Contrastive Summarization of User Reviews: An Aspect-based Abstractive Approach

Hung-Manh Hoang*, **Duc-Loc Vu***, **Huong Nguyen-Thi-Thuy***,
Duy-Cat Can and **Hoang-Quynh Le**†
VNU University of Engineering and Technology,
144 Xuan Thuy Street, Cau Giay, Hanoi, Vietnam
{21020522, 21020927, 21020467, catcd, lhquynh}@vnu.edu.vn

Abstract

Contrastive summarization involves generating summaries for two entities to highlight their differences. Although transformer-based abstractive summarization methods are powerful and effective for general summarization tasks, they often fall short in handling the diversity of aspects and viewpoints required for contrastive summarization. In this paper, we introduce a novel architecture that integrates an aspect classification method with an abstractive contrastive summarization model, allowing for comparisons based on predefined relevant aspects. Experiments conducted on the CoCoSum dataset demonstrate the effectiveness of our proposed method, achieving competitive results compared to other models that account for both common and contrastive summaries.

1 Introduction

Contrastive summarization focuses on creating summaries for two entities, such as products, with the specific goal of highlighting their differences (Lerman and McDonald, 2009). This approach has become necessary due to the demand for nuanced comparisons from various viewpoints that users encounter among numerous options. As Paul et al. (2010a) noted, diverse opinions frequently result in contrasting perspectives. A perspective, or viewpoint, is defined as "a mental position from which things are viewed" (cf. WordNet). Figure 1 shows an example of generating contrastive and common summaries from two sets of user reviews. In the context of online reviews, contrastive summarization helps users avoid visiting multiple sources, reading numerous comments, and performing time-consuming manual comparisons by summarizing entities across different viewpoints within the reviews.

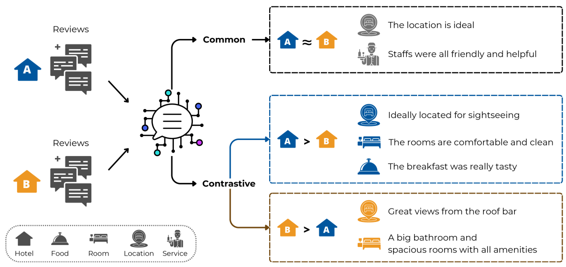

Figure 1: An example of Generating both contrastive and common summaries from two sets of user reviews

Contrastive summarization is essentially a specialized problem within document summarization. Currently, there is growing interest in abstractive summarization approaches due to their ability to produce summaries that are both concise and closely aligned with natural human language (Gupta and Gupta, 2019). Among these abstractive summarization techniques, more recent research on abstractive summarizing has been inspired by the Transformer framework (Guan et al., 2020). However, when applied to contrastive summarization, these methods exhibit certain drawbacks.

Transformer-based model often truncate long text to fit length limits, leading to fragmented context (Guan et al., 2020). This is especially problematic for multi-document, multi-opinion summarization, as it can distort or omit important viewpoints. Additionally, while effective for general summarization, transformer-based models may fail to capture subtle contrasts between viewpoints, resulting in overly broad or generalized. Finally, despite significant progress in the field of contrastive summarization, there has been limited exploration of abstractive methods (Ströhle et al., 2023), particularly those based on transformer-based models. This gap not only makes it challenging to leverage the full potential of recent advancements in natural language processing but also limits the application of contrastive summarization in more complex and

* Contributed equally
†Corresponding author

nuanced contexts.

To address this challenge, we propose a novel method that leverages transformer-based models for aspect-based contrastive summarization. Our approach aims to generate both common and contrastive summaries between entities, capturing their shared and distinct characteristics comprehensively at the aspect level. This method provides targeted insights into specific aspects and enhances understanding by clearly highlighting differences between sources. Additionally, it produces summaries that are both flexible and human-like, rephrasing and condensing information while preserving the essence of the original text. This results in a more informative, comprehensive, and comparative view for users.

The main contributions of this work are:

- We present a novel approach to summarizing reviews that leverages advanced deep learning techniques. This method offers a detailed and comparative perspective, significantly enhancing the overall understanding of the reviews.

- Our experiments conducted on the CoCoTrip dataset illustrate the effectiveness of our proposed method, achieving competitive performance relative to other models that that account for both common and contrastive summaries.

2 Related Work

Contrastive summarization was first introduced by (Lerman and McDonald, 2009). Despite growing interest in the topic, there is a lack of standardized datasets and dedicated competitive tasks, which hinders the development of new methods. Additionally, the significant advancements seen with deep learning-based language models for abstractive summarization have not yet been fully realized in the field of contrastive summarization (Ströhle et al., 2023).

Wang et al. (2013) developed a comparative extractive summarization technique, which focuses on extracting and contrasting the most distinctive sentences between comparable document groups. Similarly, (Kim and Zhai, 2009) proposed a model for summarizing contradictory opinions by generating summaries that contrast positive and negative opinion sets. (Paul et al., 2010b) further advanced this area by using a two-stage method involving topic extraction with LDA and a modified PageRank algorithm for summarizing contrasting viewpoints.

More recently, (Iso et al., 2022) expanded on these ideas with their work on Comparative Opinion Summarization, which generates two contrastive summaries and one common summary from distinct sets of reviews, using a method called co-decoding. This approach contrasts token probability distributions for the contrastive summaries while aggregating them for the common summary.

(Gunel et al., 2023) introduced STRUM, an innovative method for extractive aspect-based contrastive summarization, designed to aid in making comparative decisions without relying on human-written summaries or fixed aspect lists. It uses two fine-tuned T5-based models—one for aspect and value extraction, and the other for natural language inference—to generate structured summaries that contrast different choices.

Our work aligns with (Iso et al., 2022) as it also aims to produce contrastive and common summaries from review sets. However we uses a pipeline involving aspect classification, sentiment classification, and heuristic filtering, followed by a fine-tuned BART for summary generation. This structured method mirrors how a human would analyze and compare reviews, systematically breaking down the information and then synthesizing it into a summary.

3 Methods

In this section, we detail the multi-component approach developed to achieve high-quality summarization of the given content. The process is divided into three main components: Aspect and Sub-aspect Classification, Sentiment Classification, and Heuristic Filtering. Each of these components plays a critical role in refining the input data and ensuring that the generated summaries are both informative and contextually relevant. The overall architecture of the model is shown in the Figure 2

3.1 Aspect and sub-aspect classification

3.1.1 Dictionary Construction

The goal of this step is to construct a sub-aspect dictionary for effective perform sentence-level aspect classification of user reviews. The process involves multiple stages to ensure a comprehensive and accurate dictionary.

First, we employ SetFitABSA Aspect Model

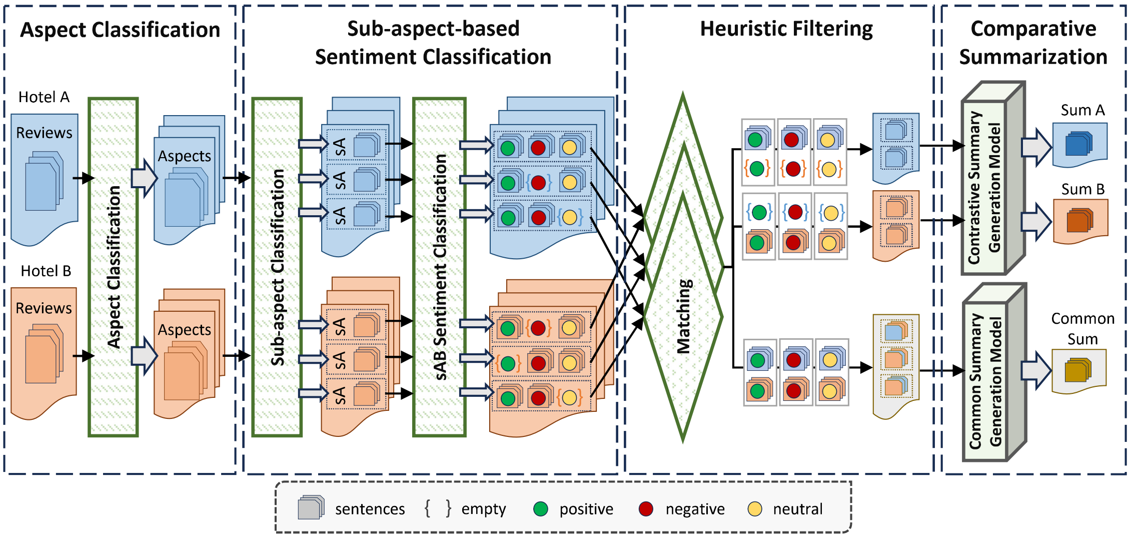

Figure 2: Overview of the architecture for contrastive summarization, comprising Aspect Classification, Sub-aspect-based Sentiment Classification, Heuristic Filtering, and Comparative Summarization. Sentences are processed through multiple stages to generate both comparative and common summaries, with sentiment analysis and heuristic matching ensuring relevance and accuracy in the final output

proposed by Tunstall et al. (2022) to identify the relevant aspects from the user reviews. This step allows us to capture the various aspects that users discuss in their reviews. Next, we utilize ChatGPT[1] with GPT-4o model to classify these extracted aspects into predefined categories automatically. This automated classification step helps in organizing the aspects into broader categories efficiently. Following this, we manually define sub-aspect categories to ensure finer granularity. We then carefully read through the dataset, selecting specific words and phrases that belong to each sub-aspect category. This manual intervention ensures that the dictionary is closely aligned with the context of the reviews. Finally, we perform data augmentation using WordNet. By expanding the dictionary with synonyms and related words from WordNet, we enhance the coverage and robustness of the sub-aspect dictionary. This ensures that the dictionary can capture variations in language and terminology across different reviews.

3.1.2 Aspect and sub-aspect classification

The goal is to assign relevant sub-aspects to each segment of the review text and prepare input for the subsequent sentiment classification. This ensures that the sentiment analysis is focused on specific aspects, leading to more precise and context-aware sentiment predictions.

Using the aspect and sub-aspect dictionary, we implement a classification method that identifies and categorizes aspects within the user reviews. This method leverages the dictionary to match review segments to corresponding sub-aspects, ensuring that the classification aligns with the predefined categories and sub-categories. Figure 3 illustrates the complete process of dictionary construction and aspect classification, outlining each step from aspect extraction and sub-aspect categorization to data augmentation, demonstrating how the dictionary is utilized for aspect classification.

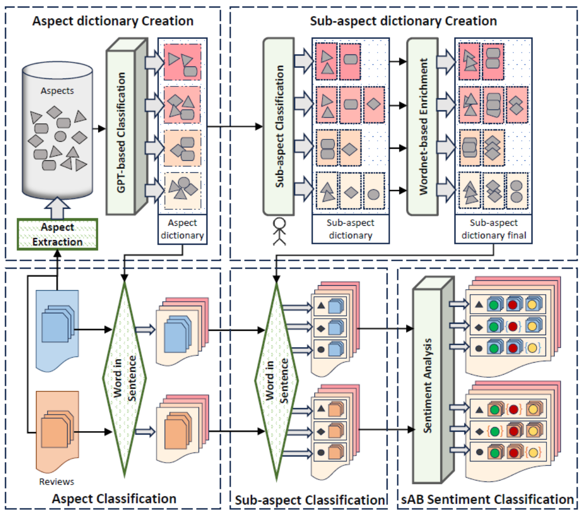

Figure 3: The process of Dictionary Construction and Aspect + Sub-aspect Classification

In the sentiment classification task, sub-aspect

[1] https://chatgpt.com/

Algorithm 1 Heuristic filtering algorithm

Let *a_rvs* and *b_rvs* are dictionaries where:
  The keys are *sub_aspect_sentimented*
  The values are lists of *sentences*
$common \leftarrow []$
$contrast\_a \leftarrow []$
$contrast\_b \leftarrow []$
for sa in sub_aspect_classified_list **do**
  if both $a\_rvs[sa]$ and $b\_rvs[sa]$ not empty **then**
    Append $a\_rvs[sa]$ to $common$
    Append $b\_rvs[sa]$ to $common$
  else if $a\_rvs[sa]$ empty and $b\_rvs[sa]$ not empty **then**
    Append $b\_rvs[sa]$ to $contrast\_b$
  else if $a\_rvs[sa]$ not empty and $b\_rvs[sa]$ empty **then**
    Append $a\_rvs[sa]$ to $contrast\_a$
  end if
end for

Figure 4: Heuristic filtering algorithm

classified sentences are analyzed to determine their emotional tone using a sentiment analysis model, such as the Twitter-roBERTa-base (Loureiro et al., 2022) (Camacho-Collados et al., 2022). This model classifies each sentence into one of three sentiment categories: positive, negative, or neutral. The sentiment classification process involves inputting the pre-processed sentences into the model, which outputs a probability distribution across the sentiment classes. Formally, let S_i be the i-th sentence, and let C be the set of sentiment categories $\{\text{positive}, \text{negative}, \text{neutral}\}$. The sentiment probability distribution for S_i is given by $p(C \mid S_i)$, where $p(c \mid S_i)$ denotes the probability of sentiment class c for the sentence S_i. The classification result is the sentiment class $\hat{c}$ that maximizes $p(c \mid S_i)$, i.e., $\hat{c} = \arg\max_{c \in C} p(c \mid S_i)$.

3.2 Heuristic Filtering

After classifying the sentences for each sub-aspect into positive, negative, and neutral categories, we apply heuristic filtering to prepare the input for the abstractive summarization model (BART). The details of the algorithm are provided in Figure 4.

3.3 Model BART

BART (Bidirectional and Auto-Regressive Transformers) (Lewis et al., 2019) is a sequence-to-sequence (seq2seq) model that combines the strengths of a bidirectional encoder, similar to BERT, and an autoregressive decoder, akin to GPT. The model undergoes pre-training through a two-step process: first, the input text is corrupted using an arbitrary noising function, and then the model is trained to reconstruct the original text from the corrupted input.

BART has demonstrated exceptional performance in tasks requiring text generation, such as summarization and translation, and is also effective in text comprehension tasks, including text classification and question answering. In this study, we utilize a specific checkpoint of BART that has been fine-tuned on the CNN/Daily Mail dataset, which comprises a large corpus of paired text and summaries, to enhance its performance on summarization tasks.

4 Results and Discussion

The experiments were performed using the CoCoTrip dataset (Iso et al., 2022), which comprises 768 reviews organized into 48 pairs of hotels. For dataset benchmarking, we employed ROUGE-1, ROUGE-2, and ROUGE-L F1 scores as automatic evaluation metrics based on reference summaries. To assess the distinctiveness of the generated summaries, we computed the average Distinctiveness Score (DS) between the generated contrastive summaries and the common summaries for all entity pairs as defined in (Iso et al., 2022).

4.1 Overal Performance and Comparisons

To evaluate the performance of our model, we compare our experimental results with those of several well-known models on the same dataset. *(i) Extractive summarization comparative models* include LexRank$_{TFIDF}$ (Erkan and Radev, 2004) and LexRank$_{BERT}$ - LexRank with Sentence-BERT embeddings (Reimers and Gurevych, 2019), two classic unsupervised opinion summarization models. (ii) *Abstractive summarization comparative models* include: MeanSum (Chu and Liu, 2019), an unsupervised model designed for single-entity opinion summarization; CopyCat (Bražinskas et al., 2020), a single-entity opinion summarization model based on leave-one-out reconstruction; BiMeanVAE (Iso et al., 2021b), an optimized variant of MeanSum for single-entity opinion summarization; and CoCoSum (Iso et al., 2022), which incorporates a few-shot learning approach with collaborative decoding and achieves state-of-the-art

Table 1: ROUGE scores for contrastive and common summaries on COCOTRIP and the distinctiveness score (DS) of generated summaries

	Contrastive summarization			Common summarization			Pair
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	DS
Extractive models							
LexRank $_{\text{TFIDF}}$ *	35.38	7.39	18.25	22.51	4.00	15.26	63.28
LexRank $_{\text{Bert}}$ *	32.65	5.67	16.67	17.91	2.95	12.60	65.56
Abstractive models							
MeanSum*	34.19	7.84	19.76	13.09	0.85	10.41	65.98
CopyCat*	35.30	8.39	18.64	36.16	11.91	25.15	40.80
BiMeanVAE*	37.44	9.41	22.02	38.47	14.17	27.46	42.55
Cocosum *	42.22	12.11	24.13	**46.80**	**20.68**	35.62	**74.02**
Our approach	**44.16**	**13.26**	**24.31**	46.74	20.34	**36.12**	63.89

**Provided by (Iso et al., 2022)*
The highest number in each column is highlighted in bold.

performance. The results of these approaches re-implemented on the CoCoSum dataset are reported in (Iso et al., 2022).

As mentioned in Table 1, our model outperforms the comparative models, including the CoCoSum model, which combines few-shot learning with collaborative decoding to generate both contrastive and common summaries. For contrastive summaries, CoCoSum achieved ROUGE-1, ROUGE-2, and ROUGE-L scores of 42.22, 12.11, and 24.13, respectively. In contrast, our model outperformed CoCoSum with ROUGE-1, ROUGE-2, and ROUGE-L scores of 44.16, 13.26, and 24.31, showing improvements of 1.94%, 1.15%, and 0.18% in each metric. This indicates our model's superior ability to capture fine-grained contrasts between entities, an essential aspect for applications that require distinguishing subtle differences between subjects.

Regarding common summaries, CoCoSum recorded ROUGE-1, ROUGE-2, and ROUGE-L scores of 46.80, 20.68, and 35.62, respectively. Our model delivered comparable ROUGE-1 and ROUGE-2 scores of 46.74 and 20.34, with a slightly higher ROUGE-L score of 36.12, reflecting differences of -0.06%, -0.34%, and +0.50%, respectively. Although the differences in ROUGE-1 and ROUGE-2 are minimal, the improved ROUGE-L score suggests our model's enhanced ability to maintain the structural coherence and fluency of the generated summaries. These results suggest that while both models are competitive, our approach offers a marginal advantage in balancing contrast extraction with the preservation of commonalities, particularly in generating cohesive and well-structured summaries. This balance is crucial in tasks where both distinctiveness and commonality need to be conveyed effectively, demonstrating the robustness of our model in varied summarization scenarios.

The Distinctiveness Score (DS), introduced by Iso et al. (2021a), measures the contrast between two contrastive summaries and a common summary based on lexical overlap. This score, scaled from 0 to 100, indicates greater contrast with lower token overlap, corresponding to higher DS values. However, it is important to note that the DS does not exclusively measure the contrast between the two contrastive summaries themselves; rather, it evaluates the relationships among all three summaries based on lexical overlap. Consequently, a high DS value may not necessarily signify a strong contrast solely between the contrastive summaries. Furthermore, since our approach emphasizes summarizing contrastiveness at the aspect level, some degree of overlap in the summaries is to be expected. As a result, despite capturing contrastiveness in our output, the DS metric may appear lower. Figure 5 illustrates an example where the DS metric may yield a low score despite the clear contrastiveness between two reviews that address the same aspect (Food) but different sub-aspects (taste and presentation). Additionally, the figure shows another case where the DS metric scores even lower when contrastiveness is conveyed through negative statements.

Common: AB : "The hotel is clean"

Contrast: A : "The food is tasty"

B : "The food is nice decorated"

DS metric: 62.5

Common: AB : "The hotel is clean"

Contrast: A : "The food is tasty"

B : "The food is not tasty"

DS metric: 50

Figure 5: Drawback of DS at contrastive summarization problem.

4.2 Model Components Contribution

We examine the impact of the main components of the proposed model on overall system performance by systematically removing each component and evaluating the model on a test set. We then compare these results with the performance of the complete system, showcasing the variations in ROUGE-1, ROUGE-2, and ROUGE-L F-scores in Figure 6. The observed changes in F-scores reveal that each component plays a role in boosting system performance, although the degree of their contributions differs across components and metrics. Notably, most components have a greater influence on common summarization than on contrastive summarization.

Aspect and Sub-aspect Classifier focuses on identifying and categorizing various aspects and sub-aspects within the input data, facilitating the creation of more structured and pertinent summaries. Excluding this component leads to a notable decline in model performance, particularly in standard summarization tasks, with reductions of 8.05%, 5.59%, and 6.78% in ROUGE-1, ROUGE-2, and ROUGE-L F-score, respectively. The decrease in performance for contrastive summarization is less pronounced compared to common summarization, with a 0.77% drop (around 2% of original results) in ROUGE-2 F-score and a 0.65% drop (about 3% of original results) in ROUGE-L F-score. However, removing this component results in a slight, though not significant, increase in ROUGE-2 F-score (0.16%).

Sentiment Classification concentrates on evaluating the sentiment (positive, negative, neutral) of the identified sub-aspects. Sentiment classification allows the model to capture differing opinions or sentiments that contribute to contrastive summaries, ensuring that the nuances of opposing views are clearly represented. Removing this component leads to the most significant decline in performance for both common and contrastive summarization tasks, excluding the ROUGE-L score for the common summary, making it the most impactful layer of the entire pipeline. Its removal underscores the critical role this component plays in maintaining the overall effectiveness of the model. The reduction in ROUGE-1, ROUGE-2, and ROUGE-L F-scores for common summarization tasks—10.51%, 7.92%, and 8.40%, respectively—underscores how critical this component is to the model's effectiveness. Even in contrastive summarization, where the performance drop is less dramatic, it still leads to the largest decline across all tested configurations, with decreases of 1.62%, 0.78%, and 1.53% in ROUGE-1, ROUGE-2, and ROUGE-L F-scores, respectively. This highlights the vital role that the component plays in ensuring the accuracy, coherence, and relevance of both common and contrastive summaries.

Heuristic Filtering applies predefined rules and heuristics to classify sentences, ensuring that the input provided to the abstractive summary model contains only the most relevant information. This step is crucial for enhancing the quality of the generated summaries by focusing on the content that matters most. In common summaries, removing heuristic filtering follows a similar trend to the previous configurations, resulting in substantial declines in ROUGE scores: 8.40% in ROUGE-1, 5.66% in ROUGE-2, and a significant 8.61% drop in ROUGE-L, the highest among the three. However, the trend shifts slightly in contrastive summaries. Interestingly, there's a minor, though not particularly notable 0.5% increase in ROUGE-1 F-score. Without heuristic filtering, contrastive summaries are generated from a single set of reviews (e.g., summarizing A without B to create a contrastive summary for A > B). This simplification turns the task into a more traditional text summarization problem, which may explain the consistent ROUGE scores, despite the lack of other components. Nevertheless, excluding this component still results in decreases in ROUGE-2 and ROUGE-L F-scores, by 0.07% and 0.82%, respectively.

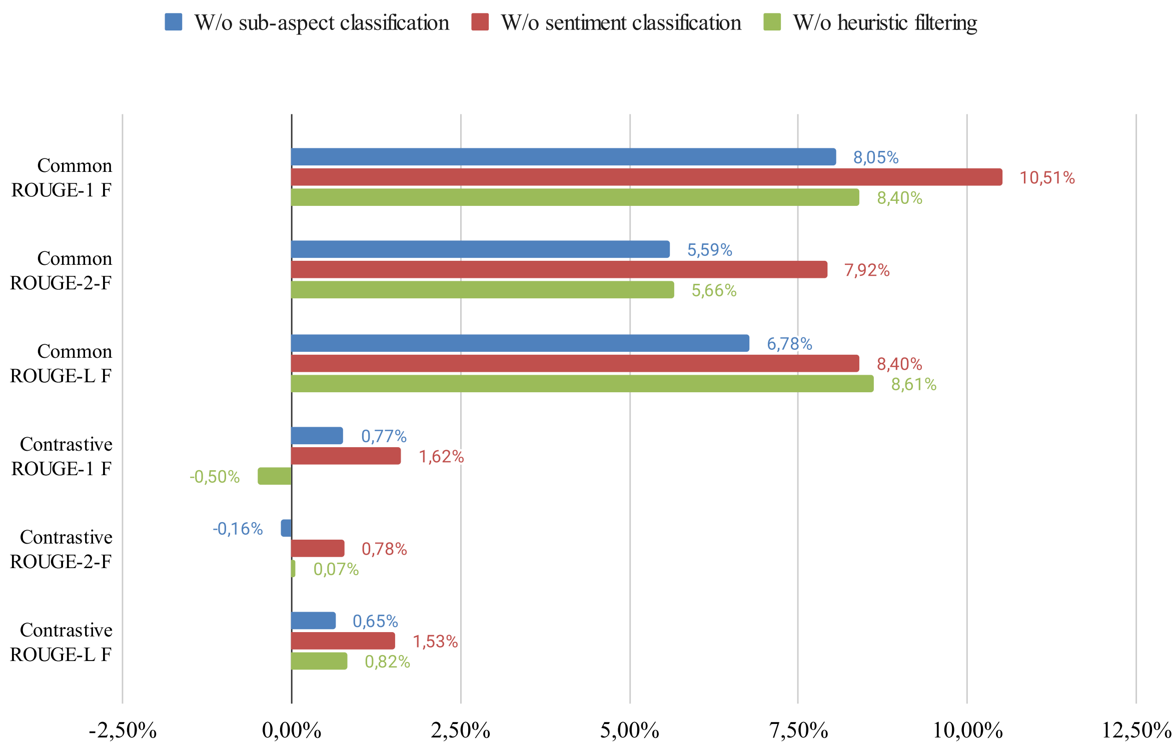

Figure 6: Impact of removing components on ROUGE Scores for common and contrastive summaries.

4.3 Qualitative Analysis

CoCoSum, leveraging the few-shot approach and collaborative decoding technique, represents the cutting edge in contrastive summarization. To thoroughly evaluate the quality of its generated summaries, we carried out two analysis, focusing on different aspects of the outputs.

4.3.1 Bias Assessment

The summary generated by our model provides a more balanced representation of both positive and negative reviewer feedback, in contrast to CoCoSum's summary, which displays a clear bias towards positive reviews. This is evident from the data in Table 2. Notably, the CoCoSum model missed negative reviews about the rooms being dark and small, as well as complaints regarding staff tipping, whereas our model effectively captured these details. This difference may be attributed to the effectiveness of our sentiment classification components.

4.3.2 Sentence Duplication Analysis

The CoCoSum model occasionally produces summaries that include repetitive sentences. In contrast, our model merges redundant information into a single sentence. This improved performance may be due to the efficiency of our aspect and sub-aspect classification components. For instance, in Table 3, CoCoSum mentions the hotel's proximity to the metro and the center of Paris twice using different phrases, while our model succinctly combines this information into one sentence.

5 Conclusion

In this paper, we introduced a novel approach to contrastive summarization that effectively integrates aspect classification, sentiment classification, and heuristic filtering with an abstractive summarization model. Our experiments on the CoCoTrip dataset demonstrated the efficacy of our method, showing competitive performance compared to existing models. Through an ablation study, we highlighted the importance of each component in our pipeline, particularly how aspect and sentiment classification significantly enhance the relevance and quality of the generated summaries. This work not only advances the field of contrastive summarization but also provides a framework that can be adapted for more complex summarization tasks. Future research could explore further refinements in aspect classification and the potential integration of more sophisticated sentiment analysis techniques to further improve summarization quality.

Table 2: Examples of sumamry generated from CoCosum and our model

Review 1:	Ideally located, within minutes of the Blue Mosque, Grand Bazaar etc and in the heart of the old City of Istanbul, close to tram and autobus and the main street in Sultanahmet. ... The bedroom was a little dark and not a lot of space if you have large suitcases, we were here for 7 nights, plenty to chose from at breakfast, although pretty much the same every day, lots of fruit, cheeses, breads, choice of eggs, teas and coffees etc. Overall a fantastic boutique style hotel, THANK YOU CANER, regards, Lyn & Shahbaz
Review 2:	We spent there 3 nights (celebrated my birthday)- and liked everything: very clean and cosy rooms, breakfast on open terrace with breathtaking views of the city&sea. ... I'm glad that we stayed at Mina hotel. Absolutely recommended!
Review 3:	Excellent situation close to blue mosque at the heart of Sultan Ahmet district. We were lucky to get a good deal in high season. ... Wonderful night view. The reception staff were polite and room service was acceptable. They also offered us free shuttle service back to the airport.
Review 4:	i liked this hotel a lot.the reception was perfect, the room good but not spacey(no 208 or 209 i think.also the management should think to put a brush in the toilet for the .you know. there was no brush.a last thing that i d like to say is that the staff of the restaurant wasnt very smiley with you if you didnt leave a tip.
Review 5:	Great location 5 minutes walk to the Grand Bazzar and many other historic and traditional places in Sultanahmet. ... The only disadvantage we faced is that there is no wireless internet in the rooms, only in the looby.
Review 6:	Lovely hotel in the heart of Sultanahmet with spectacular views from the roof terrace. ... I would recommend this hotel to anyone, and i would definately stay at the Mina hotel again.
Review 7:	This hotel is in the old quarter of Istanbul-it is short walk to all the tourist spots. ... Reception staff are ready to answer any questions you have about the area.We would definitely go back to the Hotel Mina again!
Review 8:	Not a negative to this place. Easy stay, with a breakfast fit for a king!
CoCoSum	The hotel is in an excellent situation close to the Blue Mosque, Grand Bazaar etc and in the heart of the old City of Istanbul, close to tram and autobus and the main street in Sultanahmet. Stunning views of the city andsea from the roof top terrace. Rooms are cosy but adequate for a short stay. The staff at the hotel were very pleasant and smiley. Free internet access in the reception area was an unexpected bonus.
Our Model	This hotel is ideally located in the old quarter of Istanbul and in the heart of the old City of Istanbul. It is close to the Blue Mosque, Grand Bazaar and the main street in Sultanahmet. The hotel is also close to tram and autobus and is within easy walk to all the tourist attractions in the city. The rooms are small but clean and cosy, with a cosy fireplace and comfortable beds. The room was dark and a little noisy but overall the rooms are clean and comfortable. The staff at the hotel were not very smiley if you didn't leave a tip

Table 3: Examples of sumamry generated from CoCosum and our model

CoCoSum	This hotel is an exceptional find in a lovely setting. Close to the metro station and within walking distance to the centre of Paris, it gave us a real sense of Paris without paying the high prices. The view from the room was all the way down the canal towards the rotunda. The metro station is only 5 minutes away and is a great way to get around. The staff at the hotel were really nice and didn't even say a simplesorry afterwards. Breakfast was adequate and there were hot options.
Our Model	This is a lovely hotel that is perfectly located for a romantic getaway. The hotel is close to the Stalingrad metro station and and within walking distance to the centre of Paris. The rooms in the hotel are very comfortable and the beds are big and the bathrooms are excellent. The staff at the hotel were very friendly and efficient and the room was spotless. The breakfast here at this hotel was really good but there were not enough seats in the restaurant so be ready to wait for some time during rush hour. There is a great view from the balcony overlooking the canal and the houseboats moored in the canal basin.

Acknowledgments

This work has been supported by VNU University of Engineering and Technology under project number CN24.15. We thank the anonymous reviewers for their comments and suggestions.

References

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.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez-Cámara. 2022. Tweetnlp: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. *Preprint*, arXiv:1810.05739.

G. Erkan and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Wang Guan, Ivan Smetannikov, and Man Tianxing. 2020. Survey on automatic text summarization and transformer models applicability. In *Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System*, pages 176–184.

Beliz Gunel, Sandeep Tata, and Marc Najork. 2023. Strum: Extractive aspect-based contrastive summarization. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 28–31, New York, NY, USA. Association for Computing Machinery.

Som Gupta and Sanjai Kumar Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2021a. Comparative opinion summarization via collaborative decoding. *arXiv preprint arXiv:2110.07520*.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022. Comparative opinion summarization via collaborative decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021b. Convex Aggregation for Opinion Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3885–3903, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyun Duk Kim and ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 385–394.

Kevin Lerman and Ryan McDonald. 2009. Contrastive summarization: an experiment with consumer reviews. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: Short papers*, pages 113–116.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Michael Paul, ChengXiang Zhai, and Roxana Girju. 2010a. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 66–76.

Michael Paul, ChengXiang Zhai, and Roxana Girju. 2010b. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76, Cambridge, MA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Thomas Ströhle, Ricardo Campos, and Adam Jatowt. 2023. Contrastive text summarization: a survey. *International Journal of Data Science and Analytics*, pages 1–15.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2013. Comparative document summarization via discriminative sentence selection. *ACM Trans. Knowl. Discov. Data*, 7(1).