.edu/~jaskessl/icwsm10.pdf

Kiritchenko S, Zhu X, Cherry C et al (2014) NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). Association for Computational Linguistics, Dublin, Ireland, SemEval'14, pp 437–442. https://doi.org/10.3115/v1/S14-2076

Klein A, Pereg O, Korat D et al (2022) Opinion-based relational pivoting for cross-domain aspect term extraction. In: Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics, Dublin, Ireland, WASSA'22, pp 104–112, https://doi.org/10.18653/v1/2022.wassa-1.11

Klinger R, Cimiano P (2014) The USAGE review corpus for fine grained multi lingual opinion analysis. In: Proceedings of the 9th International Conference on Language Resources and Evaluation. European Language Resources Association, Reykjavik, Iceland, LREC'14, pp 2211–2218. http://www.lrec-conf.org/proceedings/lrec2014/pdf/85_Paper.pdf

Knoester J, Frasincar F, Truşcă MM (2023) Cross-domain aspect-based sentiment analysis using domain adversarial training. World Wide Web 26(6):4047–4067

Koehn P, Knowles R (2017) Six challenges for neural machine translation. In: Proceedings of the first workshop on neural machine translation, pp 28–39. https://doi.org/10.18653/v1/W17-3204

Koolen M, Bogers T, Gäde M et al (2016) Overview of the clef 2016 social book search lab. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5–8, 2016, Proceedings 7, Springer, pp 351–370, https://doi.org/10.1007/978-3-319-44564-9

Kramer MA (1991) Nonlinear principal component analysis using autoassociative neural networks. AIChE J 37(2):233–243. https://doi.org/10.1002/aic.690370209

Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning. Morgan Kaufmann, Williamstown, MA, USA, ICML'01, pp 282–289. http://www.aladdin.cs.cmu.edu/papers/pdfs/y2001/crf.pdf

Lark J, Morin E, Peña Saldarriaga S (2018) A comparative study of target-based and entity-based opinion extraction. In: Proceedings of the computational linguistics and intelligent text processing. Springer, Budapest, CICLing'17, pp 211–223. https://doi.org/10.1007/978-3-319-77116-8_16

Lee J, Frasincar F, Trusca MM (2023) A cross-domain aspect-based sentiment classification by masking the domain-specific words. In: Proceedings of the 38th ACM/SIGAPP symposium on applied computing. association for computing machinery, Tallinn, Estonia, SAC'23, pp 1595–1602. https://doi.org/10.1145/3555776.3577633

Lester B, Al-Rfou R, Constant N (2021) The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 3045–3059. https://doi.org/10.18653/v1/2021.emnlp-main.243

Lewis M, Liu Y, Goyal N et al (2020) BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Online, ACL'20, pp 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Li X, Lei L (2021) A bibliometric analysis of topic modelling studies (2000–2017). J Inf Sci 47(2):161–175. https://doi.org/10.1177/0165551519877049

Li F, Pan SJ, Jin O et al (2012) Cross-domain co-extraction of sentiment and topic lexicons. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, ACL'12, vol 1. Association for Computational Linguistics, Jeju Island, Korea, pp 410–419. https://aclanthology.org/P12-1043/

Li Z, Li X, Wei Y et al (2019a) Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, Hong Kong, China, EMNLP-IJCNLP'19, pp 4590–4600. https://doi.org/10.18653/v1/D19-1466

Li Z, Wei Y, Zhang Y et al (2019b) Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI'19, vol 33. AAAI Press, Honolulu, HI, USA, pp 4253–4260. https://doi.org/10.1609/aaai.v33i01.33014253

Li J, Yu J, Xia R (2022) Generative cross-domain data augmentation for aspect and opinion co-extraction. In: Proceedings of the 2022 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, WA, USA, NAACL'22, pp 4219–4229. https://doi.org/10.18653/v1/2022.naacl-main.312

Liang T, Wang W, Lv F (2022) Weakly supervised domain adaptation for aspect extraction via multilevel interaction transfer. IEEE Trans Neural Netw Learning Syst 33(10):5818–5829. https://doi.org/10.1109/TNNLS.2021.3071474

Liang B, Yin R, Du J et al (2023) Embedding refinement framework for targeted aspect-based sentiment analysis. IEEE Trans Affect Comput 14(1):279–293. https://doi.org/10.1109/TAFFC.2021.3071388

Liu B (2012) Sentiment analysis and opinion mining. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-031-02145-9

Liu N, Zhao J (2022) A bert-based aspect-level sentiment analysis algorithm for cross-domain text. Comput Intell Neurosci 2022:8726621. https://doi.org/10.1155/2022/8726621

Liu Q, Gao Z, Liu B et al (2015) Automated rule selection for aspect extraction in opinion mining. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence. AAAI Press, Buenos Aires, Argentina, IJCAI'15, https://dl.acm.org/doi/10.5555/2832415.2832429

Liu J, Teng Z, Cui L et al (2021) Solving aspect category sentiment analysis as a text generation task. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, EMNLP'21, pp 4406–4416. https://doi.org/10.18653/v1/2021.emnlp-main.361

Liu J, Zhong Q, Ding L et al (2023) Unified instance and knowledge alignment pretraining for aspect-based sentiment analysis. IEEE/ACM Trans Audio Speech Language Process 31:2629–2642. https://doi.org/10.1109/TASLP.2023.3290431

Lopes E, Correa U, Freitas L (2021) Exploring bert for aspect extraction in portuguese language. In: Proceedings of the International Florida Artificial Intelligence Conference, FLAIRS'21, vol 34. Florida Online Journals, North Miami Beach, FL, USA, https://doi.org/10.32473/flairs.v34i1.128357

Machado M, Pardo TAS (2022) Evaluating methods for extraction of aspect terms in opinion texts in Portuguese - the challenges of implicit aspects. In: Proceedings of the 13th international conference on language resources and evaluation. European Language Resources Association, Marseille, France, LREC'22, pp 3819–3828. https://aclanthology.org/2022.lrec-1.407

Maharani W, Widyantoro DH, Khodra ML (2015) Sae: Syntactic-based aspect and opinion extraction from product reviews. In: Proceedings of the 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications. IEEE, Mueang Chonburi, Chonburi, Thailand, ICAICTA'15, pp 1–6. https://doi.org/10.1109/ICAICTA.2015.7335371

Majumder N, Bhardwaj R, Poria S et al (2022) Improving aspect-level sentiment analysis with aspect extraction. Neural Comput Appl 34(11, SI):8333–8343. https://doi.org/10.1007/s00521-020-05287-7

Mann B, Ryder N, Subbiah M, et al (2020) Language models are few-shot learners. arXiv preprint arXiv:2005.14165 1:3. https://doi.org/10.48550/arXiv.2005.14165

Marcacini RM, Rossi RG, Matsuno IP et al (2018) Cross-domain aspect extraction for sentiment analysis: a transductive learning approach. Decis Support Syst 114:70–80. https://doi.org/10.1016/j.dss.2018.08.009

Marquez L, Padro L, Rodriguez H (2000) A machine learning approach to pos tagging. Mach Learn 39:59–91. https://doi.org/10.1023/A:1007673816718

Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39–41. https://doi.org/10.1145/219717.219748

Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29(3):436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Ni J, Li J, McAuley J (2019) Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, Hong Kong, China, EMNLP-IJCNLP'19, pp 188–197, https://doi.org/10.18653/v1/D19-1018

Oord Avd, Li Y, Vinyals O (2019) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748. https://doi.org/10.48550/arXiv.1807.03748

Ouyang F, Shen B (2023) A mutual mean teacher framework for cross-domain aspect-based sentiment analysis. J Supercomputing pp 1–23. https://doi.org/10.1007/s11227-023-05792-1

Pak MY, Gunal S (2022) A model for cross-domain opinion target extraction in sentiment analysis. Computer Syst Sci Eng 42(3):1215–1239. https://doi.org/10.32604/csse.2022.023051

Pastore P, Iovine A, Narducci F et al (2021) A general aspect-term-extraction model for multi-criteria recommendations. In: Proceedings of the Joint KaRS and ComplexRec Workshop, CEUR-WS'21, vol 2960. CEUR-WS, Amsterdam, The Netherlands (Online), https://ceur-ws.org/Vol-2960/

Peng S, Cao L, Zhou Y et al (2022) A survey on deep learning for textual emotion analysis in social networks. Digital Commun Netw 8(5):745–762. https://doi.org/10.1016/j.dcan.2021.10.003

Pennington J, Socher R, Manning C (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Doha, Qatar, EMNLP'14, pp 1532–1543, https://doi.org/10.3115/v1/D14-1162

Pereg O, Korat D, Wasserblat M (2020) Syntactically aware cross-domain aspect and opinion terms extraction. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), COLING'20, pp 1772–1777, https://doi.org/10.18653/v1/2020.coling-main.158

Pontiki M, Galanis D, Pavlopoulos J et al (2014) SemEval-2014 task 4: Aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for Computational Linguistics, Dublin, Ireland, SemEval'14, pp 27–35, https://doi.org/10.3115/v1/S14-2004

Pontiki M, Galanis D, Papageorgiou H et al (2015) SemEval-2015 task 12: Aspect based sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics, Denver, Colorado, SemEval'15, pp 486–495, https://doi.org/10.18653/v1/S15-2082

Pontiki M, Galanis D, Papageorgiou H et al (2016) SemEval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, San Diego, California, SemEval'16, pp 19–30, https://doi.org/10.18653/v1/S16-1002

Qiu G, Liu B, Bu J et al (2011) Opinion word expansion and target extraction through double propagation. Comput Linguist 37(1):9–27. https://doi.org/10.1162/coli_a_00034

Radford A, Wu J, Child R et al (2019) Language models are unsupervised multitask learners. OpenAI 1(8):9

Raffel C, Shazeer N, Roberts A et al (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learning Res 21(140):1–67

Rajpurkar P, Zhang J, Lopyrev K et al (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, TX, USA, EMNLP'16, pp 2383–2392https://doi.org/10.18653/v1/D16-1264

Ramos DL, Fuentes FJA (2023) A model of continual and deep learning for aspect based in sentiment analysis. J Autom Mobile Robotics Intell Syst 17(1):3–12. https://doi.org/10.14313/JAMRIS/1-2023/1

Robertson S, Zaragoza H (2009) The probabilistic relevance framework: Bm25 and beyond. Found Trends Inf Retr 3(4):333–389. https://doi.org/10.1561/1500000019

Rosenthal S, Farra N, Nakov P (2017) SemEval-2017 task 4: sentiment analysis in Twitter. In: Proceedings of the 11th international workshop on semantic evaluation. Association for Computational Linguistics, Vancouver, Canada, SemEval'17. https://doi.org/10.18653/v1/S17-2088

Ruskanda FZ, Widyantoro DH, Purwarianti A (2019) Sequential covering rule learning for language rule-based aspect extraction. In: Proceedings of the 2019 International Conference on Advanced Computer Science and information Systems. IEEE, Nusa Dua, Bali, Indonesia, ICACSIS'19, pp 229–234. https://doi.org/10.1109/ICACSIS47736.2019.8979743

Santini M (2004) State-of-the-art on automatic genre identification. Tech. rep., Technical Report No. ITRI-04-03). Information Technology Research Institute

Santos BND, Marcacini RM, Rezende SO (2021) Multi-domain aspect extraction using bidirectional encoder representations from transformers. IEEE Access 9:91604–91613. https://doi.org/10.1109/ACCESS.2021.3089099

Saunders D (2022) Domain adaptation and multi-domain adaptation for neural machine translation: a survey. https://doi.org/10.48550/arXiv.2104.06951. arXiv:2104.06951

Scannavino KRF, Nakagawa EY, Fabbri SCPF et al (2017) Revisão Sistemática da Literatura em Engenharia de Software: teoria e prática. Elsevier

Shi J, Li W, Bai Q et al (2023) Soft prompt enhanced joint learning for cross-domain aspect-based sentiment analysis. Intell Syst Appl 20:200292. https://doi.org/10.1016/j.iswa.2023.200292

Sinha A, Kedas S, Kumar R et al (2022) Sentfin 1.0: entity-aware sentiment analysis for financial news. J Assoc Inf Sci Technol 73(9):1314–1335. https://doi.org/10.1002/asi.24634

Sinoara RA, Marcacini RM, Rezende SO (2021) Mineração de textos e semântica: desafios, abordagens e aplicações. Revista de Sistemas de Informação da FSMA 27:41–43. https://www.fsma.edu.br/si/edicao27/FSMA_SI_2021_1_Principal_04.html

Speer R, Chin J, Havasi C (2017) Conceptnet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. AAAI press, San Francisco, CA, USA, AAAI'17, pp 4444–4451. https://doi.org/10.5555/3298023.3298212

Sun K, Zhang R, Samuel M et al (2023) Self-training through classifier disagreement for cross-domain opinion target extraction. In: Proceedings of the International World Wide Web Conference 2023. Association for Computing Machinery, Austin, TX, USA, WWW'23, pp 1594–1603. https://doi.org/10.1145/3543507.3583325

Tarvainen A, Valpola H (2017) Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of the Advances in Neural Information Processing Systems, NeurIPS'17, vol 30. Advances in Neural Information Processing Systems, Long Beach, CA, USA, https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf

Toledo-Ronen O, Orbach M, Katz Y, et al (2022) Multi-domain targeted sentiment analysis. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, WA, USA, NAACL'22, pp 2751–2762. https://doi.org/10.18653/v1/2022.naacl-main.198

Toprak C, Jakob N, Gurevych I (2010) Sentence and expression level annotation of opinions in user-generated discourse. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Uppsala, Sweden, ACL'10, pp 575–584. https://aclanthology.org/P10-1059/

Touvron H, Lavril T, Izacard G et al (2023) Llama: Open and efficient foundation language models. arXiv eprint arXiv:2302.13971 pp 1–27. https://doi.org/10.48550/arXiv.2302.13971

Tran TU, Thanh Thi Hoang H, Dang PH et al (2021) Multidomain supervised aspect-based sentiment analysis using cnn_bidirectional lstm model. In: Proceedings of the 2021 RIVF International Conference on Computing and Communication Technologies. IEEE, Hanoi, Vietnam, RIVF'21, pp 1–6. https://doi.org/10.1109/RIVF51545.2021.9642146

van Berkum S, van Megen S, Savelkoul M et al (2022) Fine-tuning for cross-domain aspect-based sentiment classification. In: Proceedings of the IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. Association for Computing Machinery, Melbourne, VIC, Australia, WI-IAT'21, pp 524–531. https://doi.org/10.1145/3486622.3494003

Van Thin D, Quoc Ngo H, Ngoc Hao D, Luu-Thuy Nguyen N (2023) Exploring zero-shot and joint training cross-lingual strategies for aspect-based sentiment analysis based on contextualized multilingual language models. J. Inf. Telecommun. 7(2):121–143. https://doi.org/10.1080/24751839.2023.2173843

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S et al (eds) Proceedings of the Advances in Neural Information Processing Systems, NeurIPS'17, vol 30. Advances in Neural Information Processing Systems, Long Beach, CA, USA. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wallach HM (2004) Conditional random fields: An introduction. Tech. Rep. MS-CIS-04-21, University of Pennsylvania, http://www.cs.umass.edu/~wallach/technical_reports/wallach04conditional.pdf

Wang W, Pan SJ (2018) Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In: Proceedings of the 56th annual meeting of the association for computational linguistics, ACL'18, vol 1. Association for Computational Linguistics, Melbourne, Australia, pp 2171–2181. https://doi.org/10.18653/v1/P18-1202

Wang W, Pan SJ (2019a) Syntactically meaningful and transferable recursive neural networks for aspect and opinion extraction. Comput Linguist 45(4):705–736. https://doi.org/10.1162/coli_a_00362

Wang W, Pan SJ (2019b) Transferable interactive memory network for domain adaptation in fine-grained opinion extraction. In: Proceedings of the 33rd AAAI conference on artificial intelligence, AAAI'19, vol 33. AAAI Press, Honolulu, HI, USA, pp 7192–7199. https://doi.org/10.1609/aaai.v33i01.33017192

Wang Z, Guo J (2024) Self-adaptive attention fusion for multimodal aspect-based sentiment analysis. Math Biosci Eng 21(1):1305–1320. https://doi.org/10.3934/mbe.2024056

Wang W, Pan SJ, Dahlmeier D et al (2016) Recursive neural conditional random fields for aspect-based sentiment analysis. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, EMNLP'16, pp 616–626. https://doi.org/10.18653/v1/D16-1059

Wang G, Pu P, Liang Y (2018) Topic and sentiment words extraction in cross-domain product reviews. Wireless Pers Commun 102(2):1773–1783. https://doi.org/10.1007/s11277-017-5235-7

Wang JZ, Zhao S, Wu C et al (2023) Unlocking the emotional world of visual media: an overview of the science, research, and impact of understanding emotion. Proc IEEE 111(10):1236–1286. https://doi.org/10.1109/JPROC.2023.3273517

Xu H, Liu B, Shu L et al (2018) Double embeddings and CNN-based sequence labeling for aspect extraction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL'18, vol 2. Association for Computational Linguistics, Melbourne, Australia, pp 592–598. https://doi.org/10.18653/v1/P18-2094

Xu H, Liu B, Shu L et al (2019) BERT post-training for review reading comprehension and aspect-based sentiment analysis. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL'19, vol 1. Association for Computational Linguistics, Minneapolis, MN, USA, pp 2324–2335, https://doi.org/10.18653/v1/N19-1242

Xu H, Liu B, Shu L et al (2020) DomBERT: Domain-oriented language model for aspect-based sentiment analysis. In: Cohn T, He Y, Liu Y (eds) Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, EMNLP'20, pp 1725–1731. https://doi.org/10.18653/v1/2020.findings-emnlp.156

Xue J, Li Y, Li Z et al (2023) A cross-domain generative data augmentation framework for aspect-based sentiment analysis. Electronics (Switzerland) 12(13):2949. https://doi.org/10.3390/electronics12132949

Yang M, Qu Q, Shen Y et al (2020) Cross-domain aspect/sentiment-aware abstractive review summarization by combining topic modeling and deep reinforcement learning. Neural Comput Appl 32(11):6421–6433. https://doi.org/10.1007/s00521-018-3825-2

Yang M, Yin W, Qu Q et al (2021) Neural attentive network for cross-domain aspect-level sentiment classification. IEEE Trans Affect Comput 12(3):761–775. https://doi.org/10.1109/TAFFC.2019.2897093

Yang Z, Wang B, Li X et al (2023) S3 map: semisupervised aspect-based sentiment analysis with masked aspect prediction. Knowl-Based Syst 269:110513. https://doi.org/10.1016/j.knosys.2023.110513

Yelp Inc (2024) Yelp open dataset. https://www.yelp.com/dataset. Accessed 16 Aug 2024

Yu J, Jiang J (2019) Adapting bert for target-oriented multimodal sentiment classification. In: Proceedings of the 28th international joint conference on artificial intelligence. International joint conferences on artificial intelligence organization, Macao, China, IJCAI'19, pp 5408–5414. https://doi.org/10.24963/ijcai.2019/751

Yu J, Gong C, Xia R (2021) Cross-domain review generation for aspect-based sentiment analysis. In: Findings of the association for computational linguistics. association for computational linguistics, Online, ACL-IJCNLP'21, pp 4767–4777, https://doi.org/10.18653/v1/2021.findings-acl.421

Yu J, Zhao Q, Xia R (2023) Cross-domain data augmentation with domain-adaptive language modeling for aspect-based sentiment analysis. In: Proceedings of the 61st annual meeting of the association for computational linguistics, ACL'23, vol 1. Association for Computational Linguistics, Toronto, Canada, pp 1456–1470. https://doi.org/10.18653/v1/2023.acl-long.81

Zeng R, Liu H, Peng S et al (2023a) Cnn-based broad learning for cross-domain emotion classification. Tsinghua Sci Technol 28(2):360–369. https://doi.org/10.26599/TST.2022.9010007

Zeng Y, Wang G, Ren H et al (2023b) A knowledge-enhanced and topic-guided domain adaptation model for aspect-based sentiment analysis. IEEE Trans Affect Comput, pp 1–13. https://doi.org/10.1109/TAFFC.2023.3292213

Zhang W, Deng Y, Li X et al (2021) Aspect sentiment quad prediction as paraphrase generation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, EMNLP'21, pp 9209–9219. https://doi.org/10.18653/v1/2021.emnlp-main.726

Zhang W, Li X, Deng Y et al (2022) A survey on aspect-based sentiment analysis: tasks, methods, and challenges. IEEE Trans Knowl Data Eng 35(11):11019–11038. https://doi.org/10.1109/TKDE.2022.3230975

Zhang B, Fu X, Luo C et al (2023a) Cross-domain aspect-based sentiment classification by exploiting domain-invariant semantic-primary feature. IEEE Trans Affect Comput 14(4):1–15. https://doi.org/10.1109/TAFFC.2023.3239540

Zhang K, Liu Q, Qian H et al (2023b) Eatn: an efficient adaptive transfer network for aspect-level sentiment analysis. IEEE Trans Knowl Data Eng 35(1):377–389. https://doi.org/10.1109/TKDE.2021.3075238

Zhang Y, Zhu C, Xie Y (2023c) Fine-grained sentiment analysis of cross-domain chinese e-commerce texts based on skep-gram-cdnn. IEEE Access 11:74058–74070. https://doi.org/10.1109/ACCESS.2023.3296447

Zhao C, Wang S, Li D et al (2021) Cross-domain sentiment classification via parameter transferring and attention sharing mechanism. Inf Sci 578:281–296. https://doi.org/10.1016/j.ins.2021.07.001

Zhao Y, Soerjodjojo E, Che H (2022) Methods to enhance bert in aspect-based sentiment classification. In: Proceedings of the 2022 Euro-Asia Conference on Frontiers of Computer Science and Information Technology. IEEE, Beijing, China, FCSIT'22, pp 21–27. https://doi.org/10.1109/FCSIT57414.2022.00016

Zhao S, Hong X, Yang J et al (2023) Toward label-efficient emotion and sentiment analysis. Proc IEEE 111(10):1159–1197. https://doi.org/10.1109/jproc.2023.3309299

Zhao C, Wu M, Yang X et al (2024) Cross-domain aspect-based sentiment classification with pre-training and fine-tuning strategy for low-resource domains. ACM Trans Asian Low-Resour Language Inf Process. https://doi.org/10.1145/3653299

Zheng Y, Zhang R, Wang S et al (2020) Anchored model transfer and soft instance transfer for cross-task cross-domain learning: A study through aspect-level sentiment classification. In: Proceedings of the International World Wide Web Conference 2020. Association for Computing Machinery, Taipei, Taiwan, WWW'20, pp 2754–2760. https://doi.org/10.1145/3366423.3380034

Zhou J, Chen Q, Huang JX et al (2020) Position-aware hierarchical transfer model for aspect-level sentiment classification. Inf Sci 513:1–16. https://doi.org/10.1016/j.ins.2019.11.048

Zhou J, Zhao J, Huang JX et al (2021a) Masad: a large-scale dataset for multimodal aspect-based sentiment analysis. Neurocomputing 455:47–58. https://doi.org/10.1016/j.neucom.2021.05.040

Zhou Y, Zhu F, Song P et al (2021b) An adaptive hybrid framework for cross-domain aspect-based sentiment analysis. In: Proceedings of the 35th AAAI conference on artificial intelligence, AAAI'21, vol 35. AAAI Press, Online, pp 14630–14637. https://doi.org/10.1609/aaai.v35i16.17719

Zorn KM, Foil DH, Lane TR et al (2020) Machine learning models for estrogen receptor bioactivity and endocrine disruption prediction. Environ Sci Technol 54(19):12202–12213. https://doi.org/10.1021/acs.est.0c03982

Zou H, Wang Y (2025) Large language model augmented syntax-aware domain adaptation method for aspect-based sentiment analysis. Neurocomputing 625:129472. https://doi.org/10.1016/j.neucom.2025.129472