User opinion-focused abstractive summarization using explainable artificial intelligence

HYUNHO LEE, Department of Data Science, Seoul National University of Science and Technology
YOUNGHOON LEE*, Department of Industrial Engineering, Seoul National University of Science and Technology

Recent methodologies have achieved good performance in objectively summarizing important information from fact-based datasets such as XSUM and CNN/DM. These methodologies involve abstractive summarization, extracting the core content from an input text and transforming it into natural sentences. Unlike fact-based documents, opinion-based documents require a thorough analysis of sentiment and understanding of the writer's intention. However, existing models do not explicitly consider these factors. Therefore, in this study, we propose a novel text summarization model that is specifically designed for opinion-based documents. Specifically, we identify the sentiment distribution of the entire document and train the summarization model to focus on major opinions that conform to the intended message while randomly masking minor opinions. Experimental results show that the proposed model outperforms existing summarization models in summarizing opinion-based documents, effectively capturing and highlighting the main opinions in the generated abstractive summaries.

Additional Key Words and Phrases: Abstractive summarization, Opinion focused-document, Intended message, Explainable artificial intelligence, Random masking, Machine learning, Artificial intelligence

1 INTRODUCTION

Text summarization involves transforming a source text into easily understandable and valuable information. Crafting a summary that effectively reflects the original text while remaining concise and comprehensible is a crucial task. Text summarization is generally divided into two methodologies: extractive and abstractive summarization. Extractive summarization[21] involves scoring sentences for importance and assembling a summary based on these scores. However, this method is limited in expression and may compromise the flow of context, as it relies solely on the existing text in the source document.

In contrast, abstractive summarization[20], a natural language generation (NLG) methodology, generates new text rather than extracting and combining sentences. Abstractive summarization can produce summaries with a variety of words, including those that are not present in the original document, providing a more diverse and expressive summary compared to extractive summarization. Therefore, this study, aiming to achieve more refined and natural summaries, focuses on abstractive summarization.

Most existing abstractive summarization models primarily focus on objectively summarizing lengthy documents such as news articles(Xsum[22], CNNDM[13]). Indeed, language models such as bidirectional and auto-regressive

Authors' addresses: Hyunho LEE, Department of Data Science, Seoul National University of Science and Technology, 232, Gongneung-ro, Nowon-gu, Seoul, 01811; Younghoon LEE, yhoon.lee@seoultech.ac.kr, Department of Industrial Engineering, Seoul National University of Science and Technology, 232, Gongneung-ro, Nowon-gu, Seoul, 01811.

ACM 2157-6912/2024/9-ART
https://doi.org/10.1145/3696456

transformer (BART)[16], T5[26], and Pegasus[35], which effectively comprehend and summarize the key information in documents, have demonstrated high performance in fact-based document summarization, as illustrated in Table 1.

Article	BUPA is an international health and care company with bases on three continents and more than eight million customers.BUPA began as The British United Provident Association in 1947 to preserve freedom of choice in health care.It believed that with a National Health Service being introduced a year later,there would still be a need for a complimentary service enabling people from all walks of life to afford the benefits of choice in where, when and by whom they were treated. Led by the growing public demand for health care and a lack of quality private accommodation BUPA initiated the Nuffield Homes Charitable Trust - later renamed Nuffield Hospitals.....
Summary	BUPA was founded in 1947 in response to plans to establish the NHS .The company's biggest base is in the UK but has customers in three continents .BUPA care homes cater for a number of conditions, including Parkinson's

Table 1. Example of summarization of fact-based document (CNNDM dataset)

Unlike fact-based documents where objective summarization is crucial, opinion-based documents require additional considerations for effective summarization[36]. Opinion-based documents exhibit the author's writing intent, incorporating the essential opinion sentiment the author aims to express and opinion sentiments unrelated to the main intent. For instance, in a review that includes negative sentiments about the price alongside positive opinions, the negative aspect is not the primary focus but rather contributes to the review's credibility, emphasizing the overall positive sentiment intended by the author.

The price may be a bit expensive, but the food is incredibly delicious, the atmosphere is great, and the staff are also very friendly.

Illustrating this with a review text as an example, although there is a negative opinion sentiment regarding the price, this serves as a mechanism to enhance the credibility of the writing rather than an overall condemnation of the restaurant being reviewed; the user's intended message can be considered a positive opinion sentiment. This is a common expression method where the author mentions opposing sentiments to increase the credibility of the writing and emphasize the core opinion sentiment when expressing their opinion[14, 28, 31].

As confirmed by numerous studies, analyzing sentiment is a fundamental task in various natural language processing (NLP) studies. Understanding the sentiment conveyed through text is essential for businesses to gather customer insights or analyze public sentiment[12].

In essence, the process of generating a summary for opinion-based documents focuses on capturing the author's writing intent by concentrating on the core opinion sentiment. The goal is to exclude opinion sentiments that are not significantly associated with the primary intent. Consequently, as depicted in Figure 2, the correct summary (golden summary) of an opinion text includes only the essential writing intent, omitting opinion sentiments that are not closely related to the intent and incorporating content that corresponds to the core opinion sentiment.

Review	I purchased this for my 6 year old Minecraft loving son. The directions are hard to follow and as an adult I had a hard time putting them together. After I built them they came apart easily and my son could not really play with them. This toy is better suited for an older child who is only interested in building things out of paper and not playing with the figures afterward. I ended up throwing it in the trash
Summary	flimsy and difficult to build.
Sentiment	Negative

Table 2. Example of summarization of an opinion-based document

Therefore, in this paper, we propose an advanced abstractive summarization model that explicitly considers opinion sentiment to efficiently perform text summarization for opinion-based documents. We specifically utilize a classification head attached to the encoder and the Shapley additive explanation (SHAP) explainer to derive the sentiment distribution of the original text[37]. Based on the sentiment distribution information, we extract the major opinion and minor opinion parts from the text. Finally, we incorporate random masking on the

minor opinion part and integrate it into the encoder-decoder multi-attention operation during training, thereby emphasizing the major opinion sentiment.[4]

We also propose a novel random masking method to ensure effective learning when handling minor opinions. Our method involves a different probabilistic masking technique that applies varying probabilistic masks to each learning process and attention head, rather than a fixed mask. Furthermore, to facilitate the model's ability to naturally emphasize major opinion parts, another novel incremental learning strategy is proposed. The proposed method, by incrementally learning to prioritize major opinion text, facilitates the creation of a model that, during inference, neglects minor opinion segments and concentrates solely on the major opinion, even in the absence of sentiment distribution data.

To the best of our knowledge, the proposed method is the first research explicitly considering both writing intent and sentiment distribution for summarizing opinion-based text in accordance with the writing intent. Additionally, our proposed methodology is an attachable modular approach, which can be seamlessly integrated into existing summarization models, showcasing notable advantages in efficiency and reusability. Consequently, we have validated the performance improvement of the summarization task for opinion-based documents when our proposed methodology is integrated into existing summarization models.

The remainder of this paper is structured as follows. Section 2 discusses various studies on abstractive summary models using pretrained language models, and Section 3 outlines the proposed method. Section 4 describes the data used in the experiments and experimental results of our proposed method. Finally, Section 5 provides the concluding remarks and briefly discusses directions for future research.

2 LITERATURE REVIEW

2.1 Pre-trained language model (PLM) based abstractive summarization models

The most widely used abstractive summarization models are based on pretrained language models (PLMs), such as bidirectional encoder representations from transformers (BERT) and generative pre-trained transformer (GPT). BERT [30])[5] is a model comprising the encoder part of the Transformer, which is pretrained on a large corpus to effectively learn the contextual representation of text. It can understand documents by learning bidirectional representations; however, it has limitations in its direct application to summarization models because it is structured with only an encoder. Meanwhile, GPT [3] is a PLM model that uses only the decoder part of the Transformer and exhibits strengths in various natural language tasks (e.g., text generation, question answering, and text classification). However, as an auto-regressive PLM model specialized in generation, it cannot learn bidirectional information and has limitations in understanding and summarizing text.

BART [16] is a PLM that combines the strengths of the two mentioned models and assumes the form of a seq2seq model by combining the encoder and decoder. It is pretrained by reconstructing text that has been corrupted by a noise function to reduce the loss between the decoder's output and the original text. Consequently, BART employs a simultaneous approach of text infilling (replacing a text span with a single mask token) and sentence shuffling (randomly shuffling the order of sentences). This unique combination allows BART to utilize various noise types during training, surpassing the capabilities of existing auto-encoder models. Thus, various tasks, such as sequence classification, token classification, and machine translation, can be conducted through pre-learned BART, which shows excellent performance, particularly in abstractive summarization.

T5[26] is a transformer-based PLM similar to BART. It reconstructs all NLP tasks into an integrated text-to-text format in which the inputs and outputs are always text strings, unlike BERT-style models, which can output only class labels or input ranges. It differs from the existing Transformer model in that relative position encoding is used for the input token, which is popular (the Transformer uses sinusoidal embedding and BERT uses position embedding). This means that learning occurs not through assigning a uniform encoding to each token's input

position and calculating attention, but by assigning relative positional encodings to tokens within an offset boundary during self-attention calculation.

Pegasus[35] uses the gap sensation generation (GSG) method, wherein sentences are selected based on their importance determined by the ROUGE score and individually masked. This approach assumes that aligning the pre-training objective with the text summarization process will yield improved performance. Similar to the MLM [27] method, GSG learns not by masking and predicting masked tokens, but by masking and predicting important sentences. An important sentence refers to a sentence that can explain the overall context better than other sentences in a document.

However, the models described above primarily focus on fact-based documents, where objective summarization is prioritized. In contrast, summarizing opinion-based documents effectively demands attention to the document's sentiment and authorial intent, aspects often overlooked by current models.

Several studies have focused on opinion-based texts. In [9], a model that generates abstractive summaries for each aspect was proposed. In an abstract summary of product reviews using a discourse structure, as in a graph-based study [11], a model generating an effective label description to learn a text representation via two different channels[37], and a model for generating summary statements by conducting aspect-to-aspect scoring using the Page Rank algorithm [24] were proposed. However, these studies are irrelevant to the research field covered by the proposed method because they focus on identifying sentiment by aspect [8], rather than generating summaries that explicitly consider sentiment.

There are studies targeting opinion-based text that focus on the degree of augmentation. In the research by Xu et al. [32] and Cheng et al. [4], additional attention is given to historical information, while in other studies by Xu et al. [34] and Xu et al. [33], personalized features are considered as supplementary information. However, these studies also do not explicitly consider opinion sentiment.

Therefore, to the best of our knowledge, our proposed methodology is the first to explicitly consider sentiment distribution for opinion-based text summarization.

2.2 XAI methodologies utilized in our proposed study

In this study, SHAP was used to derive the sentiment distribution within a sentence. Based on this, the major opinion sentiment part on which the summarization model focuses was derived.

SHAP is the explainable artificial intelligence (XAI) [2] methodology used in this study. Figure 1 shows the entire process of SHAP, which is the original black box model f, and the process of determining g with input values that provide variations to the data, rather than putting the same input. SHAP is based on local explanations, enabling a global surrogate for the entire area of data. [18] proposed SHAP values as an "integrated measure" of characteristic importance. The conditional average of the SHAP value is calculated to define the simplified input by calculating the conditional average of f, rather than the value of the existing model f.

SHAP values quantify the extent to which a specific feature characteristic contributes to changes in model predictions under specific conditions or contexts. $E[f(z)]$ is predicted when no characteristic is known, called the base value, and SHAP values explain how the current result value $f(x)$ originates from the base value (Figure 2).

As discussed in numerous studies, SHAP has both advantages and disadvantages. Research indicates that SHAP is highly influenced by the correlations between variables when calculating Shapley values. Additionally, the computational complexity increases exponentially in high-dimensional data due to the consideration of all possible combinations of input variables in the Shapley value calculation.

However, SHAP values offer a model-agnostic approach, allowing application across different types of models. They consider both individual variable importance and overall variable importance. Despite its limitations, SHAP is widely utilized in various NLP tasks such as sentiment analysis, document classification, and document

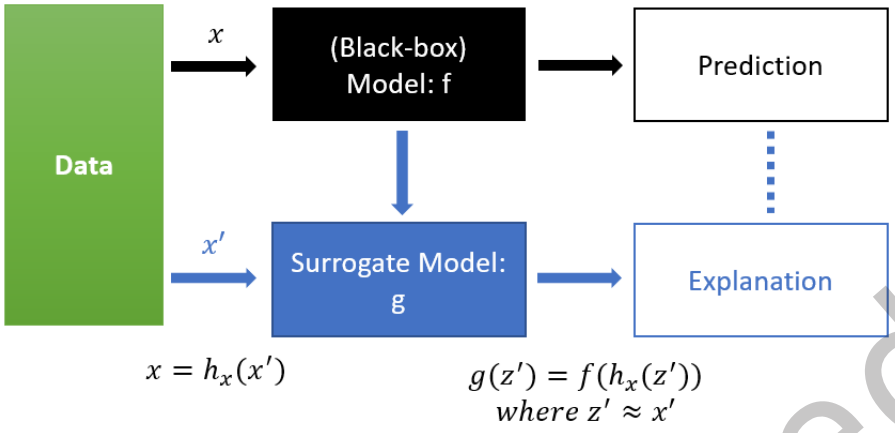

Fig. 1. SHAP framework

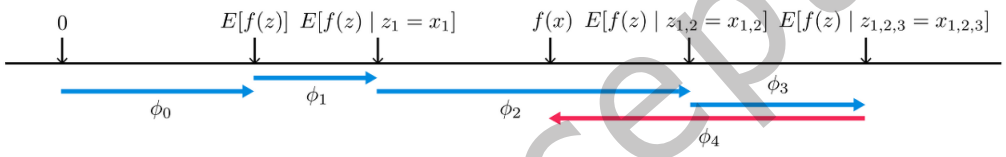

Fig. 2. SHAP base values

summarization[6, 7, 15]. In this study, we address the issue of variable correlation through a masking strategy. While the proposed research and its application objectives differ, it is noteworthy that Naretto et al. [23] also employed a masking strategy to compute the Shapley value in their study.

3 METHOD

Our proposed methodology adopts a fine-tuning approach, building upon a pretrained backbone architecture similar to existing summary models. It offers an efficient and versatile model that can be integrated into various existing summarization models. Figure 3 summarizes our proposed methodology.

The proposed methodology involves several distinct steps. First, the sentiment distribution of the input text is obtained by leveraging the classification head attached to the encoder and SHAP explainer. Second, major and minor opinion parts are extracted from the text considering the information derived from the sentiment distribution. Finally, random masking is applied to the minor opinion part and integrated into the encoder-decoder multi-attention operation during the training process. This inclusion of random masking helps prioritize and emphasize the major opinion sentiment during training. Details of each process are described in subsequent sections.

3.1 Derivation of sentiment distribution

First, the sentiment distribution of the input text is calculated using the classification head attached to the encoder and SHAP explainer. As mentioned previously, SHAP approximates a simplified input x' and a surrogate model g given an original input x and a complex black-box model f. The simplified input x' is defined by the mapping

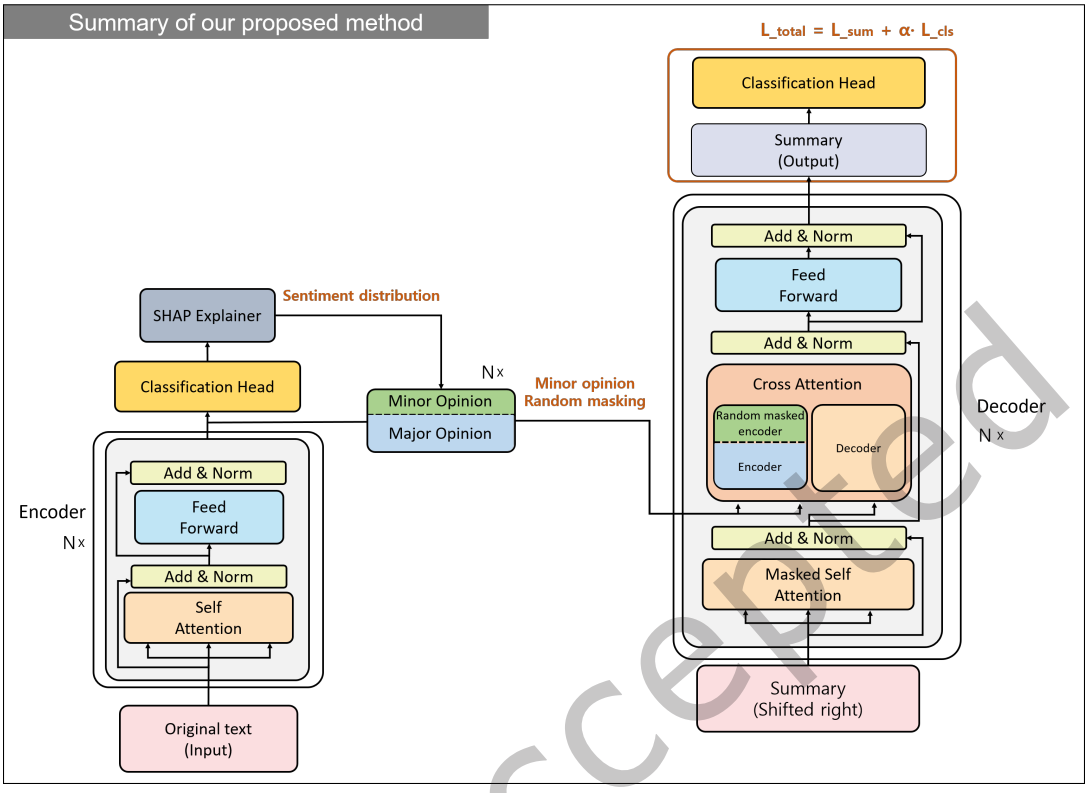

Fig. 3. Summary of the proposed method

function $x = h_x(x)$. Assuming that $z' \approx x'$, $z \in \{0,1\}^M$ is represented as a binary variable (1: feature included in the model, 0: feature excluded from the model). We then train an explainable model g such that $g(z') \approx f(h_x(x'))$. The explainable model g is approximated by a linear function of the binary variables, as shown below, where ϕ_i denotes the importance score of the i-th variable:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

In a real implementation, the result of tokenizing the dataset becomes the original input x, and the classification head becomes the black-box model f. By substituting x and f into the SHAP explainer function, we approximate the simplified input x' and surrogate model g. Finally, the sentiment distribution of each token, ϕ_i, is calculated by the result of the SHAP explainer function. Figure 4 shows examples of sentiment distribution derived using the SHAP explainer.

In this study, we employed the DeepExplainer among various SHAP explainers. SHAP explainer can be implemented in various forms, with TreeExplainer, KernelExplainer, and DeepExplainer being prominent examples. TreeExplainer is specialized for tree-based models (e.g., random forests, gradient boosting trees) and leverages the tree structure of the model for efficient and accurate computation of SHAP values. KernelExplainer is more versatile due to its minimal assumptions regarding the model, making it applicable to various models. Although it can be applied to non-linear or complex deep learning models, it suffers from high computational costs.

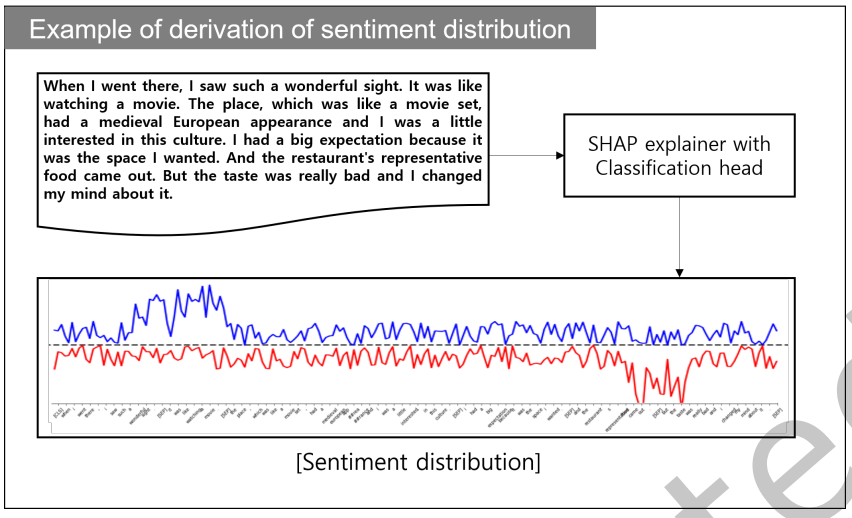

Fig. 4. Example of sentiment distribution

DeepExplainer is optimized for deep learning models and provides SHAP values by analyzing the influence propagated through the layers of neural networks. It is particularly useful for handling non-linear relationships between model inputs and outputs, making it effective even in cases with complex internal mechanisms, such as transformer models. Considering these factors, we deemed DeepExplainer most suitable for alignment with the proposed research.

3.2 Extraction of major and minor opinion parts

Based on the sentiment distribution derived above, the input text is divided into major and minor opinion parts. In our proposed methodology, the "major opinion" part denotes the text segment where the opinion intended by the user is expressed strongly and prominently. Conversely, the "minor opinion" part is the text segment where the opinion expressed is secondary and understated. Essentially, the part with an explainer value based on sentiment distribution is defined as a major opinion part.

We derive the major and minor opinion parts using two methods. First, the explainer value derived from the sentiment distribution is used as a criterion. For the SHAP explainer, the explainer value can be calculated for positive and negative polarities. Therefore, the explainer value corresponding to the predicted polarity based on the classification head is used as the criterion. If the explainer value is within the top $n\%$, it is considered a major opinion; otherwise, it is considered a minor opinion. In general, words that exhibit a polarity opposite to that intended by the user have a negative explainer value, and the explainer value of words unrelated to polarity converges to zero. Therefore, through this methodology, words expressing opposite polarities are likely to be derived as minor opinion segments. The pseudocode for the methodology expanded above is summarized in Algorithms 1 and 2.

The second method uses the absolute value of the explainer as a criterion. The rest is the same as in the methodology described above. However, if the absolute value is used as the judgement criterion, it is highly likely that a word irrelevant to the polarity where the explainer value converges to zero is derived as a minor opinion.

Algorithm 1 Splitting a set of original documents into major and minor opinion parts using explainer value

Input: Set of original documents $X_i = \{x_1^1, \ldots, x_t^t, \ldots, x_i^n\}$
Output: Major opinion part X_i^a, minor opinion part X_i^b

$S(x_i^t)$: Sentiment Score of token x_i^t corresponding to prediction of document sentiment
for $t \leftarrow 1$ to n **do**
    if $S(x_i^t)$ is in the top n percent of $S\{(x_i^1), \ldots, S(x_n^t)\}$ **then**
        $X_i^a \leftarrow x_i^t$
    else
        $X_i^b \leftarrow x_i^t$
    end if
end for

Algorithm 2 Splitting a set of original documents into major and minor opinion parts using absolute explainer value

Input: Set of original documents $X_i = \{x_1^1, \ldots, x_t^t, \ldots, x_i^n\}$
Output: Major opinion part X_i^a, minor opinion part X_i^b

$S(x_i^t)$: Sentiment score of token x_i^t corresponding to prediction of document sentiment
for $t \leftarrow 1$ to n **do**
    if $|S(x_i^t)|$ is in the top n percent of $\{|S(x_i^1)|, \ldots, |S(x_n^t)|\}$ **then**
        $X_i^a \leftarrow x_i^t$
    else
        $X_i^b \leftarrow x_i^t$
    end if
end for

3.3 Masking strategy

As previously mentioned, the proposed central learning strategy involves masking the minor opinion that deviates from the user's intention. Simultaneously, the weight or emphasis on the major opinion part is increased within the encoder-decoder attention operation. This adjustment ensures that the generated summaries prioritize and emphasize the major opinion part, aligning with the original author's intended viewpoint.

However, in this process, if a specific part of the minor opinion is collectively masked and the attention weight is given as zero, not only is part of the input text completely ignored, but the flow of the context also becomes unnatural.

Therefore, to avoid this problem, we randomly masked a certain percentage of minor opinions differently for each learning, rather than masking specific parts collectively. Additionally, the robustness of the learning process was increased using varying random masking for each attention head (Figure 5).

Our methodology aims to train a model capable of emphasizing the major opinion within a text solely based on the content, without relying on external data such as sentiment distribution, during the inference phase. Therefore, a novel strategy for the summarization model to learn how to focus on the major opinion parts incrementally and naturally is presented in Figure 6.

Specifically, the minor opinion part is initially masked at a high ratio to adapt to relatively easy problems first, and then the masking ratio is gradually reduced to learn to adapt to more difficult problems. Thus, it is possible to learn a model that effectively focuses on major opinion parts, even in the original text, without masking information.

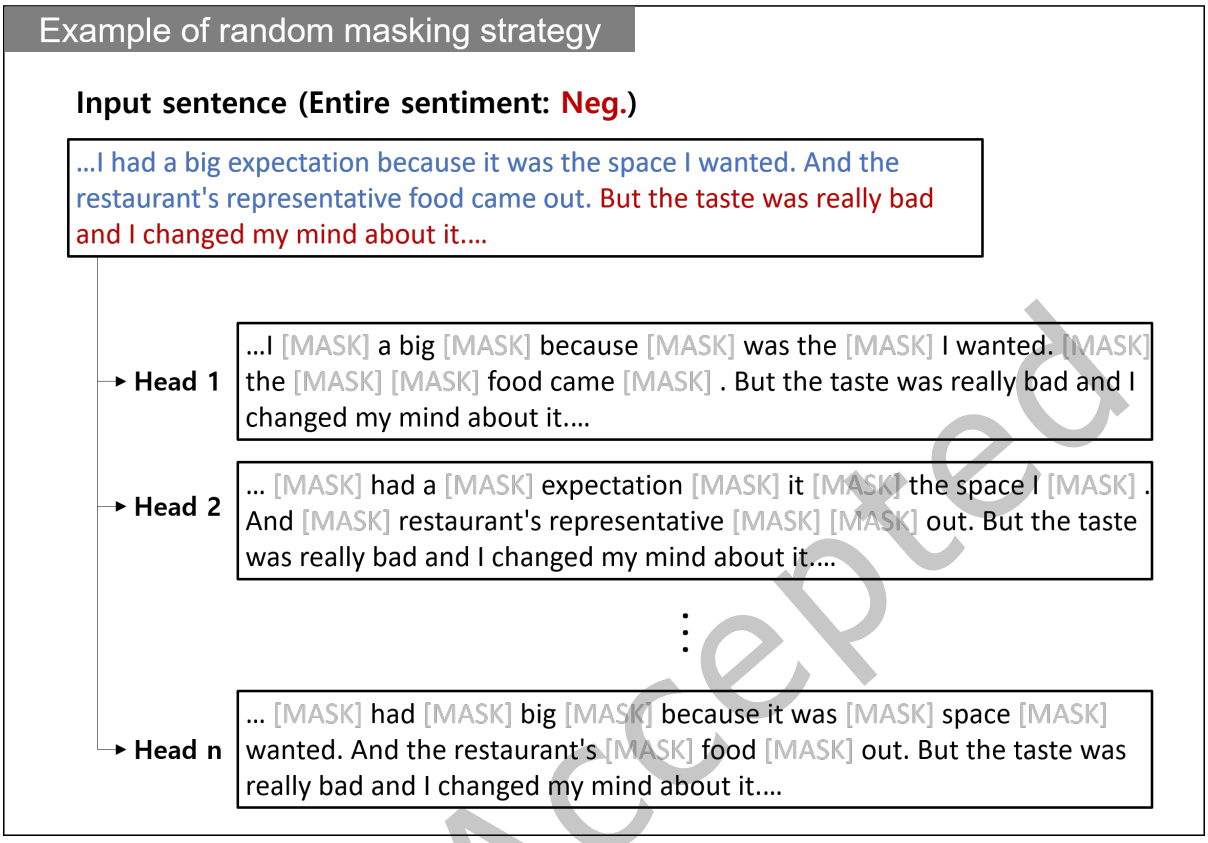

Fig. 5. Example of random masking strategy

3.4 Configuration of the loss function

Finally, in this study, the sentiment classification accuracy of the generated summary was also considered in learning to ensure that the generated summary could focus on major opinions that conform to the user's intention. Specifically, by adding a classification head to the decoder output, the classification accuracy of the generated summary and the accuracy of the generated summary are reflected in the loss function as follows: $Loss_{total} = Loss_{summaryGeneration} + \alpha Loss_{DecoderClassification}$. This aims to effectively create a summary that conforms to the user's intentions by effectively maintaining the sentiment of the input text. Furthermore, we set the loss combination ratio α to one to obtain the best performance in the experiment.

4 EXPERIMENT

4.1 Dataset

The proposed method was evaluated using the Amazon Review dataset, an opinion-based document dataset. The Amazon dataset not only contains golden summary information for the summarization task but also includes sentiment labels; therefore, it is almost the ideal dataset to verify the performance of the proposed methodology.

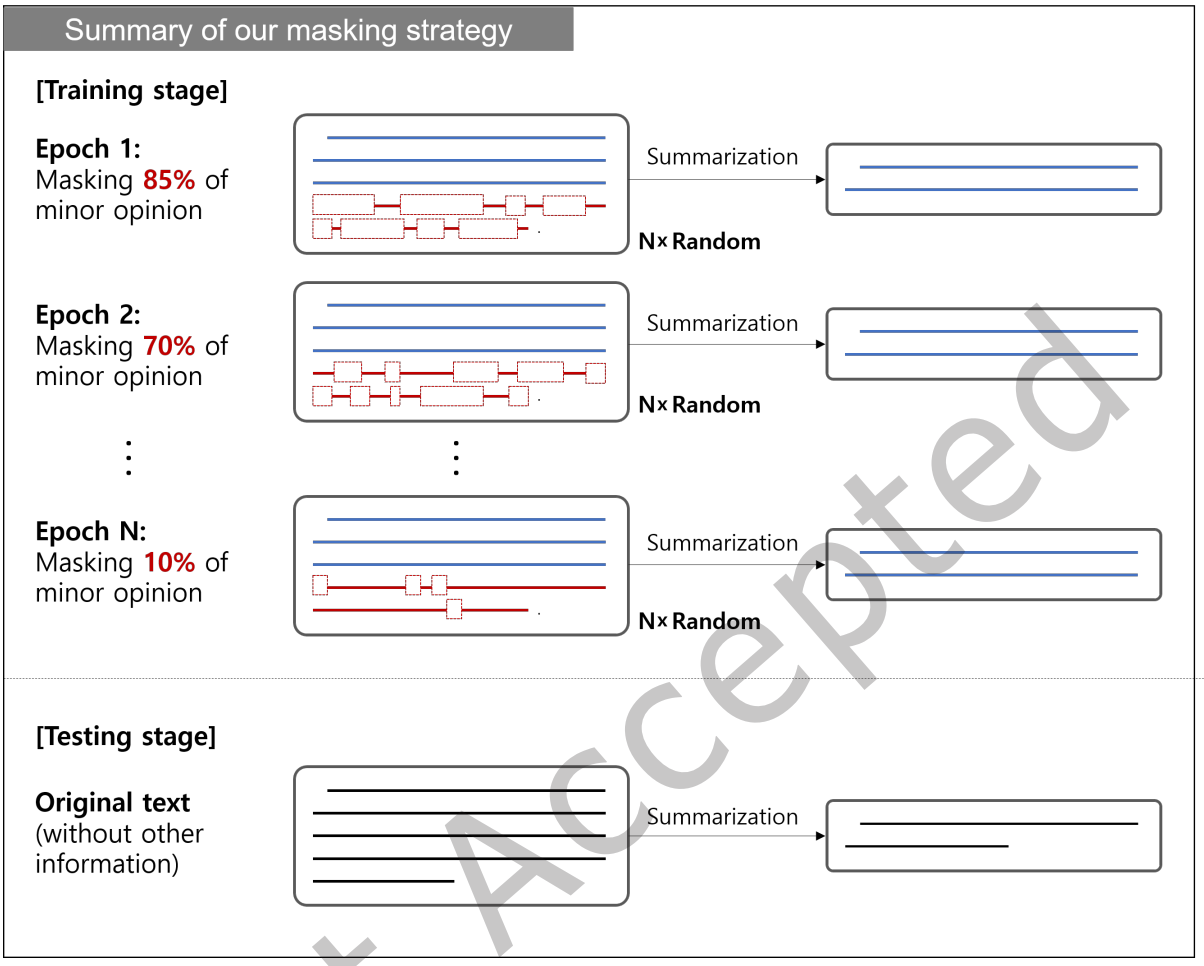

Fig. 6. Masking strategy for incremental learning

It comprises 100 categories, and the four most widely used categories were selected as the experimental dataset. The statistics of this dataset are listed in Table 3.

Dataset	Train	Validation	Test
Toy	79,000	6796	6775
Sport	140,000	7764	7734
Home	133,600	8608	8641
Movie	178,693	16,909	16,890

Table 3. Dataset statistics

Furthermore, as evident from Table 4 below, while the original text exhibits a relatively high proportion of neutral labels, the corresponding golden summary shows a decrease in neutral labels and an increase in the ratio

of positive/negative labels. The Amazon dataset effectively demonstrates the specific traits of the opinion text we have discussed, making it an excellent choice for thoroughly evaluating the effectiveness of our proposed opinion text summarization approach.

Dataset	Label	Original text	Golden summary
Toy	Pos	66%	70%
	Neu	15%	6%
	Neg	19%	24%
Sport	Pos	61%	66%
	Neu	20%	9%
	Neg	19%	25%
Home	Pos	64%	70%
	Neu	22%	11%
	Neg	14%	19%
Movie	Pos	63%	69%
	Neu	23%	10%
	Neg	14%	21%

Table 4. Ratio of positive/neutral/negative label

4.2 Experimental setup

As previously mentioned, the proposed approach offers an attachable model that can be incorporated into existing summarization models. Therefore, we validated the performance improvement of the proposed methodology by applying it to the most widely used summarization models: BART and T5. We ensured the robustness of our proposed model's performance improvement by comparing it against the most widely recognized summarization models as baselines.

In addition, we selected four benchmark models for performance comparison to demonstrate the superiority of the proposed method. The selected baselines are methodologies similar to our proposed approach, indicating that they incorporate attachable modules into existing summarization backbone architectures. By comparing these models with those that share similar improvement directions, we can objectively assess the performance enhancement of our proposed methodology. The four models we used as baselines were: 1) PGNet[29], an RNN-based abstractive summarization method with a copy mechanism; 2) C.Transformer[10], a variant of the Transformer equipped with a copy mechanism; 3) HSSC+Copy[19], a review summarization model that jointly improves review summarization and sentiment classification using a copy mechanism; 4) DualView[25], a dual-view model that improves review summarization and sentiment classification performance; 5) TRNS[32], a model that involves the historical information selectively to learn the specific representation; 6) RARS[34], a neural review-level attention model that effectively learns user preference embedding; and 7) SRSP[33], an elaborately designed multi-task fine-tuning framework that fully leverages personalized information.

The generated summaries were evaluated using the Rouge Score, a commonly used evaluation metric for text summarization that measures the ROUGE-1, ROUGE-2, and ROUGE-L [17] scores of the generated summaries. Furthermore, the proposed methodology involves splitting the major and minor parts based on the prediction results of the classification head. Consequently, we must also validate its performance. Therefore, we present the inherent classification performance of the classification head embedded in the proposed methodology.

As mentioned previously, we propose two methodologies for splitting the major and minor parts based on the explainer value and absolute explainer value. Moreover, the proposed methodology employs an approach that

induces natural learning by incrementally reducing the masking ratio. Therefore, we experimented with two masking strategies: 1) initially masking 85% of the minor opinions and decreasing the masking rate by 15% every epoch until reaching 10% and 2) initially masking 70% of the minor opinions and decreasing the masking rate by 10% every epoch until reaching 10%. We compared the performance of these options using an initial masking of 70% based on the absolute value, which showed optimal performance. The remaining options are presented in the ablation study to demonstrate their performances.

For BART-based approaches, the training parameters were set to a batch size of 256, AdamW optimizer[1] learning rate of 4e-4, and a warmup state of 0.025. The experiments were conducted on eight RTX 3090 GPUs in parallel. Similarly, for the T5-based approaches, the training parameters were set to a batch size of 128, an AdamW optimizer learning rate of 3e-4, and a warmup state of 0.025. Experiments were conducted on eight RTX 3090 GPUs in parallel, similar to the BART-based approaches.

4.3 Experimental result

4.3.1 Abstractive summarization performance. We present a comparison of the performance in the abstractive summarization task between our proposed methodology and the baselines in Table 5. As mentioned previously, the baselines included vanilla BART and vanilla T5, which represent the backbone architectures, as well as recently proposed methodologies similar to our proposed approach. This indicates that they incorporate attachable modules compatible with existing summarization backbone architectures. First, as evident from Table 5, our methodology demonstrated a significant performance improvement compared to vanilla BART and vanilla T5. This result validates the effectiveness of the attachable module designed based on our proposal. The module aims to split the major and minor opinion parts prioritizing the major opinion component during the abstractive summarization task, particularly for datasets that prioritize opinions.

Table 5 shows that the proposed methodology outperforms other recently proposed methods that incorporate attachable modules. This proves that our methodology is the most effective abstractive summarization approach for opinion-focused texts. In the experimental results, the proposed methodology achieved superior performance across all measures and datasets, except for the Movie datasets. Additionally, the observed performance of certain methodologies demonstrated poorer results compared to the backbone architecture. This demonstrates that simply incorporating an additional module does not guarantee a performance improvement for opinion-focused datasets. This result substantiates the proposed necessity to incorporate modules tailored to the unique characteristics of opinion-focused datasets.

In addition, the superiority of the proposed methodology can be evaluated through its application of the SHAP explainer. The Shapley value, which forms the basis of the SHAP explainer, is a methodology heavily influenced by the correlation between variables. The masking strategy employed in the proposed methodology reduces the correlation between tokens, partially addressing the inherent limitations of the Shapley value in this aspect. Furthermore, in the proposed methodology, using transformers to learn interactions between tokens allowed for a more accurate reflection of the correlation between variables. This factor is believed to contribute to the superior performance of the proposed methodology over existing models.

On the other hand, the performance of the proposed methodology on the Movie dataset was lower compared to some other methodologies. Since the proposed methodology involves deriving major and minor sentiment parts from reviews and utilizing them for training by masking, it necessitates appropriate coherence between the sentiment of the entire document and the sentiment of each token. However, due to inconsistencies and a significant amount of ambiguous data between the overall sentiment and the sentiment of tokens, as exemplified below, we conclude that the performance of the proposed methodology was relatively lower.

'the lucky ones is a pretty good movie. it's not great but its not bad either. i was going to give it 3 stars but it deserves 4. i like how these three soldiers stories are told at once and they don't get confusing or have some weird

Dataset	Method	R-1	R-2	R-L
Toy	PGNet	14.83	6.17	14.57
	C.Transformer	12.57	4.76	12.36
	HSSC+copy	14.70	6.18	14.46
	Dual-View	14.83	6.17	14.57
	TRNS	18.87	8.70	18.51
	RARS	18.04	5.03	17.59
	SRSP	19.05	**7.28**	18.63
	Vanilla BART	16.49	4.82	14.99
	Vanilla T5	17.61	5.86	16.66
	Ours (BART, Absolute value, 70%)	18.03	5.66	16.53
	Ours (T5, Absolute value, 70%)	**19.30**	7.10	**18.61**
Sport	PGNet	14.78	6.13	14.58
	C.Transformer	13.73	5.13	13.54
	HSSC+copy	14.64	5.95	14.43
	Dual-View	15.39	6.46	15.18
	TRNS	13.25	3.78	13.04
	RARS	14.41	5.93	16.11
	SRSP	18.26	6.66	17.87
	Vanilla BART	15.44	4.75	14.06
	Vanilla T5	17.13	5.89	16.20
	Ours (BART, Absolute value, 70%)	15.56	4.66	14.09
	Ours (T5, Absolute value, 70%)	**18.62**	**6.70**	**17.93**
Home	PGNet	14.82	6.28	14.64
	C.Transformer	13.75	5.44	13.58
	HSSC+copy	14.93	6.34	14.75
	Dual-View	15.18	6.57	15.00
	TRNS	14.60	4.90	14.32
	RARS	16.67	4.67	16.39
	SRSP	18.09	6.28	**18.08**
	Vanilla BART	14.70	4.17	13.27
	Vanilla T5	17.09	5.87	16.18
	Ours (BART, Absolute value, 70%)	15.06	4.24	13.61
	Ours (T5, Absolute value, 70%)	**18.10**	**6.67**	17.45
Movie	PGNet	12.28	5.14	12.40
	C.Transformer	12.09	4.46	11.81
	HSSC+copy	12.66	5.06	12.39
	Dual-View	12.84	5.22	12.57
	TRNS	11.72	2.85	11.23
	RARS	15.04	5.35	14.60
	SRSP	**16.08**	**5.47**	**15.57**
	Vanilla BART	12.65	3.26	11.58
	Vanilla T5	13.31	3.90	12.29
	Ours (BART, Absolute value, 70%)	13.57	3.92	12.51
	Ours (T5, Absolute value, 70%)	13.92	4.33	13.08

Table 5. Performance comparison for the abstractive summarization

thing to do with each other like some movies do. i thought they did a good job making you feel the pain each person is going through too. i was hoping for a different ending but i don't think it could have been too different at the same time. i would recommend this movie to everybody to see at least once.' (***Original label: Positive***)

'i have all of the cinematic titanic shows on DVD to date. i love most of them my favorite is war of the insects. i met the case last winter and admire them and their talents. i know humor can be subjective this one just didn't click for me. the YouTube trailer covers most of what i found funny in this movie. i kind of prefer the studio movies as opposed to the live ones but note my favorite happens to be a live DVD. i m not slamming cinematic this is just not a movie i would recommend to others. I'll still buy whatever the next release is.' (***Original label: Negative***)

4.3.2 Qualitative analysis. In Tables 6 and 7, we provide examples of the summaries generated by BART, T5, and our proposed model for comparison. Analysis of these tables reveals that, like many opinion-based texts, the original sentences combine major and minor opinions, whereas the golden summary focuses solely on the major opinion.

However, the summaries generated by the baselines include both major and minor opinion content, which is in clear contrast with the golden summary. As mentioned previously, this is because existing abstractive summarization models tend to summarize the entire text consistently without considering sentiment. Conversely, the summaries generated by our proposed model include only the major opinion segment, which is aligned with the golden summary. Based on these generated summary examples, we can demonstrate that our proposed model is effective in summarizing opinion-focused texts and that this factor contributes to achieving significantly higher Rouge scores than the baselines.

Review	The n64 was nintendo's 3rd console not counting the game boy and virtaul boy.It came out shortly after the sega saturn and sony playstation. If you do not have this system anymore or never have and collect systems and and games then i recommend this system the controller has always at first glace seemed like a gimmick and like myself probably made you think how are people supposed to use that controller.But it actually is not too bad! the system it s self is pretty simple yet cool looking and is for the most part really reliable.And having 4 controller ports is a plus too! though it would have been nice if nintendo released more mature games for the console there are some classic games I recommend the games below
Sentiment	Positive
Gold Summary	A pretty good system for it's time...and even every now and then.
BART Summary (Baseline)	Not fond of barbie but this adds to the fun.
T5 Summary (Baseline)	Nintendo 3rd console.
Our Summary	Great game system for the n64 and sony playstation.

Table 6. An example of a generated summary (positive sentiment)

4.3.3 Ablation study. In the ablation study, we performed additional comparisons of the summarization performance based on the criteria for splitting the major and minor parts (i.e., explainer value-based and absolute explainer value-based, respectively) and the ratio of the initial masking (70%/85%). Furthermore, as mentioned previously, in our proposed methodology, we applied different random masking to each head to implement a robust model. To validate the effectiveness of this approach, experiments were conducted with an option that applied the same masking to all heads. Table 8 and 9 present the results of the ablation experiments using BART and T5 as the backbone architectures, respectively.

Experimental results revealed that for the BART backbone, the performance results were significantly mixed across the datasets, whereas for the T5 backbone, relatively consistent performance comparisons were achieved. When considering T5 as the reference, using the criteria of absolute explainer value rather than explainer value achieved a relatively higher performance. Furthermore, it was evident that applying random masking consistently led to a higher overall performance than when it was not applied. However, the performance variation based on the ratio of initial masking was relatively small. Therefore, when replicating the proposed model, it is considered

Review	I thought this sounded like a pretty good game but it seemed to be for us at least more trouble than it was worth.the game controller itself is simple.it does not seem very heavy duty simply translucent blue lightweight plastic.the controller is not very comfortable in the hand like a regular video game controller is.the cartridge does not hook in very securely.to play the game you first have to load a cd into your computer.then while connected to the internet you need highspeed access you need to download more stuff.this took a looong time for us.it was not our access it was slow coming from their site.you also need to have 1 gb free...yes a whole gb! finally the game was running.while the game was running we were not able to adjust the volume on our computer.it was loud and annoying.the courses were nothing wonderful and you need the right cartridge for each car which does best on its own track.there are eight cartridges total.this was kind of annoying considering each car does not have any special chip or anything.the controller detects which cartridge is in it by which of the five pins the car triggers in the controller.so basically by buying additional cartridges you are simply buying a way to press down a couple of buttons.to top it all off the company reserves the right to terminate turbo driver online service after january 1 2010 .so you can buy the unit and all the cartridges and after a year they may be useless.the game itself is ok to play.the kids played it for a little while but became quickly bored by it.the graphics are not great the controller is not great the game can be frustrating.it is a nice idea and is not a bad toy but i would not recommend it.
Sentiment	Negative
Gold Summary	Not a great game.
BART Summary (Baseline)	A little bit of fun for the kids but not worth it
T5 Summary (Baseline)	not a bad toy.
Our Summary	Not worth the money...frustrating and slow service.

Table 7. An example of a generated summary (Negative sentiment)

Dataset	Masking Ratio	ABS R-1	ABS R-2	ABS R-L	EXP R-1	EXP R-2	EXP R-L
Toy	85%	16.23	4.50	16.70	16.47	4.65	15.00
	70%	18.03	5.66	16.53	15.95	4.50	14.42
	55%	15.04	4.01	15.48	14.90	3.90	14.20
	40%	14.95	3.95	15.01	14.84	3.85	14.55
Sport	85%	16.10	5.23	14.85	15.56	4.66	14.09
	70%	15.01	4.38	13.67	15.92	4.81	14.55
	55%	14.85	4.22	13.02	14.80	4.11	13.01
	40%	14.44	4.01	12.92	13.85	4.02	12.95
Home	85%	15.00	4.35	13.60	15.98	5.00	14.71
	70%	15.06	4.24	13.61	16.22	5.10	14.96
	55%	14.86	4.02	13.05	14.34	3.97	14.01
	40%	13.99	3.95	13.45	13.43	3.74	13.44
Movie	85%	13.40	3.90	12.33	13.27	3.87	12.27
	70%	13.57	3.92	12.51	13.00	3.50	11.90
	55%	12.94	3.45	12.11	12.44	3.30	11.40
	40%	12.87	3.24	12.01	12.32	3.10	11.50

Table 8. Ablation study (BART backbone)

more efficient to prioritize the backbone model and splitting criteria than placing excessive emphasis on the initial masking ratio.

4.3.4 Classification performance. As explained previously, the proposed methodology focuses on the major opinion part of the text to generate summaries that align with the author's intent. That is, it generates summaries that align with the sentiment of the original text. To further demonstrate the effectiveness of the proposed methodology, additional experiments were conducted to verify whether the generated summaries effectively

Dataset	Masking Ratio	ABS R-1	ABS R-2	ABS R-L	EXP R-1	EXP R-2	EXP R-L
Toy	85%	18.79	6.89	18.22	18.75	6.67	18.12
	70%	19.30	7.10	18.61	18.96	6.77	18.30
	55%	19.01	7.00	18.31	18.64	6.43	18.01
	40%	18.85	6.90	18.34	18.34	6.30	17.94
Sport	85%	18.57	6.71	17.92	18.54	6.72	17.83
	70%	18.62	6.70	17.93	18.50	6.63	17.83
	55%	18.34	6.54	17.64	18.32	6.52	17.62
	40%	18.21	6.42	17.31	18.15	6.22	17.34
Home	85%	18.00	6.61	17.43	17.97	6.40	17.31
	70%	18.10	6.67	17.45	18.06	6.50	17.43
	55%	17.94	6.54	17.30	17.64	6.32	17.32
	40%	17.74	6.32	17.01	17.21	6.21	17.11
Movie	85%	13.89	4.31	13.05	13.82	4.24	12.95
	70%	13.92	4.33	13.08	13.85	4.25	13.01
	55%	13.32	4.22	12.95	12.89	4.11	12.84
	40%	13.01	4.01	12.65	12.44	4.01	12.34

Table 9. Ablation study (T5 backbone)

Dataset	Method	Accuracy	F1-Score
Toy	Vanilla BART	85.01	91.32
	Ours (BART)	**86.93**	**92.47**
	Vanila T5	90.87	94.91
	Ours (T5)	**91.26**	**95.15**
Sport	Vanilla BART	75.90	85.21
	Ours (BART)	**77.03**	**85.90**
	Vanila T5	87.83	93.10
	Ours (T5)	**91.26**	**95.15**
Home	Vanilla BART	76.50	85.10
	Ours (BART)	**80.77**	**88.10**
	Vanila T5	85.7	91.52
	Ours (T5)	**91.26**	**95.15**
Moive	Vanilla BART	76.45	84.75
	Ours (BART)	**81.08**	**88.06**
	Vanila T5	83.38	89.80
	Ours (T5)	**91.26**	**95.15**

Table 10. Classification performance of generated summary

reflect the sentiment of the original text. Specifically, in Table 10, we conducted an experimental evaluation to assess the classification performance of the summaries generated by the proposed model. Our findings indicate notable improvements when compared with the baselines, demonstrating the enhanced quality and effectiveness of the generated summaries.

Experimental results indicate that the summaries generated by our proposed methodology were significantly superior to the baseline models across all datasets. This confirms that the summaries generated based on the proposed methodology effectively maintain the sentiment of the original text. Furthermore, this highlights the success of the proposed model in prioritizing key opinion segments.

4.3.5 Effectiveness of SHAP value. In the proposed methodology, major and minor sentiment parts are distinguished based on SHAP values and utilized for training. Therefore, this section verifies whether SHAP values are an effective measure for distinguishing major and minor sentiment parts in terms of the reliability of the proposed methodology.

Specifically, we compared the sentiment analysis results of the original sentences with the data extracted based on SHAP values, containing only tokens of major sentiment. We conducted experiments to determine whether sentiment scores are indeed more accurate and sharpened in sentences extracted using SHAP.

Dataset	Label	Original sentence (Pos:Neg)	Major sentiment part (Pos:Neg)
Toy	Positive	**0.7159**:0.3841	**0.9007**:0.0993
	Negative	0.3681:**0.6319**	0.0799:**0.9201**
Sport	Positive	**0.5640**:0.4360	**0.8715**:0.1285
	Negative	0.3383:**0.6617**	0.1015:**0.8985**
Home	Positive	**0.6817**:0.3183	**0.9247**:0.0753
	Negative	0.3218:**0.6782**	0.0882:**0.9118**
Moive	Positive	**0.5731**:0.4269	**0.8374**:0.1626
	Negative	0.3812:**0.6188**	0.1783:**0.8217**

Table 11. Comparison of sentiment prediction distributions

As shown in Table 11, when only major sentiment tokens were extracted based on SHAP values, it was observed that predictions were sharpened towards the ground truth distribution in both positive and negative cases. This indicates the effective extraction of major sentiment based on SHAP values and signifies that the SHAP values used in the proposed study are an effective methodology for distinguishing between major and minor sentiment parts.

5 CONCLUSION

Existing abstractive summary models have demonstrated state-of-the-art performance in summarizing fact-based documents, emphasizing the importance of objective information summaries. However, these models often cannot handle the unique characteristics of opinion-based documents, which requires an understanding of the intended message. Therefore, this paper proposes an abstractive summarization model that focuses on opinion-based documentation by considering the author's intended message.

Specifically, we propose a summarization model that prioritizes the core intent using the SHAP explainer to identify the sentiment of the entire document. This method enables the extraction of a major opinion segment that accurately reflects the intended message. In addition, we present an incremental learning strategy that initially addresses easier problems with high masking ratios and gradually adapts to more challenging problems with low masking ratios.

Through experiments on opinion-based documents, we have demonstrated that our proposed model achieves improved summarization performance compared with existing summarization models. Furthermore, by presenting actual examples of the generated summaries, we have confirmed the effectiveness of the proposed method. Furthermore, we determined the optimal parameters through an ablation study. Validation of the classification

performance based on the generated summaries further have demonstrated that the proposed model effectively delivers the author's intended message.

However, a limitation of our proposed research is that we solely relied on the explainer value derived from the SHAP model as a reference. Notably, the SHAP methodology employed in this study has exhibited stability issues when applied to text data, and there are limitations in the Shapley value calculation process, which forms the foundation of the SHAP methodology, due to the influence of correlations between variables. Therefore, in future studies, various improvements addressing these drawbacks of the SHAP technique could be proposed, and enhancing model robustness could be achieved by applying various XAI methodologies beyond SHAP.

Additionally, more effective end-to-end methods can be used in future studies, rather than random masking, which requires hyperparameter tuning. Furthermore, the proposed methodology yielded diminished performance on datasets where the distribution of sentiment was ambiguous. Hence, future research could propose methodologies robust to dataset variations.

ACKNOWLEDGEMENT

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00244789).

REFERENCES

[1] Omar Alqaryouti, Nur Siyam, Azza Abdel Monem, and Khaled Shaalan. 2019. Aspect-based sentiment analysis using smart government review data. *Applied Computing and Informatics* (2019).

[2] Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. 2021. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, 5 (2021), e1424.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Xin Cheng, Shen Gao, Yuchi Zhang, Yongliang Wang, Xiuying Chen, Mingzhe Li, Dongyan Zhao, and Rui Yan. 2023. Towards Personalized Review Summarization by Modeling Historical Reviews from Customer and Product Separately. *arXiv preprint arXiv:2301.11682* (2023).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Christine Dewi, Bing-Jun Tsai, and Rung-Ching Chen. 2022. Shapley Additive Explanations for Text Classification and Sentiment Analysis of Internet Movie Database. In *Asian Conference on Intelligent Information and Database Systems*. Springer, 69–80.

[7] Arwa Diwali, Kawther Saeedi, Kia Dashtipour, Mandar Gogate, Erik Cambria, and Amir Hussain. 2023. Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis. *IEEE Transactions on Affective Computing* (2023).

[8] Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications* 118 (2019), 272–299.

[9] Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6263–6273.

[10] Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792* (2018).

[11] Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1602–1613.

[12] Karthick Prasad Gunasekaran. 2023. Exploring Sentiment Analysis Techniques in Natural Language Processing: A Comprehensive Review. *arXiv preprint arXiv:2305.14842* (2023).

[13] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems* 28 (2015).

[14] Elvira Ismagilova, Emma Slade, Nripendra P Rana, and Yogesh K Dwivedi. 2020. The effect of characteristics of source credibility on consumer behaviour: A meta-analysis. *Journal of Retailing and Consumer Services* 53 (2020), 101736.

[15] Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. 16–21.

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).
[17] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
[18] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.
[19] Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. *arXiv preprint arXiv:1805.01089* (2018).
[20] Kathleen McKeown. 1992. *Text generation*. Cambridge University Press.
[21] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.
[22] Reema Narayan, Usha Y Nayak, Ashok M Raichur, and Sanjay Garg. 2018. Mesoporous silica nanoparticles: A comprehensive review on synthesis and recent advances. *Pharmaceutics* 10, 3 (2018), 118.
[23] Francesca Naretto, Roberto Pellungrini, Salvatore Rinzivillo, and Daniele Fadda. 2023. EXPHLOT: EXplainable Privacy Assessment for Human LOcation Trajectories. In *International Conference on Discovery Science*. Springer, 325–340.
[24] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab.
[25] Hou Pong Chan, Wang Chen, and Irwin King. 2020. A Unified Dual-view Model for Review Summarization and Sentiment Classification with Inconsistency Loss. *arXiv e-prints* (2020), arXiv–2006.
[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
[27] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659* (2019).
[28] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services* 64 (2022), 102771.
[29] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).
[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[31] Hui-Ju Wang. 2022. Understanding reviewer characteristics in online reviews via network structural positions. *Electronic Markets* 32, 3 (2022), 1311–1325.
[32] Hongyan Xu, Hongtao Liu, Pengfei Jiao, and Wenjun Wang. 2021. Transformer reasoning network for personalized review summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1452–1461.
[33] Hongyan Xu, Hongtao Liu, Zhepeng Lv, Qing Yang, and Wenjun Wang. 2023. Sentiment-aware Review Summarization with Personalized Multi-task Fine-tuning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2826–2835.
[34] Hongyan Xu, Hongtao Liu, Wang Zhang, Pengfei Jiao, and Wenjun Wang. 2021. Rating-boosted abstractive review summarization with neural personalized generation. *Knowledge-Based Systems* 218 (2021), 106858.
[35] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, 11328–11339.
[36] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. 2010. Extracting and ranking product features in opinion documents. In *Coling 2010: posters*. 1462–1470.
[37] Xiaofei Zhu, Zhanwang Peng, Jiafeng Guo, and Stefan Dietze. 2023. Generating effective label description for label-aware sentiment classification. *Expert Systems with Applications* 213 (2023), 119194.