Aspect-Based Opinion Summarization with Argumentation Schemes

Wendi Zhou and Ameer Saadat-Yazdi and Nadin Kökciyan
School of Informatics,
University of Edinburgh
{wendi.zhou, ameer.saadat, nadin.kokciyan}@ed.ac.uk

Abstract

Reviews are valuable resources for customers making purchase decisions in online shopping. However, it is impractical for customers to go over the vast number of reviews and manually conclude the prominent opinions, which prompts the need for automated opinion summarization systems. Previous approaches, either extractive or abstractive, face challenges in automatically producing grounded aspect-centric summaries. In this paper, we propose a novel summarization system that not only captures predominant opinions from an aspect perspective with supporting evidence, but also adapts to varying domains without relying on a pre-defined set of aspects. Our proposed framework, ASESUM, summarizes viewpoints relevant to the critical aspects of a product by extracting aspect-centric arguments and measuring their *salience* and *validity*. We conduct experiments on a real-world dataset to demonstrate the superiority of our approach in capturing diverse perspectives of the original reviews compared to new and existing methods.

1 Introduction

Online reviews are essential resources for customers to make purchase decisions, as they more authentically reflect the performance of some products or services (Boorugu and Ramesh, 2020; Amplayo et al., 2021). It is very impractical for users to go over most reviews one by one and conclude the prominent opinions discussed themselves. Ideally, users should have access to automated opinion summaries to make informed decisions.

Automatic opinion summarization offers a solution by aggregating all reviews into a concise, easy-to-read summary. Previous methods concerning opinion summarization can be mainly classified as either extractive or abstractive. We see drawbacks with both approaches. Extractive methods select the representative sentences from the input to generate the summary. Although attributable and scalable, they could encounter issues in generating concise and coherent summaries. On the other hand, abstractive methods using neural models to generate fluent and novel summaries may lead to hallucinated content that is challenging to detect without any supporting evidence. Hosking et al. (2023) implement a hybrid summarization system, HERCULES, that produces summaries reflecting the general feedback of all reviewers while abstracting away too many details. Although being abstractive and attributable, their summaries are too general for users interested in certain aspects of the entity.

We argue that an ideal summary should reflect the main opinion expressed in the reviews, be attributable with grounding evidence and include critical aspect information that is essential to assist customers while making their purchase decisions. Many attempts have been made to incorporate aspect information inside the final summary (Amplayo et al., 2021; Tang et al., 2024; Li et al., 2025); however, they either rely on the manually pre-defined aspects or they lose track of the supporting evidence with a fully automated pipeline using large language models (LLMs).

To address these existing limitations, we propose an aspect-centric review summarization framework, ASESUM, to produce high-quality opinion summaries for products. With the help of argumentation schemes and LLMs, ASESUM extracts aspect-centric arguments, where the claim is the user's sentiment towards certain aspects, and the premise is the supporting evidence mentioned by the users in the reviews. This makes the summarization model more generalisable than previous systems as it can easily adapt to new domains, does not require a pre-existing taxonomy of new aspects and can scale up with the number of reviews. By clustering claims supported by similar pieces of evidence, we define a metric to measure the salience and validity of an argument. This metric is used to rank the arguments having the critical aspects

Proceedings of the 12th Argument Mining Workshop, pages 116–125

information from which we generate our final summaries. In this paper, our main contributions can be summarized as follows:

- We develop a new automated method that can iteratively induce the aspect taxonomy within the product reviews;
- We introduce a new domain-independent argumentation scheme for aspect-centric argument extraction from customer reviews;
- We propose a novel hybrid review summarization framework (ASESUM)[1] to generate textual summaries. Our model outperforms the current state-of-the-art by 6% on average on a real-world benchmark dataset.

Our paper is organised as follows. We discuss related work on summarization and argumentation in NLP in Section 2. We introduce our review summarization framework (ASESUM) in Section 3, and Section 4 explains our experimental setup before we compare our approach to other models. In Section 5, we show that ASESUM outperforms these models, not only in terms of the amount of semantic information captured by our summaries but also in the diversity of viewpoints presented. We conclude our paper with a discussion in Section 6.

2 Related work

Earlier work on opinion summarization, or review aggregation, is either purely extractive (Mihalcea and Tarau, 2004; Rossiello et al., 2017; Alguliyev et al., 2019; Belwal et al., 2021) or abstractive (Ganesan et al., 2010; Bražinskas et al., 2020). However, both types of methods have their own shortcomings: extractive methods tend to introduce unnecessary details and struggle to cover all topics in multi-topic inputs, while abstractive methods are limited by the input length of neural models or language models and may generate hallucinated content. Hosking et al. (2023) introduce a hybrid approach, where they encode the review sentences as a hierarchy of paths and then decode the most frequent path in the hierarchy structure as the final summary. Though being unsupervised and attributable, their hierarchy encoder is domain-dependent, thus limiting its generalisability. Their approach mainly focuses on the general summary generation, neglecting aspect-relevant information.

Angelidis et al. (2021) propose an extractive method that generates aspect-specific summaries using the quantized transformer. Similarly, Amplayo et al. (2021) develop an abstractive method where they fine-tune a Pre-trained Language Model with aspect controllers for abstractive summaries generation. However, these methods extract aspects either directly from the sentence or with the assistance of humans. Recently, LLMs have demonstrated great performance across a wide range of natural language understanding tasks. Leveraging this, Tang et al. (2024) propose a fully automated aspect extraction approach through few-shot prompting. They successfully extract aspects together with users' sentiment towards that aspect from reviews; then, after clustering the <aspect, sentiment> pairs, they re-prompt LLMs to generate the aspect-specific keypoints as the final summaries. In this way, they achieve flexible aspect-centric summaries generation at scale, but this iterative prompting pipeline makes their summaries harder to validate without grounding evidence. In contrast, ASESUM framework preserves the same versatility while providing the grounding evidence by considering argumentative structure. In Li et al. (2025), they propose a more explainable and grounded summarization pipeline through prompting LLMs, which separates the tasks of aspect identification, opinion consolidation, and meta-review synthesis. However, their system requires a set of manually pre-defined aspects, while our system incorporates a flexible aspect induce approach.

Argumentation schemes have been widely studied in computational argumentation, aiming to model, extract, and generate human-like arguments. A foundational basis for this theory comes from Walton, where he defines structured patterns of common reasoning used in everyday discourse (Walton et al., 2008). Each scheme is provided with a template for constructing arguments and critical questions for evaluating their validity. More recent approaches incorporate Walton's schemes into neural models to guide argument structure prediction and improve the interpretability of human conversations (Herbets de Sousa et al., 2024).

In the context of product reviews, Wyner Adam et al. (2012) introduce a scheme for product reviews based on customer values for semi-automated review analysis. Similarly, Mumford et al. (2024) use the *Position to Know* scheme and associated critical questions to evaluate the quality of reviews. We find both these approaches limited in that they

[1]All the code is available online at: https://git.ecdf.ed.ac.uk/s2236454/asesum

Review Argument Scheme (RAS)
Claim: **A** of this product is **S**
Major Premise: **X** is a sign that **A** is **S**
Minor Premise: The user observes **X** about **A**

Table 1: Proposed argument scheme where **A**, **S**, **X** represent the aspect, sentiment and evidence respectively.

ignore the particular features (aspects) of a product that users are discussing, making the analysis too coarse-grained and the evaluation criteria difficult to apply automatically. In contrast, we base our method on a scheme based on *Argument from Characteristic Sign* which we make specific to our aspect/sentiment framework. Our approach also does not depend on critical questions and instead uses an evidence consistency measure to identify the most salient evidence to provide to a user.

3 ASESUM Framework

In this section, we introduce an aspect-centric review summarization framework, ASESUM. The framework has three stages: (i) aspect-centric argument extraction with a new argumentation scheme, *Review Argument Scheme*, (ii) argument clustering and evidence unification, and (iii) argument scoring guided by aspect-centric argument relations.

3.1 Argument extraction

Inspired by the argumentation schemes defined by Walton et al. (2008), we propose a novel argumentation scheme for product reviews as shown in Table 1. The Review Argument Scheme (RAS) consists of three variables: the aspect (**A**), the sentiment (**S**) and the evidence (**X**). In our framework, **S** takes values from {*good, bad*}.

In ASESUM, each argument is defined as an instantiation of RAS, Definition 3.1 provides a formal definition of an argument. Note that Arg_i is used to define the *i*th argument.

Definition 3.1 (Argument)**.** Arg denotes a tuple $\langle a, s, x \rangle$, where a, s, x represent the aspect, sentiment and supporting evidence respectively, as they appear in the instantiated argument scheme Arg.

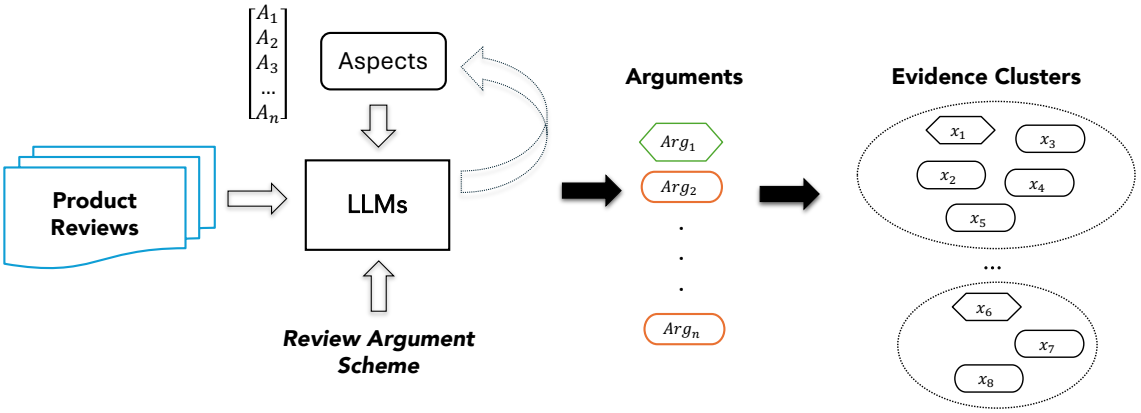
In order to instantiate RAS, we benefit from LLMs to fill in the scheme variables and generate arguments with provided user reviews. To avoid LLMs generating diverse aspect representations, we first prompt LLMs to initiate the *critical aspects* of the product given the product category information. The critical aspects represent the key evaluation factors of the product, which may greatly influence customers' purchase decisions. Then we feed them as options into the prompt to guide LLM on performing aspect-centric argument extraction (Figure 1a). However, for a small subset of reviews, LLMs fail to generate any valid arguments. As this affects only around 3% of the reviews per domain, it does not have a big influence on the final results. After obtaining all the arguments extracted by LLMs, we further unify the representations of aspects by clustering them and representing each cluster with a symbol ($A_1, A_2 ... A_n$). We will provide implementation details in Section 4.2.

3.2 Evidence-based Clustering

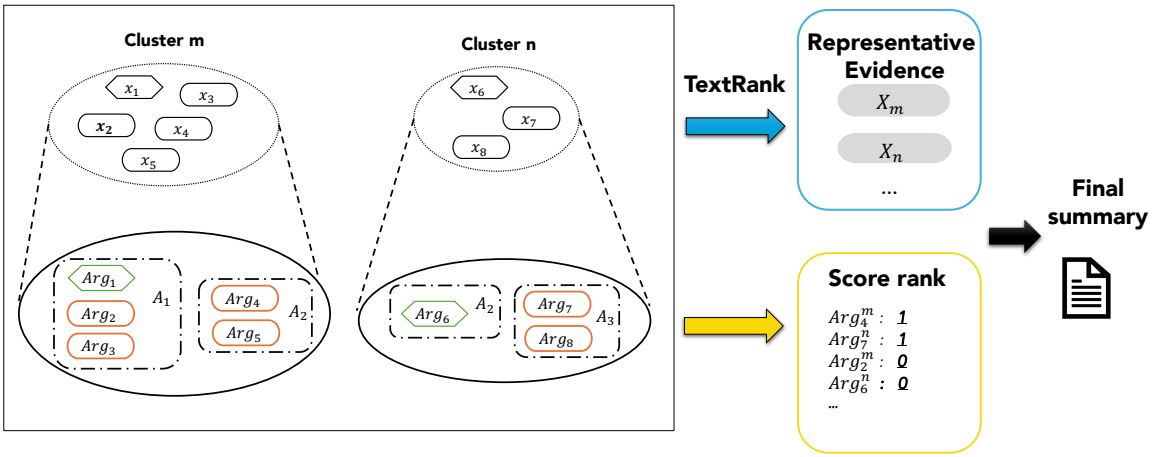
Since the evidence for each argument is extracted or slightly summarized by LLMs, it is highly unlikely they will have the same evidence even for arguments sharing the same aspect and sentiment. And so, we cluster arguments sharing semantically close evidence and then select the most representative evidence X_k for each cluster (Figure 1b). We assume the most representative evidence of the cluster is the one that entails the majority of the evidence in the cluster. To achieve this, we build a text graph where the vertices are the semantic embeddings of each sentence, and the edge weights are calculated by the cosine similarity between each node pair. Then we iterate the graph-based ranking algorithm derived from Google's PageRank (Page et al., 1999) as described in TextRank (Mihalcea and Tarau, 2004) until convergence. Finally, we select the vertex with the highest score as the most representative evidence of this cluster.

For each argument Arg_i in a cluster c, we then substitute its original evidence ($Arg_i.x$) with the representative evidence (X_c) and we rewrite the argument as Arg_i^c. In other words, the evidence of each argument in a cluster is replaced with the most representative evidence. This methodology is depicted in Figure 1b. For example, in *Cluster m*, we see five arguments, where each of them is supported with its unique evidence. X_m would be the representative evidence for *Cluster m*. If Arg_1 is represented as $\langle a_1, s_1, x_1 \rangle$, this argument would be rewritten as $Arg_1^m = \langle a_1, s_1, X_m \rangle$. All other arguments could be rewritten similarly.

After unifying similar evidence for every argument, we calculate a score for each argument based on its popularity and its validity in supporting or opposing a claim related to an aspect.



(a) A demonstration of the argument extraction, where we feed the product reviews, the defined Review Argument Scheme together with the set of aspects into LLMs to generate aspect-centric arguments (Definition 3.1). The aspect set is initiated by prompting LLMs, and is updated during the argument extraction. The arguments are then clustered based on their evidence.



(b) A diagrammatic representation of our methodology starting from clustered arguments. For each clustered argument, ASESUM selects the representative piece of evidence X_k using the TextRank algorithm. This representative evidence is then used to replace the original evidence of each argument in the cluster. Meanwhile, the system builds the relations among arguments within the same cluster based on their aspects, which are used to measure the salience and validity of the argument as defined in Equation 1. Finally, our system selects **N** unique pieces of evidence from the top-ranked arguments to generate the summary.

Figure 1: The Proposed ASESUM Framework

3.3 Aspect-centric Ranking

To quantitatively assess the salience and validity of an argument, we make use of its support and contradiction relations to other arguments in the same cluster (Section 3.2). Firstly, we provide formal definitions of the relations between arguments in Definitions 3.2 and 3.3.

Definition 3.2 (Aspect-centric Support)**.** A support relation between two arguments in the same cluster, Arg_i and Arg_j, exists if and only if both arguments have the same aspect (i.e., $Arg_i.a = Arg_j.a$) and sentiment (i.e., $Arg_i.s = Arg_j.s$).

Definition 3.3 (Aspect-centric Contradiction)**.** A contradiction relation between two arguments in the same cluster, Arg_i and Arg_j, exists if and only if both arguments have the same aspect (i.e., $Arg_i.a = Arg_j.a$) and different sentiment (i.e., $Arg_i.s \neq Arg_j.s$).

Intuitively, we consider an argument to be strengthened when a similar evidence supports the same claim from another argument, and an argument to be weakened if a similar evidence is used to support the opposite claim from another argument. For example, for a pair of shoes, a piece of evidence could be "the shoes are quite wide". If this evidence is used to support both arguments with the claim "The *fit* is good" and the claim "The *fit* is bad"; then for the aspect *fit*, "the shoes are quite wide" is a piece of controversial evidence, thus we should not include it into the final summary.

Based on Definitions 3.2 and 3.3, we measure the global validity of an argument *i* in a cluster *c*

by using Equation 1.

$$Score(Arg_i^c) = \sum_{\substack{\forall Arg_j \in c, i \neq j \\ Arg_i.a = Arg_j.a}} \hat{s}_i \times \hat{s}_j, \quad (1)$$

where $\hat{s}_i$ and $\hat{s}_j$ represent the sentiment polarity of Arg_i^c and Arg_j^c, respectively. An argument with a 'good' sentiment is assigned a polarity value of $+1.0$, while an argument with a 'bad' sentiment is assigned a polarity value of -1.0.

In ASESUM, as a final step, we assign each evidence cluster with the highest score achieved by any argument within it. The clusters are then ranked based on their scores, and the top-N representative evidence pieces are selected to generate the final summary.

4 Experimental Setup

In this section, we introduce the datasets used in our experiments (Section 4.1) and discuss the implementation details of ASESUM (Section 4.2). Then we describe other comparison systems (Section 4.3), and explain the automatic metrics for our evaluation (Section 4.4).

4.1 Dataset

We conducted our experiments by using the AmaSum dataset (Bražinskas et al., 2021), the largest abstractive opinion summarization dataset, consisting of more than 33,000 human-written summaries for Amazon products from a wide range of categories. In AmaSum dataset, each product is paired with more than 320 customer reviews and at least one reference summary. Each reference summary includes 'verdict', 'pros' and 'cons', but as the reference summaries are obtained from external resources, they are not grounded in product reviews. Similar to the work of Hosking et al. (2023), we concatenate these three sections together to construct the final reference summary. Moreover, we follow the same setting to build the test set by sampling 50 products per domain for evaluation. Detailed statistics are listed in Table 2.

Test Domain	#Reviews	Avg. Length
Electronic	568	45
Shoes	381	38
Sports & Outdoor	610	44
Home & Kitchen	680	45

Table 2: The statistics of all the domains in our sampled test set. *#Reviews* represents the average number of reviews for all the products, and *Avg. Length* represents the average number of words separated by space in reviews for a particular domain.

4.2 Implementation

In ASESUM framework, we choose one closed-source LLM *GPT-4o-mini* from OpenAI[2]and another open-source LLM *Qwen2.5-7B* (Qwen et al., 2025) as our backbone models. The prompt used for both models is shown in Appendix A.

[2]https://platform.openai.com/docs/models/gpt-4o-mini

In order to implement the aspect clustering (Section 3.1) and evidence clustering (Section 3.2), we opt for the Density-based spatial clustering of applications with noise (*DBSCAN*) algorithm (Ester et al., 1996). DBSCAN is the most ideal clustering method for ASESUM as it does not require a predefined number of clusters, thereby enhancing the generalizability of the framework. Based on a series of preliminary trials on the training set, we configure the clustering hyper-parameters as follows: the clustering metric is set to "cosine" similarity, the minimum number of sample per cluster is set to 1, and the ϵ is set to 0.5 and 0.21 for aspect clustering and evidence clustering, respectively. Additionally, we select the top 8 pieces of unique evidence to form our final summary based on our exploratory experiments.

4.3 Other Models for Comparison

As depicted in Figure 1, our proposed ASESUM framework is a hybrid summarization approach that combines *abstractive* methods (by benefiting from LLMs) and *extractive* methods (by selecting the final set of arguments for summarization with clustering and TextRank). According to this, we primarily compare our framework with the previous state-of-the-art hybrid summarization model, HERCULES (Hosking et al., 2023). Since HERCULES is domain-specific, we use their released models for the four domains (*Electronic*, *Shoes*, *Sports & Outdoor*, *Home & Kitchen*) as shown in Table 2. We evaluate the models on these four domains separately using their default configuration settings.

For comparison, we also develop an LLM-based baseline using GPT-4o-mini to evaluate the effectiveness of our ASESUM framework. In this case, we randomly sample 50 reviews (the maximum

number of reviews that would reliably fit within the context-length of gpt-4o-mini) and pass them to the model along with a simple summarization instruction[3].

4.4 Evaluation Metrics

We use various automatic evaluation metrics to compare ASESUM framework with other models, namely ROUGE-2, ROUGE-L F1 (Lin, 2004), SummaC (Laban et al., 2022). We also propose a new sentence diversity score to measure the sentence-level diversity of a summary.

We calculate the ROUGE-2, ROUGE-L F1 scores against the reference summaries of AmaSum dataset similar to the work of Hosking et al. (2023). SummaC score (Laban et al., 2022) is a popular metric for evaluating how well a summary is entailed by the input document. It segments the input document and reviews into sentences and computes the average entailment score between each pair of the input sentence and the generated summary. We calculate the SummaC score of the generated summaries against the reference (SC_{ref}) and the original input reviews (SC_{in}). Since the reference summary is built independently of the input reviews, the SummaC score computed against original reviews (SC_{in}) provides a more trustworthy indication of the summary quality.

A helpful product review summary should capture the most frequently expressed opinions from the input, but without repeating the same points redundantly. Therefore, we propose a diversity metric that evaluates the sentence-level diversity of the final summary. The idea is to segment a summary into sentences and evaluate the semantic closeness of all the sentences through clustering. As a longer summary having more sentences would result in a higher number of clusters naturally, we normalise the cluster number by the total number of sentences to obtain the final diversity score of a summary. We define this new metric in Equation 2.

$$Diversity(S) = \frac{|Clusters(S)|}{|S|}, \qquad (2)$$

where S is the set of sentences in a summary, $|Clusters(S)|$ is the number of clusters and $|S|$ is the number of sentences in S.

In our implementation, we use DBSCAN algorithm with the same parameter settings as the aspect clustering discussed in Section 4.2.

[3]Prompt: `Summarize the following list of reviews. Keep your answer concise while capturing as many diverse points of view as possible.`

5 Evaluation Results

In this section, we analyze the quantitative results based on all automatic evaluation metrics (Section 5.1) and provide a detailed qualitative discussion on the generated summaries for a randomly chosen product (Section 5.2).

5.1 Quantitative Analysis

The evaluation results are shown in Table 3. We observe that ASESUM framework with both closed-source and open-source LLMs consistently outperforms other methods on all four domains across all metrics besides ROUGE-2. Particularly for the SC_{in} score, our ASESUM achieves significantly higher SC_{in} scores across all the domains, indicating that our generated summaries are more representative of the input reviews. Surprisingly, our ASESUM framework paired with Qwen2.5-7B (ASESUM$_{\text{qwen2.5-7B}}$) achieves comparable performance with ASESUM paired with GPT-4o-mini (ASESUM$_{\text{gpt-4o-mini}}$) across all the domains and evaluation metrics, demonstrating both the robustness and the generalizability of the framework.

Across all models, the big difference between the SC_{in} and SC_{ref} score also suggests that the manually constructed reference summaries do not faithfully entail all the product reviews, as they are built separately. On the other hand, GPT-4o-mini baseline performs the worst on most of the metrics, which can be the result of the limited number of input reviews. However, it achieves higher ROUGE scores and has a smaller difference in SC_{ref} than it has in SC_{in} when compared to other methods. This indicates that summaries generated by GPT-4o-mini are more fluent and closer to manually written summaries.

In terms of the sentence-level diversity, ASESUM$_{\text{qwen2.5-7B}}$ even performs better than ASESUM$_{\text{gpt-4o-mini}}$ in most domains. Notably, the diversity of summaries generated by our ASESUM framework is greatly dependent on the diversity of unique aspects of products. For domains having products with various aspects, such as *Electronics* (on average 14 aspects per product), the diversity score of our summaries is obviously higher than other domains, such as *Shoes* (on average 10 aspects per product). While ASESUM with LLMs generate less diverse summaries for the *Shoes* and *Sports & Ourdoors* domains, it achieves higher

Models	ROUGE-2	ROUGE-L	SC_{ref}	SC_{in}	Diversity
			Electronics		
GPT-4o-mini	**2.93**	11.38	20.80	43.76	0.55
HERCULES	2.41	12.44	22.87	79.79	0.73
ASESUM$_{qwen2.5-7B}$	2.80	12.57	23.91	84.59	**0.81**
ASESUM$_{gpt-4o-mini}$	2.68	**12.80**	**24.18**	**85.28**	0.80
			Shoes		
GPT-4o-mini	**3.75**	**13.23**	21.46	42.73	0.47
HERCULES	1.80	12.06	24.35	84.45	0.72
ASESUM$_{qwen2.5-7B}$	2.14	11.41	25.30	92.72	**0.75**
ASESUM$_{gpt-4o-mini}$	2.01	11.09	**27.09**	**95.28**	0.72
			Sports & Outdoors		
GPT-4o-mini	**2.98**	12.68	20.69	44.68	0.47
HERCULES	1.72	11.45	**24.85**	86.22	**0.86**
ASESUM$_{qwen2.5-7B}$	2.20	12.67	24.79	87.27	0.82
ASESUM$_{gpt-4o-mini}$	2.65	**12.95**	24.81	**89.15**	**0.86**
			Home & Kitchen		
GPT-4o-mini	**2.74**	12.07	20.62	43.62	0.55
HERCULES	2.26	11.35	23.31	83.24	0.81
ASESUM$_{qwen2.5-7B}$	2.45	12.59	**24.10**	87.10	**0.87**
ASESUM$_{gpt-4o-mini}$	**2.74**	**12.80**	23.66	**87.38**	0.86
			Average		
GPT-4o-mini	**3.10**	12.34	20.89	44.68	0.51
HERCULES	2.05	11.83	23.85	83.43	0.78
ASESUM$_{qwen2.5-7B}$	2.40	12.31	24.53	87.92	0.81
ASESUM$_{gpt-4o-mini}$	2.52	**12.41**	**24.94**	**89.27**	**0.81**

Table 3: Results for automatic evaluation on review summarization. ROUGE-2 and ROUGE-L F1 scores are computed against the reference summaries. SC_{ref} and SC_{in} indicate the consistency (measured using SummaC) of generated summaries against reference summaries and input reviews, respectively. Our proposed *Diversity* measures the sentence-level diversity of the final generated summaries. Bold denotes the best score per domain.

SummaC scores on these domains compared to the others. This reveals that a summary could attain a high SummaC score by repeating opinions closely aligned with the input documents, even if such a summary may not be considered helpful in a real-life setting.

5.2 Qualitative Analysis

In addition to the numerical results in Table 3, we randomly select one example product from the *Home & Kitchen* domain to discuss qualitative aspects of our generated summaries. As shown in Figure 2, we notice that our generated results are significantly more faithful to the original reviews. This is because HERCULES decodes the summary from a hierarchical discrete latent embedding space, which strongly relies on its pre-trained codebook that performs the mapping from the discrete code to continuous embeddings (Hosking et al., 2023). However, since the codebook is pre-trained on the training set, for an uncommon product in the training set, their model would struggle to encode the reviews properly and decode the relevant information accordingly. This is also justified by the unsatisfying performance of HERCULES in the *Electronics* domain, where the types of products are more diverse than in other domains. On the contrary, our summaries maintain the topic at hand and minimise the likelihood of hallucination as we only apply abstractive summarization in the

Great Peeler. This product is a joke. Love this *ice crusher*. Not too heavy, not too light. *Easy to peel off*. *Keeps my coffee hot for hours*. This *ice crusher* works great. The lids fit snug. The plastic is very thin and flimsy. *Crushes ice* very well. *Love this water bottle*!

(a) HERCULES

One tray shattered the first time we used it. I like the fact these have lids. Very easy to pop out the ice cubes. cubes end up being a little small. Lids don't stay closed at all. Lids are nice to help keep the water in the trays when transferring from the sink to the freezer and for stacking while they make ice. Ice cubes are small. trays are very small, not easy to use as ice is hard to remove and there is only enuf ice per tray for one small glass. They stack great.

(b) ASESUM$_{\text{qwen2.5-7B}}$

Very easy to pop out the ice cubes. one tray broke. I like the fact these have lids. the lids do not stay on. the size of the cubes, they seem much smaller than a standard ice cube tray. cubes end up being a little small. Cubes could be a little larger. Each one comes with a lid so it's easy to stack. the silicone bottom makes them pop out with absolutely no effort.

(c) ASESUM$_{\text{gpt-4o-mini}}$

Figure 2: Example generated summaries from HERCULES and ASESUM with LLMs, for a randomly selected product (ice-tray).

initial aspect-centric argument extraction step. In addition, by comparing the textual summary from ASESUM$_{\text{qwen2.5-7B}}$ and ASESUM$_{\text{gpt-4o-mini}}$, we observe that evidence extracted by GPT-4o-mini is summarized to be more concise, which may lead to a lower diversity score for some domains.

6 Conclusion

This paper presents a novel summarization framework that integrates aspect-based sentiment analysis with argument mining to extract aspect-centric arguments for generating diverse yet faithful summaries. Although evaluating arguments based on their controversy level may not be the most ideal solution, our approach obtains strong performance on a benchmark dataset in both numerical and qualitative evaluations. Furthermore, by combining both extractive and abstractive summarization techniques, we demonstrate strong generalisability of our framework through automated aspect generation, the incorporation of multiple LLMs and domain-independent summarization.

Our approach relies on the dynamic extraction of relevant aspects and sentiments towards these aspects. We are planning to use these aspects to generate summaries as part of our future work. We will also conduct user studies to find meaningful ways to present the summary together with this aspect sentiment structure. Future research should also focus on finding ways to automatically evaluate structured summaries, which remains as a challenging problem for the community.

Limitations

ASESUM framework can be easily adapted to other domains and incorporated with other language models; however, we have a number of hyperparameters set to run the clustering algorithm. The consistent performance of our framework across four domains suggests the generalisability of this set of chosen parameters, but it may require more adjustments when adapting to new datasets. Besides, since our summaries are generated by concatenating pieces of evidence from different arguments, they may lack coherence in general.

Acknowledgment

This work was supported by the University of Edinburgh-Huawei Joint Lab grants CIENG4721 and CIENG8329.

References

R. M. Alguliyev, R. M. Aliguliyev, N. R. Isazade, A. Abdi, and N. Idris. 2019. COSUM: Text summarization based on clustering and optimization. *Expert Systems*, 36(1):e12340.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Domini-

can Republic. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

R. C. Belwal, S. Rai, and A. Gupta. 2021. A new graph-based extractive text summarization using keywords or topic modeling. *Journal of Ambient Intelligence and Humanized Computing*, 12:8975–8990.

Ravali Boorugu and G. Ramesh. 2020. A survey on nlp based text summarization for summarizing product reviews. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 352–356.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Luis Henrique Herbets de Sousa, Guilherme Trajano, Analúcia Schiaffino Morales, Stefan Sarkadi, and Alison R. Panisson. 2024. *Using Chatbot Technologies to Support Argumentation*. SciTePress.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. *Preprint*, arXiv:2305.11603.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Miao Li, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025. Aspect-aware decomposition for opinion summarization. *Preprint*, arXiv:2501.17191.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Jack Mumford, Stefan Sarkadi, Katie Atkinson, and Trevor Bench-Capon. 2024. Applying Argument Schemes for Simulating Online Review Platforms. In Chris Reed, Matthias Thimm, and Tjitze Rienstra, editors, *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21, Valencia, Spain. Association for Computational Linguistics.

An Tang, Xiuzhen Zhang, Minh Dinh, and Erik Cambria. 2024. Prompted aspect key point analysis for quantitative review summarization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10691–10708, Bangkok, Thailand. Association for Computational Linguistics.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.

Wyner Adam, Schneider Jodi, Atkinson Katie, and Bench-Capon Trevor. 2012. Semi-Automated Argumentative Analysis of Online Product Reviews. In *Frontiers in Artificial Intelligence and Applications*. IOS Press.

A Appendix A

We provide the prompt used in our paper in Table 4.

Fill the scheme with the provided review. {**Review Argumentation Scheme**} Note: 1. Identify the aspects mentioned in the review. Then provide a new scheme with the relevant evidence for each identified aspect. 2. The most mentioned aspects are {**aspect**}. 3. Only generate a new aspect when there is no matching one above. 4. Do NOT provide scheme having aspect wasn't mentioned in the text. 5. Do NOT include too much details in the evidence. Please return the values in JSON format: [{"aspect": "the property / feature of the product", "sentiment": "positive/negative", "evidence": "support from the argument"}, ...]

Table 4: Prompt provided to $\text{ASESUM}_{\text{gpt-4o-mini}}$ and $\text{ASESUM}_{\text{qwen2.5-7B}}$, where "Review Argumentation Scheme" is the placeholder to fit in the RAS (Table 1) and the "aspect" is the placeholder to interactively input the most popular aspects we have in the current aspect set.