Artificial Intelligence Review (2024) 57:17
https://doi.org/10.1007/s10462-023-10633-x

Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: a comprehensive survey

Hao Zhang[1,2] · Yu-N Cheah[1] · Osamah Mohammed Alyasiri[1,3] · Jieyu An[1]

Published online: 20 January 2024

Abstract
In contrast to earlier ABSA studies primarily concentrating on individual sentiment components, recent research has ventured into more complex ABSA tasks encompassing multiple elements, including pair, triplet, and quadruple sentiment analysis. Quadruple sentiment analysis, also called aspect-category-opinion-sentiment quadruple Extraction (ACOSQE), aims to dissect aspect terms, aspect categories, opinion terms, and sentiment polarities while considering implicit sentiment within sentences. Nonetheless, a comprehensive overview of ACOSQE and its corresponding solutions is currently lacking. This is the precise gap that our survey seeks to address. To be more precise, we systematically reclassify all subtasks of ABSA, reorganizing existing research from the perspective of the involved sentiment elements, with a primary focus on the latest advancements in the ACOSQE task. Regarding solutions, our survey offers a comprehensive summary of the state-of-the-art utilization of language models within the ACOSQE task. Additionally, we explore the application of ChatGPT in sentiment analysis. Finally, we review emerging trends and discuss the challenges, providing insights into potential future directions for ACOSQE within the broader context of ABSA.

Keywords Quadruple sentiment analysis · Aspect sentiment quadruple extraction · Implicit aspect and opinions · ACOSQE · Survey · Pre-trained language models · Aspect-based sentiment analysis (ABSA) · BERT · BART · T5 · ChatGPT

✉ Hao Zhang
zhanghaousm@gmail.com

✉ Yu-N Cheah
yncheah@usm.my

Osamah Mohammed Alyasiri
osama.alyasiri@atu.edu.iq

Jieyu An
anjieyu@student.usm.my

[1] School of Computer Sciences, Universiti Sains Malaysia, 11800 USM Penang, Malaysia

[2] School of Computer Science and Engineering, Cangzhou Normal University, Cangzhou 061001, China

[3] Karbala Technical Institute, Al-Furat Al-Awsat Technical University, Karbala 56001, Iraq

1 Introduction

Aspect-based sentiment analysis (ABSA) is a significant area of research within the field of fine-grained sentiment analysis. One of the critical tasks in ABSA is aspect category opinion sentiment Quadruple Extraction (ACOSQE), which involves extracting four elements of information from text, including category, aspect, opinion, and sentiment polarity. ACOSQE has practical applications in business evaluation and public opinion analysis. It has recently become a popular research topic, with numerous scholars devoting attention to it. The process of ACOSQE is illustrated in Fig. 1.

Several review papers have been published summarizing the relevant tasks, methods, and challenges in the ABSA field. Zhang et al. (2022) provided a comprehensive overview of ABSA research, including different sub-tasks introductions, commonly used datasets, and evaluation metrics, and also discussed the development process and research trends of ABSA methods, including traditional machine learning methods and deep learning methods.

Zhang et al. (2018), Zhou et al. (2019), and Zhu et al. (2022) had conducted extensive research in the development process and research trends of ABSA methods based on deep learning, including convolutional neural networks, recurrent neural networks, attention mechanisms, and pre-training language models, are discussed. It was believed in these works that the proposal and widespread application of deep learning methods provided strong support for the research on ACOSQE.

Implicit aspect detection is an important problem in ABSA because it can identify aspects and opinions not explicitly mentioned in the text, thus improving the accuracy of sentiment analysis (SA). Soni and Rambola (2022) surveyed implicit aspect detection and emphasized the significance of identifying aspects not explicitly mentioned in the text to improve the accuracy of aspect extraction and polarity classification in SA. Discussions on the terminology, issues, and scope of implicit aspects were covered, and recent techniques proposed for detecting them were reviewed. Several implicit sentiment datasets were also introduced, including those related to ACOSQE.

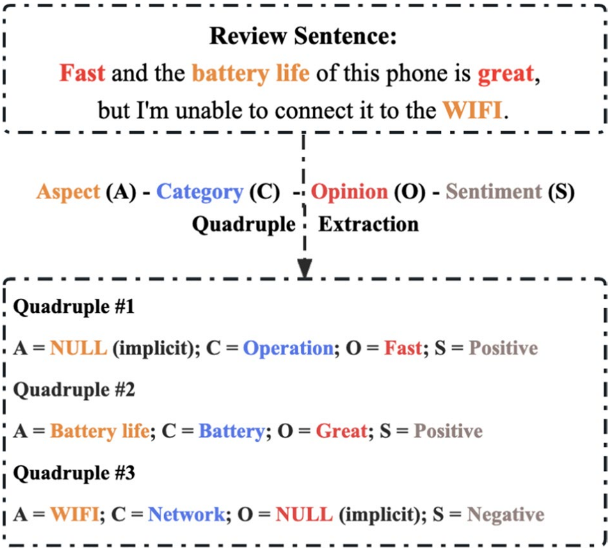
Review Sentence:
Fast and the battery life of this phone is great,
but I'm unable to connect it to the WIFI.
Aspect (A) - Category (C) - Opinion (O) - Sentiment (S)
Quadruple Extraction
Quadruple #1
A = NULL (implicit); C = Operation; O = Fast; S = Positive
Quadruple #2
A = Battery life; C = Battery; O = Great; S = Positive
Quadruple #3
A = WIFI; C = Network; O = NULL (implicit); S = Negative

Fig. 1 An example of the aspect-category-opinion-sentiment quadruple extraction task with implicit aspects and opinions

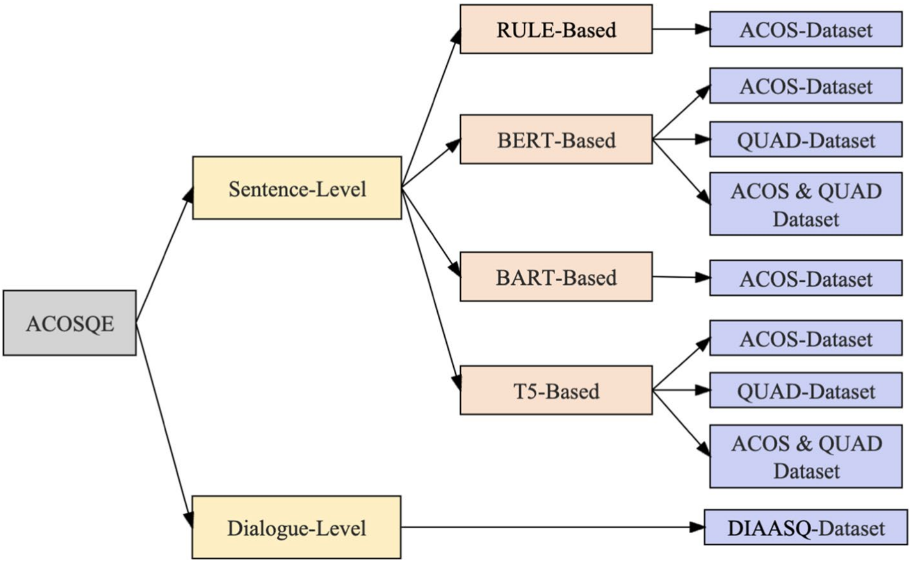

Fig. 2 The research subareas of ACOSQE from the perspective of different approaches and datasets

By analyzing these review papers, ABSA is a challenging task that needs to be explored from multiple perspectives by researchers. However, a systematic review of ACOSQE and their corresponding solutions still needs to be completed. This paper conducted a comprehensive review of research on ACOSQE with implicit aspects and opinions, offering new ideas and directions for research in the ABSA field.

During the course of this investigation, Google Scholar served as the primary search engine to retrieve relevant papers containing keywords such as “quadruple sentiment analysis,” “aspect-based sentiment analysis (ABSA),” “pre-trained language models,” “implicit aspects and opinions,” and “ChatGPT.” These encompassed works from top-tier conferences in Artificial Intelligence (AAAI) and Natural Language Processing (ACL, NAACL, EMNLP), journals, relevant arXiv papers, book chapters, and doctoral theses. The results were assessed, and a classification of subareas was devised based on the predominant research directions in this field (see Fig. 2).

This paper can be classified as a comprehensive survey of ACOSQE with implicit aspects and opinions in the field of ABSA. It first introduces the four sentiment elements of ABSA and discusses the research background related to ACOSQE problems, including concept definitions, sub-task introductions, and dataset characteristics. The paper then explores the research progress of ACOSQE from both sentence-level and dialogue-level perspectives, as shown in Fig. 2. Following that, the paper presents the research on rule-based methods at the sentence level, explicitly using the ACOS-Dataset. In the subsequent sections, various research methods based on pre-trained language models such as BERT, BART, and T5 are introduced, along with their performance on ACOS-Dataset (Cai et al. 2021a), QUAD-Dataset (Zhang et al. 2021b), or a combination of both. Additionally, the paper explores ACOSQE at the dialogue level using the DiaASQ-Dataset (Li et al. 2022b).

Furthermore, the paper discusses the latest research on ChatGPT in the context of SA. It explores the prospects and limitations of large language models like ChatGPT in addressing ACOSQE problems while highlighting emerging trends and unresolved challenges.

This comprehensive review paper provides valuable insights into the current research status of ACOSQE in ABSA.

The paper aims to address the following research questions related to ACOSQE with implicit aspects and opinions in ABSA:

- What is the latest progress in ACOSQE?
- How can methods based on pre-trained language models and datasets be applied in ACOSQE with implicit aspects and opinions?
- What are the techniques for constructing more practical ACOSQE in generative large language models?
- How can advanced language models like ChatGPT improve the accuracy and efficiency of ACOSQE?
- What are the current challenges and future directions in ACOSQE research?
- What are the implications and opportunities for further research in the field of ACOSQE in ABSA?

2 Background

In recent years, the innovative developments of social networking, travel, transportation, and e-commerce platforms such as Facebook, Amazon, eBay, JD.com, Airbnb, and Uber have enriched online content, giving rise to a continuous influx of textual information encompassing user sharing, interactions, and product evaluations. Much of this information is authentic and trustworthy, characterized by its richness, diversity, extensive coverage, and rapid generation. Simultaneously, these texts carry the sentiments and attitudes of individuals towards products, services, organizations, and more.

In the realm of business, SA aids companies in comprehending consumer concerns and product shortcomings, facilitating product enhancements, improved services, and precise marketing strategies. In the political sphere, SA assists governments in understanding public emotions, managing public opinion, refining methodologies, and making informed decisions. However, the manual analysis of such extensive textual data demands significant time and effort, emphasizing the urgent need for automated methods to extract viewpoints and attitudes from the text. SA has gradually become a focal point for researchers, emerging as one of the most active research directions in natural language processing (NLP). SA is typically categorized into document-level, sentence-level, and aspect-level analysis. Early SA predominantly concentrated on document and sentence levels. Over the years, various relevant corpora like “Twitter” have emerged, propelling the advancement of ABSA. Simultaneously, breakthroughs in deep learning technology have led to a shift towards deep learning in NLP, yielding significant accomplishments. The standardization of corpora and the application of deep learning techniques have attracted more attention to ABSA, fostering the development of numerous exceptional systems and models.

Nevertheless, ABSA remains a challenging task. Textual SA can be divided into explicit SA and implicit SA. Much of the research is concentrated in the domain of explicit SA. Implicit SA, due to the absence of explicit sentiment words as cues, introduces complexity to sentiment expression, making it one of the challenges in SA. Early research primarily focused on predicting singular elements, such as extracting aspect terms. In recent years, the ABSA domain introduced pair and triplet prediction tasks, foreseeing sentiment elements in a multi-tuple format. In the past couple of years, the emergence of ACOSQE has gradually shifted focus towards implicit SA, providing comprehensive support for ABSA. The following subsections will explore definitions of ACOS, ABSA, relevant datasets, and evaluation approaches.

2.1 ACOS (aspect-category-opinion-sentiment)

According to Liu (2012), SA is a field of study in which computers analyze people's expressions of opinions, emotions, and attitudes in text. It is an interesting problem with significant implications for both business and society.

ABSA is a sub-task of SA that focuses on aspect-level sentiment analysis, identifying the following elements (ACOS) in each sentence.

- Aspect Category $\boldsymbol{c} \in \boldsymbol{V}_{\mathrm{x}}$ (Pre-Defined)
- Aspect Term $\boldsymbol{a} \in \boldsymbol{V}_{\mathrm{x}} \cup \{NULL\}$
- Opinion Term $\boldsymbol{o} \in \boldsymbol{V}_{\mathrm{x}} \cup \{NULL\}$
- Sentiment Polarity $\boldsymbol{s} \in \{\mathrm{POS, NEU, NEG}\}$

These four sentiment elements constitute the core of ABSA research. ACOSQE aims to extract all aspect-category-opinion-sentiment polarity quadruples from the text to more comprehensively and accurately reflect sentiment information.

Aspect terms *(a)* refers to specific characteristics of the entity or topic being evaluated or discussed, which can be explicitly or implicitly mentioned and can be associated with different categories or domains. For example, for a mobile phone, possible aspects include taking pictures, making calls, browsing the internet, and attributes such as color, size, and weight.

Categories *(c)* refers to the broader field or domain to which an aspect belongs. It can be predefined or extracted from the text, and it helps contextualize the SA results. For example, categories could be design or brand for the laptop domain.

Opinion terms *(o)* can be conveyed either explicitly or implicitly. Explicit expressions involve words or phrases expressing a positive, negative, or neutral attitude toward a specific aspect or category. It can include adjectives, adverbs, or verbs that indicate a particular evaluation or attitude. Implicit expressions do not directly state a positive or negative attitude but imply a particular evaluation or attitude. It can include more subtle language, such as humor, irony, slang, or other nuanced phrasing that allows for multiple interpretations. Considering not only explicit but also implicit expressions of opinion terms is essential for gaining a comprehensive understanding of people's attitudes, beliefs, and opinions. By analyzing both types of expression used, researchers can uncover deeper insights into how individuals evaluate and feel about various aspects and categories.

Sentiment polarity *(s)* refers to the direction of the sentiment towards an aspect or category, including *positive*, *negative*, and *neutral*. Sentiment polarity can be binary or multiclass and can be inferred from the expression of the opinion and its context. For example, in a movie review, sentiment polarity could be positive (such as "this is a very good movie") or negative (such as "this movie is very boring").

The four elements of ACOS sentiment analysis provide a comprehensive framework for analyzing and extracting sentiment information from text, helping us capture the diversity and complexity of people's opinions and attitudes towards products and services.

2.2 ABSA definition

Based on the combination of the a, c, o, s, there exist several sub-tasks in ABSA. We summarize these sub-tasks in Table 1. Specifically, their definitions are as follows in the context of the example sentence: "The battery life of this phone is great."

2.2.1 Aspect term extraction (ATE)

ATE is a pivotal and fundamental task in ABSA, aiming to identify and extract aspect terms mentioned in sentences, specifically focusing on aspects of products or services being discussed. In this example, the aspect term is "battery life."

Early ATE relied on heuristic rules and sentiment lexicons and fell within the unsupervised techniques category. The most commonly used publicly available lexicons for this task include WordNet-Affect (WordNet-Affect 2004), Senti-WordNet (Senti-WordNet 2006), and SenticNet (SenticNet 2010). Generally, domain-specific aspect words cluster around certain nouns or noun phrases. Therefore, high-frequency nouns or noun phrases often function as explicit aspect expressions. Hu and Liu (2004) initially established a known set of seed words with sentiment labels. They then expanded this seed word set based on word relationships such as synonyms, antonyms, and other lexical relations available in WordNet(WordNet 2010). Finally, they organized and compiled a comprehensive sentiment lexicon. They pioneered aspect extraction by using part-of-speech information to identify nouns and noun phrases, subsequently filtering out high-frequency terms as aspects. While this approach is straightforward, it has limitations as the extracted aspect words can carry substantial noise.

To enhance accuracy, Popescu and Etzioni (2007) aimed to exclude non-descriptive aspects from lists of high-frequency nouns and noun phrases through pointwise mutual information calculations between candidate aspects (e.g., "iPhone12") and automatically generated discriminative phrases (e.g., "Apple is a phone").

In addition to leveraging the noun-centric aspects, certain studies have also explored the connection between aspects and sentiments. Given that sentiment, expressions are inherently directed at objects, aspects, and corresponding sentiments often manifest together. Hence, this relationship can be exploited for aspect extraction.

Hu and Liu (2004) utilized this link to extract non-high-frequency aspects. The fundamental concept is that if a comment lacks high-frequency aspect words but contains sentiment terms, the nearest noun or noun phrase to the sentiment term is extracted as an aspect. Similar methodologies have been applied by Blair-Goldensohn et al. (2008) for building a sentiment summarizer for local service reviews. Zhuang et al. (2006) employed dependency parsers to recognize relationships between opinions and aspects, facilitating aspect extraction. Qiu et al. (2011) extended this notion by introducing a double-propagation algorithm based on dependency trees, enabling concurrent extraction of sentiment terms and aspects.

Zhang et al. (2022) have classified ATE methods into three categories: supervised (Liu et al. 2015; Yin et al. 2016; Wang et al. 2016a; Li and Lam 2017; Wang et al. 2017; Li et al. 2018b; Xu et al. 2018; Ma et al. 2019; Yang et al. 2020; Yin et al. 2020; Al-Janabi et al. 2022), semi-supervised (Li et al. 2020; Chen and Qian 2020a; Wang et al. 2021) and unsupervised methods (He et al. 2017; Luo et al. 2019; Liao et al. 2019) based on the availability of labeled data.

Table 1 A list of various ABSA sub-tasks. The example input sentence is: "The battery life of this phone is great"

Type	Abbr	Task Name	Input	Output	Method
Single	ATE	Aspect Term Extraction	S	a (battery life)	Extraction
	OTE	Opinion Term Extraction	S	o (great)	Extraction
	ACD	Aspect Category Detection	S	c (battery)	Classification
	AOCE	Aspect Opinion Co-Extraction	S	a (battery life), o (great)	Extraction
	AOOE	Aspect-Oriented Opinion Extraction	S + a (battery life)	o (great)	Extraction
	ABSC	Aspect-Based (Aspect-level) Sentiment Classification	S + a (battery life)	s (positive)	Classification
	COSC	Category-Oriented Sentiment Classification	S + c (battery)	s (positive)	Classification
Pair	AOPE	Aspect Opinion Pair Extraction	S	(a, o) (battery life, great)	Extraction
	ASPE	Aspect Sentiment Pair Extraction	S	(a, s) (battery life, positive)	Extraction & Classification
	CSPE	Category Sentiment Pair Extraction	S	(c, s) (battery, positive)	Extraction & Classification
Triplet	ACSTE (TASD)	Aspect Category Sentiment Triplet Extraction or Target Aspect Sentiment Detection	S	(a, c, s) (battery life, battery, positive)	Extraction & Classification
	AOSTE (ASTE)	Aspect Opinion Sentiment Triplet Extraction or Aspect Sentiment Triplet Extraction	S	(a, o, s) (battery life, great, positive)	Extraction & Classification
Quad	ACOSQE	Aspect Category Opinion Sentiment Quadruple Extraction	S	(a, c, o, s) (battery life, battery, great, positive)	Extraction & Classification

S, a, c, o, and s represent the sentence, aspect term, aspect category, opinion term, and sentiment polarity, respectively

2.2.2 Opinion term extraction (OTE)

This task involves extracting opinions in a review sentence. In this example, the opinion term is “great”. OTE involves identifying expressions that convey opinions toward a specific aspect or category. As opinion terms and aspect terms typically co-occur, extracting the opinion term without considering its associated aspect is useless. Therefore, most existing OTE methods can be decomposed into two sub-tasks: Aspect-Opinion Co-Extraction (AOOE) and Aspect-Oriented Opinion Extraction (AOCE).

Kobayashi et al. (2007) first utilizes dependency trees to find candidate aspect and opinion word pairs and then employs tree structure classification methods to learn and categorize these word pairs. Aspect extraction is a specific case of information extraction. Therefore, sequence learning models such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF) can be used for aspect extraction. HMM is a statistical model used to model sequential data, and it can be applied to aspect and sentiment extraction. CRF is a statistical model used for sequence labelling.

Jin et al. (2009) employed a tokenized HMM model for extracting aspects and their sentiment. Li et al. (2010) extended the linear-chain CRF model to propose Skip-chain, Tree CRF, and Skip-tree CRF models for aspect extraction. Skip-chain CRF, Tree CRF, and Skip-tree CRF are variations of linear-chain CRF designed for aspect extraction tasks. These models can employ rich features to extract object features and positive and negative opinions simultaneously.

Jakob and Gurevych (2010) performed aspect extraction using CRF in single-domain and cross-domain settings, employing multiple feature templates. They devised various feature templates to support these tasks, including word features, part-of-speech features, dependency relation features, word distance features, and sentiment features. In cross-domain aspect extraction tasks, it emphasizes the challenges in performing cross-domain aspect extraction. Opinion words in different domains may exhibit differing sentiment tendencies, and the substantial differences in aspect vocabularies across domains make cross-domain aspect extraction particularly complex.

2.2.3 Aspect category detection (ACD)

This task involves identifying the category to which the aspect term belongs. In this example, the aspect category is “battery.”

Zhang et al. (2022) have categorized ACD methods into two primary types: supervised ACD and unsupervised ACD. The presence or absence of annotated labels determines this categorization.

- **Supervised ACD**: RepLearn (Zhou et al. 2015) presents a representation learning approach for ACD in user-generated reviews, achieving state-of-the-art performance by automatically learning features from noisy labelled data using semi-supervised word embeddings and neural networks. TAN (Movahedi et al. 2019) introduces a topic-attention network, a deep neural network method with an attention mechanism, for ACD in user-generated reviews, demonstrating superior performance on restaurant domain datasets from the SemEval workshop and effective topic-based word identification through attention weight visualization. LICD (Ghadery et al. 2019) is a language-independent approach for ACD in SA of customer reviews. It employs text matching and semantic similarity measures to identify aspect categories, showcasing superior performance compared to baseline methods across multilingual SemEval-2016 (Pontiki 2016) datasets in the restaurant domain. Proto-AWATT (Hu et al. 2021) introduces a multi-label few-shot learning approach for ACD. It leverages prototypical networks and two attention mechanisms to enhance accuracy and outperform baseline methods.

- **Unsupervised ACD**: CAt (Tulkens and van Cranenburgh 2020) introduces a straightforward unsupervised method called Contrastive Attention for ACD. It requires only word embeddings and a POS tagger, demonstrating significant performance improvement and interpretability. It doesn't rely on syntactic features or complex neural models. SSCL (Shi et al. 2021) introduces self-supervised contrastive learning, featuring an attention-based model with a novel smooth self-attention (SSA) module, high-resolution selective mapping (HRSMap) for efficient aspect assignment, and knowledge distillation techniques to enhance aspect detection.

2.2.4 Aspect-opinion co-extraction (AOCE)

This task involves extracting opinions and aspects of a product or service separately. In this example, the opinion is "great," and the aspect of "battery life."

AOCE task is often tackled as a *tokenclass* problem (Yu et al. 2018), where either two label sets are utilized to extract aspect and opinion terms separately, or a unified label set (Wu et al. 2020a; Wang and Pan 2018) is utilized to extract both sentiment elements simultaneously. However, a limitation of this task is that the aspects and opinions are not paired, leading to incomplete information. Given the strong correlation between aspects and opinions, the primary research question in AOCE revolves around modeling this dependency. Several models have emerged to address the aspect-opinion relationship, encompassing dependency-tree-based approaches: UWDPE (Yin et al. 2016), RNCRF (Wang et al. 2016a), attention-based approaches: MIN (Li and Lam 2017), CMLA (Wang et al. 2017), HAST (Li et al. 2018a, b), and approaches that incorporate syntactic structures to impose explicit constraints on predictions: GMTCMLA (Yu et al. 2018) and DeepWMaxSAT (Wu et al. 2020a).

UWDPE (Unsupervised Word Dependency Path Embeddings) (Yin et al. 2016) primarily employs unsupervised methods to learn distributed representations of words and dependency paths. These representations are used as features for aspect term extraction. The approach involves connecting words with dependency relationships based on dependency path information in the embedding space. This method effectively distinguishes words with similar contexts but different syntactic functions.

RNCRF (Recursive Neural Network with Conditional Random Fields) (Wang et al. 2016a) is a recursive neural network based on sentence-level dependency trees. Its purpose is to learn high-level feature representations for each word in a sentence within the context and to understand the association between aspect and opinion terms based on the dependency structure. The most significant advantage of RNCRF lies in its ability to capture the underlying dual propagation between aspect and opinion terms.

MIN (Memory Interaction Network) (Li and Lam 2017) is an attention-based method and a deep multi-task learning framework. This paper employs two LSTMs with extended memory for aspect and opinion extraction.

CMLA (Coupled Multi-Layer Attentions) (Wang et al. 2017) consists of a multi-layer attention network, with each layer comprising two attention structures that incorporate tensor operators. One attention is designed to extract aspect terms, while the other focuses on extracting opinion terms. These attentions interact and facilitate bidirectional information flow between them. The multi-layer architecture enables the exploration of indirect associations between terms, leading to more precise information extraction, including less prominent terms. A notable highlight of this paper is that it doesn't rely on any parsers or linguistic prior knowledge.

HAST (History-Aware Self-Attention Model with Two-Level Cooperative Learning) Li et al. (2018b) consists of two main parts: Truncated History Network (THA) and Selective Transformation Network (STN). THA utilizes historical predictions of aspects to generate feature vectors for the current aspect, while STN generates opinion summary vectors. The importance of opinion information for aspect extraction is evident, but the author's perspective that content without opinion information should not be considered aspects may warrant further discussion.

GMTCMLA (Yu et al. 2018) enhances opinion mining by implicitly capturing task relations through multi-task learning and explicitly modeling syntactic constraints, leading to consistent improvements over base models. DeepWMaxSAT (Wu et al. 2020a) addresses this limitation by incorporating logic rules and MaxSAT (maximizing the number of satisfiable clauses) to represent these relationships. It combines them with deep neural networks through a unified framework for logical reasoning, resulting in improved performance in aspect-based sentiment extraction tasks.

2.2.5 Aspect-oriented opinion extraction (AOOE)

This task involves extracting opinions on specific aspects of a product or service. In this example, the opinion is "great" and relates to the aspect of "battery life." Fan et al. (2019) initially introduced this sub-task and proposed its datasets (FAN 2019). A target-fused sequence labelling neural network is designed for this task, encoding target information into context using an Inward-Outward LSTM and combining left and right contexts with the global context to identify opinion words. Experimental results demonstrate the superiority of the proposed model over other methods, potentially benefiting SA and pair-wise opinion summarization.

2.2.6 Aspect-based sentiment classification (ABSC)

This task involves classifying the sentence's sentiment towards a particular product or service aspect. In this example, the sentiment is positive towards the aspect of "battery life."

In the past, ABSC systems relied on feature-based approaches (Brun et al. 2014). However, deep learning-based models have become more popular. LSTM (Tang et al. 2016a) can effectively capture sequential dependencies, making them suitable for tasks involving ordered data. Nevertheless, they may encounter difficulties in capturing long-range dependencies within lengthy sequences. Attention-based LSTM models (Wang et al. 2016b; Liu and Zhang 2017; Ma et al. 2019; Tay et al. 2018) incorporate attention mechanisms, allowing them to focus on specific parts of the input sequence. These models excel at handling tasks where different input parts contribute differently to the output. Nevertheless, they can be computationally intensive. CNNs (Li et al. 2018a; Xue and Li 2018) are particularly proficient in handling spatial information, making them well-suited for tasks involving grid-like data such as images. However, their ability to capture sequential dependencies may be limited compared to RNN-based models. Gated neural networks (Zhang et al. 2016; Xue and Li 2018), like LSTMs, can model sequential dependencies effectively and are generally more computationally efficient. However, they may not perform as well as attention-based models in tasks requiring selective attention. Memory neural networks (Tang et al. 2016b; Chen et al. 2017) are designed to store and retrieve information over long sequences, making them suitable for tasks with extensive context. However, they can be complex to train and require significant computational resources. Pre-trained language models have emerged as the predominant foundation for the ABSC task. Sun et al. (2019) innovatively refrains the ABSC task as a sentence pair classification problem by introducing an auxiliary sentence. This approach harnesses the enhanced sentence pair modeling capabilities inherent in BERT, thereby improving the effectiveness of ABSC.

2.2.7 Category-oriented sentiment classification (COSC)

This task involves classifying the sentence's sentiment towards a particular product or service category. In this example, the sentiment is positive towards the category of "battery."

Recent studies in COSC have explored various techniques. For instance, Liu et al. (2021) proposed a BART-based generation method that takes a more direct approach by using pre-trained language models to transform the COSC task into a natural language generation task, representing outputs with natural language sentences. This approach closely adheres to the task's settings during pre-training, enabling a more direct utilization of pre-trained knowledge within seq2seq language models. It leverages the advantages of BART without introducing additional model parameters, thereby allowing for semantic-level summarization of input data.

The BART-based approach performs better than traditional sentiment classification methods, especially in zero-shot and low-shot learning scenarios. Aspect-aware graphs, as introduced in the work by Liang et al. (2021), represent a novel approach for COSC. This method leverages external knowledge to construct aspect-aware graphs. By assigning aspect-aware weights to sentiment-related words, this method effectively captures aspect-related contextual sentiment dependencies and outperforms existing baseline methods on six benchmark datasets. An approach that utilizes BERT (Sun et al. 2019) for constructing auxiliary sentences in ABSA is introduced. This method transforms ABSA into a sentence-pair classification task, thereby improving the fine-grained identification of opinion polarity towards specific aspects.

Additionally, other approaches include aspect-aware LSTM models (Xing et al. 2019), attentive LSTM models that embed commonsense knowledge (Saeidi et al. 2016; Ma et al. 2018) introducing the SentiHood dataset (Saeidi 2016), and hierarchical models (Ruder et al. 2016).

2.2.8 Aspect opinion pair extraction (AOPE)

This task involves extracting pairs of aspects and their corresponding opinions from a sentence. In this example, the aspect-opinion pair is "battery life-great."

There are different methods for AOPE (Yan et al. 2021; Zhang et al. 2022). The pipeline approach breaks down the task into smaller sub-tasks. In contrast, the MRC approach (Gao et al. 2021) uses a model to extract all aspect terms and then creates a question for another MRC model to identify the corresponding opinion.

Unified approaches for Aspect-Opinion Pair Extraction (AOPE) aim to extract aspect-opinion pairs jointly and address the risk of error propagation associated with the pipeline approach. For instance, GTS (Wu et al. 2020b) involves the model predicting whether word pairs belong to the same aspect, opinion, the aspect-opinion pair, or none of the above. This transforms the original pair extraction task into a unified *TokenClass* problem. Another approach is the span-based multi-task learning framework known as SpanMlt (Zhao et al. 2020), which allows for the simultaneous extraction of aspect/opinion terms and pair relations. Similarly, the Two-channel model (Chen et al. 2020) is designed for the separate extraction of aspect/opinion terms and relations. Additionally, it incorporates two synchronization mechanisms to facilitate information exchange between these two channels. Simultaneously, a model incorporates rich syntactic and linguistic knowledge through a syntax fusion encoder to enhance extraction performance (Wu et al. 2021b). This model utilizes label-aware graph convolutional networks (LAGCN) and local-attention modules to encode syntactic features and POS tags, thereby improving term boundary detection. Additionally, this model employs Biaffine and Triaffine scoring for high-order pairing of aspect-opinion terms, leveraging syntax-enriched representations from LAGCN.

2.2.9 Aspect sentiment pair extraction (ASPE)

This task involves extracting pairs of aspects and their corresponding sentiment from a sentence. In this example, the aspect-sentiment pair is “battery life-positive.”

To address the challenge in ASPE, researchers have developed various approaches, including pipeline methods, unified tagging schemas (Mitchell et al. 2013; Zhang et al. 2015; Li et al. 2019), multi-task learning (Hu et al. 2021; Chen and Qian 2020b), and span-based (Hu et al. 2019a) techniques. In recent works, additional methods such as few-shot learning (Hosseini-Asl et al. 2022), zero-shot ABSA (Shu et al. 2022) cross-lingual ABSA (Zhang et al. 2021c), Machine Reading Comprehension (MRC) (Yu et al. 2021), and structured SA (dependency graph parsing) (Barnes et al. 2021) have been explored to improve performance. Moreover, new datasets have been introduced to enhance the evaluation of ABSA models (Orbach et al. 2021).

2.2.10 Category sentiment pair extraction (CSPE)

This task involves extracting pairs of categories and their corresponding sentiment from a sentence. In this example, the category-sentiment pair is “battery-positive.”

CSPE can predict category-sentiment pairs regardless of whether the aspect is explicitly mentioned or implicit in the sentence (Bu et al. 2021). CSPE uses a pipeline approach. However, detecting a subset of aspect categories is challenging, and errors in the first step can impact performance. The relationship between the two steps is essential but often ignored (Hu et al. 2019b). Multi-task learning benefits both tasks (Hu et al. 2019b; Ma et al. 2018; Dai et al. 2020).

Unified methods for CSPE treat ACD as a multi-label classification and ABSC as a multi-class classification. Four unified methods include Cartesian product, add-one-dimension, hierarchy classification, and Seq2Seq modelling. Cartesian product (Wan et al. 2020) generates all category-sentiment pairs, resulting in a more extensive training set and higher cost. Add-one-dimension (Schmitt et al. 2018) adds an extra dimension to aspect category prediction. Hierarchical (Cai et al. 2020) and shared sentiment prediction (Liu et al. 2021) capture relations between aspect categories and sentiments. Seq2Seq modeling (Liu et al. 2021) benefits few-shot and zero-shot settings.

2.2.11 Aspect category sentiment triplet extraction (ACSTE)

This task involves extracting a sentence's aspect, category, and sentiment triplets. In this example, the triplet is "battery life-battery-positive."

Wan et al. (2020) proposed the ACSTE task, leveraging the TAS-BERT method based on the pre-trained language model BERT to capture dependencies on both aspect categories and aspects for sentiment prediction. They demonstrated superior performance on the SemEval-2015 (Pontiki et al. 2015) and SemEval-2016 (Pontiki et al. 2016) restaurant datasets, even in cases with implicit targets, surpassing state-of-the-art methods in related subtasks. This work was subsequently improved upon by MEJD (Wu et al. 2021a), which presents a novel end-to-end multiple-element joint detection model (MEJD) for ACSTE. MEJD utilizes BERT and bidirectional LSTM to extract (aspect category, aspect, sentiment) triples effectively. Additionally, GAS-T5 (Zhang et al. 2021d) introduces a unified generative framework for ACSTE. The proposed approach achieves state-of-the-art results across various ABSA tasks and datasets, demonstrating its versatility and effectiveness without requiring task-specific model design.

2.2.12 Aspect opinion sentiment triplet extraction (AOSTE)

This task involves extracting a sentence's aspect, opinion, and sentiment triplets. In this example, the triplet is "battery life-great-positive."

Peng et al. (2020) introduced the AOSTE task, which is addressed using a two-stage framework. In the first stage, aspect terms, opinion terms, and sentiment polarity are extracted, transforming the task into two sequence labelling tasks: one for aspect terms and their corresponding sentiment polarities and another for opinion terms. In the second stage, aspect and opinion terms are paired to construct the triplets. Jet-BERT (Xu et al. 2020b) introduced an end-to-end model with a novel position-aware tagging scheme for AOSTE, achieving improved performance by jointly capturing target aspects, associated sentiments, and opinion spans in triplets. Dual-MRC (Mao et al. 2021) presents an end-to-end solution by jointly training two BERT-Machine Reading Comprehension (MRC) models to address aspect term extraction, opinion term extraction, and aspect-level sentiment classification. This approach achieves superior performance compared to existing methods. One model predicts aspect terms and then opinion terms, while the other model first predicts the opinion and then the aspect. BMRC, as presented in the study by Chen et al. (2021), transformed AOSTE into a bidirectional Machine Reading Comprehension (MRC) problem. This approach addressed the intricate correspondence between aspects and opinions by employing context-specific bidirectional queries. This method effectively facilitated the mutual extraction of information, resulting in improved sentiment prediction across various contexts. To enhance the handling of the ACOSTE task, BMRC combined tokenization with exclusive classifiers and improved span matching by introducing priority rules for combining probability and positional relationships. Furthermore, BMRC optimized probability generation to prevent one-sided reductions.

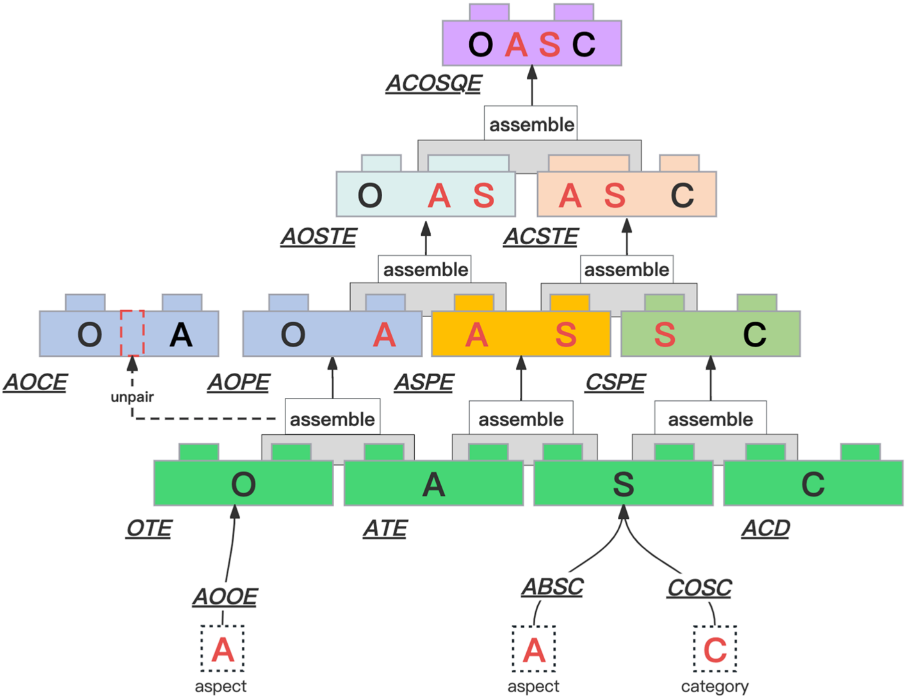

Fig. 3 Assembling the ABSA task like building with Lego blocks

2.2.13 Aspect category opinion sentiment quadruple extraction (ACOSQE)

This task involves extracting a sentence's quadruples of aspect, category, opinion, and sentiment. In this example, the quadruple is "battery life-battery-great-positive."

Cai et al. (2021b) and Zhang et al. (2021a) introduced ACOSQE for extracting sentiment quadruples from laptop and restaurant reviews, respectively. The ACOS (Aspect-Category-Opinion-Sentiment) task, as defined by Cai et al. (2021b), and aspect-based sentiment quad predication (ASQP), as defined by Zhang et al. (2021a), are both definitions for extracting quadruples from review sentence. To ensure consistency, this paper will refer to the ACOS and ASQP tasks as ACOSQE task.

Notably, Cai et al. (2021b) proposed that while pair or triplet extraction tasks primarily focus on explicit aspects and opinions expressed in sentences, ACOSQE places greater emphasis on implicit aspects and opinions. Researchers have also recently extended ABSA to the dialogue level and proposed the dialogue ACOSQE task (DiaASQ) (Li et al. 2022a). It involves extracting aspect terms, categories, corresponding sentiment polarities, and opinion terms from dialogues.

Inspired by Gao et al. (2022), these tasks can be combined flexibly like Lego bricks to form various compound tasks in Fig. 3. This Lego-style assembly provides a more flexible solution for ABSA. This review primarily focuses on ACOSQE research, as presented in Table 1.

2.3 Dataset

To broaden the scope of ACOSQE research, some researchers have constructed their datasets based on SemEval and proposed new tasks and challenges. This section introduces some standard datasets and provides type, language, domain, source, and URL information for each dataset in Table 2. In the following sections, this paper will further explore the research progress and challenges based on different pre-trained models and datasets in Sec 2.2.13.

The datasets provided by SemEval-2014 (Pontiki et al. 2014), SemEval-2015 (Pontiki et al. 2015), and SemEval-2016 (Pontiki et al. 2016), as part of shared tasks, are widely regarded as the most extensively used benchmarks in the literature. Annotated datasets are crucial for the development of ACOSQE task. However, traditional ABSA research has mostly been limited to single text snippets, with datasets such as SemEval-2014 (Pontiki et al. 2014), SemEval-2015 (Pontiki et al. 2015), and SemEval-2016 (Pontiki et al. 2016) providing only sentence-level annotations. These datasets comprise user-generated reviews spanning two domains, laptops and restaurants, and come with detailed annotations, including aspect categories, aspect terms, and sentiment polarities. These datasets find direct applicability in numerous ABSA tasks, such as aspect term extraction and aspect sentiment classification, although not all required information is encompassed. SentiHood (Saeidi et al. 2016) is a widely employed dataset for ACOSTE. It includes 5,215 English sentences, with 3,862 focusing on a single aspect, while the others involve multiple aspects. Each sentence is annotated with tuples containing the aspect, associated category, and corresponding sentiment polarity, covering both positive and negative sentiments. Nevertheless, these datasets lack annotations for opinion terms, a deficiency addressed by the dataset introduced in FAN (2019), designed for target-oriented opinion word extraction (TOWE) tasks. Xu et al. (2020b) further consolidated these annotations, with slight refinements, to create the ASTE-Data-V1, V2 datasets (Xu et al. 2020a), where each sample sentence contains triples of aspect terms, opinion terms, and sentiment polarities.

More recently, for the ACOSQE task, two new datasets, ACOS (Cai et al. 2021a) and QUAD (Zhang et al. 2021b), have been introduced, with each data instance annotated with four sentiment elements. To expand the dataset's capacity, the ASQP (Zhang et al. 2023) has released two novel datasets, en-Phone and zh-FoodBeverage. En-Phone is an English dataset for ACOSQE in the cell phone domain collected from various e-commerce platforms. In contrast, zh-FoodBeverage represents the first Chinese dataset for ACOSQE, encompassing multiple sources within the food and beverage domain. Compared to existing datasets for ACOSQE, the ASQP dataset offers more samples, with an increase ranging from 1.75 to 4.19 times, and features a higher quadruple density, amplified by 1.3 to 1.8 times. MEMD (Cai et al. 2023b) has developed a multi-element, multi-domain dataset spanning five domains: books, clothing, hotels, restaurants, and laptops. This dataset comprises nearly 20,000 review sentences, making it four to five times larger than previous SemEval ABSA datasets. Furthermore, it boasts annotations for nearly 30,000 quadruples, supporting multi-element extraction tasks that involve both explicit and implicit aspects and opinions, making it suitable for ACOSQE research. Li et al. (2022a) has constructed the extensive DiaASQ dataset (Li et al. 2022b) by collecting comments about electronic products from Chinese social media. This dataset includes 1000 dialogues and 7452 utterances. The data have also

Table 2 An overview of ABSA benchmark datasets

Type	Dataset Paper	Language	Domain	Source
Sentence	SemEval-2014 (Pontiki 2014)	English	Laptops,Restaurants	GSNY (Ganu et al. 2009)
Sentence	SemEval-2015 (Pontiki 2015)	English	Laptops,Restaurants	GSNY (Ganu et al. 2009) Amazon.com
Sentence	SemEval-2016 (Pontiki 2016)	multilingual	Electronics, Hotels,Restaurants	GSNY (Ganu et al. 2009) Amazon.com Loukachevitch et al. (2015)
Sentence	SentiHood (Saeidi 2016)	English	Urban Neighborhoods	Yahoo Question Answering
Sentence	TOWE (FAN 2019)	English	Laptops,Restaurants	SemEval-14 Restaurants SemEval-15 Restaurants SemEval-16 Restaurants
Sentence	MAMS(Jiang 2019)	English	Restaurant	CNSY (Ganu et al. 2009)
Sentence	ARTS(Xing 2020)	English	Laptop,Restaurant	SemEval-14 Restaurants SemEval-14 Laptops
Sentence	ASTE-Data-V1,V2 (Xu et al. 2020a)	English	Laptop,Restaurant	SemEval-14 Restaurants SemEval-15 Restaurants SemEval-16 Restaurants SemEval-14 Laptops
Sentence	ASAP(Bu 2021)	Chinese	Restaurant	O2O E-commerce Sites
Sentence	ACOS (Cai et al. 2021a)	English	Laptop,Restaurant	SemEval-2016 Restaurants Amazon.com
Sentence	QUAD (Zhang et al. 2021a)	English	Restaurant	SemEval-2015 Restaurants SemEval-2016 Restaurants
Sentence	ASQP (Zhou et al. 2023)	English	Phone, FoodBeverage	E-commercial platforms Multiple sources under the categories of Food and Beverage

Table 2 (continued)

Type	Dataset Paper	Language	Domain	Source
Sentence	MEMD (Cai et al. 2023a)	English	Books, Clothing, Hotel Restaurant, Laptop	Ni et al. (2019) Boston dataset of Yelp Boston dataset of Airbnb Amazon platform
Dialogue	DiaASQ (Li et al. 2022b)	English	Electronic Products	Chinese social media

Table 3 Data statistics for the ACOS-Dataset

	Restaurant-ACOS	Laptop-ACOS
#Categories	13	121
#Sentences	2286	4076
#Quads	3661	5773
#Quads/Sentences	1.60	1.42
#EA & EO	2429 (66.40%)	3269 (56.77%)
#IA & EO	530 (14.49%)	910 (15.80%)
#EA & IO	350 (9.57%)	1237 (21.48%)
#EA & IO	349 (9.54%)	342 (5.94%)
#POS	2503	3578
#NEU	151	316
#NEG	1007	1879
#Train	1531	2934
#Dev	170	326
#Test	585	816
#Train (Quads)	2484	4172
#Dev (Quads)	261	440
#Test (Quads)	916	1161

#Denotes the number of corresponding elements. EA, EO, IA, and IO denote explicit aspect, explicit opinion, implicit aspect, and implicit opinion, respectively

been translated into English, revealing that each dialogue typically features around five speakers, with approximately 22.2% of the quadruples spanning multiple utterances.

In addition to datasets designed for various ABSA tasks, specialized datasets have been introduced to explore specific aspects. Jiang et al. (2019) introduced a multi-aspect multi-sentiment (MAMS) dataset (Jiang 2019), where each sentence in MAMS contains at least two aspects with different sentiment polarities, making the dataset more challenging. Xing et al. (2020) constructed an aspect robustness dataset (ARTS) (Xing 2020) based on the SemEval-2014 dataset to investigate the robustness of ABSA models. More recently, Bu et al. (2021) released a large-scale Chinese dataset called ASAP (Bu 2021), representing aspect category SA and rating prediction. Each sentence in ASAP has been annotated with sentiment polarities for 18 predefined aspect categories, making it suitable for studying the relationship between coarse-grained and fine-grained sentiment analysis tasks.

Since the current focus of ACOSQE research predominantly centers around the ACOS (Cai et al. 2021a) and QUAD (Zhang et al. 2021b) datasets, the following two sections will primarily delve into these datasets.

2.3.1 ACOS-dataset

Cai et al. (2021b) created two new datasets, Restaurant-ACOS and Laptop-ACOS, for the ACOSQE task. Restaurant-ACOS was derived from the SemEval 2016 Restaurant (Pontiki et al. 2016) and its expansions, while Laptop-ACOS was collected from Amazon (2017-2018) and contained 4,076 review sentences covering ten laptop types under six brands. The SemEval 2016 Restaurant dataset (Pontiki et al. 2016) was annotated with explicit and implicit aspects, categories, and sentiment with opinion annotations added by Fan et al.

(2019) and Xu et al. (2020b). The annotations were used to construct aspect-category-opinion-sentiment quadruples and annotate implicit opinions. For Laptop-ACOS, the researchers annotated the four elements and their corresponding quadruples.

The basic statistics of the two datasets are reported in Table 3. The Restaurant-ACOS dataset comprises 2,286 sentences and 3,661 quadruples, while the Laptop-ACOS dataset contains 4,076 sentences and 5,773 quadruples. As previously mentioned, a significant proportion of the quadruples in both datasets contain implicit aspects or opinions. However, upon comparing the two datasets, it is evident that the Laptop-ACOS dataset has a higher percentage of implicit opinions than the Restaurant-ACOS dataset. The Laptop-ACOS dataset is also larger than the Restaurant-ACOS dataset regarding the number of samples, aspect categories, and quadruplets.

Moreover, the table shows that the Laptop-ACOS dataset has a higher density of quadruplets per sample compared to the Restaurant-ACOS dataset. The metrics related to the types of quadruplets (EA &EO, EA &IO, IA &EO, IA &IO) indicate that both datasets have a mix of explicit and implicit aspects and opinions. Similarly, the metrics related to the sentiment of the quadruplets (#NEG, #NEU, #POS) reveal that both datasets have a mix of negative, neutral, and positive sentiments. The original dataset will be divided into a training set, a validation set, and a testing set.

Cai et al. (2021b) compared two ACOS datasets with existing ABSA datasets in Table 4. Restaurant 2014/2016 and Laptop 2014/2016 are SemEval 2014/2016 Restaurant and Laptop datasets with different category definitions. Laptop 2014 has aspect and sentiment annotations, while Laptop 2016 has category and sentiment annotations.

Fan et al. (2019) proposed Restaurant-2014-AO and Restaurant-2016-AO, removing sentences with implicit aspects and adding opinion annotations. Xu et al. (2020b) added the sentiment to create Restaurant-2014-AOS and Restaurant-2016-AOS. These two aspect-opinion-sentiment triple datasets were originally included in Restaurant 2014/2016 to Restaurant-2014/2016-AO. Cai et al. (2021b) integrated these annotations to construct ACOS quadruples in Rest-ACOS, keeping sentences with implicit aspects and annotating implicit opinions. Rest-ACOS is 1.6 times larger than Restaurant-2016-AO and Restaurant-2016AOS. Laptop-ACOS has 4076 review sentences and 5758 ACOS quadruples, nearly twice the size of Restaurant-ACOS.

2.3.2 QUAD-dataset

Zhang et al. (2021a) created two new datasets, REST15 and REST16, for the ACOSQE task in the restaurant domain. These datasets were built upon the SemEval Shared Challenges as a basis, with annotations for aspect category and opinion term from Peng et al. (2020) and Wan et al. (2020), respectively. Zhang et al. (2021a) merged the annotations with the same aspect term in each sentence and added additional annotations for sentences without explicit aspect terms. Quadruples with implicit opinion expressions were discarded, and for cases where the same aspect term was associated with multiple aspect categories or opinion terms, the merged result will contain over four sentiment elements for each quadruple. Subsequently, the author manually reviews these instances to rectify labels, ensuring alignment between aspect category and corresponding opinion term within the same quadruple. Two human annotators assess each sample, with conflict cases undergoing verification. The resulting REST15 and REST16 datasets contain review sentences with one or multiple sentiment quadruples.

Table 4 The comparison between the sizes of ACOS-Dataset and existing representative ABSA datasets

Dataset	Sentence	Aspect	Category	Opinion	Sentiment	AS Pair	AO Pair	AOS Triple	ACS Triple	ACOS Quad
Restaurant-2014 (Pontiki et al. 2014)	3841	4827	4738	–	4534	4827	–	–	–	–
Laptop-2014 (Pontiki et al. 2014)	1910	3012	–	–	3012	3012	–	–	–	–
Restaurant-2016 (Pontiki et al. 2016)	2295	3122	3001	–	3122	3182	–	–	3364	–
Laptop-2016 (Pontiki et al. 2016)	2612	–	3705	–	3705	–	–	–	–	–
Laptop-2014-AO (Fan et al. 2019)	2125	3503	–	3610	–	–	4092	–	–	–
Restaurant-2016-AO (Fan et al. 2019)	1407	1968	–	2146	–	–	2294	–	–	–
Restaurant-2014-AOS (Xu et al. 2020b)	2068	3399	–	3443	3399	3399	3908	3908	–	–
Restaurant-2016-AOS (Xu et al. 2020b)	1393	1946	–	2101	1946	1946	2247	2247	–	–
Restaurant-ACOS (Cai et al. 2021b)	2286	3110	2967	3335	3110	3155	3571	3575	3335	3658
Laptop-AOS (Cai et al. 2021b)	4076	4958	4992	5378	4958	5035	5726	5731	5227	5758

Table 5 Data statistics for the Rest15 (QUAD-Dataset)

Dataset	Rest15					
	#Sentence	#Quads	#S/Q	#POS	#NEU	#NEG
Train	834	1354	1.62	1005	34	315
Dev	209	347	1.66	252	14	81
Test	537	795	1.48	453	37	305
All	1080	2496	2.31	1710	85	701

# POS, # NEU, and # NEG denote the number of sentences and the number of positive, neutral, and negative quads, respectively

Table 6 Data statistics for the Rest16 (QUAD-Dataset)

Dataset	Rest16					
	#Sentence	#Quads	#S/Q	#POS	#NEU	#NEG
Train	1264	1989	1.57	1369	62	558
Dev	316	507	1.60	341	23	143
Test	544	799	1.47	544	40	176
All	2124	3295	3.05	2254	125	877

# POS, # NEU, and # NEG denote the number of sentences and the number of positive, neutral, and negative quads, respectively

Table 7 Confusion matrix

	True label	
	Positive	Negative
Predicted label		
Positive	True Positive (T_P)	False Positive (F_P)
Negative	False Negative (F_N)	True Negative (T_N)

Tables 5 and 6 provide the basic statistics for these two data sets. The Rest-15 data set consists of 1,080 sentences and 2,496 quadruples, whereas the Laptop-16 data set contains 2,124 sentences and 3,295 quadruples. The resulting REST15 and REST16 data sets contain review sentences with one or multiple sentiment quadruples. Unlike the emphasis on analyzing implicit combinations of aspects and opinions in the ACOS dataset by Cai et al. (2021b), Zhang et al. (2021a) did not emphasize the analysis of implicit combinations of aspects and opinions in the ASQP-Dataset.

2.4 Evaluations

This section introduces corresponding evaluation metrics. For ACOSQE, the quadruple (aspect, category, opinion, sentiment) must match the annotated quadruple to be considered correct. For ACOS-Dataset, exact-match evaluation requires the extracted quadruple to include implicit aspects and opinions in addition to the explicit aspects and opinions. This strict metric reflects the overall performance of the Language Model in identifying and extracting all aspects and opinions, both explicit and implicit.

The evaluation metrics for ACOSQE are precision (P), recall (R), F1-score (F1), and accuracy (Acc). The formulas for these metrics are (see Table 7 for definition of T_P, F_P, F_N and T_N):

- Precision (P): The proportion of correctly predicted positive aspects among all predicted positive aspects.

$$P = \frac{T_P}{T_P + F_P} \tag{1}$$

where T_P is the number of true positives and F_P is the number of false positives.
- Recall (R): The proportion of correctly predicted positive aspects among all true positive aspects.

$$R = \frac{T_P}{T_P + F_N} \tag{2}$$

where T_P is the number of trule positives and F_N is the number of false negatives.
- F1-score (F1): The harmonic mean of precision and recall.

$$F1 = \frac{2PR}{P + R} \tag{3}$$

- Accuracy (Acc): The proportion of correctly predicted aspects among all aspects.

$$Acc = \frac{T_P + F_N}{T_P + T_N + F_P + F_N} \tag{4}$$

where T_P is the number of true positives, T_N is the number of true negatives, F_P is the number of false positives, and F_N is the number of false negatives.

3 Approach using rules

Cai et al. (2021b) first described the evaluation of the ACOSQE task, which created four baseline methods: Double-Propagation-ACOS, JET-ACOS, TAS-BERT-ACOS, and Extract-Classify-ACOS. These systems were adapted from existing AOPE, ACSTE, or AOSTE approaches to be compatible with the ACOSQE task.

Following their research, a new series of works have been proposed in this field. Table 8 shows a comparison of methods in Rule-based, BERT-based, Bart-based, and T5-based benchmark for the ACOSQE task based on ACOS-Dataset. This section will cover the pipeline method with rule-based (Double-Propagation-ACOS) on the ACOS-Dataset (Cai et al. 2021b). The following section will describe the other three methods with BERT (JET-ACOS, TAS-BERT-ACOS, and Extract-Classify-ACOS).

The Double Propagation (DP) method (Qiu et al. 2011) is one of the four baseline systems that Cai et al. (2021b) used to evaluate the ACOSQE task. It builds on the Double-Propagation approach and extends it by leveraging the relationships between extracted aspects and opinions to expand their coverage in the text.

The first step in the proposed methodology is to extract aspect-opinion-sentiment triples using the DP algorithm. It is done by iteratively extracting aspects and opinions

Table 8 Comparison of methods in Rule-based, BERT-based, BART-based, and T5-based benchmark for the ACOSQE task based on ACOS-Dataset

Methods	REST-ACOS			LAPTOP-ACOS			REST-ACOS (F1.)				LAPTOP-ACOS (F1.)			
	P.	R.	F1.	P.	R.	F1.	EA & EO	IA& EO	EA & IO	IA & IO	EA & EO	IA & EO	EA & IO	IA & IO
Rule-based														
DP-ACOS (Cai et al. 2021b)	34.67	15.08	21.04	13.0	5.70	8.0	26.0	N/A	N/A	N/A	9.8	N/A	N/A	N/A
BERT-based														
JET-ACOS (Cai et al. 2021b)	59.81	28.94	39.01	44.52	16.25	23.81	52.3	N/A	N/A	N/A	35.7	N/A	N/A	N/A
TAS-BERT-ACOS (Cai et al. 2021b)	26.29	46.29	33.53	47.15	19.22	27.31	33.6	31.8	14.0	39.8	26.1	41.5	10.9	21.2
Extract-Classify-ACOS (Cai et al. 2021b)	38.54	52.96	44.61	45.56	29.48	35.80	45.0	34.7	23.9	33.7	35.4	39.0	16.8	18.6
BART-based														
PARAPHRASE-BART (Xiong et al. 2023)	43.62	36.19	39.56	36.36	29.63	32.65	38.6	37.8	16.7	38.5	31.3	38.9	21.1	35.6
GEN-NAT-SCL-BART (Xiong et al. 2023)	48.93	40.51	44.32	37.13	32.44	34.63	46.9	30.5	20.5	37.6	35.9	40.7	20.9	30.2
BART-CRN (Xiong et al. 2023)	50.84	47.10	48.90	48.16	31.83	38.32	54.1	50.6	18.9	42.9	38.9	**54.3**	24.5	**40.7**
BARTABSA(Hoang) (Hoang et al. 2022)	55.77	50.66	53.09	35.80	38.01	36.88	N/A				N/A			
BARTABSA(split) (Hoang et al. 2022)	56.80	51.09	53.45	41.06	37.89	39.41	58.5	43.9	20.0	42.9	39.9	52.8	23.4	29.8
BARTABSA (Bao) (Bao et al. 2022)	56.62	55.35	55.98	41.65	40.46	41.05	N/A				N/A			
T5-based														
Seq2Path (Mao et al. 2022)	N/A	N/A	58.41	N/A	N/A	42.97	N/A				N/A			

Table 8 (continued)

Methods	REST-ACOS			LAPTOP-ACOS			REST-ACOS (F1.)				LAPTOP-ACOS (F1.)			
	P.	R.	F1.	P.	R.	F1.	EA & EO	IA& EO	EA & IO	IA & IO	EA & EO	IA & EO	EA & IO	IA & IO
GAS (Bao) (Zhang et al. 2021d)	60.69	58.52	59.59	41.60	42.75	42.17	N/A				N/A			
Multi-Task Instruction Tuning (Wang et al. 2022)	N/A	N/A	60.60	N/A	N/A	42.58	N/A				N/A			
Special_Symbols (Hu et al. 2022)	59.98	58.40	59.18	43.58	42.72	43.15	N/A				N/A			
Special_Symbols + UAUL (Hu et al. 2023)	61.22	59.87	60.53	44.38	43.65	44.01	N/A				N/A			
DLO (Hu et al. 2022)	60.02	59.84	59.93	43.40	43.80	43.60	N/A				N/A			
DLO + UAUL (Hu et al. 2023)	61.03	60.55	60.78	43.78	43.53	43.65	N/A				N/A			
ILO (Hu et al. 2022)	58.43	58.95	58.69	44.14	44.56	44.35	N/A				N/A			
ILO + UAUL (Hu et al. 2023)	59.46	59.12	59.29	43.92	43.46	43.69	N/A				N/A			
MvP (multi-task) (Gou et al. 2023)	N/A	N/A	60.36	N/A	N/A	43.84	N/A				N/A			
PARAPHRASE (Peper) (Zhang et al. 2021a)	N/A	N/A	60.97	N/A	N/A	44.08	65.4	53.3	45.6	45.6	45.7	51.0	33.0	39.6
PARAPHRASE (Gou) (Zhang et al. 2021a)	N/A	N/A	61.16	N/A	N/A	43.51	N/A				N/A			
MvP (Gou et al. 2023)	N/A	N/A	61.54	N/A	N/A	43.92	N/A				N/A			

Table 8 (continued)

Methods	REST-ACOS			LAPTOP-ACOS			REST-ACOS (F1.)				LAPTOP-ACOS (F1.)			
	P.	R.	F1.	P.	R.	F1.	EA & EO	IA& EO	EA & IO	IA & IO	EA & EO	IA & EO	EA & IO	IA & IO
GEN-SCL-NAT (Peper and Wang 2022)	N/A	N/A	62.62	N/A	N/A	45.16	**66.5**	**56.5**	**46.2**	**50.7**	**45.8**	54.0	**34.3**	39.6
Opinion Tree (Bao et al. 2022)	**63.96**	**61.74**	**62.83**	**46.11**	**44.79**	**45.44**	N/A				N/A			

The best results are in bold. EA, IA, EO, and IO denote explicit aspect, implicit aspect, explicit opinion, and implicit opinion, respectively based on their syntactic relations in the review sentence and assigning sentiments (*positive*, *negative*, or *neutral*) using a sentiment lexicon.

The second step is to identify the aspect category for each extracted triple. If the aspect is present in the training set, the most co-occurred aspect category is assigned as the final aspect category. Otherwise, the aspect category of the nearest aspect in the input review is adopted as the final aspect category. Based on these two steps, ACOS quadruples can be extracted from each review sentence. Table 8 shows that DP-ACOS achieves the lowest performance. The limitations of DP-ACOS are as follows:

- **Dependence on rules and lexicons:** DP-ACOS relies on rules and lexicons for its performance, which may limit its accuracy if the rules and lexicons do not cover all cases.
- **Limited suitability for new domains:** DP-ACOS requires pre-labelled aspect category information in the training set. This means that for new domain data, the aspect category information needs to be relabeled, or the model needs to be retrained, which can increase workload and time costs.
- **Sensitive to text structure:** DP-ACOS relies on syntactic dependency relations for extracting aspects and opinions, making it highly dependent on text structure and grammar. Complex or ungrammatical text structures may negatively impact the extraction performance.
- **Unable to handle polysemy and ambiguity:** DP-ACOS uses a lexicon-based SA method to assign sentiment to aspects and opinions. However, SA may not be accurate for words with multiple meanings or ambiguous words.

In summary, the rule-based approach provides a baseline for the ACOSQE task but may have limitations. This paper will offer a comprehensive overview of the rule-based approach, considering it from both the pros and cons perspectives.

- **Pros**
  - Rule-based approach in ACOSQE provides explicit control over the extraction process, allowing researchers to define specific rules and patterns for identifying aspects, opinions, and sentiment polarity.
  - Rule-based approach can be effective in scenarios where the domain-specific knowledge is well-defined and can be easily translated into extraction rules.
  - Rule-based approach can be computationally efficient and require less training data than machine learning-based approaches.
- **Cons**
  - The rule-based approach, while offering control and efficiency in ACOSQE, is constrained by predefined rules, limiting coverage, adaptability, and the handling of unknown aspects due to its inability to capture all language variations and complexities, ultimately leading to reduced accuracy. Additionally, its determinism presents challenges in dealing with underlying data uncertainty and managing unknown or ambiguous aspects not explicitly covered by predefined rules.
  - Rule-based approach may require manual effort and expertise to design and maintain the extraction rules, making them less scalable and adaptable to new domains or languages.

Table 9 Comparison of methods in BERT-based and T5-based benchmark for the ACOSQE task based on QUAD-Dataset

Type	Methods	REST15			REST16		
		P.	R.	F1.	P.	R.	F1.
BERT-based							
Pipeline	HGCN-BERT + BERT-Linear (Zhang et al. 2021a)	24.43	20.25	22.15	25.36	24.03	24.68
Pipeline	HGCN-BERT + BERT-TFM (Zhang et al. 2021a)	25.55	22.01	23.65	27.40	26.41	26.90
Unified	TASO-BERT-Linear (Zhang et al. 2021a)	41.86	26.50	32.46	49.73	40.70	44.77
Unified	TASO-BERT-CRF (Zhang et al. 2021a)	44.24	28.66	34.78	48.65	39.68	43.71
Pipeline	Extract-Classify-ACOS (Cai et al. 2021b)	35.64	37.25	36.42	38.40	50.93	43.77
T5-based							
Unified	GAS (Zhang et al. 2021d)	45.31	46.70	45.98	54.54	57.62	56.04
Unified	LEGO-ABSA(multi-task) (Gao et al. 2022)	N/A	N/A	46.10	N/A	N/A	57.60
Unified	LEGO-ABSA(separate) (Gao et al. 2022)	N/A	N/A	45.80	N/A	N/A	57.70
Unified	Text (Varia et al. 2022)	N/A	N/A	46.79	N/A	N/A	57.41
Unified	IT (Varia et al. 2022)	N/A	N/A	46.59	N/A	N/A	57.48
Unified	IT-MTL (Varia et al. 2022)	N/A	N/A	46.59	N/A	N/A	57.61
Unified	PARAPHRASE (Zhang et al. 2021a)	46.16	47.72	46.93	56.63	59.30	57.93
Unified	Opinion Tree (Lee) (Bao et al. 2022)	N/A	N/A	47.60	N/A	N/A	58.07
Unified	SENER-orig (Lee and Kim 2023)	N/A	N/A	48.45	N/A	N/A	58.46
Unified	SENER-syn (Lee and Kim 2023)	N/A	N/A	47.19	N/A	N/A	59.40
Unified	Special_Symbols (Hu et al. 2022)	48.24	48.93	48.58	58.74	60.35	59.53
Unified	Special_Symbols+UAUL (Hu et al. 2023)	49.12	50.39	49.75	59.24	61.75	60.47
Unified	DLO (Hu et al. 2022)	47.08	49.33	48.18	57.92	61.80	59.79
Unified	DLO+UAUL (Hu et al. 2023)	48.03	50.54	49.26	59.02	62.05	**60.50**
Unified	ILO (Hu et al. 2022)	47.78	50.38	49.05	57.58	61.17	59.32
Unified	ILO+UAUL (Hu et al. 2023)	46.84	49.53	48.15	58.23	61.35	59.75
Unified	MvP (Gou et al. 2023)	N/A	N/A	51.04	N/A	N/A	60.39
Unified	MvP (multi-task) (Gou et al. 2023)	N/A	N/A	**52.21**	N/A	N/A	58.94

The best results are in bold. EA, IA, EO, and IO denote explicit aspect, implicit aspect, explicit opinion, and implicit opinion, respectively

4 Approach using BERT

BERT (Bidirectional Encoder Representations from Transformers) (Kenton and Toutanova 2019) is a pre-trained language model developed by Google. It utilizes the Transformer architecture and has gained significant popularity in NLP tasks, including SA. BERT's key feature is its ability to understand the contextual meaning of words within a sentence by considering both the preceding and succeeding words.

By training on large-scale corpora, BERT learns comprehensive word representations using masked language modelling and next-sentence prediction tasks. It can be fine-tuned for various NLP tasks, such as text classification and question-answering, by adding task-specific layers. With its contextual understanding and effective pre-training, BERT has achieved state-of-the-art results on benchmark datasets, making it widely used in industry and academia for SA and other language-processing tasks.

This section explores applying different BERT-based models to the ACOSQE task and divides it into three parts based on the sentence-level dataset used, as shown in Tables 8 and 9. Section 4.1 introduces the method using the ACOS-Dataset, Sect. 4.2 introduces the method using the QUAD-Dataset, Sect 4.3 concludes the methods using both the ACOS and QUAD Dataset and Sect. 4.4 summarizes this approach.

4.1 Approach based on ACOS-dataset

JET (Xu et al. 2020b) is an end-to-end pipeline method that combines the identification of aspects, their corresponding opinions, and their sentiment polarities with a position-aware tagging scheme.

To adapt JET to the ACOSQE task, Cai et al. (2021b) first extracted the aspect-opinion-sentiment triples using JET and then predicted the aspect category for each extracted triple. A BERT-based model was used to get the aspect category of the extracted triples.

In the training stage, the standard binary cross-entropy loss function was used for optimization. In the inference stage, Xu et al. (2020b) combined the extracted aspect-opinion-sentiment triples from JET and their predicted aspect categories to get all the quadruples from each review sentence.

The average vectors of words in the aspect and the opinion were used to obtain the representation of the aspect and the opinion. Then, the aspect and opinion vectors were concatenated and fed into a fully connected layer with the Sigmoid function for each category. The output of the Sigmoid function indicates whether a quadruple is valid or invalid.

Comparative experiments in Table 8 show that JET-ACOS performs better than Double-Propagation-ACOS in the task of ABSA. The main reasons are:

- JET-ACOS adopts an end-to-end framework with a position-aware labeling scheme to recognize aspect, opinion, and sentiment polarity, which can more accurately extract ACOS information.
- JET-ACOS uses a BERT-based model to obtain aspect categories, which can better capture the semantic information of the text compared to DP-Propagation-ACOS, which uses manual feature extraction.
- JET-ACOS uses the sigmoid function for classification prediction, which can better handle classification problems.

This method has the following drawbacks:

- The processing effect for long texts may not be ideal, as the method is based on sentence-level extraction and may not capture the aspect, opinion, and sentiment information well for long texts containing multiple sentences.
- The handling of unknown aspects may not be optimal. This method needs to predict the aspect category for each extracted triple, and if an unknown aspect appears, it cannot be predicted correctly.
- This method requires pre-extraction of triples and prediction of aspect categories, so two models need to be trained, which may increase the complexity and training difficulty of the model.
- This method uses simple average vector representation for aspects and opinions, which may need to better capture their complex semantic information. Therefore, in some cases, the representative ability of this method may be limited.

As shown in Table 8, DP-ACOS and JET-ACOS can't discover implicit aspects and opinions.

4.2 Approach based on QUAD-dataset

Zhang et al. (2021a) integrated two pipeline methods for the ACOSQE task. HGCN-BERT+BERT-Linear utilized Heterogeneous Graph Convolutional Networks (HGCN) to detect the aspect category and sentiment polarity. On the other hand, BERT was employed to extract the aspect and opinion terms. A linear layer is then used to combine the extracted features. The second model, HGCN-BERT+BERT-TFM, is similar but replaces the final stacked layer with a transformer block (BERT-TFM).

Zhang et al. (2021a) introduced these two baseline models for comparison with his proposed TASO-BERT-Linear, TASO-BERT-CRF, GAS (Zhang et al. 2021d) and PARAPHRASE (Zhang et al. 2021a) model, aiming to demonstrate the superiority of the PARAPHRASE model. Later, Hu et al. (2022) compared these models with dataset-level order (DLO) and instance-level order (ILO) methods and demonstrated that DLO and ILO achieve better performance in ACOSQE task. These models will be introduced in the following sections. The statistics are summarized in Table 9.

4.3 Approach based on ACOS-dataset and QUAD-dataset

4.3.1 TAS-BERT

TAS-BERT (Wan et al. 2020) is a unified method for extracting aspect-category-sentiment triples. This method integrates aspect category-based sentiment classification and aspect extraction in a unified framework by attaching the aspect category and the sentiment polarity to the review sentence and using it as the input of BERT.

To adapt TAS-BERT to the ACOSQE extraction task on ACOS-Dataset, Cai et al. (2021b) proposed TAS-BERT-ACOS (Pipeline) to adopt the input transformation strategy in TAS-BERT to perform category-sentiment conditional aspect-opinion co-extraction, followed by filtering out the invalid aspect-opinion pairs to form the final quadruples. By comparing the F1 score in Table 8, JET-ACOS showed better results on the REST-ACOS, and TAS-BERT-ACOS performed better on the LAPTOP-ACOS.

To adapt TAS-BERT to the ACOSQE extraction task on QUAD-Dataset, Zhang et al. (2021a) changed its tagging schema to predict aspect and opinion terms simultaneously for constructing a unified model to predict the quad, denoted as TASO-TAS with Opinion (Unified). TASO can be split into two models. The first model, TASO-BERT-Linear, uses BERT for feature extraction and a linear layer for classification. The second model, TASO-BERT-CRF, extends the previous model by adding a Conditional Random Field (CRF) layer on top of the TASO-BERT-Linear model. By comparing the F1 score in Table 9, TASO-BERT-CRF shows better results on the REST15 dataset, and TASO-BERT-CRF performs better on the REST16 dataset.

4.3.2 Extract-classify-ACOS

Cai et al. (2021b) proposed Extract-Classify-ACOS for the task of ACOSQE. The method involves first-extract-then-classify two steps: aspect-opinion co-extraction and category-sentiment prediction.

The first step involves inserting two [CLS] tokens at the beginning and end of the review sentence and feeding it to BERT to obtain context-aware token representations. The explicit aspect-opinion co-extraction is based on a CRF layer with the modified BIO tagging scheme. The second step involves applying two binary classification tasks on the [CLS] tokens to predict whether there is an implicit aspect or implicit opinion. It helps obtain the potential aspect set and opinion set and perform Cartesian Product on the aspect set and opinion set to obtain a set of candidate aspect-opinion pairs. Finally, the category-sentiment classification is modelled as a multiple multi-class classification problem. For each category c, the average vectors of each aspect-opinion pair a-o are concatenated and fed to a fully connected layer with SoftMax function to obtain the sentiment given current a-o and c or indicate an invalid quadruple.

In summary, the method involves using BERT as the backbone to obtain the input text's first and last CLS tokens, which are utilized to detect implicit aspects or opinions. Two binary classification tasks are performed on these tokens, and the middle part of the BERT output is utilized for token classification. The results of the first step are then subjected to multi-label classification, with BERT used again as the backbone.

Compared to TAS-BERT in Table 8, the Extract-Classify-ACOS method first extracts the aspects and opinions and predicts the implicit ones before performing ACOS classification. Extract-Classify-ACOS enables it better to identify implicit aspects and opinions in the text. Therefore, for these implicit aspects and opinions, the method can classify them correctly into their corresponding ACOS categories, thereby improving the overall accuracy. Hu et al. (2022) also employed the Extract-Classify-ACOS method in the REST15 and REST16 Dataset, and it achieved better results compared to the previous BERT-based framework as shown in Table 9.

Although this method was proposed as a baseline, there is still room for improvement. For example, since BERT is the backbone model in both steps without modification, merging the two steps into an end-to-end process or using only one BERT for feature extraction in industrial applications to save computational resources may be possible. The ACOS-Dataset discussed in Extract-Classify-ACOS has an imbalanced distribution across the four combinations of explicit/implicit and aspect/opinion. Various training strategies for imbalanced datasets can be tried. Extracting implicit aspects and opinions from comments still offers significant opportunities for improvement.

4.4 Summary

BERT-based approaches offer a unified framework for ACOSQE but may require extensive data and face challenges in complex scenarios. Further research is needed to address these limitations and improve the efficiency and performance of BERT-based methods in ACOSQE task. In this section, this paper will offer a comprehensive overview of the BERT-based approach, considering it from both the pros and cons perspectives.

- **Pros**
  - BERT approach, such as TAS-BERT and TASO-BERT, provides a unified framework for ACOSQE task, integrating aspect extraction and sentiment classification in a single model.
  - The BERT approach offers several advantages over other deep learning methods for ACOSQE task. It effectively utilizes extensive unlabeled data for pre-training, reducing the need for task-specific data and enhancing model generalization.

BERT's bidirectional context handling enables it to handle ambiguity and complexity more efficiently than unidirectional models. Moreover, BERT's adaptability to various tasks and domains by fine-tuning specific datasets increases its flexibility and applicability.
  - BERT approach has shown promising results in various datasets, such as ACOS-Dataset and QUAD-Dataset, individually or in combination.

- **Cons**
  - BERT is primarily used for natural language understanding tasks such as text classification and named entity recognition, but it cannot generate text. In contrast, BART and T5 are designed for generative tasks such as text summarization and machine translation, making them more advantageous in text generation.
  - The BERT approach typically employs a pipeline framework. In SA tasks, it can encounter the problem of gradient vanishing. This occurs when the gradients become exceedingly small during the backpropagation process, resulting in slow convergence or, in some cases, no convergence at all. The pipeline framework breaks down the task into smaller sub-tasks, and errors in the early stages can propagate and affect the final results, leading to suboptimal performance. The relationship between the different steps in the pipeline framework is essential but often ignored, resulting in a lack of coherence and consistency in the extracted sentiment information.

5 Approach using BART

BART (Bidirectional and Auto-Regressive Transformer) (Lewis et al. 2020) is a pre-trained language model based on the Transformer architecture proposed by the Facebook AI Research (FAIR) team in 2019. BART can perform both forward and backward auto-regressive generation and bidirectional generation, making it suitable for various NLP tasks such as text summarization, machine translation, and question-answering systems.

Compared to other language models, BART's distinguishing feature is its use of a bidirectional encoder and an auto-regressive decoder. The encoder can process both forward and backward text flows simultaneously, while the decoder is used for auto-regressive generation of text sequences. Additionally, BART employs pre-training tasks such as masked language modelling and denoising auto-encoding to enhance model performance. Through pre-training, BART can learn universal language representations that can be fine-tuned for specific NLP tasks. BART has performed excellently in multiple NLP tasks and is widely used in various natural language processing applications.

Section 5.1 explores applying different BART-based methods to the ACOSQE task based on the ACOS-Dataset, as shown in Table 8, and Sect. 5.2 summarizes this approach.

5.1 Approach based on ACOS-dataset

Yan et al. (2021) first proposed a unified generative framework (BARTABSA) using the BART model to solve seven sub-tasks (ATE, OTE, ABSC, AOOE, AOPE, ASPE, AOSTE) of ABSA in an end-to-end manner shown in Table 1, resulting in substantial performance gain and providing a real unified solution.

Inspired by BARTABSA (Li et al. 2019) and GEN-NAT-SCL methods (Peper and Wang 2022; Xiong et al. 2023) proposed the BART-based Contrastive and Retrospective Network (BART-CRN) to extract all ACOS quadruples from a given sentence. The model utilized a machine reading comprehension-based supervised contrastive and retrospective learning module to establish connections among all quadruples and determine context-related generative quadruples end-to-end. As Xiong et al. (2023) used BART-based as the basis, BART-CRN was compared with the baseline methods PARAPHRASE (Zhang et al. 2021a) and GEN-NAT-SCL (Peper and Wang 2022), where PARAPHRASE used T5 as the backbone, and GEN-NAT-SCL was based on the PARAPHRASE model.

For comparison, the T5-based PARAPHRASE and GEN-NAT-SCL-BART models were converted to BART-based models, resulting in PARAPHRASE-BART and GEN-NAT-SCL-BART. GEN-NAT-SCL-BART was used as a baseline model. Experimental results demonstrated that the proposed method significantly outperformed the baselines. The BART-CRN model achieved an F1 score increase of 4.29% and 2.52% compared to the Extract-Classify-ACOS model on the REST-ACOS and LAPTOP-ACOS, respectively. TAS-BERT-ACOS and Extract-Classify-ACOS had overall F1 scores that were not competitive.

BART-CRN overcame the error propagation problem in two-stage models and could jointly model aspects, options, categories, and sentiments. Compared to the PARAPHRASE-BART and GEN-NAT-SCL-BART models, BART-CRN typically had a higher F1 score, which indicates that the MRC-based supervised contrastive and retrospective module can effectively improve the performance of generative models. The outcomes of identifying implicit expressions are presented in Table 8. Compared to the previous baseline models, BART-CRN performs the best in most cases, indicating the effectiveness of their model in implicit quads extraction. Compared with Extract-Classify-ACOS and GEN-NAT-SCL-BART, BART-CRN performs better on EA &IO in LAPTOP-ACOS but worse in REST-ACOS. One reasonable explanation is that the proportion of EA &IO in REST-ACOS is much lower than that in LAPTOP-ACOS, which limits the proposed model's ability to fully capture the features of EA &IO.

According to the research by Yan et al. (2021), Hoang et al. (2022) modified the BARTABSA method to adapt it for the ACOSQE task. This paper refers to the modified method as BARTABSA(Hoang). The method does not change the core algorithm or add other models from BARTABSA. However, it unifies the previous ABSA sub-tasks (i.e., aspect-opinion pair extraction and aspect-opinion-sentiment triple extraction) and extends them to the ACOSQE task.

To address the task of ACOSQE, BARTABSA(Hoang) modified both the data and the model by adding category labels to the data and incorporating a category prediction layer and an opinion prediction layer into the model. At the same time, a version of Hoang's BARTABSA model that used the original sets of categories for classification was tested to demonstrate the effectiveness of dividing categories into two sets of classes. This paper refers to the modified method as BARTABSA(split). The BARTABSA(split) model outperformed the BARTABSA(Hoang). Table 8 shows a significant performance gap between the BARTABSA(split) and Extract-Classify-ACOS models. However, the BARTABSA(split) model performs worse than the previous model Extract-Classify-ACOS on the EA &IO sub-task, with a decrease of about 4% in REST-ACOS as shown in Table 8. Hoang et al. (2022) suggests that this issue could be caused by various factors, such as the skewed data distribution for REST-ACOS, which favours the EA &EO quadruples by a significant margin, and the trade-offs between the model's ability to detect each type of quadruple, as other types have shown a minimum increase of 9% compared to the previous best result in

Table 8. Another comparison with BART-CRN (Xiong et al. 2023) in Table 8 shows that BARTABSA(Hoang) performs better on EA &IO in REST-ACOS and LAPTOP-ACOS. However, it performs worse on IA &EO in REST-ACOS and LAPTOP-ACOS and on IA &IO in LAPTOP-ACOS.

Bao et al. (2022) further improved the ACOSQE task by adapting the BARTABSA (Yan et al. 2021) model. This review refers to the modified method as BARTABSA(Bao). BARTABSA(Bao) achieved a better F1 score than BARTABSA (split). However, BARTABSA(Bao) did not analyze the implicit expressions.

5.2 Summary

BART-based approaches offer the advantage of bidirectional generation and have shown promise in ACOSQE tasks. In this section, this paper will offer a comprehensive overview of the BART-based approach, considering it from both the pros and cons perspectives.

- **Pros**
  - The advantage of the BART-based approach lies in its initial application in generation tasks, especially in the ACOSQE task. This has paved the way for exploring new approaches to tackle SA problems.
- **Cons**
  - BART-based approach may have limitations regarding prompt sensitivity, entity, relation classification, and optimization for ACOSQE. Further research is needed to address these limitations and enhance the performance of BART-based methods in ACOSQE task
  - There is a relatively limited research literature on the BART-based approach in the ACOSQE task. Subsequent studies have found that the T5 model performs better than BART in SA. Nevertheless, it should be noted that BART remains a valuable research tool widely used in other generation tasks and domains. In the future, researchers may explore further improvements and optimizations of BART to meet the specific requirements of SA and consider combining the strengths of different models to achieve better performance.

6 Approach using T5

The T5 model (Raffel et al. 2020) from Google AI Language is a robust transformer-based language model that can perform several natural language processing tasks, including text classification, question answering, summarization, and translation. T5 processes input text in parallel using the transformer architecture and learns contextual relationships between words and phrases, making it effective in understanding natural language meaning and context. It has achieved state-of-the-art performance in several natural language processing benchmarks and is widely used in research and industry applications. T5 has also been utilized to generate natural language text, such as news articles and product descriptions, with the potential to revolutionize the field of natural language generation.

This section explores applying different T5-based models to the ACOSQE task and divides it into three parts based on the sentence-level dataset used, as shown in Tables 8

Table 10 Examples of the target text construction methods depending on different methods

Input sentence: the pizza is delicious	
Method	Target sentence
GEN-NAT	food | the pizza is delicious | positive
SENER	aspect is pizza, opinion is delicious, category is food, sentiment is positive
GAS-A	The [pizza | delicious | food | positive] is delicious.
GAS-E	(Pizza, delicious, food, positive)
PARAPHRASE	food is delicious because pizza is delicious
Opinion Tree	(Root, (Quad, (food, pizza), (positive,delicious)))
Special_Symbols	[AT]pizza[OT]delicious[AC]food[SP]positive

and 9. Section 6.1 introduces the method using ACOS-Dataset, Sect. 6.2 introduces the methods using QUAD-Dataset, Sect. 6.3 concludes the methods using both ACOS and QUAD Dataset, Sect. 6.4 summarizes this approach, and Sect. 6.5 discusses the comparison of methods used in the best results of each approach based on ACOS and QUAD Dataset. Different target text construction methods will be shown in Table 10.

6.1 Approach based on ACOS-dataset

6.1.1 Multi-task instruction tuning

Wang et al. (2022) proposed a general-purpose ABSA framework (UNIFIED-ABSA) based on multi-task instruction tuning, which can uniformly model various tasks and capture the inter-task dependency with multi-task learning. They designed unified sentiment instructions (USI) for each task to prompt the T5 model and provide more semantic information. It only requires the annotated data from the ACOSQE task to derive the annotation for all the eleven ABSA tasks, which shows its advantage in terms of data efficiency.

UNIFIED-ABSA achieved highly competitive performance on 11 tasks, improving the average performance by 1–2% in the fully supervised setting and by 5% F1 score in an extremely low resource setting (with only 32 shots) compared to dedicated models. The effectiveness and superiority of UNIFIED-ABSA over models designed specifically for each task were demonstrated in both fully supervised and low-resource settings in experiments on two public datasets. However, there is still room for improvement in UNIFIED-ABSA, such as automating instruction design, considering SA scenarios across domains, languages, and modalities, and exploring the potential of this framework in other natural language processing tasks.

6.1.2 Contrastive learning and target text construction—NAT

Peper and Wang (2022) proposed a generative model GEN-SCL-NAT that combines two novel techniques: supervised contrastive learning (SCL) and target text construction called natural annotation tree (NAT) for ACOSQE task. GEN-SCL-NAT was based on the T5-Large model, which refers to a specific variant of the T5 model that has a more significant number of parameters than the base T5 model.

- **GEN-SCL** This method distinguishes between input aspects, such as implicit and explicit aspects and opinions.
- **GEN-NAT** This method adapts to a new structured generation format that can better capture the quadruple information in Table 10. Based on the mapping of the four elements (c, a, o, s) to their corresponding semantic values (x_c, x_a, x_o, x_s), the generation format for GEN-NAT is as follows::

$$x_c \mid the\, x_a\; is\, x_o \mid x_s \tag{5}$$

The GEN-SCL-NAT model achieved promising results on the ACOS-Dataset, especially on sentences with implicit sentiment expressions, as shown in Table 8. However, when dealing with longer and more complex examples, output structure validity issues may arise for the GEN-SCL-NAT and PARAPHRASE (Zhang et al. 2021a) models. The evaluation of GEN-SCL was limited to generative sequence prediction models and did not consider other possible forms. Xiong et al. (2023) transferred GEN-NAT-SCL to GEN-NAT-SCL-BART as a baseline model BART-CRN.

Peper and Wang (2022), and Xiong et al. (2023) utilized contrastive learning in their proposed models. However, the differences are Xiong et al. (2023) proposed a machine reading comprehension (MRC)-based supervised contrastive and retrospective learning module, which aims to learn the associations among different types of quadruples and determine the context-related generative quadruples through an end-to-end way. Peper and Wang (2022) proposed a new structured generation format-NAT, which aims to adapt autoregressive encoder-decoder models better to extract quadruples generatively. Table 8 reveals a considerable gap in performance between BART-CRN and GEN-NAT-SCL. GEN-NAT-SCL performs better in most cases, both in overall metrics and the evaluation of implicit expressions identification, indicating the effectiveness of their model in implicit quads extraction. However, it performs slightly worse on IA &EO and IA &IO in the LAPTOP dataset than BART-CRN.

6.1.3 Seq2Path with beam search, pruning and data augmentation

Mao et al. (2022) proposed a novel approach (seq2path) that can generate tree paths of sentiment tuples, where sentiment tuples consist of elements such as aspect words, opinion words, aspect categories, and sentiment polarities as shown in Fig. 4.

The innovation of this approach is that it can better represent the "1-to-n"[1] relationship between aspect words and opinion words, avoiding the arbitrary ordering problem of sentiment tuples in previous methods.

Previous methods used sequence-to-sequence models to generate tuples as sequences, but this might introduce unnecessary order among tuples and ignore the relationship between aspect words and opinion words. In addition, Mao et al. (2022) also introduced a discriminative token and a data augmentation technique for selecting valid paths during inference. The seq2Path method was evaluated on five ABSA tasks and four benchmark datasets, achieving a better F1 score in ACOSQE than the baseline method Extract-Classify-ACOS (Cai et al. 2021b).

[1] The term "1-to-n" refers to a relationship where one element is associated with multiple elements. In this case, a single aspect word can be linked to multiple opinion words within the sentiment tuples.

6.2 Approach based on QUAD-dataset

This section will be divided into three parts to describe the research progress of ACOSQE based on the QUAD-Dataset. The first part will cover multi-task training like assembling Lego bricks; the second part will focus on multi-task instruction tuning for few-shot ABSA; the third part will discuss temple-order data augmentation.

6.2.1 Multi-task training like assembling lego bricks

Gao et al. (2022) proposed a unified generative multi-task framework (LEGO-ABSA) that can solve multiple ABSA tasks (AOPE, ASPE, CSPE, ACSTE, AOSTE, ACOSQE) by controlling the type of task prompts consisting of multiple element prompts.

LEGO-ABSA is a Lego-like method that can transform basic tasks (AOPE, ASPE, CSPE) into advanced tasks (ACSTE, AOSTE, ACOSQE) by assembling task prompts like assembling Lego bricks as shown in Fig. 3. Gao et al. (2022) evaluated the LEGO-ABSA approach by conducting two models.

- **LEGO-ABSA (multi-task)** It involves mixing the training dataset of the individual task and shuffling the order.
- **LEGO-ABSA (separate)** Each task is trained with only one dataset.

Gao et al. (2022) compared TASO-BERT-CRF, PARAPHRASE (Zhang et al. 2021d), and GAS (Zhang et al. 2021a), with his proposed LEGO-ABSA (multi-task) and LEGO-ABSA (separate) and demonstrate that the LEGO-ABSA models outperform them, as shown in Table 9. LEGO-ABSA conducts experiments on multiple benchmark datasets, demonstrating that its multi-task framework achieves new state-of-the-art results in almost all tasks and competitive results in task transfer scenarios.

The approach has also shown competitive results in cross-domain scenarios, demonstrating its effectiveness in handling ABSA tasks in different domains. However, there is still room for improvement in the approach's performance in cross-domain scenarios, and further research is needed to address this challenge.

6.2.2 Multi-task instruction tuning for few-shot ABSA

Varia et al. (2022) proposed a unified instruction Tuning framework for few-shot ABSA tasks. This framework reformulates the ABSA task and its sub-tasks as PARAPHRASE generation problems, fine-tunes a T5 model with instructional prompts in a multi-task learning fashion, and demonstrates that the proposed multi-task prompting approach improved the performance of ABSA in few-shot scenarios on REST15 and REST16 datasets. The influence of all components in the approach is evaluated through three model ablations in Table 11.

- **Text**: Uses $TEXT as input without any transformation. The model must infer the task goal and output format directly from the original text, which may require more data and complex models to achieve good results.

Table 11 Illustration of input prompts to multi-task instruction tuning

Ablation	Input prompt
Text	$TEXT
IT	What are the aspect terms in the text: $TEXT?
IT-MTL	
ATE	What are the aspect terms in the text: $TEXT?
ASPE	What are the aspect terms and their sentiments in the text: $TEXT?
ACSTE	What are the aspect terms, sentiments and categories in the text: $TEXT??
AOSQE	What are the aspect terms, opinion terms and sentiments in the text: $TEXT??
ACOSQE	What are the aspect terms, opinion terms, sentiments and categories in the text: $TEXT ?

$TEXT is the placeholder for the actual text

- **IT**: Transforms $TEXT into instructions by adding natural language descriptions of the task before the text. This method reduces the model's training difficulty and data requirements while improving its generalization ability and flexibility.
- **IT-MTL**: Multiple related tasks are trained simultaneously based on multi-task learning. This method allows the model to obtain context information from multiple tasks (ATE, ASPE, ACSTE, AOSTE, ACOSQE) in Table 1 and perform knowledge transfer between these tasks, enhancing the model's generalization ability and flexibility. The IT-MTL method can leverage multiple instructions to increase the model's generalization ability and flexibility.

Varia et al. (2022) compared the performance of three approaches and discovered that IT-MTL outperforms both IT and Text in most few-shot scenarios in REST15 and REST16, indicating that multi-task learning and instruction tuning can enhance the model's generalization ability in Table 9. Additionally, IT was found to perform better than Text, suggesting that instruction tuning can aid the model in understanding the task's objective and output format. Based on these findings, it was concluded that the trend is IT-MTL > IT > Text. Furthermore, when trained on the whole training data, IT-MTL achieved a comparable or better F1 score than PARAPHRASE (Zhang et al. 2021a) in the REST15 dataset. Overall, it was found that IT-MTL effectively utilizes context from multiple tasks, improving the generalization of the seq-to-seq model for all ABSA tasks in few-shot settings.

6.2.3 Target text construction—named entity recognition

Lee and Kim (2023) proposed the sentiment element named entity recognition (SENER), which significantly outperforms previous works on several ABSA tasks, including ACSTE, AOSTE, and ACOSQE in Table 1. SENER integrates the concepts of named entity recognition (NER) and generative ABSA to retrieve the sentiment entities with predefined sentiment element names (see Table 10), leading to better semantic and sentiment structure understanding. The SENER proposed in the paper has two variants: SENER-orig and SENER-syn.

The main difference between SENER-orig and SENER-syn lies in including specific tokens in the pre-trained vocabulary. In the original version, SENER-orig, the tokens "aspect," "opinion," "category," and "sentiment" are already part of the T5 vocabulary. These tokens and their corresponding embeddings are pre-trained along with the T5 model.

On the other hand, in SENER-syn, four additional tokens, namely “e_a,” “e_o,” “e_c,” and “e_s,” have been introduced as synthetic tokens. This brings the total number of tokens in the vocabulary to 32,132. The specific content of these synthetic tokens can be designated arbitrarily since only their token indices are essential during the tokenization process. The embeddings for these synthetic tokens are randomly initialized and trained during the fine-tuning phase of the model.

In summary, SENER-orig utilizes the existing tokens in the T5 vocabulary, while SENER-syn incorporates additional synthetic tokens to extend the vocabulary size and introduce new elements for SA.

6.3 Approach based on ACOS-dataset & QUAD-dataset

This section focuses on the research progress in ACOSQE, specifically targeting the ACOS and QUAD Dataset using the prompt method. The section is divided into six parts, with the first three subsections dedicated to different methods for constructing target texts. The fourth subsection focuses on template-order data augmentation. The fifth subsection revolves around optimization for regenerating content to reduce the impact of negative noise. The final subsection will centre around the multi-view prompting method.

6.3.1 Target text construction—GAS method

In the realm of ABSA, discriminative approaches are often used to predict specific task outcomes by designing task-specific classification networks that use class indexes as labels for training. However, this approach disregards the semantic richness present in labels, and the need for multiple classification models for different ABSA tasks makes it challenging to generalize the model.

To overcome these challenges, Zhang et al. (2021d) proposed a novel approach called Generative-based SA (GAS) that utilizes a unified generative approach to handle various ABSA tasks. By encoding natural language labels into the target output, the model leverages rich semantic information in labels and can adapt seamlessly to multiple tasks without requiring additional model design.

To train the T5 model, Zhang et al. (2021d) designed a flat sequence including two paradigms, annotation-style (GAS-A) and extraction-style (GAS-E) in Table 10, that frame each ABSA task as a text generation problem. The labelled paradigm annotates the target sentence with label information, while the extraction paradigm directly uses the expected natural language label of the input sentence as the target. Both paradigms produce an original sentence and a target sentence, paired as training instances for the generative model. Additionally, a prediction normalization strategy is proposed in GAS to address the problem of disconnected generated sentiment elements from the corresponding label vocabulary set.

GAS is innovative for the following reasons:

- Novel generative approaches are proposed to handle various ABSA tasks.
- Two paradigms are introduced to formulate each task as a generative problem and a predictive normalization strategy to optimize the generated output.
- Experiments are conducted on multiple benchmark datasets for four ABSA tasks, and the proposed method outperforms the baseline method in almost all cases.

6.3.2 Target text construction—PARAPHRASE method

Zhang et al. (2021a) were among the first to propose the ACOSQE task, which aims to detect quadruples for a given opinionated sentence jointly. A modelling paradigm called PARAPHRASE was introduced to tackle the ACOSQE task, transforming the ACOSQE task into a paraphrase generation problem. PARAPHRASE is approached end-to-end by "re-writing" sentences into structured target sequences, allowing easy decoding of quadruples and making it mainstream in ACOSQE.

$$x_c \; is \; x_s \; because \; x_a \; is \; x_o \tag{6}$$

Using pre-defined rules by Zhang et al. (2021a), the four elements (c, a, o, s) are initially mapped to semantic values (x_c, x_a, x_o, x_s), which are then input into a fixed-order template Eq. 6 to generate a natural language target sequence, as shown in Table 10.

It was argued that PARAPHRASE could provide a more comprehensive and complete aspect-level sentiment structure than existing ABSA tasks. PARAPHRASE leveraged the pre-trained language model T5 to generate natural language sentences that contain all the desired sentiment elements and then recover the sentiment quadruples from the generated sentences. The results demonstrated that PARAPHRASE outperformed the baseline methods regarding performance on QUAD-Dataset, ACOS-Dataset, and other ABSA tasks. Additionally, its unified framework facilitated knowledge transfer across different tasks.

6.3.3 Target text construction—opinion tree method

Bao et al. (2022) proposed a new task called Opinion Tree generation, which aims to jointly detect all sentiment elements in a tree node for a given review sentence as shown in Table 10. Opinion Tree shows that the more complex the tree structure, the better the performance, as it can discover relationships between nodes. The Opinion Tree can reveal a more comprehensive and complete aspect-level sentiment structure for generating sentiment elements.

It was also shown that T5 is introduced to integrate syntax and semantic features for Opinion Tree generation, and the current basis for SA is large generative models that are optimized by constructing target text from the input sentence and output. As shown in Fig. 4, through a comparison between the Opinion Tree and Seq2Path, it was observed that Mao et al. (2022) proposed a method for converting the generation sequence of sentiment tuples into numerical paths. This approach effectively addresses the issue of one-to-many relationships, such as when one aspect entity corresponds to multiple opinion words. It allows for the independent generation of each path without mutual dependence. The average loss of the Seq2Seq method on the path was computed during training. A constrained beam search was applied for inference, and an additional token was introduced to select effective paths automatically.

On the other hand, Opinion Tree Generation (Bao et al. 2022) is a technique used to visualize the results of ABSA by representing them in the form of a tree structure. The root node of the tree represents the overall sentiment of the text. In contrast, the child nodes represent different aspects or features of the entity mentioned in the text and their associated sentiment scores.

Bao et al. (2022) applied the GAS (Zhang et al. 2021d), PARAPHRASE (Zhang et al. 2021a) and BARTABSA (Yan et al. 2021) to the ACOS-Dataset. In this paper, the modified

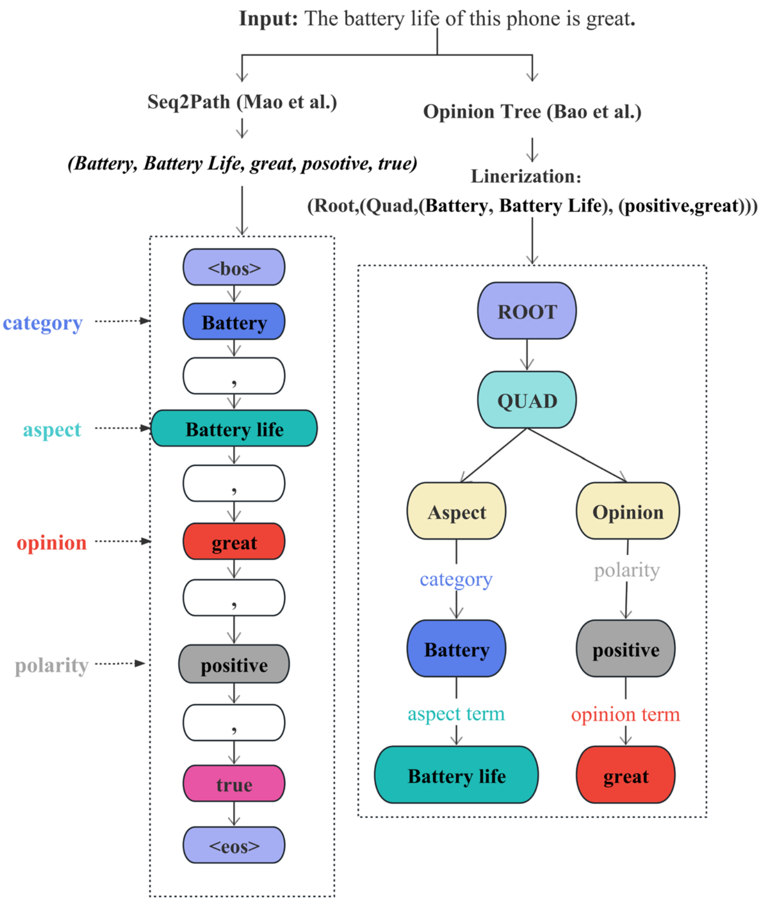

Fig. 4 Compared with the proposed Seq2Path method (Mao et al. 2022) and Opinion Tree method (Bao et al. 2022)

method is referred to GAS(Bao), PARAPHRASE(Bao), and BARTABSA(Bao) to adapt to the ACOSQE task. Subsequently, Bao et al. (2022) compared their methods to the other methods with T5 backbone as shown in Table 8, and the Opinion Tree method performs better on the ACOS-Dataset. Lee and Kim (2023) applied the Opinion Tree (Zhang et al. 2021d) to the QUAD-Dataset. In this paper, the modified method is called Opinion Tree (Lee).

The Opinion Tree method is innovative in that it:

- Introduced a new task and structure for ABSA, which can capture the semantic relations between aspect terms and opinion words more effectively than existing methods as paths of a tree.
- Proposed a constrained decoding algorithm, which can guide the generation process using opinion schemas and ensure the validity of the opinion tree.
- Explored joint learning of several pre-training tasks to integrate syntax parsing and semantic feature parsing, which are very helpful for forming the Opinion Tree structure.

6.3.4 Template-order data augmentation

Hu et al. (2022) argued that the existing methods PARAPHRASE by Zhang et al. (2021a) that used a fixed template order to generate the target sequence are suboptimal because different orders may provide different views of the quadruplet. ACOSQE is not a conventional generation task, and it is not necessary to fix the element order of the quadruplet as long as it can be accurately extracted. So, a novel data augmentation method for the task of ACOSQE was introduced.

The proposed method leverages the order-free property of quadruples. It generates multiple target sequences with different template orders to increase data diversity and provide more information perspectives for pre-trained models.

Given an input sentence x and its quadruples (c, a, o, s), Hu et al. (2022) followed Zhang et al. (2021a) to convert them into semantic values $\{(x_c, x_a, x_o, x_s)\}$. There will be 4! = 24 permutations. They construct all 24 target sequences with multiple order mapping functions S_{O_i}, where i ∈ [1, 24]. An example O_i is shown below. O_i is an abbreviation that denotes "Order i." The Eq. 7 is from Hu et al. (2022).

$$O_i(x_c, x_a, x_o, x_s) = x_a x_c x_o x_s \tag{7}$$

The proposed method is composed of two stages. The first stage aims to select template orders via pre-trained T5. The second stage constructs training samples with the selected orders and fine-tunes T5. They designed the following two strategies for selecting templates.

- **Dataset-level order (DLO)** To choose the dataset-level orders, Hu et al. (2022) compute a score for each order on the training set. The Eq. 8 is from Hu et al. (2022).

$$\mathcal{S}_{O_i} = \frac{1}{|\mathcal{T}|} \sum_{\mathcal{T}} \mathcal{E}(\boldsymbol{y}_{O_i} \mid \boldsymbol{x}) \tag{8}$$

  where S_{O_i} is the average entropy of all instances for the template order O_i, T is the training set, E is the entropy function, y_{O_i} is the predicted quadruplet for the template order O_i, and x is the input sentence. Hu et al. (2022), then rank the scores and choose the template orders with smaller values.
- **Instance-level order (ILO)**, which involves choosing template orders at the instance level based on the context and semantics of each instance.

Multiple template orders can provide different perspectives on a quadruplet during fine-tuning with selected orders. However, the challenge is how to train them jointly. Special markers ([C], [A], [O], [S]) can be used to represent the information structure (x_c, x_a, x_o, x_s) and overcome the difficulty of identifying the value type during inference when concatenating values with a comma or blank space. The target sequence can be constructed using these markers as Eq. 9.

$$\begin{aligned}\mathcal{Y}_{O_i} &= O_i([C]x_c, [A]x_a, [O]x_o, [S]x_s) \\ &= [A]x_a[C]x_c[S]x_s[O]x_o\end{aligned} \tag{9}$$

Hu et al. (2022) conducted joint training across multiple orders and achieved the recovery of quadruplets using these special markers during inference. Unlike data augmentation methods that create multiple inputs for one label, this approach generates multiple labels for one input sequence. It is advantageous for ACOSQE tasks that use generation-based models.

To facilitate the joint training of models with multiple templates, four special markers (T1, T2, T3, T4) were designed for the four elements. The use of various symbols was examined to differentiate the type of element in each position. Specifically, the templates selected for the comparison were T2, T3, and T4, which employed specific words to annotate the type of information. Through experimentation, it was found that T1 achieved the best performance. T1 was denoted as Special_Symbols in Table 10 and designed for comparison with ILO and DLO. It is worth highlighting that both ILO and DLO employ special symbol templates, but they further enhance data augmentation by combining multiple template orders.

- T1: [AT] x_a [OT] x_o [AC] x_c [SP] x_s
- T2: aspect term: x_a opinion term: x_o aspect category x_c sentiment polarity: x_s
- T3: Aspect Term: x_a Opinion Term: x_o Aspect Category: x_c Sentiment Polarity: x_s
- T4: x_a, x_o, x_c, x_s

The results showed that ILO and DLO significantly improve, particularly in low-resource settings. DLO outperforms the previous models on F1 score in REST15, while ILO outperformed the previous models on F1 score in REST16 as shown in Table 9.

While achieving better performance, the template-order approach still has limitations that can guide future research, including exploring alternative criteria for template order selection, developing more sophisticated strategies for selecting orders, and considering augmenting both input and output sequences for further performance improvement.

6.3.5 Uncertainty-aware unlikelihood learning (UAUL)

In their recent review of the template-order data augmentation method, Hu et al. (2023) shed light on the new challenges and proposed a method called Uncertainty-Aware Unlikelihood Learning (UAUL) to improve ACOSQE. UAUL is a template-agnostic method based on T5 that effectively handles negative noise and enhances prediction accuracy. Traditional methods primarily focused on modifying input or output and determining what to generate, which presented challenges in addressing the influence of negative noise while neglecting considerations for what not to generate.

The proposed UAUL method employed multiple iterations of uncertainty-aware sampling to obtain crucial positive and negative samples and controlled their optimization to enhance the discrimination between noise and errors. It aimed to model the influence of negative noise. UAUL consisted of two main components: Monte Carlo dropout (MC dropout) (Gal and Ghahramani 2016) and marginalized unlikelihood learning (MUL). MC dropout was used to randomly drop the last layer parameters of the decoder, obtaining uncertainty-aware samples. Then, UAUL utilized MUL to increase the probability of positive samples and reduce the impact of negative sampling.

By employing MC dropout to obtain crucial positive and negative samples (i.e., noise words) and applying MUL to increase the probability of positive samples while reducing the influence of negative samples, UAUL effectively balanced the optimization process. Additionally, entropy minimization was used to balance the impact of MUL.

The proposed method was applied to GAS (Zhang et al. 2021d), PARAPHRASE (Zhang et al. 2021a), Special_symbols (Hu et al. 2022), DLO (Hu et al. 2022), and ILO (Hu et al.

2022) as GAS+UAUL (Hu et al. 2023), PARAPHRASE+UAUL (Hu et al. 2023), Special_symbols+UAUL (Hu et al. 2023), DLO+UAUL (Hu et al. 2023), and ILO+UAUL (Hu et al. 2023). This paper compared PARAPHRASE+UAUL, Special_symbols +UAUL, DLO+UAUL, and ILO+UAUL with their baseline methods. The experiments demonstrated the effectiveness of the UAUL method across various templates, as shown in Table 8 and Table 9. UAUL, combining MC dropout, MUL, and entropy minimization, achieved outstanding performance in the generative ACOSQE task.

The effectiveness of the UAUL method was validated on four public datasets, showcasing its applicability for learning across various templates. While UAUL performed well in balancing the weights of positive and negative samples, there are still challenges in handling implicit information.

6.3.6 Multi-view prompting (MVP)

Gou et al. (2023) proposed the Multi-view Prompting (MVP) method, which incorporates the DLO method (Hu et al. 2022) and employs element markers to represent the information structure (Paolini et al. 2021). MVP improves upon existing generative methods by aggregating sentiment elements generated in different orders. Existing studies usually predict sentiment elements in a fixed order, which ignores the effect of the interdependence of the elements in a sentiment tuple and the diversity of language expression on the results. MVP introduces element order prompts to guide the language model to generate multiple sentiment tuples, each with a different element order, and then selects the most reasonable tuples by voting. This approach aligns training and inference with multi-view prompt learning, improving the model's effectiveness, flexibility, and cross-task transferability.

MVP can naturally model multi-view and multi-tasks as permutations and combinations of elements, respectively, outperforming previous task-specific designed methods on multiple ABSA tasks with a single model. MVP demonstrates a significant improvement in the state-of-the-art performance on 10 datasets for 4 benchmark tasks included in QUAD-Dataset and ACOS-Dataset, as presented in Table 8 and Table 9. Additionally, it performs effectively in low-resource settings.

6.4 Summary

Based on the research presented in this paper, it is evident that the T5-based approach currently dominates the ACOSQE task field. Numerous studies have emerged based on T5, incorporating techniques such as various templates for target text construction stemming from prompt engineering, data augmentation, contrastive learning, multitask learning, extensions to both the T5 encoder and decoder components, as well as instruction tuning, and more. These methods have contributed to enhancing ACOSQE. In this section, this paper will offer a comprehensive overview of the T5-based approach, considering it from both the pros and cons perspectives.

- **Pros**
  - In the field of ACOSQE, it is evident that many current endeavours are built upon the foundation of the T5-based approach. The T5 approach closely follows the original encoder-decoder architecture of the transformer model, which has become a cornerstone in ACOSQE research. This architecture enables the T5 approach to excel in various generative tasks, making it a versatile and influential model in the domain.

  - As ACOSQE research advances, the foundational framework of T5 plays a crucial role as a reference point, enabling ongoing progress and innovations in the field. Some notable developments include prompt engineering, data augmentation, contrastive learning, multitask learning, T5 encoder and decoder component extensions, instruction tuning, and more.

- **Cons**
  - Fine-tuning T5 for specific ACOSQE tasks can be complex and require substantial labelled data, which might not always be readily available.
  - Like many deep learning models, T5's inner workings can be challenging to interpret, making it difficult to understand why certain predictions are made, particularly in complex sentiment analysis task like ACOSQE.

6.5 Summary of the best method for each approach

The comparison of methods used in the best results of each approach for the ACOSQE task based on ACOS-Dataset and QUAD-Dataset is presented in Table 12.

- For ACOS-Dataset, the best-performing method in the rule-based approach is DP-ACOS (Cai et al. 2021b), while in the BERT-based approach, Extract-Classify-ACOS (Cai et al. 2021b) performs the best. BARTABSA (Bao) (Bao et al. 2022) shows the best results in the BART-based approach. For the T5-based approach, the opinion tree method performs the best.
- For QUAD-Dataset, the focus is mainly on the BERT and T5 approaches. Based on the BERT approach, the Extract-Classify-ACOS (Cai et al. 2021b) method performs best. In the T5 approach, the MVP method (Gou et al. 2023) shows the best results.

These methods underscore the pivotal role of transformer architecture-based pre-trained language models in driving the rapid advancements in ACOSQE. Notably, the T5-based approach consistently outperforms both the BART-based and BERT-based approaches. This emphasizes the robustness, versatility, and dependability of the T5-based model when tackling complex SA tasks, solidifying its position as a front-runner in the field.

A noteworthy insight from these results is the substantial improvement observed in generation-based methods compared to their pipeline-based counterparts. This improvement is particularly striking because pipeline-based methodologies often fall prey to error propagation, where inaccuracies introduced at one stage can ripple through subsequent stages, ultimately compromising overall performance. In contrast, generation-based methods excel by considering the entire context when generating sentiment-related content, resulting in more accurate and coherent sentiment predictions.

Unfortunately, it's disappointing that research on implicit sentiment analysis within methodologies based on BERT, BART, and the T5 approach has remained relatively limited. Despite the remarkable performance of T5 in the ACOSQE task, the exploration of implicit sentiment within the T5 framework remains somewhat underdeveloped.

Implicit sentiment analysis presents a unique and intricate set of challenges, demanding a keen understanding of context, nuanced language usage, and the ability to unveil sentiments subtly interwoven within the text. While models such as BERT, BART, and T5 have showcased their strengths in explicit sentiment analysis, a wealth of untapped potential is still waiting to be explored in uncovering concealed sentiments.

Table 12 Comparison of best results of each approach for the ACOSQE task based on ACOS-Dataset and QUAD-Dataset

Methods	REST-ACOS			LAPTOP-ACOS			REST15-QUAD			REST16-QUAD		
	P.	R.	F1.	P.	R.	F1.	P.	R.	F1.	P.	R.	F1.
Rule-based												
DP-ACOS	**34.67**	**15.08**	**21.04**	**13.0**	**5.70**	**8.0**	N/A			N/A		
BERT-based												
Extract-Classify-ACOS	**38.5**	**52.96**	**44.61**	**45.56**	**29.48**	**35.80**	**35.64**	**37.25**	**36.42**	**38.4**	**50.93**	**43.77**
BART-based												
BARTABSA (Bao)	**56.62**	**55.35**	**55.98**	**41.65**	**40.46**	**41.05**	N/A			N/A		
T5-based												
MVP	–			–			N/A	N/A	**51.04**	N/A	N/A	**60.39**
MvP (multi-task)	–			–			N/A	N/A	**52.21**	N/A	N/A	**58.94**
Opinion tree	**63.96**	**61.74**	**62.8**	**46.11**	**44.79**	**45.4**	–			–		

The best results are in bold

Addressing this research gap is of paramount importance. Expanding our knowledge and capabilities in implicit sentiment analysis within BERT, BART, and T5 could lead to more robust and comprehensive SA systems. As the field continues to evolve, it is imperative to encourage and support further investigations into implicit sentiment analysis within these frameworks, ultimately enhancing our ability to decode complex emotions within textual data.

7 Dialogue-level quadruple analysis

Song et al. (2022) introduced a new task called conversational aspect sentiment analysis (CASA), which aims to provide fine-grained sentiment information for dialogue understanding and planning.

CASA extended the standard ABSA to the conversational scenario with several significant adaptations. 3,000 chit-chat dialogues (27,198 sentences) were annotated with fine-grained sentiment information, including all sentiment expressions, their polarities, and the corresponding target mentions. An out-of-domain test set of 200 dialogues is also included for robustness evaluation. Multiple baselines are developed based on pre-trained BERT. CASA can help select knowledge for knowledge-driven dialogue response generation, improving memory efficiency and performance.

Although CASA is centered on ABSA at the dialogue level, omitting crucial elements such as aspects could render it incapable of comprehensively depicting the opinion status. Li et al. (2022a) proposed a task of conversational ACOSQE (DiaASQ) to detect quadruples from dialogues.

DiaASQ is a novel task that combines ABSA and conversational opinion mining. It constructs a high-quality, large-scale dataset in Chinese and English, proposes a benchmark model that can effectively perform end-to-end quadruple prediction, and leverages rich dialogue-specific and discourse feature representations to improve cross-utterance quadruple extraction.

The model consists of several layers, including:

- **Base encoding:** Li et al. (2022a) uses BERT to encode the dialogue utterances. The encoding is done separately for each utterance using the [CLS] and [SEP] tokens to separate each utterance. The contextual representation of each word is obtained using the PLM.
- **Multi-view interaction layer:** Multi-view interaction is a layer in the DiaASQ model that captures the relationships between the speakers and their utterances in dialogue. This layer aggregates dialogue-specific feature representations, such as the threads, speakers, and replying information, to improve the model's ability to ACOSQE. The multi-view interaction layer is built upon the multi-head self-attention mechanism and uses attention masks to control the interactions between tokens based on the prior feature information.
- **RoPE layer:** RoPE stands for Rotary Position Embedding, a technique used in the DiaASQ model to add relative dialogue distance information to the input sequence. This helps the model better understand the conversation's discourse structure and improves its ability to ACOSQE across utterances.
- **Prediction layer:** This layer predicts the sentiment quadruples based on the grid-tagging labels.

Table 13 Main results of the DiaASQ task

Language	Model	Quadruple (F1)	
		Micro	Iden.
ZH	CRF-Extract-Classify	8.81	9.25
	SpERT	13.00	14.19
	PARAPHRASE	23.27	27.98
	Span-ASTE	27.42	30.85
	DiaASQ	**34.94**	**37.51**
EN	CRF-Extract-Classify	11.59	12.80
	SpERT	13.07	13.38
	PARAPHRASE	24.54	26.76
	Span-ASTE	26.99	28.34
	DiaASQ	**33.31**	**36.80**

The best results are in bold

The model was optimized during training using the cross-entropy loss function, computed based on the predicted sentiment quadruples and the ground truth labels. The model was trained end-to-end using backpropagation and stochastic gradient descent. *micro F1* and *identification F1* were used for measurements, where *micro F1* measures the entire quadruples, including sentiment polarity, while *identification F1* (Barnes et al. 2021) did not differentiate polarity.

Li et al. (2022a) compared DiaASQ with four baseline models. The first model is CRF-Extract-Classify, proposed by Cai et al. (2021b) as mentioned earlier in Sect. 4.3.2 and was retrofitted to support target term extraction for DiaASQ. The second model is SpERT, proposed by Eberts and Ulges (Eberts and Ulges 2019) for joint extraction of entity and relation based on a span-based transformer. It was modified for triple-term extraction and polarity classification for DiaASQ. The third model is Span-ASTE, a span-based approach for triplet ABSA extraction proposed by Xu et al. (2021), which was edited to enumerate triplets and made compatible with DiaASQ. The fourth model is PARAPHRASE (Zhang et al. 2021a), which outputs were modified to adapt to the DiaASQ task. All baseline models used the same model-BERT, except for PARAPHRASE,[2] which utilized mT5-base (Xue et al. 2021). Through comparison, it was found that the DiaASQ model achieved the best results on both English and Chinese Datasets, as shown in Table 13.

The DiaASQ model is designed to extract sentiment quadruples from dialogues by incorporating rich dialogue-specific and discourse feature representations. The model is trained to minimize the cross-entropy loss and is optimized using backpropagation and stochastic gradient descent. The task is challenging because it requires cross-utterance extraction and dialogue-specific features. To facilitate follow-up research in this direction, Li et al. (2022a) suggests several potential future directions:

[2] When Zhang et al. (2021a) first proposed the PARAPHRASE model, they used the T5 model (Raffel et al. 2020).

Table 14 Comparison of parameters and pretraining data size among the large language models

Released date	Institution	Large language model name	Model parameters	Pre-training data size
June 2018	OpenAI	GPT	110 M	5GB
Dec 2018	Google	BERT	330 M	16GB
February 2019	OpenAI	GPT-2	1.5B	40GB
Dec 2020	Facebook	BART	406 M	160GB
Dec 2020	Google	T5	11B	750GB
May 2020	OpenAI	GPT-3	175B	45T
Mar 2023	OpenAI	GPT-4	Unpublished	Unpublished

- **Making better use of the dialogue discourse structure information:** exploring how to leverage better the dialogue discourse structure information, such as the relationships between sentences in the dialogue, to improve the model's performance.
- **Enhancing coreference resolution:** using more complex coreference resolution models to accurately identify pronouns and noun phrases in the dialogue, thus further improving the model's performance.
- **Extracting overlapped quadruples:** exploring how to extract overlapped sentiment quadruples in the dialogue to comprehensively understand the sentiment in the conversation.
- **Transferring well-learned sentiment knowledge from existing systems:** exploring how to apply sentiment knowledge learned from existing systems to the DiaASQ task to improve the model's performance.
- **Multi/cross-lingual dialogue ABSA:** expanding the DiaASQ task to multi and cross-lingual contexts to adapt to a wider range of application scenarios.

8 Exploring ChatGPT for sentiment analysis

ChatGPT[3] is a language model that utilizes cognitive computing and artificial intelligence. It's built on the Transformer architecture (Vaswani et al. 2017) and generative pre-training technique, GPT (Radford et al. 2018), to train the model and predict the probability distribution of the next word, generating natural language text.

OpenAI's language models have continuously improved from GPT-1 (Radford et al. 2018) to GPT-3 (Brown et al. 2020), with an increase in model parameters and improvements in self-supervision, enhancing their language processing and generation capabilities as shown in Table 14. InstructGPT, based on the RLHF (reinforcement learning from human feedback) (Ouyang et al. 2022)method, significantly reduced the probability of harmful, untruly, and biased outputs. ChatGPT was launched as a sister model of InstructGPT, with chat-specific attributes and a public testing version. Apart from generative pre-training, ChatGPT's success relies on several core technologies.

[3] https://chat.openai.com/.

- **RLHF** The RLHF method is a reinforcement learning approach that improves the model's responses by utilizing human evaluations of the dialogue agent's answers.
- **Instruction fine-tuning (IFT)** ChatGPT uses IFT (Wei et al. 2022a) to simulate human chat behavior. This technology can track, learn, and reproduce chat history and apply it to natural language modeling and inference in real-time conversations. In addition to fine-tuning the model using classical NLP tasks like SA, text classification, and summarization, IFT demonstrates various written instructions and their outputs to the base model in diverse task sets, achieving fine-tuning of the model.
- **The chain-of-thought (CoT)** The CoT is a few-shot prompting technology (also known as in-context learning, ICL) proposed by Google (Wei et al. 2022b) to help large language models better understand language requests by providing contextual information. It can improve the model's accuracy and flexibility, enhancing its processing capability.

8.1 Evaluation of ChatGPT

Wang et al. (2023) presented a preliminary evaluation of ChatGPT's performance in SA tasks, including standard evaluation, polarity flipping evaluation, open-domain evaluation, and sentiment reasoning evaluation. Comparing ChatGPT's performance on 18 benchmark datasets, it was discovered that ChatGPT performs well. The key findings include:

- **Zero-shot performance**[4] ChatGPT demonstrates remarkable zero-shot performance in some tasks like ASPE and can compete with fine-tuned BERT. However, it lags slightly behind the state-of-the-art fully-supervised models in domain-specific settings.
- **Few-shot performance**[5] ChatGPT's ability can be significantly improved through few-shot prompting across various tasks, datasets, and domains, allowing it to outperform fine-tuned BERT in some cases but still lag behind state-of-the-art models.
- **Human evaluation** Exact matching evaluation hinders ChatGPT's performance. ChatGPT is less accurate in sentiment information extraction tasks but can still generate reasonable answers and perform well in human evaluation despite not strictly matching the textual expression.
- **Polarity flipping evaluation** ChatGPT outperforms fine-tuned BERT in making accurate predictions for SA tasks involving polarity shift phenomena such as negation and speculation
- **Open-domain evaluation** Compared to domain-specific models that struggle when applied to unseen domains, ChatGPT shows strong open-domain SA abilities overall. However, its performance may be limited in certain specific domains.
- **Sentiment inference evaluation** ChatGPT has a remarkable ability to infer sentiment, showing similar performance on the emotion-cause extraction task or emotion-cause pair extraction task as fully supervised SOTA models.

Zhang et al. (2023) conducted further exploration of ChatGPT in SA. The performance of Large Language Models (LLMs) in SA tasks was evaluated, and a comparison was made with Small Language Models (SLMs) in zero-shot and few-shot scenarios. The question of whether current evaluation practices for SA are still applicable in the era of LLMs was

[4] Zero-shot refers to the learning process of a model without annotated data.

[5] Few-shot refers to learning with only a small amount of labeled data.

also raised. The performance was evaluated across 13 tasks on 26 datasets. The LLMs and SLMs were as follows:

- **Large language models (LLMs)** LLMs refer to large-scale language models, such as GPT-3. These models utilize the Transformer architecture, possess many parameters and pretraining data, and perform excellently in various natural language processing tasks. Zhang et al. (2023) selected two models from the Flan model family, namely Flan-T5 (XXL version, 13B)(Chung et al. 2022) and Flan-UL2 (20B)(Tay et al. 2023). Additionally, two models from the GPT−3.5 family were employed, including ChatGPT (gpt−3.5-turbo) and text-davinci-003 (text-003, 175B)(Brown et al. 2020; Ouyang et al. 2022). These models were directly used for inference in SA tasks without specific training.
- **Small language models (SLMs)** SLMs, which are smaller-scale language models, are typically trained on specific datasets. T5 (large version, 770 M) was selected as the SLM by Zhang et al. (2023). The models were trained using either the entire training set or a few-shot approach by sampling a subset of the data.

In zero-shot Performance, LLM demonstrates impressive zero-shot performance in simple SA tasks. The experimental results show that without pretraining, LLM performs well in Sentiment Classification (SC) and Multifaceted Analysis of Subjective Text (MAST) tasks. However, some slightly more challenging tasks, such as Yelp-5 with increased categories, reveal that LLM performs relatively poorly compared to fine-tuned models. It is important to note that larger models do not necessarily guarantee better performance.

According to the experimental results, LLM excels in SC and MAST tasks without pretraining. However, it can be observed that LLM struggles with more complicated tasks like Yelp-5 with increased categories, where it underperforms compared to fine-tuned models. LLM faces challenges in extracting fine-grained structured sentiment and opinion information. The experiments also show that Flan-T5 and Flan-UL2 are unsuitable for ABSA tasks, while text-003 and ChatGPT achieve better results. However, they are still weaker than fine-tuned small language models (SLMs). On the QUAD-Dataset, the T5-Large outperforms ChatGPT significantly.

RLHF may lead to unexpected phenomena. The experimental results indicate that ChatGPT performs poorly in detecting hate speech, irony, and offensive language. Even compared to text-003, which performs similarly well in many other tasks, ChatGPT's performance is notably worse in these three tasks. This finding highlights the necessity for further research and improvement in these domains, as it suggests that ChatGPT's RLHF process excessively aligns with human biases and preferences.

In few-shot Performance, LLM performs superior to SLM in different few-shot settings. Across the 1-shot, 5-shot, and 10-shot setups, LLM consistently outperforms SLM in almost all cases. The advantage of LLM is particularly prominent in ABSA tasks, which require structured sentiment information, an area where SLM falls behind LLM. This could be attributed to the increased difficulty of learning such patterns in limited data scenarios. By increasing the number of shots, SLM achieves consistent performance improvements in most tasks, indicating its ability to effectively utilize more examples for better performance.

The experiments also reveal the complexity of tasks. While the performance of the T5 model gradually stabilizes for sentiment classification tasks, it continues to improve for ABSA and MAST tasks, suggesting the need for more data to capture their underlying patterns. On the QUAD-Dataset, the performance gap between the T5large model and

ChatGPT diminishes as the number of shots increases. The impact of increasing shots on LLM varies across different tasks. The influence of increasing shot quantities on LLM is task-dependent. Increasing shots for relatively simple tasks like SC does not noticeably improve performance. Additionally, for datasets like MR and Twitter, as well as stance and comparison tasks, performance even deteriorates with the increase in shots. It could be due to handling more extended contexts misleading LLM's results. However, increasing the number of shots enhances LLM's performance for ABSA tasks requiring more profound and precise output formats. It suggests that more examples are not a universal solution for all tasks but should be determined based on task complexity.

Zhang et al. (2023) also calls for a more comprehensive evaluation, a more natural model interaction, and sensitivity to prompt design when discussing the potential of ChatGPT. A comprehensive evaluation of LLM across various SA tasks has been conducted to enhance our understanding of its capabilities within the SA field. The experiment results indicate that while LLM performs well on simple tasks under zero-shot conditions, it faces difficulty handling more complex tasks. LLM consistently outperforms SLM in the few-shot setting, demonstrating its potential when labelled resources are scarce. The limitations of current evaluation practices are also highlighted, and the SENTIEVAL benchmark is proposed as a more comprehensive and realistic evaluation tool.

In conclusion, regarding training SA systems for specific domains, ChatGPT has demonstrated the ability to function as a universal and well-performing sentiment analyzer. However, it's essential to consider both its advantages and limitations. This paper will delve into these aspects from both the pros and cons perspectives.

- **Pros**
  - ChatGPT is good at solving language tasks such as ACOSQE with simple prompts.
  - Compared to fine-tuning BERT, ChatGPT is more effective in addressing the polarity shift issue in SA and performs well in open-domain scenarios.
  - It can adapt to various domains and scenarios without specific datasets or pre-trained models, using its powerful generation ability and common-sense knowledge to handle diverse expressions of sentiment.
  - ChatGPT also benefits from a larger corpus, comprising extensive unlabeled pre-training data and a training set that includes over 23 million dialogue records spanning multiple languages, such as English and Chinese. It encompasses around 70 million lines, including sentences generated by many real users.
  - ChatGPT can interact with users conversationally to refine the results of quadruple extraction.
- **Cons**
  - ChatGPT's dependency on prompts and access to relevant information may limit its effectiveness in ACOSQE task.
  - ChatGPT's performance may vary depending on the quality and relevance of the training data used.
  - ChatGPT may lack sufficient domain-specific depth, potentially resulting in less reasonable generated content. Additionally, there is the possibility of bias in ChatGPT since it is trained on a large amount of data, which biases could influence in the training data.
  - ChatGPT still encounters various technical challenges, including significantly high training expenses, system intricacy, and increased testing costs. In cases where com-

putational resources are inadequate, ChatGPT may not be utilized efficiently. Nonetheless, the most pressing concern is the possibility of its filtering system being circumvented or “breached.”

8.2 T5 vs ChatGPT

The previous discussions show that the current approaches to solving ACOSQE primarily rely on the T5-base and T5-large models. In this section, this review will discuss the differences between the T5 and ChatGPT.

ChatGPT and T5 (Raffel et al. 2019) are two powerful natural language processing models that can be used for ACOSQE tasks. However, they differ significantly in quadruple extraction performance.

The T5-based method is more robust and produces higher quality results, while ChatGPT has stronger generality and generalization ability, giving it an advantage in dialogue tasks. The T5-based method uses sequence-to-sequence mapping, making it flexible and scalable. This method usually involves pre-training and fine-tuning stages. The pre-trained model learns universal text representations and language patterns from large-scale text data, while the fine-tuning model trains on specific SA datasets to improve the model’s generalization ability and performance. In some datasets, the T5-based method has achieved good results. Overall, this method has advantages in performance and flexibility, but the impact of factors such as training datasets and model parameters must be considered.

However, compared with the methods based on the T5 model, ChatGPT has some limitations or challenges in the ACOSQE task. Firstly, ChatGPT requires complex prompts to define the structure and format of the quadruples, which may affect the user experience and model efficiency. Secondly, ChatGPT’s performance is unstable when dealing with a long tail and complex scenarios, and there is a significant performance gap compared with specially designed models.

In addition, ChatGPT is very sensitive to different prompt styles, which may lead to inconsistent or incorrect results. The T5 model is trained to extract quadruples by supervised learning with a large amount of quadruple data, which can improve performance using a large amount of annotated data. While ChatGPT has yet to be trained and optimized for the quadruple task directly, its quadruple extraction ability mainly relies on dialogue understanding and question answering, and it needs multiple rounds of dialogue to get all the quadruple information. It does not explicitly give the entity and relation classification results.

In summary, ChatGPT and T5 have advantages and disadvantages in the ACOSQE task, and the specific choice should be based on the actual task and needs. Now that GPT-4 is being developed OpenAI (2023), it possesses a powerful neural network prediction capability and builds upon the foundation of ChatGPT to transition from a single modality to a multi-modal approach, elevating the safety, creativity, and logic of generative AI models to new heights. Information received by humans from the real world often originates from diverse modalities, and this shift from single to multi-modal capabilities enhances the model’s versatility and generalization ability. Currently, GPT-4 demonstrates a high level of maturity in text generation and interactive communication, yet certain technical limitations persist. For instance, it lacks proficiency in handling inputs and outputs beyond textual formats, leading to the potential generation of inaccurate or irrelevant responses.

While GPT-4 has progressively approached and even exceeded human capabilities in perception and judgment, it still falls short of human-level logical reasoning and decision analysis. Furthermore, GPT-4 heavily relies on imitating responses based on training datasets and struggles with verifying information sources and authenticating textual content.

Given that the design objective of large language models (LLMs) is to enhance natural language understanding and generation, assessing the potential positive and negative impacts of introducing new LLM versions is crucial.

- **Pros**
  - Positive impacts could encompass improved text generation accuracy, enhanced coherency, and better contextual relevance, potentially leading to more effective communication, refined content creation, and novel possibilities in creative writing and human-computer interaction.
- **Cons**
  - Conversely, potential negative effects must be considered. As LLMs become more complex, there is a risk of generating misleading or biased content. Ethical concerns related to inaccurate information, privacy breaches, and the potential reinforcement of harmful stereotypes could become more pronounced. Additionally, the increasing complexity of LLMs might impede their interpretability and debugging, posing challenges in ensuring transparency and accountability.
  - While LLMs don't depend on domain-specific annotated data for training, there is still a need to assess their open-domain SA performance across various domains. Consequently, a significant large-scale, multi-domain, multi-element dataset for ABSA tasks is lacking.

To balance these positive and negative impacts, rigorous testing, continuous monitoring, and robust fine-tuning and updating mechanisms for LLMs are essential. Developing and deploying new LLM versions involving diverse stakeholders, including linguists, ethicists, and the broader public, is paramount.

9 Challenges and future directions

ACOSQE has made significant progress in recent years, but some challenges and issues still need to be addressed. Rule-based approach relies on predefined rules and patterns to identify aspects, opinions, and sentiment polarity in text, which may not capture the complexity and nuances of natural language. The BERT-based approach marks the first use of a transformer encoder-based pre-trained language model in ACOSQE. However, it exhibits limitations when combined with traditional machine learning methods like CRF. These limitations underscore the challenges of employing a pipeline approach in SA tasks, including issues such as gradient vanishing, error propagation, and better coordination among different stages. The BART-based approach was the first to introduce an encoder-decoder model for ACOSQE. However, it had limitations, including a lack of extensive fine-tuning method research. Most researchers simply adapted fine-tuning methods based on T5 and applied them to BART for comparison. Research focusing on implicit sentiment in ACOSQE remains relatively scarce. T5-based approach has limitations in terms of model parameters compared to LLMs (large language models). These limitations persist, although there have been significant developments in this area, including prompt engineering, data augmentation, contrastive learning, multitask learning, and extensions to the T5 encoder and decoder components.

With the growing amount of text data from social media and online comments, it is crucial to increase the accuracy and efficiency of ACOSQE. This section will mainly focus on these challenges and potential solutions.

9.1 Richer quadruple datasets

A significant challenge with ACOSQE is the lack of large-scale and diverse datasets covering various fields, scenarios, and languages. Currently, most datasets for this task are small, limited in scope, and focus only on explicit aspects and perspectives. For example, the SemEval (Pontiki et al. 2014, 2015; Hercig et al. 2016) dataset contains only about a few thousand sentences per domain (restaurants and laptops) and contains no implied aspects and perspectives. The ACOS-Dataset is more extensive and more affluent than the SemEval dataset and has implicit aspects and opinions but is still insufficient to fully capture the complexity and diversity of ACOSQE. Therefore, more datasets are needed to provide comprehensive and fine-grained annotations for explicit and implicit aspects, categories, opinions, and sentiments. These datasets can assist in evaluating and enhancing the performance of current methods, as well as developing new models capable of handling more challenging scenarios.

9.2 Ambiguity, implicit sentiment & consistency

ACOSQE is challenged not only by the scarcity and diversity of datasets but also by the ambiguity, implicitness, and complexity of aspects and opinions and the relationships and consistency among quadruples. Ambiguity arises when the same word or phrase can convey different meanings or sentiments depending on the context or domain. For example, “small” may be positive or negative for different products or attributes. Implicitness and complexity occur when some aspects or opinions are not explicitly mentioned in the sentence and require common sense or background knowledge to infer. For example, “this restaurant is clean” suggests the aspect of “service” and the opinion of “good,” but they are not overtly expressed. Relationships and consistency involve the logical or semantic connections among aspects or opinions, which must be preserved or differentiated. For example, “screen” and “monitor” may refer to the same or different aspects depending on the product or attribute. These challenges make ACOSQE a difficult and complex task that demands more advanced and robust methods.

9.3 ChatGPT’s potential in ACOSQE

To comprehensively evaluate ChatGPT’s potential in ACOSQE, especially in implicit sentiment extraction, we need to conduct extensive experiments and compare the results using evaluation metrics to evaluate its performance comprehensively. Before testing, we must preprocess the datasets to fit ChatGPT’s input format. After completing training, we can compare ChatGPT’s performance on ACOS-Dataset, Quad-Dataset, and DiaASQ Dataset with other classic models to better understand its performance in the ACOSQE extraction task.

Zhang et al. (2023) propose that current evaluations based on ChatGPT primarily focus on specific SA tasks and datasets, limiting a comprehensive understanding of LLM capabilities. This limitation reduces the reliability of evaluation results and restricts the model's adaptability to different SA scenarios. Therefore, there is a need for comprehensive evaluations across a wide range of SA tasks. Zhang et al. (2023) also advocates for more natural model interactions. Traditional SA tasks typically pair a sentence with a corresponding sentiment label. While this format helps learn the mapping between text and sentiment, it may not be suitable for LLMs as they are often generative models. In practice, different writing styles lead to different ways LLMs approach SA tasks, so considering diverse expressions in the evaluation process is crucial to reflect realistic use cases. It ensures that evaluation results reflect real-world interactions and provides more reliable insights. Additionally, the sensitivity of prompt design poses a challenge as different prompts may have varying effects on different models, making fair comparisons complex.

9.4 Future directions

Several directions are worth exploring to address these challenges, Firstly, building larger and more diverse quadruple datasets that cover various domains, scenarios, and languages is crucial. Additionally, handling implicit, complex, and ambiguous aspects and opinions and addressing issues related to relationships and consistency between quadruples need attention. Secondly, developing more effective and robust quadruple extraction methods using the advanced pre-trained model and API interface, such as ChatGPT, Sparkdesk,[6] ERNIE Bot,[7] Bard[8] can significantly improve the accuracy and efficiency of ACOSQE. Thirdly, exploring additional applications for ACOSQE, such as dialogue systems, and integrating them with other natural language processing tasks, such as text summarization and generation, can demonstrate the value and importance of ACOSQE. Furthermore, handling multilingual and cross-domain sentiment data remains a significant challenge that requires ongoing research and solutions in ACOSQE.

10 Conclusion

This paper provides a comprehensive and in-depth examination of various aspects of ACOSQE, including task definition, dataset characteristics, method categorization, challenge analysis, and future prospects. The paper begins by explaining the research background of ACOSQE based on the four elements of ABSA. It further elucidates the concept of the task, dataset features, and evaluation metrics. Subsequently, the paper analyzes and compares research methods across different branches of pre-trained language models and datasets, specifically focusing on a novel dialogue-based ACOSQE approach. Additionally, the paper investigates the capabilities of ChatGPT in ABSA tasks and compares the advantages and disadvantages of specific pre-trained models in the context of ACOSQE. Following this, the paper summarizes the issues and challenges in the ACOSQE domain, such as data scarcity, expression ambiguity, and implicit sentiment recognition. It proposes

[6] https://xinghuo.xfyun.cn/.

[7] https://yiyan.baidu.com/welcome.

[8] https://bard.google.com/.

potential solutions and suggestions to address these challenges and enhance ChatGPT's capabilities in SA. Finally, the paper outlines future research directions.

Future research should be dedicated to finding larger and more challenging datasets to improve model generalization. Additionally, while many current ACOSQE studies rely on simple template constructions, formalizing the ACOSQE problem as a generation task or machine reading comprehension problem, future research should emphasize rich model architecture design. For instance, a deeper extension of the Transformer architecture could facilitate more effective extraction of implicit sentiment features and sentiment-semantic relationships. As research progresses, attention should be focused on the accuracy of the ACOSQE task and its application in various contexts, such as cross-domain ABSA, cross-lingual ABSA, lifelong ABSA, and multimodal ABSA. These methods can potentially bring about new challenges and opportunities for ACOSQE, thereby contributing to this field's continuous development and innovation.

Author contributions BV performed the selection process and wrote the main manuscript text. All authors read and discussed the selected studies, and all authors reviewed the manuscript.

Declarations

Conflict of interest The authors declare no competing interests.

References

Al-Janabi OM, Ibrahim MK, Kanaan-Jebna A et al (2022) An improved bi-LSTM performance using DT-we for implicit aspect extraction. 2022 International Conference on Data Science and Intelligent Computing (ICDSIC). IEEE, pp 14–19

Bao X, Wang Z, Jiang X et al (2022) Aspect-based sentiment analysis with opinion tree generation. IJCAI 2022:4044–4050

Barnes J, Kurtz R, Oepen S, et al (2021) Structured sentiment analysis as dependency graph parsing. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, vol 1: Long Papers, pp 3387–3402

Blair-Goldensohn S, Hannan K, McDonald R, et al (2008) Building a sentiment summarizer for local service reviews. WWW2008 workshop on NLP challenges in the information explosion era

Brown T, Mann B, Ryder N et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

Brun C, Popa DN, Roux C (2014) Xrce: hybrid classification for aspect-based sentiment analysis. In: SemEval@ COLING, Citeseer, pp 838–842

Bu (2021) Asap. https://github.com/Meituan-Dianping/asap

Bu J, Ren L, Zheng S, et al (2021) Asap: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. In: Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: human language technologies, pp 2069–2079

Cai H, Tu Y, Zhou X, et al (2020) Aspect-category based sentiment analysis with hierarchical graph convolutional network. In: Proceedings of the 28th international conference on computational linguistics, pp 833–843

Cai H, Xia R, Yu J (2021a) Acos-dataset. https://github.com/NUSTM/ACOS

Cai H, Xia R, Yu J (2021b) Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, vol 1: Long Papers), pp 340–350

Cai H, Song N, Wang Z, et al (2023a) Memd-absa: a multi-element multi-domain dataset for aspect-based sentiment analysis. arXiv preprint arXiv:2306.16956

Cai H, Song N, Wang Z, et al (2023b) Memd-dataset. https://github.com/NUSTM/ACOS

Chen P, Sun Z, Bing L, et al (2017) Recurrent attention network on memory for aspect sentiment analysis. In: Proceedings of the 2017 conference on empirical methods in natural language processing. association for computational linguistics, Copenhagen, Denmark, pp 452–461. https://doi.org/10.18653/v1/D17-1047. https://aclanthology.org/D17-1047

Chen S, Liu J, Wang Y, et al (2020) Synchronous double-channel recurrent network for aspect-opinion pair extraction. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 6515–6524

Chen S, Wang Y, Liu J, et al (2021) Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In: Proceedings of the AAAI conference on artificial intelligence, pp 12666–12674

Chen Z, Qian T (2020a) Enhancing aspect term extraction with soft prototypes. In: Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp 2107–2117. https://doi.org/10.18653/v1/2020.emnlp-main.164. https://aclanthology.org/2020.emnlp-main.164

Chen Z, Qian T (2020b) Relation-aware collaborative learning for unified aspect-based sentiment analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp 3685–3694. https://doi.org/10.18653/v1/2020.acl-main.340. https://aclanthology.org/2020.acl-main.340

Chung HW, Hou L, Longpre S, et al (2022) Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416

Dai Z, Peng C, Chen H, et al (2020) A multi-task incremental learning framework with category name embedding for aspect-category sentiment analysis. In: Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing (EMNLP), pp 6955–6965

Eberts M, Ulges A (2019) Span-based joint entity and relation extraction with transformer pre-training. arXiv preprint arXiv:1909.07755

FAN (2019) Towe-dataset. https://github.com/NJUNLP/TOWE

Fan Z, Wu Z, Dai X, et al (2019) Target-oriented opinion words extraction with target-fused neural sequence labeling. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (Long and Short Papers), pp 2509–2518

Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: International conference on machine learning, PMLR, pp 1050–1059

Ganu G, Elhadad N, Marian A (2009) Beyond the stars: improving rating predictions using review text content. In: WebDB, pp 1–6

Gao L, Wang Y, Liu T, et al (2021) Question-driven span labeling model for aspect–opinion pair extraction. In: Proceedings of the AAAI conference on artificial intelligence, pp 12875–12883

Gao T, Fang J, Liu H, et al (2022) LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In: Proceedings of the 29th international conference on computational linguistics. International committee on computational linguistics, Gyeongju, Republic of Korea, pp 7002–7012. https://aclanthology.org/2022.coling-1.610

Ghadery E, Movahedi S, Faili H, et al (2019) Mncn: a multilingual ngram-based convolutional network for aspect category detection in online reviews. In: Proceedings of the AAAI conference on artificial intelligence, pp 6441–6448

Gou Z, Guo Q, Yang Y (2023) Mvp: multi-view prompting improves aspect sentiment tuple prediction. arXiv preprint arXiv:2305.12627

He R, Lee WS, Ng HT, et al (2017) An unsupervised neural attention model for aspect extraction. In: Proceedings of the 55th annual meeting of the association for computational linguistics, vol 1 (Long Papers). Association for computational linguistics, Vancouver, pp 388–397. https://doi.org/10.18653/v1/P17-1036. https://aclanthology.org/P17-1036

Hercig T, Brychcín T, Svoboda L, et al (2016) UWB at 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, pp 342–349. https://doi.org/10.18653/v1/S16-1055. https://aclanthology.org/S16-1055

Hoang CD, Dinh QV, Tran NH (2022) Aspect-category-opinion-sentiment extraction using generative transformer model. In: 2022 RIVF international conference on computing and communication technologies (RIVF), IEEE, pp 1–6

Hosseini-Asl E, Liu W, Xiong C (2022) A generative language model for few-shot aspect-based sentiment analysis. Findings of the Association for Computational Linguistics: NAACL 2022:770–787

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, pp 168–177

Hu M, Peng Y, Huang Z, et al (2019a) Open-domain targeted sentiment analysis via span-based extraction and classification. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, Italy, pp 537–546. https://doi.org/10.18653/v1/P19-1051. https://aclanthology.org/P19-1051

Hu M, Zhao S, Zhang L, et al (2019b) Can: constrained attention networks for multi-aspect sentiment analysis. In: Conference on empirical methods in natural language processing and international joint conference on natural language processing, association for computational linguistics

Hu M, Zhao S, Guo H, et al (2021) Multi-label few-shot learning for aspect category detection. In: Joint conference of the annual meeting of the association for computational linguistics and the international joint conference on natural language processing. Association for Computational Linguistics (ACL)

Hu M, Wu Y, Gao H, et al (2022) Improving aspect sentiment quad prediction via template-order data augmentation. arXiv preprint arXiv:2210.10291

Hu M, Bai Y, Wu Y, et al (2023) Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction. arXiv preprint arXiv:2306.00418

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, MA, pp 1035–1045. https://aclanthology.org/D10-1101

Jiang (2019) Mams-dataset. https://github.com/siat-nlp/MAMS-for-ABSA

Jiang Q, Chen L, Xu R, et al (2019) A challenge dataset and effective models for aspect-based sentiment analysis. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 6280–6285

Jin W, Ho HH, Srihari RK (2009) A novel lexicalized hmm-based learning framework for web opinion mining. In: Proceedings of the 26th annual international conference on machine learning. Citeseer

Kenton JDMWC, Toutanova LK (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp 4171–4186

Kobayashi N, Inui K, Matsumoto Y (2007) Extracting aspect-evaluation and aspect-of relations in opinion mining. In: Proceedings of the 2007 Joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 1065–1074

Lee SK, Kim JH (2023) Sener: Sentiment element named entity recognition for aspect-based sentiment analysis. ICASSP 2023–2023 IEEE international conference on acoustics: speech and signal processing (ICASSP). IEEE, pp 1–5

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, pp 7871–7880

Li B, Fei H, Wu Y, et al (2022a) Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. arXiv preprint arXiv:2211.05705

Li B, Fei H, Wu Y, et al (2022b) Diaasq-dataset. https://github.com/unikcc/DiaASQ

Li F, Han C, Huang M, et al (2010) Structure-aware review mining and summarization. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), pp 653–661

Li K, Chen C, Quan X, et al (2020) Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 7056–7066. https://doi.org/10.18653/v1/2020.acl-main.631. https://aclanthology.org/2020.acl-main.631

Li X, Lam W (2017) Deep multi-task learning for aspect term extraction with memory interaction. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2886–2892

Li X, Bing L, Lam W, et al (2018a) Transformation networks for target-oriented sentiment classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol. 1

(Long Papers). Association for Computational Linguistics, Melbourne, pp 946–956. https://doi.org/10.18653/v1/P18-1087. https://aclanthology.org/P18-1087

Li X, Bing L, Li P, et al (2018b) Aspect term extraction with history attention and selective transformation. In: Proceedings of the 27th international joint conference on artificial intelligence, pp 4194–4200

Li X, Bing L, Li P, et al (2019) A unified model for opinion target extraction and target sentiment prediction. In: Proceedings of the AAAI conference on artificial intelligence, pp 6714–6721

Liang B, Su H, Yin R, et al (2021) Beta distribution guided aspect-aware graph for aspect category sentiment analysis with affective knowledge. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 208–218. https://doi.org/10.18653/v1/2021.emnlp-main.19. https://aclanthology.org/2021.emnlp-main.19

Liao M, Li J, Zhang H, et al (2019) Coupling global and local context for unsupervised aspect extraction. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 4579–4589. https://doi.org/10.18653/v1/D19-1465. https://aclanthology.org/D19-1465

Liu B (2012) Sentiment analysis and opinion mining. Synth Lectures Human Lang Technol 5(1):1–167

Liu J, Zhang Y (2017) Attention modeling for targeted sentiment. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, Short Papers. Association for Computational Linguistics, Valencia, Spain, pp 572–577. https://aclanthology.org/E17-2091

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, pp 4406–4416. https://doi.org/10.18653/v1/2021.emnlp-main.361. https://aclanthology.org/2021.emnlp-main.361

Liu P, Joty S, Meng H (2015) Fine-grained opinion mining with recurrent neural networks and word embeddings. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1433–1443

Loukachevitch N, Blinov P, Kotelnikov E, et al (2015) Sentirueval: testing object-oriented sentiment analysis systems in Russian. In: Proceedings of international conference dialog, pp 3–13

Luo L, Ao X, Song Y, et al (2019) Unsupervised neural aspect extraction with sememes. In: IJCAI, pp 5123–5129

Ma D, Li S, Wu F, et al (2019) Exploring sequence-to-sequence learning in aspect term extraction. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, Italy, pp 3538–3547. https://doi.org/10.18653/v1/P19-1344. https://aclanthology.org/P19-1344

Ma Y, Peng H, Cambria E (2018) Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In: Proceedings of the AAAI conference on artificial intelligence

Mao Y, Shen Y, Yu C, et al (2021) A joint training dual-MRC framework for aspect based sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence, pp 13543–13551

Mao Y, Shen Y, Yang J et al (2022) Seq2path: generating sentiment tuples as paths of a tree. Findings of the Association for Computational Linguistics: ACL 2022:2215–2225

Mitchell M, Aguilar J, Wilson T, et al (2013) Open domain targeted sentiment. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Seattle, Washington, pp 1643–1654. https://aclanthology.org/D13-1171

Movahedi S, Ghadery E, Faili H, et al (2019) Aspect category detection via topic-attention network. arXiv preprint arXiv:1901.01183

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 (EMNLP-IJCNLP), pp 188–197

OpenAI (2023) Gpt-4 technical report. arXiv:2303.08774

Orbach M, Toledo-Ronen O, Spector A, et al (2021) Yaso: A targeted sentiment analysis evaluation dataset for open-domain reviews. In: Conference on empirical methods in natural language processing

Ouyang L, Wu J, Jiang X et al (2022) Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 35:27730–27744

Paolini G, Athiwaratkun B, Krone J, et al (2021) Structured prediction as translation between augmented natural languages. In: International conference on learning representations

Peng H, Xu L, Bing L, et al (2020) Knowing what, how and why: a near complete solution for aspect-based sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence, pp 8600–8607

Peper JJ, Wang L (2022) Generative aspect-based sentiment analysis with contrastive learning and expressive structure. arXiv preprint arXiv:2211.07743

Pontiki M (2014) Semeval-2014. https://alt.qcri.org/semeval2014/task4/

Pontiki M (2015) Semeval-2015. https://alt.qcri.org/semeval2015/task12/

Pontiki M (2016) Semeval-2016-dataset. https://alt.qcri.org/semeval2016/task5/

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, pp 27–35. https://doi.org/10.3115/v1/S14-2004. https://aclanthology.org/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), pp 486–495

Pontiki M, Galanis D, Papageorgiou H, et al (2016) Semeval-2016 task 5: aspect based sentiment analysis. In: ProWorkshop on semantic evaluation (SemEval-2016). Association for Computational Linguistics, pp 19–30

Popescu AM, Etzioni O (2007) Extracting product features and opinions from reviews. Natural language processing and text mining pp 9–28

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

Radford A, Narasimhan K, Salimans T, et al (2018) Improving language understanding by generative pre-training

Raffel C, Shazeer N, Roberts A, et al (2019) Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR abs/1910.10683. arXiv:1910.10683

Raffel C, Shazeer N, Roberts A et al (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(1):5485–5551

Ruder S, Ghaffari P, Breslin JG (2016) A hierarchical model of reviews for aspect-based sentiment analysis. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 999–1005

Saeidi (2016) Sentihood. https://huggingface.co/datasets/bhavnicksm/sentihood

Saeidi M, Bouchard G, Liakata M, et al (2016) Sentihood: targeted aspect based sentiment analysis dataset for urban neighbourhoods. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, pp 1546–1556

Schmitt M, Steinheber S, Schreiber K, et al (2018) Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for Computational Linguistics, Brussels, Belgium, pp 1109–1114. https://doi.org/10.18653/v1/D18-1139. https://aclanthology.org/D18-1139

Senti-WordNet (2006) Senti-wordnet. https://github.com/aesuli/SentiWordNet

SenticNet (2010) Senticnet. https://github.com/senticnet

Shi T, Li L, Wang P, et al (2021) A simple and effective self-supervised contrastive learning framework for aspect detection. In: Proceedings of the AAAI conference on artificial intelligence, pp 13815–13824

Shu L, Xu H, Liu B, et al (2022) Zero-shot aspect-based sentiment analysis. arXiv preprint arXiv:2202.01924

Song L, Xin C, Lai S et al (2022) Casa: conversational aspect sentiment analysis for dialogue understanding. J Artif Intell Res 73:511–533

Soni PK, Rambola R (2022) A survey on implicit aspect detection for sentiment analysis: terminology, issues, and scope. IEEE Access 10:63932–63957

Sun C, Huang L, Qiu X (2019) Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, pp 380–385. https://doi.org/10.18653/v1/N19-1035, https://aclanthology.org/N19-1035

Tang D, Qin B, Feng X, et al (2016a) Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee, Osaka, pp 3298–3307. https://aclanthology.org/C16-1311

Tang D, Qin B, Liu T (2016b) Aspect level sentiment classification with deep memory network. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, Austin, pp 214–224. https://doi.org/10.18653/v1/D16-1021. https://aclanthology.org/D16-1021

Tay Y, Tuan LA, Hui SC (2018) Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence

Tay Y, Dehghani M, Tran VQ, et al (2023) Ul2: unifying language learning paradigms. In: The eleventh international conference on learning representations

Tulkens S, van Cranenburgh A (2020) Embarrassingly simple unsupervised aspect extraction. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 3182–3187

Varia S, Wang S, Halder K, et al (2022) Instruction tuning for few-shot aspect-based sentiment analysis. arXiv preprint arXiv:2210.06629

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30

Wan H, Yang Y, Du J, et al (2020) Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence, pp 9122–9129

Wang Q, Wen Z, Zhao Q, et al (2021) Progressive self-training with discriminator for aspect term extraction. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 257–268. https://doi.org/10.18653/v1/2021.emnlp-main.23. https://aclanthology.org/2021.emnlp-main.23

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, vol 1 (Long Papers), pp 2171–2181

Wang W, Pan SJ, Dahlmeier D, et al (2016a) Recursive neural conditional random fields for aspect-based sentiment analysis. arXiv preprint arXiv:1603.06679

Wang W, Pan SJ, Dahlmeier D, et al (2017) Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: Proceedings of the AAAI conference on artificial intelligence

Wang Y, Huang M, Zhu X, et al (2016b) Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, Austin, pp 606–615. https://doi.org/10.18653/v1/D16-1058. https://aclanthology.org/D16-1058

Wang Z, Xia R, Yu J (2022) Unifiedabsa: a unified ABSA framework based on multi-task instruction tuning. arXiv preprint arXiv:2211.10986

Wang Z, Xie Q, Ding Z, et al (2023) Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv:2304.04339

Wei J, Bosma MP, Zhao V, et al (2022a) Finetuned language models are zero-shot learners

Wei J, Wang X, Schuurmans D, et al (2022b) Chain-of-thought prompting elicits reasoning in large language models. In: Advances in neural information processing systems

WordNet (2010) Wordnet. https://wordnet.princeton.edu/download/current-version

WordNet-Affect (2004) Wordnet-affect. https://wndomains.fbk.eu/wnaffect.html

Wu C, Xiong Q, Yi H et al (2021) Multiple-element joint detection for aspect-based sentiment analysis. Knowl-Based Syst 223:107073

Wu M, Wang W, Pan SJ (2020a) Deep weighted maxsat for aspect-based opinion extraction. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 5618–5628

Wu S, Fei H, Ren Y, et al (2021b) Learn from Syntax: improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. IJCAI international joint conference on artificial intelligence, pp 3957–3963. https://doi.org/10.24963/ijcai.2021/545. arXiv:2105.02520

Wu Z, Ying C, Zhao F et al (2020) Grid tagging scheme for aspect-oriented fine-grained opinion extraction. Findings of the Association for Computational Linguistics: EMNLP 2020:2576–2585

Xing (2020) Arts-dataset. https://github.com/zhijing-jin/ARTS_TestSet

Xing B, Liao L, Song D, et al (2019) Earlier attention? Aspect-aware LSTM for aspect-based sentiment analysis. arXiv preprint arXiv:1905.07719

Xing X, Jin Z, Jin D, et al (2020) Tasty burgers, soggy fries: probing aspect robustness in aspect-based sentiment analysis. arXiv preprint arXiv:2009.07964

Xiong H, Yan Z, Wu C, et al (2023) Bart-based contrastive and retrospective network for aspect-category-opinion-sentiment quadruple extraction. Int J Mach Learn Cybern:1–13

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, vol 2 (Short Papers), pp 592–598

Xu L, Li H, Lu W, et al (2020a) Aste-data-v1,v2. https://github.com/xuuuluuu/SemEval-Triplet-data/tree/master

Xu L, Li H, Lu W, et al (2020b) Position-aware tagging for aspect sentiment triplet extraction. arXiv preprint arXiv:2010.02609

Xu L, Chia YK, Bing L (2021) Learning span-level interactions for aspect sentiment triplet extraction. arXiv preprint arXiv:2107.12214

Xue L, Constant N, Roberts A, et al (2021) mT5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics, pp 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41. https://aclanthology.org/2021.naacl-main.41

Xue W, Li T (2018) Aspect based sentiment analysis with gated convolutional networks. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol 1 (Long Papers). Association for Computational Linguistics, Melbourne, pp 2514–2523. https://doi.org/10.18653/v1/P18-1234. https://aclanthology.org/P18-1234

Yan H, Dai J, Qiu X, et al (2021) A unified generative framework for aspect-based sentiment analysis. arXiv preprint arXiv:2106.04300

Yang Y, Li K, Quan X, et al (2020) Constituency lattice encoding for aspect term extraction. In: Proceedings of the 28th international conference on computational linguistics. International Committee on Computational Linguistics, Barcelona, pp 844–855. https://doi.org/10.18653/v1/2020.coling-main.73. https://aclanthology.org/2020.coling-main.73

Yin Y, Wei F, Dong L, et al (2016) Unsupervised word and dependency path embeddings for aspect term extraction. arXiv preprint arXiv:1605.07843

Yin Y, Wang C, Zhang M (2020) PoD: Positional dependency-based word embedding for aspect term extraction. In: Proceedings of the 28th international conference on computational linguistics. International Committee on Computational Linguistics, Barcelona, pp 1714–1719. https://doi.org/10.18653/v1/2020.coling-main.150. https://aclanthology.org/2020.coling-main.150

Yu G, Li J, Luo L, et al (2021) Self question-answering: Aspect-based sentiment analysis by role flipped machine reading comprehension. In: Findings of the association for computational linguistics: EMNLP 2021. Association for Computational Linguistics, Punta Cana, pp 1331–1342. https://doi.org/10.18653/v1/2021.findings-emnlp.115. https://aclanthology.org/2021.findings-emnlp.115

Yu J, Jiang J, Xia R (2018) Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. IEEE/ACM Trans Audio Speech Lang Process 27(1):168–177

Zhang L, Wang S, Liu B (2018) Deep learning for sentiment analysis: a survey. Wiley Interdisc Rev 8(4):e1253

Zhang M, Zhang Y, Vo DT (2015) Neural networks for open domain targeted sentiment. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 612–621. https://doi.org/10.18653/v1/D15-1073. https://aclanthology.org/D15-1073

Zhang M, Zhang Y, Vo DT (2016) Gated neural networks for targeted sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence

Zhang W, Deng Y, Li X, et al (2021a) Aspect sentiment quad prediction as paraphrase generation. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 9209–9219

Zhang W, Deng Y, Li X, et al (2021b) Quad-dataset. https://github.com/IsakZhang/ABSA-QUAD

Zhang W, He R, Peng H, et al (2021c) Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 9220–9230

Zhang W, Li X, Deng Y, et al (2021d) Towards generative aspect-based sentiment analysis. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, vol 2 (Short Papers), pp 504–510

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

Zhang W, Deng Y, Liu B, et al (2023) Sentiment analysis in the era of large language models: a reality check. arXiv preprint arXiv:2305.15005

Zhao H, Huang L, Zhang R, et al (2020) Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 3239–3248

Zhou J, Huang JX, Chen Q et al (2019) Deep learning for aspect-level sentiment classification: survey, vision, and challenges. IEEE Access 7:78454–78483

Zhou J, Yang H, He Y, et al (2023) Asqp-dataset. https://github.com/NUSTM/ACOS

Zhou X, Wan X, Xiao J (2015) Representation learning for aspect category detection in online reviews. In: Proceedings of the AAAI conference on artificial intelligence

Zhu L, Xu M, Bao Y et al (2022) Deep learning for aspect-based sentiment analysis: a review. PeerJ Comput Sci 8:e1044

Zhuang L, Jing F, Zhu XY (2006) Movie review mining and summarization. In: Proceedings of the 15th ACM international conference on Information and knowledge management, pp 43–50