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

Emil Häglund
Department of Computing Science
Umeå University, Sweden
emilh@cs.umu.se

Johanna Björklund
Department of Computing Science
Umeå University, Sweden
johanna@cs.umu.se

Abstract

We introduce *opinion units*, a contribution to the field Aspect-Based Sentiment Analysis (ABSA) that extends aspect-sentiment pairs by including substantiating excerpts. The goal is to provide fine-grained information without sacrificing succinctness and abstraction. Evaluations on review datasets demonstrate that large language models (LLMs) can accurately extract opinion units through few-shot learning. The main types of errors are providing incomplete contexts for opinions and and mischaracterising objective statements as opinions. The method reduces the need for labelled data and allows the LLM to dynamically define aspect types. As a practical evaluation, we present a case study on similarity search across academic datasets and public review data. The results indicate that searches leveraging opinion units are more successful than those relying on traditional data-segmentation strategies, showing 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 paraphrased to only include relevant text, 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.

Initial ABSA research focused on classifying reviews into predefined aspect- and sentiment categories (Zhang et al., 2022). Over time, this came to include the extraction of aspect- and sentiment keywords (Zhang et al., 2022; Gao et al., 2021). While the reduction of a text to keywords is helpful for many applications, it also lead to information loss. In contrast, opinion units offer a structured representation that retains more of the original nuance. The emergence of generative LLMs, with their capacity for longer sequence-to-sequence outputs, enable the flexible extraction of phrases required for creating opinion units. For concisely expressed opinions, as in the short-sentence examples used in keyword-extraction benchmarks like SEMEVAL Res-15 & 16 (Pontiki et al., 2016), opinion units closely resembles keyword extraction. However, in real-world reviews, customer opinions often involve descriptions and motivations spanning longer passages. Phrase extraction provides a more natural and expressive method for capturing these nuanced opinions. For instance keyword extraction would overlook subtlety in a sentence like: “The outdoor area is delightful, especially in the evening, with its soft lighting and comfortable chairs creating a cozy atmosphere”. Moreover, phrases provide better interpretability for end users, allowing them to identify which sections of the raw text influenced the decisions made by downstream applications.

The extraction of opinion units can serve as a standalone chunking strategy for applications requiring detailed information. However, it can also be made as preprocessing step before keyword extraction (Siddiqi and Sharan, 2015), because the

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 230–240

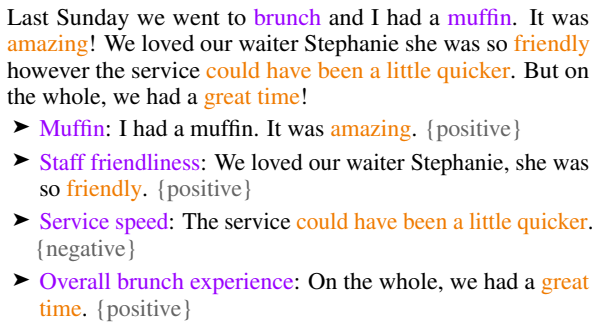

Figure 1: Four opinion units extracted from a review, each representing an opinion in the text and consisting of an aspect label, an excerpt from the text, and a sentiment label. The colour purple indicates aspects, and orange indicates sentiment terms.

atomic nature of opinion units—each representing a single opinion about one aspect—simplifies analysis. This is advantageous compared to analysing "raw text", which often contains intertwined opinions and unrelated non-opinionated content.

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., 2024). 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. 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.

In the following sections, we first investigate the ability of LLMs in generating opinion units by evaluating GPT-4-turbo, GPT-3.5-turbo, and Llama2-70B. This evaluation is conducted on subsets of SEMEVAL restaurant review sentence dataset (Pontiki et al., 2016) as well as a Yelp dataset (Yelp, 2015) containing complete restaurant reviews. Furthermore, we categorize the errors produced by the LLMs, where providing incomplete context, 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 the competing chunking strategies of sentence and passage chunking. 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 reveal which topics customers discuss in reviews.

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

RQ1. To what extent can LLMs extract accurate opinion units?

RQ2. What are the types and frequencies of errors made by the LLMs in this process?

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

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). The main challenge in these tasks is the accurate pairing of aspect-sentiment elements (Zhang et al., 2022).

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).

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 is needed 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 also 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., 2024). Chen et al. (2024) 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., 2024). Corpus segmentation using propositions is described as an orthogonal strategy that can be used in conjunction with other methods for improving dense retrieval such as supervised retrievers (Chen et al., 2024), data augmentation (Wang et al., 2022) or mixed-strategy retrieval (Ma et al., 2023).

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. However, sentences can include multiple aspect and lack necessary context when dependencies span multiple sentences (Yang et al., 2019).

3 Opinion units

As stated in Section 1, an opinion unit is composed of three elements: i) an aspect label, ii) a text ex-

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.

cerpt 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 are inspired by the factual propositions of Chen et al. (2024) described in Section 2.3, but are tailored for the ABSA domain. The principles 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 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 evaluate the performance of three LLMs (GPT-4 turbo, GPT-3.5 turbo, and Llama2-70B) in generating well-formed opinion units. Second, we perform a case study on opinion retrieval, comparing data segmentation based on opinion units to traditional chunking strategies.

4.1 Generation of Opinion Units

We generate opinion units using LLMs in a few-shot approach. The prompt template, provided in Figure 3, instructs the LLM to perform ABSA, extracting the three components of an opinion unit. An example review with opinion units is provided in the template. The examples are 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—for instance, by producing an incorrect JSON object— the generation is repeated (this happens approximately 5% of the time). For all LLMs we use a temperature of 1.0.

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 or positive.

Aspect-sentiment pairs should not mix opinions on different aspects. Make sure to include all aspects. An aspect should be independent and not have to rely on other aspects to be understood.

If an opinion in the review is about the restaurant or experience in general then label this aspect as "overall experience". Opinions not related to the restaurant should not be included.

Example input: I just left Mary's with my lovely wife. The gorgeous outdoor patio seating was fantastic with a nice view of the ocean. We came for brunch and were blown away! We split dozen oysters. They were the best I had in my life! FRESH! Delicious! The avocado toast was excellent as were the crab cakes. 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 with a nice view of the ocean", "positive"],
["View", "a nice 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"],
["Avocado toast", "the avocado toast was excellent", "positive"],
["Crab cakes", "the crab cakes were excellent", "positive"],
["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 3: Prompt template: opinion unit generation

4.2 Opinion Unit Evaluation

To assess the correctness of the generated opinion units, we conduct evaluations on subsets of SEMEVAL Res15 and Res16, which consist of restaurant-review sentences (Pontiki et al., 2016), as well as full Yelp restaurant reviews (Yelp, 2015). We compare the performance of GPT-3.5-turbo, GPT-4-turbo and Llama2-70B. For these subsets, we created solution keys of correct opinion units by manually identifying aspects and their sentiments in each text. For the SEMEVAL subset, sentiment labels followed the ASTE annotations provided by (Zhang et al., 2021). In the solution keys, we selected approved LLM-generated aspect labels and excerpts to serve as examples of correct opinion unit components. For the SEMEVAL subset we select reviews from the Res15 and Res16 test sets that, according to (Zhang et al., 2021)'s annotations, include multiple aspects. The subset used for SEMEVAL evaluation consists of 565 opinion units in the solution key, stemming from 238 review sentences. A similar size subset was randomly subsampled from the Yelp dataset, constituting 505 opinion units from 96 reviews.

We evaluate opinion units according to the principles outlined in Section 3. These principle include, ensuring that each unit reflects a single opinion, provides enough context to motivate its sentiment and that the sentiment classification and identified aspects align with the solution key. We classify errors into the categories listed below; an opinion unit is considered correct only if it avoids all these errors. The evaluation was conducted by the two authors and was not blind to which LLM generated the opinion units. Disagreements that arose during the evaluation were revisited and resolved through careful re-examination in accordance with the established error and evaluation guidelines.

Atomicity error. An opinion unit lacks *atomicity*, 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-opinion. A non-opinionated excerpt from the text is incorrectly classified as an opinion.

Sentiment error. The sentiment label is incorrect.

Aspect-label error. The aspect label does not adequately describe the opinion.

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

To quantify the results, we use three metrics: Precision, the ratio of correct generated units to total generated units; recall, the ratio of correct generated units to total opinion units in the solution key; and F1-score, the harmonic mean of precision and recall. In the scoring, certain cases were handled with special consideration. For the short SEMEVAL reviews, the LLMs in addition to individual aspects, sometimes created instances of "overall experience" which combined multiple aspects as a characterization of the overall experience.

When considered reasonable reflections of overall sentiment, these were excluded from scoring and did not impact the precision and recall values.

Our evaluation inherently involves a degree of subjectivity. For example, differing human assessments may arise about whether an extracted phrase provides full context or if an aspect label is descriptive enough to capture the opinion. This subjectivity, though typical for many NLP annotations (Röttger et al., 2021) and perhaps especially for unstructured generative LLM outputs, makes the evaluation unsuitable as a strict benchmark, like ABSA benchmarks for classification and keyword extraction (Pontiki et al., 2016). Despite these limitations, we believe this evaluation to be crucial for understanding the performance of opinion unit generation in isolation and not just through its impact on downstream tasks. Additionally, the error classification offers important insights for future work on using LLMs for opinion extraction.

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. For this opinion retrieval task, opinion units were generated using GPT-3.5-turbo, selected for its balance of performance (as demonstrated in Section 5.1) and cost-efficiency.

Retrieval Tasks. We designed 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. We categorized the 50 tasks 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. Task 4 seeks reviews that highlight staff friendliness, using the query: "The staff were very friendly. 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, and the remaining a negative sentiment. The returned reviews were assessed by a team of 4 evaluators who were blind to the chunking strategies used. On average, each returned review received 2.3 annotations. Conflicts were resolved through majority voting; in cases of equal votes, an additional evaluator was consulted for final assessment. The reviews were presented in a randomized order to eliminate a potential source of bias. The full list of review tasks, including queries and task descriptions are available online. Implementations of opinion unit generation, retrieval and passage and sentence chunking are also provided[1].

Evaluation Groups. We compare dense retrieval based on opinion units to the conventional approaches of passage- and sentence chunking (Chen et al., 2024). 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 (avg. 28.2 words in Yelp dataset) compared to sentence chunking (avg. 12.9 words) and opinion units (14.9 words). 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 is used to evaluate results by measuring the percentage of relevant reviews among the top k returned reviews for each task.

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 ask for

[1] https://github.com/emilhagl/Opinion-Units

increasingly specific topics. When evaluating on the SEMEVAL dataset, we omit the 20 most specific tasks (i.e., Task 31–50) because the scope of the dataset is so limited that these fine-grained tasks do not contribute to the evaluation in a meaningful way. 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).

5 Results and Discussion

5.1 Opinion Unit Evaluation

We evaluate the opinion units generated for the SEMEVAL and YELP subsets with respect to the methodology described in Section 4.1. Our analysis reveals that GPT-4-turbo achieves the best performance across datasets (YELP: Precision = 85.3, Recall = 87.4; SEMEVAL: Precision = 89.3, Recall = 92.7). GPT-3.5-turbo shows slightly lower performance (YELP: Precision = 87.0, Recall = 82.2; SEMEVAL: Precision = 87.5, Recall = 89.6), while Llama2 exhibits a more pronounced drop in performance (see Table 1). Notably, recall values are lower for the YELP dataset, where longer reviews result in a greater number of overlooked aspects. Overall, the strong performance of the GPT-models is promising for downstream tasks.

Furthermore, we categorize the errors according to the classification described in Section 4.1, to understand the types of problems the LLMs encounters when generating opinion units. The frequency of these errors is presented in Figure 4. The most common errors are missing context or categorizing non-opinion statements like “we went to sit at the bar” as opinions (see Figure 4). For the Yelp dataset with long text reviews, missing aspects were a frequent error. Although issues like missing context, injectivity, or atomicity are less than ideal in terms of error severity, an opinion unit could still function reasonably well as a retrieval unit. In contrast, missing aspects and the characterizing non-opinions as opinions have a more certain negative impact on downstream tasks.

	Yelp			SEMEVAL		
	P	**R**	**F1**	**P**	**R**	**F1**
GPT-4-turbo	85.3	87.4	86.3	89.3	92.7	91.1
GPT-3.5-turbo	87.0	82.2	84.6	87.5	89.6	88.5
Llama2-70B	76.9	74.5	75.7	75.6	88.8	81.6

Table 1: Precision (P), Recall (R) & F1-scores for evaluation on Yelp and SEMEVAL subsets.

A few hallucinations were identified, primarily produced by Llama2, where the LLM invented an excerpt not present in the review. These mostly occurred when the LLM added an “overall experience" label with an invented excerpt, an artefact of the prompt template’s instructions for “overall experience.", (see Figure 3).

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. The increased retrieval precision stems from the ability to address 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 senti-

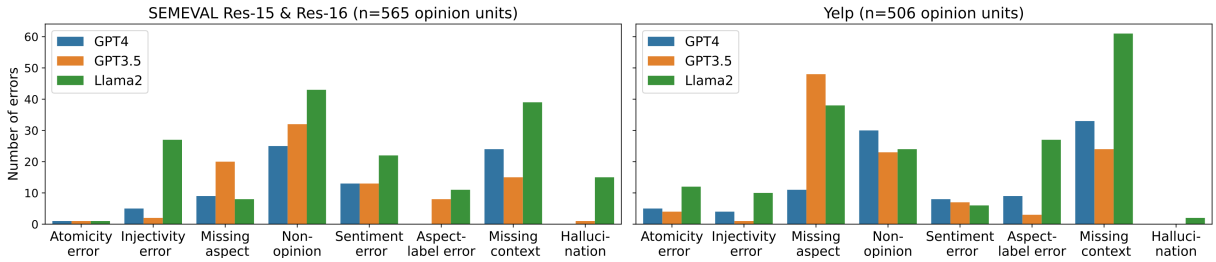

Figure 4: Error type frequency in generated opinion units for SEMEVAL and Yelp subsets.

(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

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

ment 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).

6 Summary and Conclusion

We have presented opinion units as a structured representation for subjective viewpoints, enhancing traditional aspect-sentiment pairs by incorporating substantiating excerpts that retain detailed information. Opinion units can function as an independent chunking strategy for applications that require detailed information or be utilized as a preprocessing step that allows for further abstractions such as category classification or keyword extraction. Our findings demonstrate the ability of LLMs to accurately extract opinion units from restaurant review datasets datasets. The most frequent errors were insufficient excerpt context and misclassifying non-opinion statements as opinions. Furthermore, a case study showcased the effectiveness of opinion units in opinion retrieval using dense embeddings, outperforming traditional segmentation methods.

The few-shot approach allows the LLM to identify aspects without annotated data or predefined aspect categories. Each opinion unit represents a single opinion, consisting of an aspect label, a text excerpt that provides context, and a sentiment label that conveys the expressed sentiment. These units facilitate downstream applications, e.g., clus-

tering and retrieval. The excerpt generation handles difficulties such as intertwined opinions, where discussions interleave opinions with other topics, and multi-sentence opinions. Furthermore, the sentiment label allows 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).

7 Limitations and Future Work

In this study, we did not fine-tune the LLMs for the opinion unit generation task. While demonstrating that LLMs can perform well on this task without the requiring additional training data is a strength in itself, fine-tuning has the potential to improve accuracy and enable the use of smaller, more efficient models. Exploring the potential improvements in performance through fine-tuning, particularly with regard to specific error, is an intriguing avenue for future research.

Our study implemented a baseline dense retrieval system to isolate the impact of opinion units on retrieval performance. However, we do not demonstrate the effectiveness of opinion units in refined downstream applications. A more refined implementation could integrate various techniques. For instance, sentiment refined word embeddings (Yu et al., 2017), supervised retrievers (Chen et al., 2024), 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 be synergistic with opinion units, where the segmentation of the retrieval corpus into structured opinion is a separate pre-processing step. 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.

The next group of limitations stem from the need for a larger labelled ABSA dataset. The current SEMEVAL datasets are restricted not only by the number of reviews, but primarily by the brevity and inauthenticity of these reviews, as they consist of individual sentences rather than complete review texts. A larger annotated dataset would facilitate the evaluation of opinion units with reduced reliance on custom annotation and assessment. Such a dataset 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., 2024). Such datasets could serve as a direct benchmark or foundational basis for evaluation.

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., 2024) 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@k-based evaluation we currently employ.

Finally, our evaluation of opinion units as a structure for opinions focused on 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).

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.

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.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Sifatullah Siddiqi and Aditi Sharan. 2015. Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications*, 109(2).

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. 2021. 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.

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.