Opinion Units:
Concise and Contextualized Representations for Aspect-Based Sentiment Analysis

Anonymous ACL submission

Abstract

We introduce *opinion units*, a novel approach to Aspect-Based Sentiment Analysis (ABSA) that extends traditional aspect-sentiment pairs by including substantiating excerpts derived through hybrid abstractive-extractive summarisation. This reduces the information loss inherent in traditional ABSA methods, and the structured format facilitates downstream processing tasks. Experiments on benchmark datasets for ABSA demonstrate that large language models (LLMs) can accurately extract opinion units using a few-shot approach. The main types of errors are overlooking aspects in the text, and characterising objective statements as opinions. The method eliminates the need for labelled data and allows the LLM to dynamically define aspect types. Additionally, we present a case study on similarity search for opinions in academic datasets and public review data. Our results indicate that searches based on opinion units are more successful than those using traditional data-segmentation strategies, demonstrating robustness across datasets and embeddings.

1 Introduction

We propose *opinion units* as a representation for subjective viewpoints in text. An opinion unit consists of (i) an aspect such as price, quality, or location, (ii) an excerpt, which may be lightly summarised or paraphrased, that contextualises the opinion, (iii) and a sentiment such as positive, negative or neutral. The structured nature of opinion units makes them suitable for applications requiring fine-grained *aspect-based sentiment analysis* (ABSA), such as the mining and retrieval of opinions. ABSA goes beyond the surface level of traditional sentiment analysis. Instead of assigning a sentiment to an entire text, ABSA identifies opinions expressed about particular features of, for instance, a product, service or event. This multi-faceted analysis provides valuable insights for those seeking to understanding public opinion on a particular topic. For example, for retailers, ABSA of customer reviews or interactions can suggest areas for improvement, personalise marketing strategies, and gauge overall customer satisfaction.

Previous work on ABSA has focused on classifying reviews into predefined aspect- and sentiment categories (Zhang et al., 2022). However, recent studies improve on the approach by extracting aspect- and sentiment keywords using sequence-to-sequence models (Zhang et al., 2022; Gao et al., 2021). This is a step forward as the category types are no longer set in advance, but they are still limited to the terms used in the analysed text.

In this article, we explore how opinion units can be extracted from subjective commentary, specifically customer reviews, by large language models (LLMs). The models are prompted in a way that allows them to dynamically generate aspect categories not explicitly mentioned in the text, and to choose and paraphrase motivating text excerpts that retain only the most relevant information. An example of how opinion units are formed is given in Figure 1 and a formal definition is provided in Section 3. The main benefit opinion units is that they provide a structured representation of the opinions expressed in a text, while retaining much of the nuance through the supportive excerpt.

Language models excel at many of the tasks involved in the generation of opinion units, including information extraction, text summarization, entity recognition, and sentiment analysis. Previous work has successfully applied LLMs to extract *propositions*, that is, atomic factual statements, to facilitate question answering in a dense retrieval setting where both the query and documents are transformed into embeddings (Chen et al., 2023). We transfer this method to the ABSA domain, demonstrating that LLMs can effectively identify opinion aspects, extract concise snippets of text expressing the opinion, and accurately classify the sentiment of the excerpt.

Last Sunday we went to brunch and I had a muffin. It was amazing! We loved our waiter Stephanie she was so friendly however the service could have been a little quicker. But on the whole, we had a great time!

- Muffin: I had a muffin. It was amazing. {positive}
- Staff friendliness: We loved our waiter Stephanie, she was so friendly. {positive}
- Service speed: The service could have been a little quicker. {negative}
- Overall brunch experience: On the whole, we had a great time. {positive}

Figure 1: Four opinion units extracted from an example review. Each unit represents an opinion expressed in the text and consists of an aspect label, an excerpt from the text, and a sentiment label. The colour purple indicates aspects, and orange indicates sentiment terms.

An important advantage of extracting opinion units with LLMs stems from the few-shot approach. Unlike traditional ABSA methods that often rely on pre-defined categories or require labeled training data, LLMs can extract opinion units without such constraints. This opens doors for broader application across diverse domains and allows for more efficient and scalable analysis. Another advantage is the atomic nature of opinion units, each of which contains only one opinion about a single aspect. In contrast, in "raw" review texts, multiple aspects may be discussed in the same sentence, or a single aspect may be discussed over multiple sentences. This makes applications like opinion retrieval and opinion mining challenging. Keyword-based extraction approaches are an alternative, but these invariably lead to information loss since they fail to capture nuances.

In the following sections, we first demonstrate the ability of LLMs to produce high-quality opinion units through an evaluation on benchmark datasets frequently used in ABSA research. Furthermore, we categorize the errors produced by the LLMs, where missing aspects and the conflation of objective statements with opinions turn out to be the most serious sources of error. Finally, we demonstrate the effectiveness of opinion units in dense similarity search, where words are represented by embeddings. In particular, we show that opinion units outperform competing chunking strategies such as sentence and passage chunking on real-world review datasets, as well as on ABSA benchmark datasets. These positive results suggest that opinion units are potentially useful also for dense retrieval, retrieval-augmented generation and clustering applications. For example, in topic modeling, opinion units can help reveal which topics customers focus on in reviews, and how these correlate with overall ratings and reactions.

The experiments conducted in this article serve to answer the following research questions:

RQ1. To what extent can LLMs generate opinion units?

RQ2. What are the types and frequencies of errors that LLMs make when generating opinion units?

RQ3. How does the performance of opinion units in dense similarity search for opinions compare to other data-segmentation strategies like passage- and sentence chunking?

2 Related Work

This section recalls related work on ABSA, summarisation, and information retrieval.

2.1 Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis is a specialized area within the broader field of sentiment analysis. Its focus is on identifying and extracting sentiment in relation to specific aspects in a given text (Zhang et al., 2022). The analysis typically involves establishing some or all of the following sentiment elements: The aspect category c which is the general concept to which the sentiment pertains; the aspect term a which is the entity being referred to; the opinion term o which conveys the aspect sentiment; and the sentiment polarity p which is the valance of the emotion expressed (Zhang et al., 2022). Given the sentence "the tiramisu was amazing", these elements could be mapped accordingly: $c =$ 'dessert', $a =$ 'tiramisu', $o =$ 'amazing', and $p =$ 'positive'. We note that the construction of opinion units involves all four sentiment elements: The opinion label corresponds to the aspect category, although in our case it is generated on the fly by the LLM rather than chosen from a set of predefined categories. The excerpt in opinion units includes both aspect and opinion terms. Finally, each opinion unit includes a sentiment polarity.

Earlier works concentrated on solutions for isolated sentiment elements, such as aspect term extraction (Liu et al., 2015; Li and Lam, 2017) or aspect category detection (Zhou et al., 2015; Luo et al., 2019). Later studies extract several factors at once, capturing both the opinion aspect and expression (Peng et al., 2020; Gao et al., 2021). Pipeline methods offer a modular approach, decomposing the overall task into sequential sub-tasks (Peng et al., 2020). While this strategy can leverage existing solutions and achieve good performance on each sub-task, they are prone to error propagation where mistakes made in earlier stages cascade and negatively impact the final outcome. (Peng et al., 2020; Chen et al., 2020).

We are now seeing significant advancements in the implementation of multifaceted analysis tasks. A salient example is sequence-to-sequence models which output the result of the analysis as a natural-language statement. This approach has been shown to outperform classification methods and exhibits particular strengths in scenarios with limited training data thanks to few-shot and zero-shot learning (Ma et al., 2019; Zhang et al., 2022).

A comprehensive understanding of sentiment is achieved through the prediction of all four sentiment elements, a process known as Aspect Quad Prediction. The task is deemed challenging, with the primary hurdle being the accurate pairing of various sentiment elements (Zhang et al., 2022). Recent works, employing pre-trained language models, achieve about 60-70% F1 scores on benchmark datasets (Zhang et al., 2021a,b).

In a recent study, Zhang et al. (2023) compare several LLMs, including ChatGPT, against smaller but fine-tuned models on ABSA. They find that the LLMs struggle with fine-grained sentiment analysis and are out-performed by the smaller models. Since method introduced here encourages the LLMs to produce supporting excerpts that add detail, they could help mitigate this problem.

2.2 Summarisation

Opinion mining benefits from both extractive and abstractive summarization (Anand Babu and Badugu, 2023). The former produces a summarisation by concatenating informative segments from the source document, whereas the latter generates a summary based on the semantics of the source, which at a superficial level can be very different from the original text. Extractive summarisation relevant because it provides evidence in the source material for the generated opinion units (Priya and Umamaheswari, 2020), but to keep the excerpts short and self-contained, a degree of abstractive summarisation is necessary.

Yang et al. (2019) evaluate ChatGPT on abstractive summarization. Even with a zero-shot approach, the model performs on par with smaller LMs fine-tuned for the task. This stands in contrast to the case for aspect-based sentiment analysis discussed above, where the smaller, fine-tuned models were more successful (Zhang et al., 2023). A related task is key-point extraction (Bar-Haim et al., 2020a,b, 2021), where the objective is to extract salient viewpoints from a text. Also here LLM-enabled aspect-based approaches have been successfully applied (Tang et al., 2024) and reduce the number of partially overlapping key points.

2.3 Information Retrieval

Dense retrievers are a common type of modern retrieval systems where a dual-encoder architecture transforms documents and queries into dense embeddings for similarity comparison (Ni et al., 2022). These similarity functions, also used for embedding-based clustering (Chandrasekaran and Mago, 2021), have limitations in understanding complex semantics and can be misled by irrelevant information (Chen et al., 2023). Chen et al. (2023) explored using propositions, factual statements distilled from text using LLMs (GPT-4), as retrieval units for Wikipedia passage retrieval and retrieval-augmented LLM question answering. Using propositions to segment and index the retrieval corpus outperformed traditional methods like sentence or fixed-length passage chunking. In their context of fact retrieval, each proposition represented a single atomic fact with relevant context, phrased concisely in natural language (Chen et al., 2023). The authors describe corpus segmenting using propositions as an orthogonal strategy that can be used in conjunction with other methods for improving dense retrieval such as supervised retrievers (Chen et al., 2023), data augmentation (Wang et al., 2022), hybrid sparse-dense retrieval (Luan et al., 2021) or mixed-strategy retrieval (Ma et al., 2023).

Compared to traditional chunking methods, propositions offers a high information density with complete context. Comparatively, passage chunking constitutes a coarse information unit, often containing unrelated and multiple aspects. This lack of conciseness can distract downstream applications such as retrieval relying on similarity comparison (Yu et al., 2023). Sentence chunking provides more fine-grained information and is appropriate when each aspect is treated in a separate sentence. However, sentences can include multiple aspect and lack necessary context when dependencies span multiple sentences (Yang et al., 2019).

3 Opinion units

As explained in Section 1, an opinion unit is composed of three elements: i) an aspect label, ii) a text excerpt substantiating a subjective viewpoint on the aspect, and iii) a sentiment label that quantifies the sentiment expressed according to some set scale. Additionally, we outline four key principles that together characterize opinion units. These principles are inspired by the factual propositions of Chen et al. (2023) (see Section 2.3), but are tailored for the ABSA domain. They are as follows:

Atomicity. Every opinion unit should represent exactly one opinion (i.e., aspect-sentiment pair).

Injectivity. No two opinion units should represent the same opinion.

Completeness. Collectively, the set of extracted opinion units should encompass all the opinions expressed in the text.

Contextuality. The excerpt associated with each opinion unit should explicity name the target aspect and give sufficient contextual information to motivate the inferred sentiment. If needed, the excerpt may refer to other aspects or sentiments.

When used for data segmentation in applications such as customer-satisfaction surveys or brand studies, LLM-enabled generation of opinion unit overcomes a number of challenges (see Figure 2). First of all, opinion units can handle sentences and passages with multiple opinions, and as well as opinions spanning multiple sentences. In these cases, traditional segmentation strategies such as sentence and passage chunking (which we benchmark against in Section 4), create irrelevant or uninformative chunks. Opinion units, in contrast, isolate opinions and adapt the excerpt length to match the coverage of the aspect in the source text.

Another benefit is that the aspect label generated by the LLM facilitate the clustering of opinion units that refer to the same concept, even though the terms and wording used in the source text may vary. Similarly, the sentiment label can be used to filter opinion units based on sentiment polarity. This approach leverages the LLM's high performance in sentiment analysis (Zhang et al., 2023) while ensuring efficient inference (see Section 5.2). Incorporating other metadata than sentiment, or a finer sentiment scale would also possible and could be beneficial for specific applications. For chunking strategies like passage- or sentence chunking, the presence of multiple opinions or non-opinionated text within a single chunk can make sentiment labeling less straightforward and precise.

Finally, the LLM can be prompted to disregard sections of the source text that do not express opinions, which is valuable because also subjectively written texts can have strictly objective passages. For example, in the context of restaurant reviews, as statement such as "I went with my two friends and sat in a corner booth" may not have much bearing on the writer's assessment of the food. In passage- or sentiment chunking, these non-opinionated texts cannot be avoided and add noise to the analysis process.

4 Method

The experimental evaluation of opinion units comprises two parts. First, we assess the ability of LLMs to generate well-formed opinion units using the SEMEVAL ABSA benchmark datasets. Second, we conduct a case study on opinion retrieval, where data segmentation based on opinion units is compared to traditional chunking strategies.

4.1 Generation of Opinion Units

We generate opinion units using OpenAI's GPT-3.5 in a few-shot approach. GPT-3.5 was selected for its balance of performance and cost-efficiency; although GPT-4 might offer superior results, the simpler model is sufficient to show the strengths of the approach. The prompt template for generating opinion units is detailed in Appendix A. This template instructs the LLM to perform ABSA, extracting the three components of an opinion unit. An example review with opinion units formatted as a bullet list is provided in the template. This example is designed to address issues discussed in Section 3, such as non-opinionated text and opinions spanning multiple sentences. If the generated opinion units deviate from the format defined in the prompt template, the generation is rerun (this happens approximately 5% of the time).

For hyperparameters, we opt for a relatively high temperature of 1.3. This value is found effective in distinguishing separate aspects in texts and providing insightful opinion labels.

4.2 Opinion Unit Evaluation

To assess the correctness of the generated opinion units, we conduct evaluations using the benchmark datasets SEMEVAL Rest15 and Rest16, which consist of restaurant-review sentences (Pontiki et al.,

Challenge	Example of review and extracted opinion units	Benefits of opinion units
Passages expressing multiple opinions	*The food is great but the drinks sucked.* ➤ Food: The food is great {positive} ➤ Drinks: The drinks sucked {negative}	Unlike passage and sentence chunking, opinion units separate aspects which avoids noisy and non-concise segments.
Opinions spanning multiple sentences	*We had margaritas. They tasted absolutely wonderful!* ➤ Margaritas: We had margaritas. They tasted absolutely wonderful. {positive}	Opinion units provide full context spanning several sentence. Sentence chunking provides incomplete context and passage chunking could be incomplete or include noise, depending on the length of the relevant passage.
Lack of contextual information	*The restroom was not ADA compliant.* ➤ Disabled persons accessibility: The restroom was not ADA compliant. {negative }	The opinion label generated by the LLM provides helpful context for later processing steps. In the example, ADA stands for Americans with Disabilities Act which ensures equal access for people with disabilities.
Insufficient sentiment understanding and filtering	*The portion size was perfect... for an ant.* ➤ Portion size: The portion size was perfect... for an ant. {negative}	LLMs are more adept at understanding sentiments or irony compared to word embeddings at inference time. Opinion units can be filtered by sentiment.

Figure 2: Examples and summary of four challenges when segmenting opinionated texts for downstream applications where opinion units provide advantages compared to passage- and sentence chunking.

2016). We compare the generated opinion units against the annotations for the Aspect Sentiment Triplet Extraction (ASTE) task, provided by Zhang et al. (2021a). These annotations include the correct opinion and aspect terms as well as sentiment polarities. For example, a review sentence "The fish was good however the service was terribly slow and it took forever to get our food." would correspond to the ASTE labels: (fish, good, positive) and (service, slow, negative), and the opinion units: ("Fish: the fish was good", positive) and ("Service speed: the service was terribly slow and it took forever to get our food.", negative).

Since the tasks of opinion-unit generation and ASTE serve different ends—the latter involves extracting keywords and the former generating excerpts—we formulate the following, adapted, evaluation criteria:

1. The generated opinion units' aspect and sentiment labels should correspond to the ASTE aspect-sentiment pairs (Zhang et al., 2021a).
2. Each opinion unit's excerpt:
   (a) Should be consistent with the unit's aspect and sentiment labels.
   (b) Should not include other aspect or sentiment terms, except as needed for motivation.[1]

Condition 1 tests for *injectivity* and *completeness*, while conditions 2a and 2b correspond to *contextuality* and *atomicity*, respectively (Sec. 3).

For our evaluations, we use the test sets of Res15 and Res16, selecting only the sentences that, according to (Zhang et al., 2021a), include multiple different aspects. Extracting an opinion unit excerpt from a single-aspect sentence is a trivial task, so to make the most of our annotation efforts, we focus on the more complex cases. In total, the evaluation consists of 239 review sentences, yielding 591 opinion units.

4.3 Case Study: Opinion Retrieval

Whereas the experiment just described tests the viability of LLM-extracted opinion units, the following case study evaluates the method's usefulness.

Retrieval Tasks. We provide 50 similarity search tasks for restaurant reviews. The goal of the retrieval system is to return reviews that contain opinions that are similar to the opinion provided as the query. The 50 tasks are broken down into 10 general tasks and 40 detailed tasks. General tasks correspond to common and overarching opinions found in restaurant reviews, such as overall experience, value for money, and staff friendliness. For instance, Task 1 has the query: "All in all, we had a great time." For returned reviews to be considered correct, they must express satisfaction with the overall experience. Detailed tasks focus on specific aspects mentioned in fewer reviews. For example, the query for Task 24 is: "The food was cold when we received it." Returned reviews must detail negative experiences related to receiving cold food at the restaurant. Out of the 50 tasks, half entail a positive sentiment, while the other half reflect a negative sentiment. The full list of review tasks, including queries and task descriptions is found in supplementary material to this paper. Example tasks are provided in Appendix B.

[1] For example, the opinion unit (lamb, "the steak was good, and so was the lamb", positive) involves the aspect 'steak' which does not have any explanatory value for why the experience of the lamb was positive, and the mentioning of which could be avoided by the LLM through paraphrasing.

The returned reviews were assessed by a team of 4 evaluators who were blind to the chunking strategies used. Additionally, the reviews were presented in a randomized order to eliminate a potential source of bias.

Evaluation Groups. We compare dense retrieval based on opinion units to the conventional approaches of passage- and sentence chunking (Chen et al., 2023). In sentence chunking, each sentence serves as a retrievable unit, whereas in passage chunking, we employ Langchain's `RecursiveCharacterTextSplitter` with parameters `size=200` and `overlap=20`. The retrievable units in passage chunking are on average longer compared to sentence chunking and opinion units, as detailed in Table 1. In addition to standard opinion units, we also use opinion units with sentiment filtering as a retrieval unit (denoted *opinion + sf* in results tables). In this approach, only opinion units labeled with the specific sentiment demanded by the task are considered by the retrieval system. For each retrieval strategy, we extract 20 unique reviews. Precision @5, 10, and 20 are used to evaluate the results.

The primary dataset used for evaluating the opinion retrieval case study is the Yelp dataset (Yelp, 2015), which contains millions of authentic reviews. We refine this dataset to include only restaurant reviews, extracting the first 20 000 reviews of restaurants located in California to serve as our retrieval corpus. As a secondary dataset, we use a concatenation of the SEMEVAL Res15 train and test datasets and the Res16 test dataset (excluding the Res16 train dataset, as it duplicates the Res15 train and test reviews). This dataset is considerably smaller than the Yelp dataset, containing 2 280 reviews. On average, each review spans approximately 14.49 words and 1.75 opinion units. In contrast, the average Yelp review contains 92.7 words and 5.5 opinion units, with the 95th percentile extending to 257 words and 10.0 opinion units. The 50 retrieval tasks, are designed to become increasingly specific. Due to the limited scope of the SEMEVAL dataset, we omit tasks 31-50 for this dataset. For similar reasons we only report Precision @5 and @10 as our evaluation metrics.

To ascertain the robustness of retrieval results we perform the evaluation using two different embedding models from the sentence-transformers framework: `all-mpnet-base-v2` and `all-MiniLM-L6-v2` (Transformers, 2024). Both embedding models are optimized for general tasks, including sentiment analysis, however `all-mpnet-base-v2` is a considerably larger model (80MB vs. 420MB). For our dense retrieval implementation, we used the Faiss package and its function `similarity_search` (Langchain, 2024).

	Yelp		SEMEVAL	
	Units	Avg. Words	Units	Avg. Words
Passage	69 890	28.2	2 155	14.3
Sentence	144 039	12.9	2 280	13.5
Opinion Units	110 245	14.8	3 716	10.1

Table 1: The number of units and average number of words per unit for each combination of dataset and chunking strategy.

5 Results and Discussion

The generated opinion units as well as evaluation annotations used in Section 5.1 are available as supplementary material.

5.1 Opinion Unit Evaluation

We evaluate the opinion units generated for the Rest15 and Rest16 test sets with respect to the criteria listed in Section 4.1. We find that 540 out of 591 opinion units (90.9%) rate as fully correct, exhibiting 55 errors (an opinion unit can include multiple errors). The high level of performance is promising for downstream tasks. It is important to note that comparisons with other ASTE and ASQP benchmarks (Zhang et al., 2021a) should be avoided, as they are different tasks with different evaluation criteria. The opinion unit evaluation is excerpt-based, allowing for multiple small variations to be correct, whereas ASTE & ASQP-tasks are strict keyword extraction benchmarks.

Furthermore, we categorize the errors to understand the types of problems GPT-3.5 encounters when generating opinion units. The frequency of these errors is presented in Figure 3. The error categories are as follows:

Atomicity error. An opinion unit lacks *atomicity*, representing or unnecessarily providing context for multiple opinions.

Injectivity error. Collectively, opinion units are redundant, lacking *injectivity*.

Missing aspect. Collectively, the opinion units lack *completeness*, meaning that not all opinions in the review were captured.

Missing context. An opinion unit is not *contextualized*, i.e., does not provide sufficient contextual information to motivate the inferred sentiment.

Non-opinon. A non-opinionated excerpt from the text is incorrectly classified as an opinion.

Sentiment error. The sentiment label of an opinion unit is incorrect.

Aspect-term error. The excerpt for a particular aspect corresponds to another aspect than the opinion label.

Hallucination. The LLM invents aspects or excerpts that are not part of the review.

Sentiment errors are the most common type of error observed. However, 14 out 18 errors are failures to recognise neutrality in language. These instances are often ambiguous; for example, the phrase “the food was nothing more than average” might be interpreted as either neutral or negative.

More serious are errors such as missing aspects or categorizing non-opinion statements like “we went to sit at the bar” as opinions. These issues reflect the central challenge in ABSA which is isolating opinion pairs from text (Zhang et al., 2022), and several of the sentences in the benchmark datasets are designed to trigger this type of error. Conversely, longer reviews can introduce more errors due to context-length strain and longer dependencies, leading to missing aspects or inclusion of non-opinionated text.

In our test set, the model did not invent aspects or excerpts. However, in the expanded datasets used for opinion retrieval, hallucinations were observed in a few cases, specifically adding an “overall experience” label with an invented excerpt corresponding to the overall sentiment of the review. This issue stems from the prompt template used to generate opinion units and the handling of “overall experience” (see Appendix A). Nonetheless, hallucination does not seem to be a significant problem in opinion-unit generation.

Potential remedies to mitigate errors in opinion unit generation include larger and more sophisticated language models, fine-tuning models, or increasingly relying on abstractive rather than extractive summarization. While abstraction could improve the isolation of opinion aspects, it may reduce the ability to point to specific text segments that support the extracted opinions, which is important for transparency and trust.

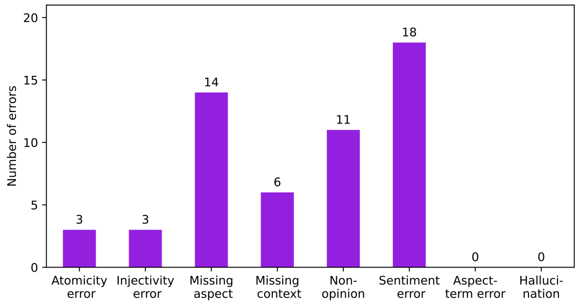


Figure 3: Frequency of error types in opinion units based on the subset of test-data from Rest15 and Rest16 involving more than one aspect.

5.2 Case Study: Opinion Retrieval

In our case study we compare the performance of alternative chunking strategies on 50 different retrieval tasks, each of which consists in retrieving reviews which include some specific opinions (see Section 4.3). The retrieval results, presented in Table 2, delineate the performance across datasets (Yelp and SEMEVAL-Rest) and the two different word embedding models. The larger embedding model, `all-mpnet-base-v2`, leads to better results than the smaller `all-MiniLM-L6-v2`.

Consistently, across all experimental conditions, opinion units outperform passage- and sentence chunking, with sentence chunking being most competitive. This implies that opinions in reviews are often expressed within a single sentence. The results show the benefit of the opinion units ability to provide a concise and structured representation in opinion retrieval. Behind the increased retrieval precision lies the ability to solves the challenges highlighted in Section 3 such as passages with intertwined opinions and opinion spanning multiple sentences detailed.

It is worth noting the large performance gap between standard opinion units and opinion units with sentiment filtering (opinion unit + sf). In our evaluation tasks, the objective is to retrieve reviews with certain combinations of aspects and sentiments. Filtering by the LLM-generated sentiment labels thus contributes towards an important subgoal. The resulting gains in precision also highlights the limitations of word embeddings in sentiment comprehension (Yu et al., 2017), where words with similar vector representations can exhibit contrasting sentiment polarities, e.g., “friendly” and “unfriendly”. Refining word embeddings to better reflect both semantics and sentiment is therefore an important avenue for future work (Yu et al., 2017).

Table 2: Precision results for different combinations of dataset and embedding model

(a) Yelp Restaurant, `all-mpnet-base-v2`

Tasks	**Chunking strategy**	**Precision**		
		@5	**@10**	**@20**
All (Task 1-50)	Passage	61.6	54.4	56.0
	Sentence	76.4	70.6	63.3
	Opinion unit	81.6	74.4	69.5
	Opinion unit + sf	88.0	82.2	77.9
General (Task 1-10)	Passage	78.0	76.0	70.5
	Sentence	90.0	86.0	81.5
	Opinion unit	94.0	90.0	86.0
	Opinion unit + sf	96.0	92.0	89.5
Detailed (Task 11-50)	Passage	57.7	54.0	52.4
	Sentence	73.0	66.8	58.8
	Opinion unit	78.5	70.5	65.4
	Opinion unit + sf	86.0	79.8	75.0

(b) Yelp Restaurant, `all-MiniLM-L6-v2`

Tasks	**Chunking strategy**	**Precision**		
		@5	**@10**	**@20**
All (Task 1-50)	Passage	54.4	53.6	49.3
	Sentence	65.6	62.8	54.6
	Opinion unit	70.8	65.0	61.1
	Opinion unit + sf	82.0	80.4	76.1
General (Task 1-10)	Passage	68.0	68.0	63.5
	Sentence	78.0	74.0	70.0
	Opinion unit	78.0	78.0	76.5
	Opinion unit + sf	84.0	89.0	88.5
Detailed (Task 11-50)	Passage	51.0	50.0	45.8
	Sentence	62.5	60.0	50.8
	Opinion unit	69.0	61.7	57.2
	Opinion unit + sf	81.5	78.2	73.0

(c) SEMEVAL Res15+Res16, `all-mpnet-base-v2`

Tasks	**Chunking strategy**	**Precision**	
		@5	**@10**
All (Task 1-30)	Passage	53.3	41.7
	Sentence	53.3	42.0
	Opinion unit	67.3	56.7
	Opinion unit + sf	74.0	60.3
General (Task 1-10)	Passage	78.0	63.0
	Sentence	78.0	64.0
	Opinion unit	80.0	81.0
	Opinion unit + sf	84.0	85.0
Detailed (Task 11-30)	Passage	41.0	31.0
	Sentence	41.0	31.8
	Opinion unit	61.0	44.5
	Opinion unit + sf	69.0	48.0

(d) SEMEVAL Res15+Res16, `all-MiniLM-L6-v2`

Tasks	**Chunking strategy**	**Precision**	
		@5	**@10**
All (Task 1-30)	Passage	46.0	42.3
	Sentence	46.0	42.3
	Opinion unit	54.7	46.7
	Opinion unit + sf	72.0	62.3
General (Task 1-10)	Passage	58.0	55.0
	Sentence	60.0	54.0
	Opinion unit	68.0	64.0
	Opinion unit + sf	78.0	77.0
Detailed (Task 11-30)	Passage	40.0	36.0
	Sentence	39.0	36.5
	Opinion unit	48.0	38.0
	Opinion unit + sf	69.0	55.0

6 Summary and Conclusion

In summary, we have introduced opinion units as a concise and contextualised representation for subjective viewpoints in text, demonstrated that these can be automatically extracted with modern language technology, and that they lead to improved performance for opinion retrieval tasks.

The chosen few-shot approach allows the LLM to identify aspects without the need for annotated data or rigid, predefined, aspect categories. Each opinion unit captures a single opinion and is composed of an aspect label, a text excerpt that contextualises the opinion on the aspect, and a sentiment label that captures the expressed sentiment. These units are designed to facilitate downstream applications, e.g., clustering and retrieval. By balancing abstractive and extractive summarization in the excerpt generation, the approach handles difficulties such as intertwined opinions, where discussions interleave opinions with other topics, and opinions that span multiple sentences. Furthermore, the sentiment label is helpful for filtering at inference time, mitigating the issue with word embeddings where words with contrasting sentiment polarities have similar vector representations (Yu et al., 2017).

Our findings demonstrate the ability of LLMs to accurately extract opinion units from benchmark datasets for aspect-based sentiment analysis. Furthermore, a case study involving 50 tasks showcased the effectiveness of opinion units in opinion retrieval using dense embeddings on a large, real-world, review dataset. Our approach outperformed the traditional methods of corpus indexing of sentence- and passage-level chunking.

While our study implemented a baseline dense retrieval system to isolate the impact of opinion units, a more refined implementation could integrate various techniques. For instance, sentiment refined word embeddings (Yu et al., 2017), supervised retrievers (Chen et al., 2023), data augmentation (Wang et al., 2022), hybrid sparse-dense retrieval (Luan et al., 2021) or mixed strategy retrieval (Ma et al., 2023). These methods should however be compatible with, and complementary to, opinion units. Additionally, it would be interesting to cluster opinions based on the corresponding opinion units, to learn how groups of aspects and sentiments correspond to overall ratings or buying decisions, and how the principles of atomicity and contextuality (see Section 3) affect the results.

7 Limitations

The first group of limitations stems from the need for a larger labelled benchmark ABSA dataset. The current SEMEVAL datasets are restricted not only by the number of reviews, but also by the brevity and authenticity of these reviews, as they consist of individual sentences rather than complete review texts. A larger dataset would enable a more realistic evaluation of opinion unit generation. This should ideally include a significant amount of non-opinionated texts and of opinions that require multi-hop reasoning to understand, challenges that LLMs are known to struggle with (Chen et al., 2023).

Another dataset-related limitation is the absence of annotated retrieval datasets specifically for opinion mining. To address this, we designed 50 custom retrieval tasks to simulate opinion retrieval and evaluated the top-ranked reviews returned by these tasks. Annotated datasets, akin to those used in the QA domain (Chen et al., 2023) or TREC challenges (Grossman et al., 2016), contain pre-annotated relevant documents for each task and would facilitate a more comprehensive assessment using recall and F1 metrics. Such datasets would provide a more holistic understanding of retrieval performance, complementing the precision-based evaluation we currently employ.

Secondly, our opinion-retrieval system is not optimised. The advantage of a a simpler implementation is that we can isolate the effect of opinion-units on retrieval performance. However, we do not demonstrate the effectiveness of opinion units in refined downstream applications. While similarity comparison using dense embeddings presents many advantages in finding similar textual passages it can also produce undesired outcomes. For instance, in Retrieval Task 6, which required retrieving negative opinions on value for money, reviews containing phrases like "we had a bad experience" were returned because the word embedding model deemed them semantically similar to "bad value for money". A simple keyword search for "price" could, for this specific task, have returned reviews more aligned with the intended results, at least from an aspect-matching perspective. Additionally, while word embeddings capture sentiment to some extent, terms such as "accessible" and "inaccessible" with contrasting sentiment polarities can have similar vector representation (Yu et al., 2017). A more sophisticated implementation could include training of a supervised retriever (Chen et al., 2023) (also requiring labelled relevance data), refining word embeddings for sentiment analysis (Yu et al., 2017), or data augmentation (Wang et al., 2022). Hybrid sparse-dense retrieval (Luan et al., 2021) or mixed strategy retrieval (Ma et al., 2023) could also be beneficial. These methods should be synergistic with opinion units, where the segmentation of the retrieval corpus into structured opinion is a separate pre-processing step.

Finally, our evaluation of opinion units as a structure for opinions was limited to customer reviews. Other opinionated texts, such as longer political writings, could present additional challenges. These texts may make it more difficult to extract excerpts that contextualize an opinion, and they may require a greater degree of abstractive summarization to accurately capture the context.

References

G. L. Anand Babu and Srinivasu Badugu. 2023. A survey on automatic text summarisation. In *Proceedings of the Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022*, pages 679–689. Springer.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021. Every bite is an experience: Key point analysis of business reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3376–3386. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 39–49. Association for Computational Linguistics.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Computing Surveys*, 54(2).

Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6515–6524.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Dong Yu, and Hongming Zhang. 2023. Dense X retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*.

Lei Gao, Yulong Wang, Tongcun Liu, Jingyu Wang, Lei Zhang, and Jianxin Liao. 2021. Question-driven span labeling model for aspect–opinion pair extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12875–12883.

Maura R. Grossman, Gordon V. Cormack, and Adam Roegiest. 2016. Trec 2016 total recall track overview. In *Proceedings of the 25th Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA*, volume 500-321. National Institute of Standards and Technology (NIST).

Langchain. 2024. Faiss. `https://python.langchain.com/v0.2/docs/integrations/vectorstores/faiss/`. Accessed: 2024-04-20.

Xin Li and Wai Lam. 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*, pages 2886–2892.

Pengfei Liu, Shafiq Joty, and Helen Meng. 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*, pages 1433–1443.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Ling Luo, Xiang Ao, Yan Song, Jinyao Li, Xiaopeng Yang, Qing He, and Dong Yu. 2019. Unsupervised neural aspect extraction with sememes. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5123–5129.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3538–3547.

Kaixin Ma, Hao Cheng, Yu Zhang, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. 2023. Chain-of-skills: A configurable model for open-domain question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1599–1618, Toronto, Canada. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

V. Priya and K. Umamaheswari. 2020. Aspect-based summarisation using distributed clustering and single-objective optimisation. *Journal of Information Science*, 46(2):176–190.

An Tang, Xiuzhen Zhang, and Minh Dinh. 2024. Aspect-based key point analysis for quantitative summarization of reviews. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1419–1433, St. Julian's, Malta. Association for Computational Linguistics.

Sentence Transformers. 2024. Pretrained models. `https://www.sbert.net/docs/sentence_transformer/pretrained_models.html`. Accessed: 2024-04-20.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Yelp. 2015. Yelp open dataset. Dataset.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark. Association for Computational Linguistics.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *Preprint*, arXiv:2305.15005.

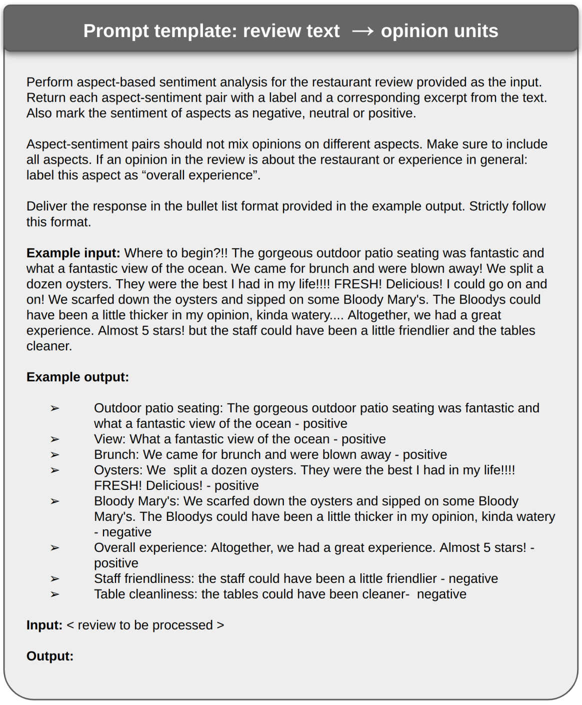
Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short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

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

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015. Representation learning for aspect category detection in online reviews. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Appendix A: Prompt template for opinion unit generation

In Figure 4 we provide our prompt use to create opinion units for restaurant reviews.



Prompt template: review text → opinion units

Perform aspect-based sentiment analysis for the restaurant review provided as the input. Return each aspect-sentiment pair with a label and a corresponding excerpt from the text. Also mark the sentiment of aspects as negative, neutral or positive.

Aspect-sentiment pairs should not mix opinions on different aspects. Make sure to include all aspects. If an opinion in the review is about the restaurant or experience in general: label this aspect as “overall experience”.

Deliver the response in the bullet list format provided in the example output. Strictly follow this format.

Example input: Where to begin?!! The gorgeous outdoor patio seating was fantastic and what a fantastic view of the ocean. We came for brunch and were blown away! We split a dozen oysters. They were the best I had in my life!!!! FRESH! Delicious! I could go on and on! We scarfed down the oysters and sipped on some Bloody Mary's. The Bloodys could have been a little thicker in my opinion, kinda watery.... Altogether, we had a great experience. Almost 5 stars! but the staff could have been a little friendlier and the tables cleaner.

Example output:

➢ Outdoor patio seating: The gorgeous outdoor patio seating was fantastic and what a fantastic view of the ocean - positive
➢ View: What a fantastic view of the ocean - positive
➢ Brunch: We came for brunch and were blown away - positive
➢ Oysters: We split a dozen oysters. They were the best I had in my life!!!! FRESH! Delicious! - positive
➢ Bloody Mary's: We scarfed down the oysters and sipped on some Bloody Mary's. The Bloodys could have been a little thicker in my opinion, kinda watery - negative
➢ Overall experience: Altogether, we had a great experience. Almost 5 stars! - positive
➢ Staff friendliness: the staff could have been a little friendlier - negative
➢ Table cleanliness: the tables could have been cleaner- negative

Input: < review to be processed >

Output:

Figure 4: Prompt template for input review text to opinion units.

Appendix B: Case Study Task Examples

In this appendix we provide six examples of tasks for the case study on opinion retrieval. The full details of the 50 tasks are provided in the supplementary material. The 50 tasks consist of 10 general tasks and 40 detailed tasks. General tasks correspond to common and overarching opinions found in hotel reviews, such as overall experience, value for money, and staff friendliness. Tasks 11-50 focus on more specific aspects mentioned in fewer reviews. Each task has 5 data fields:

1. *ID:* the task’s ID
2. *Title:* the task’s title
3. *Task description:* a description describing criteria for which reviews that should be annotated as relevant.
4. *Sentiment:* the sentiment (positive or negative) that reviews should have towards the task aspect. This sentiment field is used for the retrieval method: opinion units + sentiment filter.
5. *Query:* the query that is used to retrieve reviews.

Task 1: Overall Experience (positive)

- ID: 1
- Title: Overall Experience (positive)
- Description: Reviews should include positive opinions from customers expressing satisfaction or enjoyment of their overall experience. Positive opinions about specific aspects, even if their are several, are not enough.
- Sentiment: positive
- Query: On the whole, it was an excellent experience

Task 6: Value for Money (negative)

- ID: 6
- Title: Value for Money (negative)
- Description: Assessment of whether the prices are justified by the quality, service, food, experience etc. The review should mention EXPENSIVE prices/ bad value for money etc.
- Sentiment: negative
- Query: It was expensive and bad value for money

Task 9: Restaurant Atmosphere (positive)

- ID: 9
- Title: Restaurant atmosphere (positive)
- Description: Reviews should include positive opinions from customers expressing satisfaction with the atmosphere, ambiance or overall vibe of the restaurant.
- Sentiment: positive
- Query: The restaurant had a great atmosphere

Task 22: Portion sizes (negative)

- ID: 22
- Title: Portion sizes (negative)
- Description: Reviews should include negative opinions about the portion sizes of the dishes served at the restaurant. Customers should express dissatisfaction with the quantity or value of the food portions.
- Sentiment: negative
- Query: I was dissatisfied with the portion size

Task 37: Accomodating for food allergies (positive)

- ID: 37
- Title: Accomodating for food allergies (positive)
- Description: The review should highlight or provide examples of how well the restaurant handled accommodations for food allergies and special dietary needs such as lactose, gluten etc.
- Sentiment: positive
- Query: They were very accommodating of my food allergies

Task 48: Too hot in restaurant (negative)

- ID: 48
- Title: Too hot in restaurant (negative)
- Description: Reviews should include negative opinions/experiences about the temperature being too hot for comfort inside the restaurant.
- Sentiment: negative
- Query: It was too hot inside the restaurant

Appendix C: Computation details

All steps of the computational experiments were conducted on a laptop. The most computationally intensive task was embedding the review chunks. Using the larger embedding model, `all-mpnet-base-v2`, this process took approximately 5 hours per chunking strategy for the dataset consisting of 20,000 Yelp reviews. In contrast, the smaller model, `all-MiniLM-L6-v2`, completed the same task in about 20 minutes.

Appendix D: Opinion Retrieval Evaluation Details

The reviews returned for the opinion retrieval tasks were assessed by a team of four evaluators. This team included one of the article's authors, along with friends and associates, none of whom were compensated for their participation. The evaluators were blind to the chunking strategies used, and the reviews were presented in a randomized order to eliminate potential sources of bias.

The evaluators' task was to determine the relevance of the reviews to the provided retrieval task descriptions. Examples of these task descriptions can be found in Appendix B, with the full list available in the supplementary materials. Detailed instructions for the evaluation were provided orally. More practical instructions for the evaluation were provided orally.