ISBN: 978-981-94-4198-3
2025 the 15th International Workshop on Computer Science and Engineering (WCSE 2025)
doi: 10.18178/wcse.2025.06.010

Product Review Summarization System Extracting Evaluation Criteria and Visualizing with Supporting Sentences

Tanzan Rikuya [1], Eiji Kamioka [1+], Chanh Minh Tran [1] and Phan Xuan Tan [1]

[1] Graduate School of Engineering and Science, Shibaura Institute of Technology, Japan

Abstract. With the widespread use of review websites, consumers increasingly rely on product reviews. However, when reviews are excessively long or numerous, the useful information obtained per unit of reading time decreases, increasing the burden on users. To address this issue, this study proposes a new product review summarization system comprising the Evaluation Criteria Extraction Module, the Polarity Analysis Module, and the Supporting Sentence Extraction Module. First, morphological analysis is applied to reviews to extract evaluation criteria with AI assistance. Next, the sentiment polarity of sentences containing these criteria is determined, and the proportion of positive and negative opinions is visualized using bar graphs. Additionally, relevant sentences are displayed below the graphs to retain essential information while improving review readability. Performance evaluation demonstrated that the proposed system effectively reduces the burden of reading reviews without significantly compromising the amount of information or user interest in the product.

Keywords: Product review, Natural language processing, LLM, BERT

1. Introduction

With the widespread use of review websites and online shopping, consumers increasingly rely on user-generated product reviews on the Internet when purchasing. Reviews serve as a crucial source of information for consumers, influencing their choices. According to a survey [1], 93% of users reported that online reviews influenced their purchase decisions, highlighting the importance of review sites.

One of the key advantages of review websites is that they allow users to gather a large amount of product-related information quickly if they have Internet access. However, this advantage can also become a drawback. When reviews are too long or a single product has too many reviews, the amount of key information obtained per unit of reading time decreases, resulting in a higher cognitive burden on users. Therefore, a review summarization system is needed to extract essential information and reduce the burden on users reading those reviews.

To address these issues, this study proposes a system that extracts key elements for product evaluation from information on review websites and presents them to users in an easily understandable format. In this paper, the construction methods of the proposed system are clarified, and whether the system can reduce the burden on users when obtaining necessary information is discussed.

2. Related Work

Various studies exist on visualization of product reviews. Joshi et al. [2] proposed a model to extract product features and opinions from Amazon reviews and determine their polarity. The proposed model eventually develops a UI using bubble charts and bar graphs, making Amazon reviews more intuitive and easier to understand.

Kamal [3] proposed a method that combines machine learning and natural language processing to analyse and visually represent user ratings of individual features of a product or service. This study uses the Google Chart API to visualize reviews by displaying ratings for each product feature in a bar chart or pie chart.

Sauper et al. [4] proposed a model that uses machine learning to estimate aspects such as “food” and “service” from restaurant reviews and the related sentiment. This study finally considers the presentation of reviews related to each aspect.

+ Corresponding author.
E-mail address: kamioka@sic.shibaura-it.ac.jp

These studies show that various methods for visualizing reviews exist. However, prioritizing visual clarity often results in insufficient information retention, while excessive organization of summarized reviews can lead to user fatigue from reading similar sentences repeatedly. Therefore, a system that effectively balances information retention and reduces cognitive load is necessary.

3. Proposed System

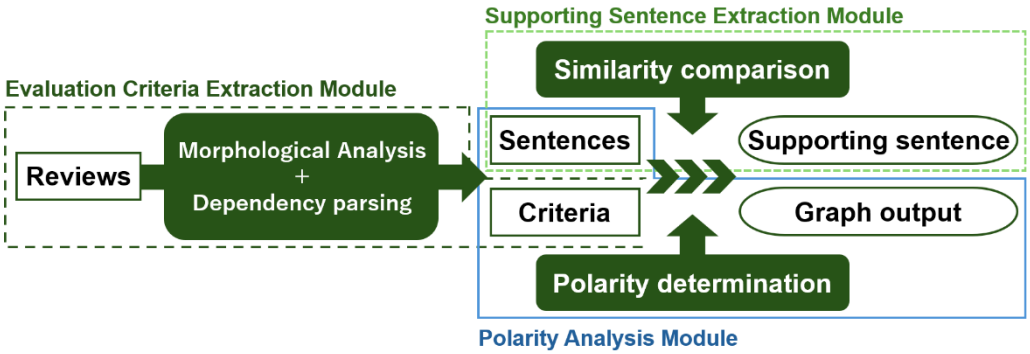


Fig. 1: Overview of Proposed System

This paper proposes a new system for summarizing product review websites to reduce the reading load for product purchasers. The system automatically extracts key evaluation criteria from review information, visualizes the summary, and provides minimal explanatory sentences as supporting evidence.

Figure 1 illustrates the overview of the proposed system. This proposed system comprises three main modules: the Evaluation Criteria Extraction Module, the Polarity Analysis Module, and the Supporting Sentence Extraction Module. The following parts provide detailed explanations of each module.

(1) Evaluation Criteria Extraction Module

The Evaluation Criteria Extraction Module breaks down into individual sentences, treating each as an evaluation sentence. Then, Morphological analysis and dependency parsing are applied to the subject of each evaluation sentence to extract evaluation criteria such as “durability” and “price.” In this study, MeCab [5] and GINZA [6] are utilized to perform Morphological analysis and handle dependency parsing, respectively. This decision comes from the finding that words related to evaluation criteria in reviews are often the subjects of sentences. However, relying solely on this approach results in many extraction errors and omissions. To solve this problem, the following techniques are introduced. First, a stop-word list is introduced, including words such as "shipping fee" and "review" which appear as sentence subjects but are not associated with product evaluation. Filtering out these words results in the avoidance of incorrect extractions. Second, AI-assisted evaluation criteria extraction is incorporated. Specifically, the ChatGPT API [7], fine-tuned through instruction tuning, enhances the extraction accuracy.

(2) Polarity Analysis Module

The Polarity Analysis Module analyzes the polarity of evaluation sentences to show the overall evaluation trend of the review. The polarity of a sentence is determined based on the total count of positive and negative words within it. In this study, Oseti [8], which is a sentiment analysis library that leverages a Japanese polarity dictionary, is used in this module. However, since Oseti relies on a dictionary-based approach, it cannot account for contextual nuances when determining polarity. To address this limitation, in this study, the program code is modified to enable context-aware evaluation for specific expressions commonly found in reviews. These expressions include sentences related to price and phrases such as "be easy to" or "be difficult to," which require contextual interpretation. Once polarity analysis is completed for all sentences within a product's reviews, the extracted evaluation criteria are used to aggregate evaluations for each criterion. The total counts of positive, negative, and neutral evaluations are then visualized using bar charts.

(3) Supporting Sentence Extraction Module

The Supporting Sentence Extraction Module selects a supporting sentence for each result of the polarity analysis. To do this, Sentence-BERT [9] is first used to calculate the cosine similarity between sentences in reviews that share the same evaluation criterion.

BERT is a natural language processing model proposed by Devlin et al. [10] based on Transformer architecture. Transformer is proposed by Vaswani et al. [11], which is a model for deep learning using only the Attention layer. The Transformer architecture enables a bi-directional understanding of word meaning by pre-training with a mask language model and a next-sentence prediction task. The Sentence BERT is an optimized model for computing sentence similarity based on the BERT.

This module groups particularly similar sentences based on the similarity scores it computes. After grouping, one sentence from each group is randomly selected and displayed below the bar chart. This means that an equal number of sentences as groups are shown for each evaluation criterion. This approach provides users with detailed information and its associated sentiment, offering insights that cannot be grasped from the bar charts alone. This method ensures that the research goal of reducing the cognitive load of reading reviews is achieved while still providing minimal but sufficient information.

Figure 2 illustrates the final output generated by the proposed system. In the bar chart, red represents positive reviews, blue represents negative reviews, and green represents neutral reviews.

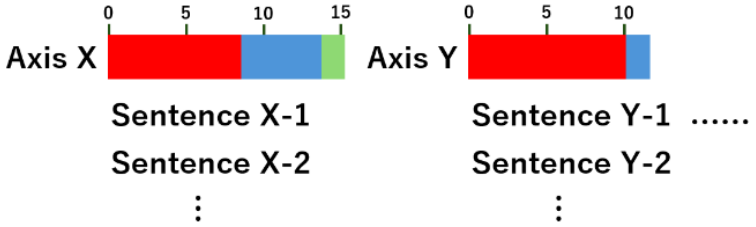
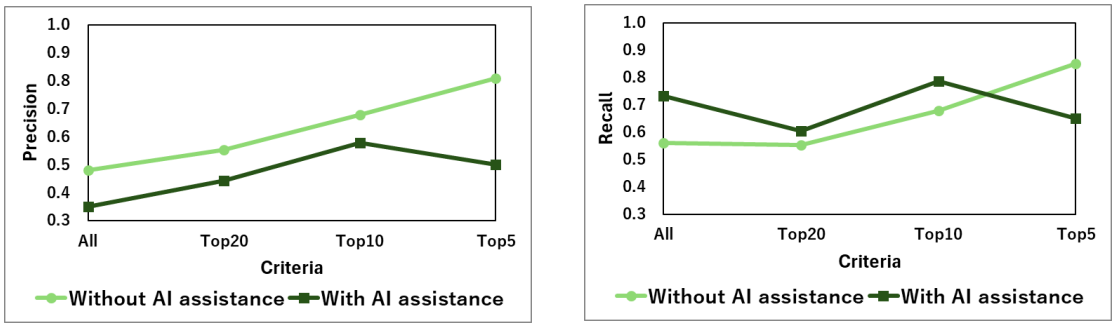


Fig. 2: Final Output Generated by Proposed System

4. Evaluation

First, to validate the effectiveness of AI assistance in the Evaluation Criteria Extraction Module, an evaluation experiment was conducted using a dataset of 22 air cleaner reviews. Figures 3 (a) and (b) show the Precision and Recall of the evaluation criteria extraction. The green and black lines show the results without AI assistance and with AI assistance, respectively. In both graphs, the horizontal axis represents the ranking of the extracted evaluation criteria up to which the top items are adopted. Figure (a) representing the Precision shows that there are fewer false detections without AI assistance compared to those with AI assistance. On the other hand, Figure (b) representing the Recall indicates that for ranks lower than the top 5, there are fewer undetected cases with AI assistance compared to those without AI assistance. From the perspective of users viewing product reviews, it is preferable not to overlook any evaluation criteria, and having more criteria makes it easier to decide on purchasing a product. Therefore, the effectiveness of the Evaluation Criteria Extraction Module with AI assistance is demonstrated.



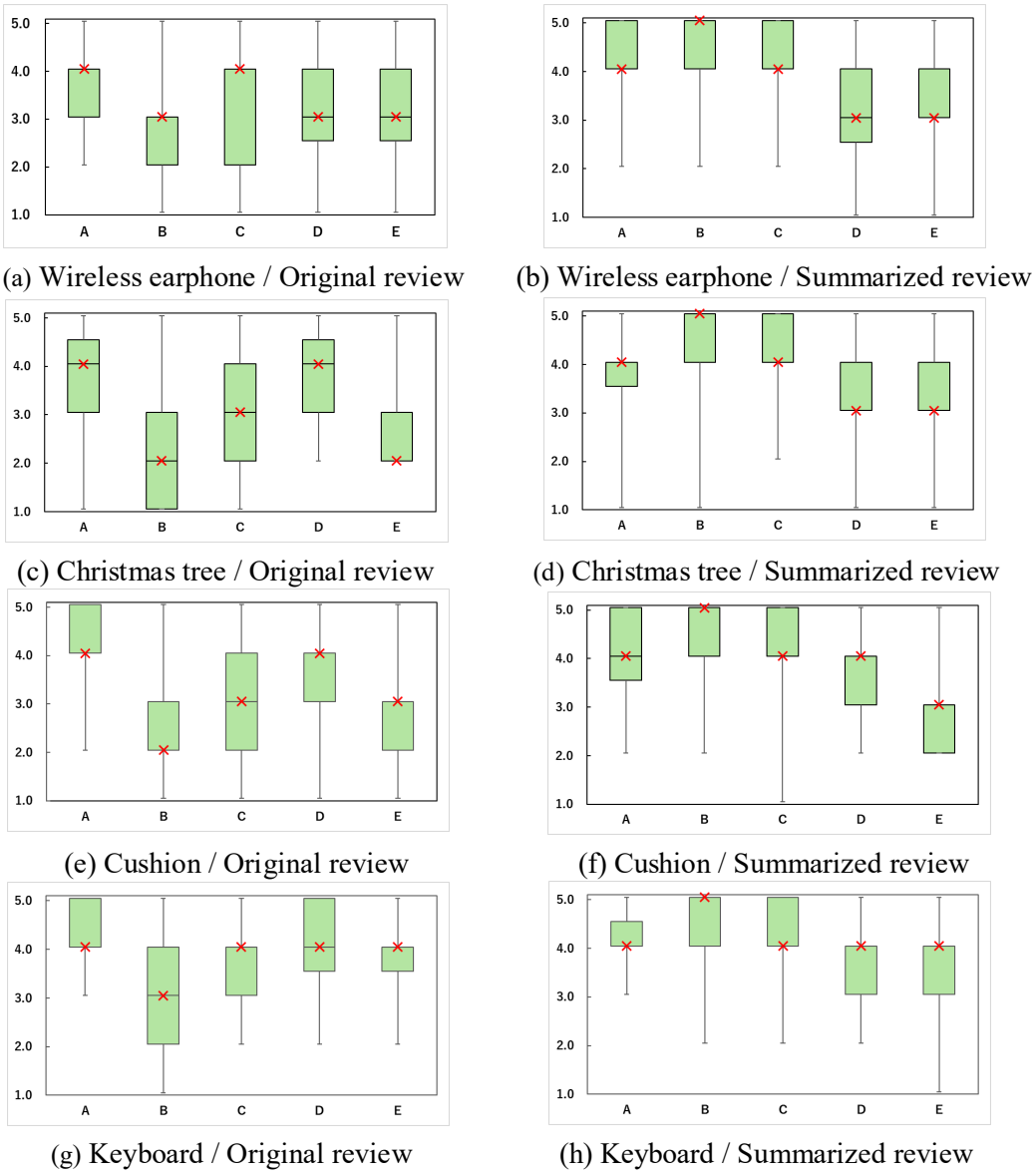
(a) Precision (b) Recall

Fig. 3: Comparison of Recall and Precision with and without AI Assistance

Second, to demonstrate the effectiveness of the proposed system, a subjective evaluation experiment was conducted using a questionnaire survey. Participants were asked to read both the original reviews on the website and the summarized reviews generated by the proposed system for four different products, alternating between them. For each case, they rated five questions on a five-point Likert scale: (A) “Were you able to understand the details and features of the product?” (B) “Did you find it easy to read without feeling burdened?” (C) “Were you able to grasp the overall trend of the reviews?” (D) “Did you find the amount of information

sufficient?" (E) "Did the content increase your interest in the product?" Here, a rating of 1 indicates the most negative response, while a rating of 5 indicates the most positive response.

The number of reviews used in this experiment was as follows: 20 for wireless earphones, 19 for Christmas trees, 20 for cushions, and 19 for keyboards. 27 university students participated in this experiment. Each graph in Figure 4 shows the experimental results as a box plot.



(a) Wireless earphone / Original review (b) Wireless earphone / Summarized review

(c) Christmas tree / Original review (d) Christmas tree / Summarized review

(e) Cushion / Original review (f) Cushion / Summarized review

(g) Keyboard / Original review (h) Keyboard / Summarized review

Fig. 4: Results of Subjective Evaluation Experiment

Figure 4 shows that, for all products, the interquartile range of the summarized reviews generated by the proposed system is larger for almost all questions compared to the original reviews, with almost no questions having a decreasing median.

To obtain more precise results, a statistical test was conducted. The Wilcoxon signed-rank test, which is a non-parametric test, was used, and a two-sided test with a significant level α of 5% was performed for all questions across all products. Table 1 shows the results.

Table 1: Results of Wilcoxon Signed-rank Test

	Question A	Question B	Question C	Question D	Question E
Wireless earphone	0.01738	0.00006	0.02509	0.57000	0.41445
Christmas tree	0.88709	0.00006	0.00455	0.25863	0.03285
Cushion	0.62326	0.00011	0.00126	0.94339	0.01128
Keyboard	0.41445	0.00066	0.07829	0.06487	0.72989

Question A	Were you able to understand the details and features of the product?
Question B	Did you find it easy to read without feeling burdened?
Question C	Were you able to grasp the overall trend of the reviews?
Question D	Did you find the amount of information sufficient?
Question E	Did the content increase your interest in the product?

In this table, the cells where the p-value falls below the significance level of 0.05 are shaded in gray. They indicate a statistically significant difference between the original review ratings and those generated by the proposed system. Figures 3 and Table 1 show that for Question B, the medians for the summarized reviews are larger than those for the original reviews across all products, and the p-value was below the significance threshold, suggesting that the proposed system effectively reduced the reading burden. For Question D, while the interquartile ranges for the summarized reviews are smaller than those for the original reviews for some products, the p-value did not fall below the significance level across all products. This implies that there were relatively few respondents who strongly felt that the output of the proposed system lacked sufficient information compared to the original reviews.

These results demonstrate that the proposed system contributes to reducing the reading burden while maintaining a minimum level of information adequacy.

5. Conclusion

This study aims to reduce the burden of reading product reviews on websites by summarizing the reviews while preserving the original information content. In this paper, a new product review summarization system, which reduces the reading load for product purchasers while preserving sufficient information, was proposed. The proposed system comprises the Evaluation Criteria Extraction Module, the Polarity Analysis Module, and the Supporting Sentence Extraction Module.

The evaluation of the Evaluation Criteria Extraction Modul showed the effectiveness of the AI assistance as well as the introduction of a stop-word list. Furthermore, a user survey showed that the system's output enables users to intuitively grasp overall review trends and reduces the cognitive load of reading compared to the original text.

However, some challenges still exist. When reviews contain many technical terms, the accuracy of morphological analysis and sentiment classification tends to decrease. This highlights the need for dictionary updates and sentiment model optimization. Other issues include better accounting for context in sentiment analysis and refining the selection of final output sentences based on similarity comparisons.

Future work will address these challenges to further improve the performance of the review visualization system. Additionally, in the proposed system, the final output sentences were randomly selected from each group of similar sentences. However, this approach is not necessarily optimal, hence, developing a more effective sentence selection method remains an important topic for future research. For extracting useful supporting sentences, there are existing methods like the one proposed by Gamzu et al. [12], which utilizes a dataset of Amazon reviews labeled with usefulness scores and employs a fine-tuned BERT model for prediction. Another method by Kim et al. [13] extracts useful sentences by training a Support Vector Machine (SVM) using features such as sentence structure, vocabulary, and semantics. Moving forward, it will be essential to build upon these existing studies to develop a more accurate supporting sentence extraction method.

6. References

[1] EXPLODING TOPICS, " 81 Online Review Statistics (New 2024 Data)" 2024. https://explodingtopics.com/blog/online-review-stats?utm_source

[2] Achyut Joshi, Andrew Giannotto, Ishika Arora, and Sumedha Raman, "Aspect and Opinion Extraction for Amazon Reviews", 2023.

[3] Ahmad Kamal, "Review Mining for Feature Based Opinion Summarization and Visualization," International Journal of Computer Applications, Vol. 119, No. 17, 2015, pp. 24-31.

[4] Christina Sauper and Regina Barzilay, "Automatic Aggregation by Joint Modeling of Aspects and Values," Journal of Artificial Intelligence Research, Vol. 46, 2013, pp. 89-127.

[5] MeCab, https://taku910.github.io/mecab/

[6] GINZA, https://megagonlabs.github.io/ginza/

[7] OpenAI API, https://platform.openai.com/docs/overvie

[8] Oseti, https://github.com/ikegami-yukino/oseti

[9] Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, November 2019, pp. 3982–3992.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Human Language Technologies, Vol.1, Minneapolis, Minnesota, June 2-7, 2019, pp.4171-4186.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems, Vol.30, 2017, pp.5998-6008.

[12] Iftah Gamzu, Hila Gonen, Gilad Kuitel, Ran Levy, and Engene Agichtein, “Identifying Helpful Sentences in Product Reviews,” Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Human Language Technologies, Virtual, Online, June 6-11, 2021, pp.678-691.

[13] Soo-Min Kim, Partrick Pantel, Tim Chklovski, and Marco Pennacchiotti, “Automatically assessing review helpfulness,” Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), Sudney Australia, July 22-23, 2006, pp.423-430.