**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 7.056**

***IJCSMC, Vol. 15, Issue. 2, February 2026, pg.1 – 8***

# Impact of Synthetic Review Filtering on the Performance of Sentiment Analysis: A Category-wise Analysis Using BERT+CRF

**Dae Su Ha[1]; Taek Lee[2]**

[1,2]Division of Computer Science and Engineering, Sun Moon University, Republic of Korea
[1] gkeotn66@sunmoon.ac.kr; [2] comtaek76@sunmoon.ac.kr

(Corresponding Author: Taek Lee)

**DOI: https://doi.org/10.47760/ijcsmc.2026.v15i02.001**

**Abstract: This study examines the impact of synthetic reviews, which have proliferated with the rapid adoption of generative AI, on the performance of sentiment analysis models in online review environments. Although AI-generated reviews often exhibit fluent and coherent language, they may fail to reflect genuine user experiences, potentially distorting training data distributions and degrading model performance. To investigate this issue, we conduct a comparative experiment using an AI Hub(https://aihub.or.kr/) aspect-based sentiment analysis dataset composed of real user reviews. Synthetic reviews generated under fixed conditions are selectively included or excluded from the training data to evaluate the effect of synthetic review filtering. Experiments are performed across five review categories—home appliances, IT devices, daily goods, cosmetics, and fashion—using a BERT+CRF-based model as the primary architecture, with an LSTM-based model as a baseline. The results show that removing synthetic reviews consistently improved sentiment analysis performance, with a notable increase in Macro-F1 scores. Moreover, the magnitude of improvement varies by review category, with larger gains observed in information-oriented domains such as home appliances and IT devices. These findings suggest that domain characteristics play a crucial role in the effectiveness of synthetic review filtering. Overall, this study provides empirical evidence that data refinement strategies are essential for enhancing the reliability and robustness of sentiment analysis systems in the era of generative AI.**

**Keywords: Synthetic Review, Sentiment Analysis, Aspect-Based Sentiment Analysis, BERT+CRF, Review Filtering**

 

## I. INTRODUCTION

With the rapid growth of online shopping, consumers have become increasingly reliant on reviews written by other users during the product purchasing process. Reviews provide both direct and indirect information regarding product quality, usage experience, and overall satisfaction, and they play a crucial role in shaping platform credibility and influencing purchase decisions. In this context, text-based sentiment analysis has emerged as a core technology in a wide range of applications, including e-commerce analytics, service monitoring, and customer relationship management systems. In particular, the use of aspect-based sentiment analysis (ABSA), which classifies users' positive, negative, or neutral sentiments toward specific product attributes, has gained increasing attention in recent years[1].

In this study, reviews generated by generative artificial intelligence are collectively referred to as Synthetic Reviews, and this term is used consistently throughout the paper. With the recent advancement of large language models (LLMs), AI-generated reviews can be easily produced and disseminated on online platforms. Although such reviews often exhibit grammatically fluent and contextually coherent expressions, they do not necessarily reflect genuine user experiences, which may undermine the informational reliability of review data. Furthermore, characteristics such as repetitive stylistic patterns, exaggerated emotional expressions, and atypical distributions of attribute mentions may distort the predictions of sentiment analysis models. These issues pose a potential threat to the overall accuracy and stability of review-based sentiment analysis systems.

The objective of this study is to investigate how AI-generated reviews affect the training distribution and performance of sentiment analysis models. To this end, we compare sentiment analysis performance using training datasets that either include or exclude Synthetic Reviews, and we examine whether the effects of Synthetic Review filtering vary across different review categories, namely home appliances, IT devices, daily goods, cosmetics, and fashion. In addition, while the analysis primarily focuses on the performance of a BERT+CRF-based sentiment analysis model, a traditional recurrent neural network model (LSTM) is employed as a baseline to verify that the observed effects of Synthetic Review filtering are not limited to a specific high-capacity model architecture[1][2].

The main contributions of this study are summarized as follows. First, we quantitatively evaluate the impact of including Synthetic Reviews in training data on sentiment analysis performance, thereby empirically examining the effectiveness of data filtering as a preprocessing strategy. Second, by comparing multiple model architectures, including BERT+CRF and LSTM, we demonstrate that the observed effects are not confined to a particular model structure. Building upon these contributions, this study provides practical insights into ensuring the reliability of review-based data processing in environments increasingly influenced by generative AI, and it offers meaningful implications for improving the quality of e-commerce platforms and user-generated content-based analysis systems.

## II. RELATED WORK

Text-based sentiment analysis and AI-generated text detection have been continuously evolving research areas in the field of natural language processing, supported by extensive empirical studies based on diverse model architectures and datasets. This section reviews major studies related to sentiment analysis, aspect-based sentiment analysis, Transformer-based models, and AI-generated review detection, and discusses how this study differs from the existing work.

### *A. Trends in Sentiment Analysis*

Sentiment analysis is a technique that automatically classifies users' sentiments expressed in text into categories such as positive, negative, or neutral. Early approaches primarily relied on sentiment lexicon-based methods or traditional machine learning algorithms, including Support Vector Machines (SVM) and Naive Bayes classifiers. Subsequently, recurrent neural network (RNN)-based models that account for word order were introduced, leading to the widespread adoption of Long Short-Term Memory (LSTM) architectures for text classification tasks. However, LSTM-based models suffer from structural limitations, such as difficulties in capturing long-range contextual dependencies and limited parallelization capability, which constrain learning efficiency when applied to large-scale datasets.

### *B. Aspect-Based Sentiment Analysis*

Aspect-based sentiment analysis (ABSA), which classifies sentiments with respect to fine-grained attributes of products or services, requires more detailed and nuanced analysis than general sentiment classification. Previous studies have typically focused on identifying specific aspects within a sentence and predicting the corresponding sentiment polarity. In this context, BERT-based models have been widely adopted due to their superior contextual representation capability[1][2]. More recent research has reported that models combining a BERT encoder with a Conditional Random Field (CRF) layer achieve strong performance by performing sequence labeling at the aspect level[1][3]. The CRF layer enables the optimization of label sequences by

 

modeling transition dependencies between token-level predictions generated by BERT, thereby improving contextual consistency in sentiment labeling.

*C. Transformer-Based Sentiment Analysis Models*

The Transformer architecture overcomes the limitations of RNN-based models by leveraging a self-attention mechanism that enables parallel learning of contextual information. As a result, models such as BERT, RoBERTa, and ELECTRA have been successfully applied to a wide range of natural language processing tasks, including text classification, sentence similarity measurement, and sentiment analysis[2][6]. BERT-based sentiment analysis models, in particular, demonstrate strong performance in both contextual understanding and representation learning, and have achieved high accuracy in Korean sentiment analysis tasks. Among these approaches, the BERT+CRF architecture has been frequently adopted for aspect-based sentiment analysis, as it provides stable performance across diverse text domains through fine-tuning.

*D. AI-Generated Text Detection*

With the rapid advancement of generative AI models, research on distinguishing between human-written and machine-generated text has also expanded significantly. Large language models such as GPT and LLaMA produce text that appears fluent and coherent on the surface; however, prior studies have identified differences in emotional expression patterns, sentence structure repetition, and abnormal word usage frequencies when compared to human-authored text[4][5]. Based on these characteristics, various AI-generated text detection models have been proposed. Among them, ELECTRA has shown strong performance in AI-generated text detection due to its pretraining objective, which explicitly discriminates between real and artificially generated tokens[6]. Previous studies have demonstrated the effectiveness of ELECTRA-based architectures for detecting AI-generated text, suggesting that generative AI can introduce structural distributional shifts in textual data. Although the present study does not implement an automated synthetic review detection model as its primary focus, prior work on ELECTRA-based detection provides indirect evidence that Synthetic Reviews may alter the statistical properties of review datasets.

*E. Originality of This Study*

Most prior studies have focused either on sentiment analysis or on AI-generated text detection as separate research problems, and relatively few have examined both within the same review data environment. In contrast, this study adopts a comparative experimental design that directly contrasts training datasets with and without AI-generated reviews, thereby empirically analyzing the impact of generative AI on review-based sentiment analysis systems. This integrated perspective distinguishes this study from existing research and contributes to a more comprehensive understanding of how Synthetic Reviews influence sentiment analysis performance.

## III. METHODOLOGY

We adopt a comparative experimental design based on the inclusion or exclusion of Synthetic Reviews in order to analyze their impact on the stability and accuracy of review-based sentiment analysis models. The methodology consists of three main components: dataset construction, model design, and experimental procedure.

The primary objective of this study is not to compare different generative AI models or prompt engineering strategies, but rather to examine how the presence of Synthetic Reviews in the training data distribution affects sentiment analysis performance. Accordingly, the Synthetic Review generation process was conducted under fixed conditions throughout all experiments, ensuring fairness in comparison and consistency in result interpretation.

*A. Dataset Construction*

For sentiment analysis experiments, we utilized the shopping mall review–based aspect-based sentiment analysis dataset provided by AI Hub[7]. To conduct Synthetic Review filtering experiments, additional Synthetic Reviews were generated by applying a single large language model (LLM)–based generative AI to real user reviews included in the original AI Hub dataset, modifying their writing style and expression while preserving the original content.

The generative AI model and prompt conditions used for Synthetic Review generation were kept identical across all experiments. The prompt design followed two core principles: first, preserving the original review's attribute and sentiment polarity, and second, maintaining semantic content while altering only the surface-level expression. This design ensured that experimental outcomes reflected differences in training data composition rather than variations in generation strategy.

To verify generation quality, randomly sampled Synthetic Reviews were manually inspected to confirm consistency between the original and generated reviews in terms of attributes and sentiment polarity, thereby minimizing unintended label changes. Each generated Synthetic Review was assigned an *is_ai* label, which was

 

used solely for constructing training datasets with and without Synthetic Reviews. Model performance was then compared under both conditions using an identical real-only test dataset.

*1) Data Splitting Strategy:* To fix the evaluation data distribution, only real user reviews were used to construct the test dataset. The training data were divided into two conditions based on the inclusion or exclusion of Synthetic Reviews. This design enabled a direct comparison of sentiment analysis performance under different training data compositions while holding the test distribution constant.

*B. Model Design*

This study compares two types of sentiment analysis models: a traditional LSTM-based model and a Transformer-based BERT+CRF model, focusing on their performance under different Synthetic Review filtering conditions.

*1) LSTM-Based Sentiment Analysis Model (Baseline) :* The Long Short-Term Memory (LSTM) model extends recurrent neural networks by capturing sequential dependencies in textual data. The baseline model consists of an embedding layer, one or more LSTM layers, and a fully connected (dense) output layer.

This model is widely used in sentiment analysis tasks and was adopted as a baseline to facilitate performance comparison with Transformer-based models.

*2) BERT+CRF–Based Aspect-Based Sentiment Analysis Model :* BERT (Bidirectional Encoder Representations from Transformers) is a self-attention–based model that learns bidirectional contextual representations and has demonstrated strong performance across various natural language processing tasks. In this study, review text and attribute information were provided as inputs to the BERT encoder to generate contextual embeddings. A Conditional Random Field (CRF) layer was applied at the output stage to perform sequence-level prediction by modeling transition probabilities between labels. This architecture is known to be effective for aspect-based sentiment analysis by maintaining contextual consistency in sentiment labelling.

*C. Experimental Design*

A unified experimental environment and quantitative evaluation metrics were applied to ensure fair comparison across models and data conditions.

*1) Evaluation Metrics :* The following evaluation metrics were used for both the BERT+CRF and LSTM models: Accuracy, Macro-F1, Micro-F1, and Aspect-level F1. Among these, Macro-F1 was selected as the primary evaluation metric, as it effectively reflects performance across imbalanced sentiment classes.

*2) Hyperparameter Settings :* Batch size, learning rate, and the number of training epochs were determined based on hyperparameter configurations validated in prior BERT fine-tuning studies[2][8]. The BERT+CRF model was fine-tuned using pretrained models provided by the HuggingFace Transformers library. Early stopping was applied to prevent overfitting during training.

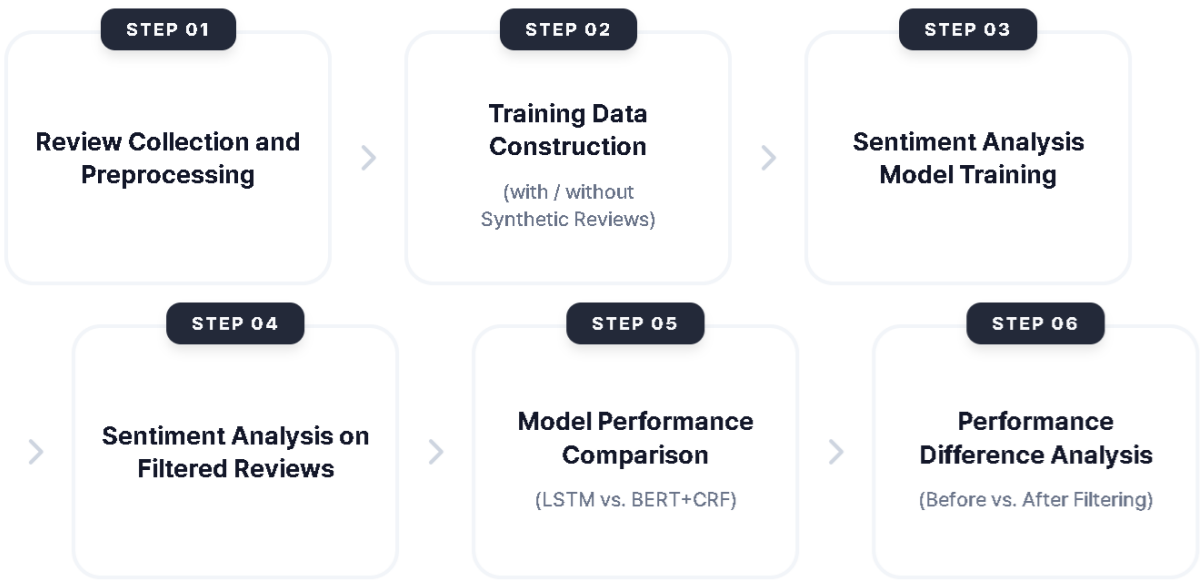


Fig. 1 Overall experimental pipeline for sentiment analysis with and without synthetic review filtering.

As shown in Fig. 1, our proposed experimental design enables a systematic evaluation of how AI-generated review filtering affects the performance and stability of sentiment analysis models.

 

## IV. ANALYSIS AND RESULTS

This section analyzes performance changes and category-wise differences resulting from training data configurations before and after Synthetic Review filtering, with a primary focus on the BERT+CRF–based sentiment analysis model. Through this analysis, we examine the overall trend of Synthetic Review filtering effects as well as domain-specific characteristics.

In this study, we formulate a research hypothesis that the effect of Synthetic Review filtering on sentiment analysis performance varies across review categories. Specifically, we hypothesize that the filtering effect of Synthetic Reviews differs between categories dominated by information-oriented reviews and those characterized by subjective and affective expressions. This hypothesis is examined through the subsequent experimental analysis for checking its validity.

### *A. Experimental Settings*

All experiments were conducted using the same BERT+CRF model architecture and identical hyperparameter settings, while varying only the composition of the training data. The evaluation dataset was fixed to real user reviews (real-only) to ensure that observed performance differences were attributable to changes in training data distribution rather than differences in test data composition. To maintain comparability across categories, an equal number of test samples was used for each category.

### *B. Overall Performance Comparison Before and After Synthetic Review Filtering*

Experimental results comparing training datasets with and without Synthetic Reviews indicate that sentiment analysis performance improved consistently when Synthetic Reviews were removed. In particular, an improvement of approximately 0.17 in the Macro-F1 score was observed as shown in Fig. 4. Because the evaluation dataset was fixed to real user reviews and the category-wise distributions in the training data were preserved, this improvement can be interpreted as resulting from differences in training data distribution rather than from changes in data volume.

This performance improvement was consistently observed across 10,000 iterations of paired bootstrap resampling under identical experimental settings, suggesting that the inclusion or exclusion of Synthetic Reviews significantly influenced model performance. To further assess the statistical robustness of the observed performance difference, a paired bootstrap–based confidence interval analysis was conducted on the fixed real-only test dataset. The mean Macro-F1 difference (After − Before) was estimated as 0.1675, with a 95% confidence interval of [0.1385, 0.1958].

As illustrated in Fig. 4, the bootstrap distribution of the Macro-F1 performance difference (After − Before) is centered well above zero, and the entire 95% confidence interval lies in the positive range. This indicates that the observed performance improvement after Synthetic Review removal is statistically robust and consistently positive across resampled test sets, rather than being driven by random sampling variation.

A similar performance improvement trend was also observed for the LSTM-based model after Synthetic Review removal; however, both the absolute performance level and the magnitude of improvement were more pronounced in the BERT+CRF model.

### *C. Category-wise Performance Change Analysis*

As shown in Fig. 2, category-wise analysis revealed that Macro-F1 performance improved across all categories after Synthetic Review filtering, although the magnitude of improvement varied. As illustrated in Fig. 3, the largest performance gain was observed in the home appliance category, followed by IT devices, daily goods, and cosmetics, while the fashion category exhibited a relatively smaller improvement.

These results suggest that the effectiveness of Synthetic Review filtering is influenced by linguistic characteristics and information density inherent to each review category. Categories dominated by objective and information-rich descriptions benefited more substantially from filtering, whereas those characterized by subjective and affective expressions showed comparatively limited improvements.

 

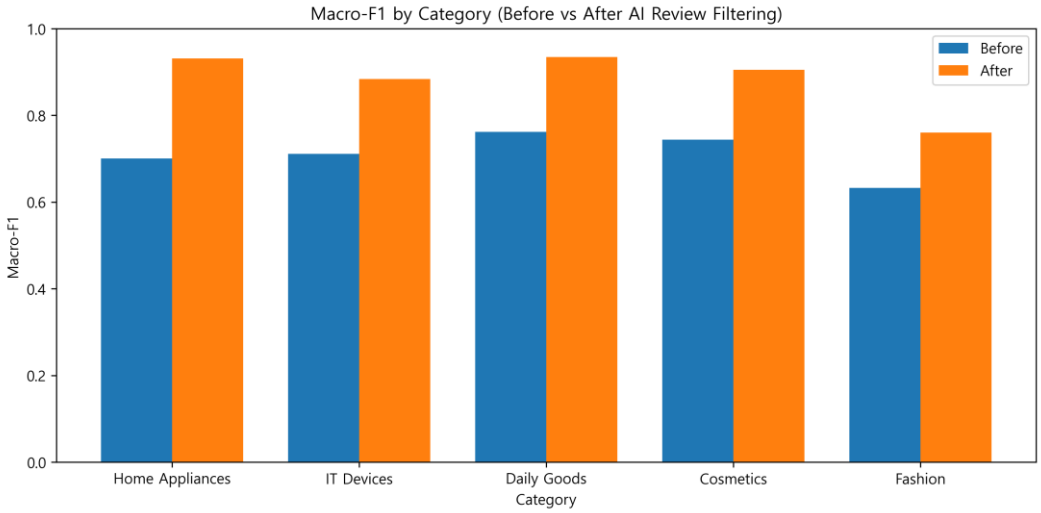


Fig. 2 Macro-F1 performance by category before and after synthetic review filtering (BERT+CRF).

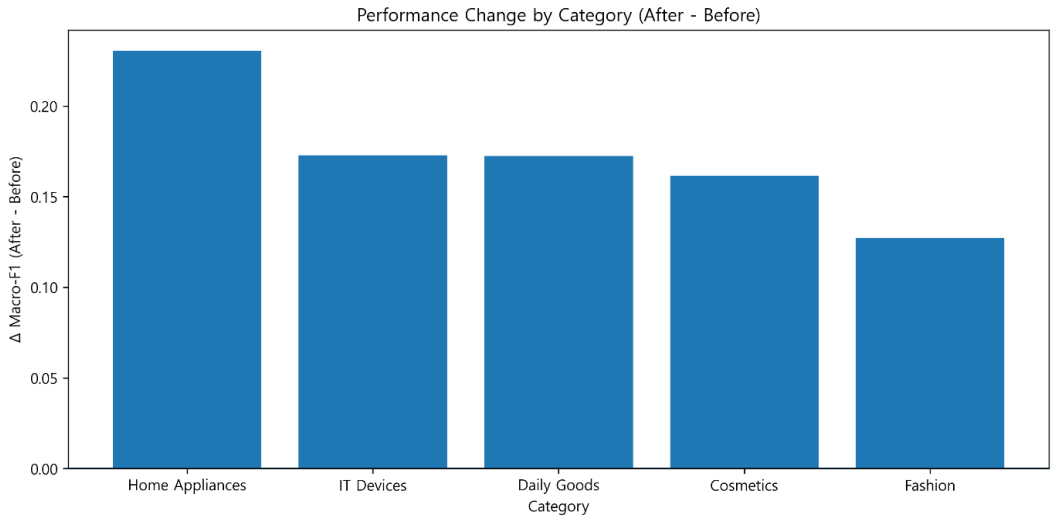


Fig. 3 Category-wise performance change (Δ Macro-F1) after synthetic review filtering

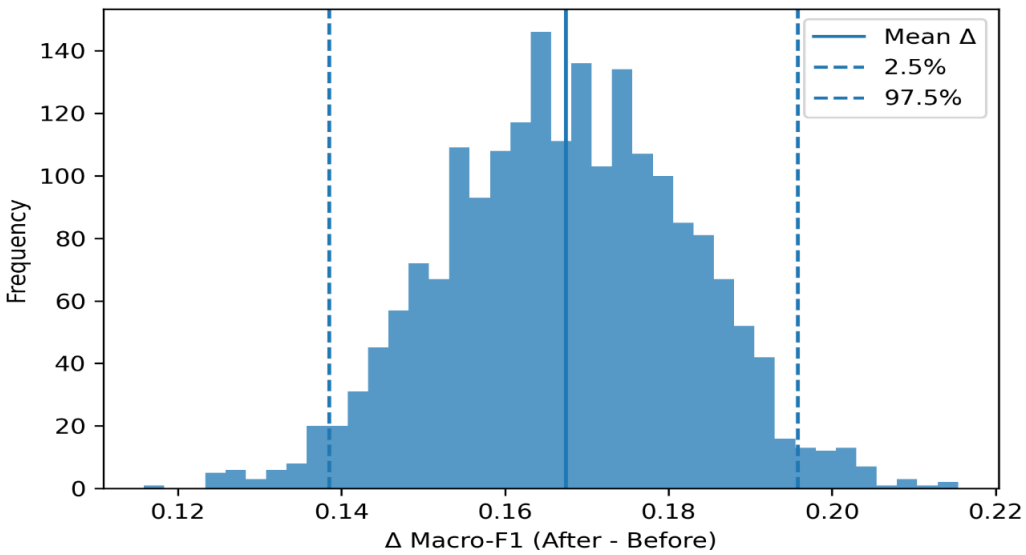


Fig. 4 Bootstrap distribution of Δ Macro-F1 (After − Before) on the fixed real-only test set. Dashed lines indicate the 95% confidence interval.

 

Fig. 2 visually compares category-wise Macro-F1 performance of the BERT+CRF model before and after Synthetic Review filtering, while Fig. 3 illustrates the magnitude of performance change following filtering. These figures confirm that the effect of Synthetic Review filtering varies substantially across review categories.

*D. Discussion*

The results indicate that generative Synthetic Reviews can distort the training data distribution of sentiment analysis models, and that the extent of this distortion varies across review domains. In other words, Synthetic Review filtering functions not merely as a means of improving overall performance, but as an important preprocessing strategy for modulating domain-specific performance characteristics.

Notably, the bootstrap-based confidence interval analysis showed that the 95% confidence interval for the Macro-F1 performance difference after Synthetic Review removal did not include zero, providing quantitative evidence that the observed performance improvement follows a consistent positive trend. However, since this analysis is based on resampling within a fixed test dataset, future work should further validate the generalizability of these findings through alternative data splits or repeated experimental settings.

## V. CONCLUSION

This study examined how the degradation of data reliability caused by the proliferation of generative AI in online review environments affects the performance of sentiment analysis models. To this end, we employed a comparative experimental design using training datasets with and without Synthetic Reviews and quantitatively evaluated performance changes, with a primary focus on a BERT+CRF–based sentiment analysis model.

The experimental results demonstrate that training data from which Synthetic Reviews were removed led to a consistent improvement in overall sentiment analysis performance. In particular, a clear improvement was observed in the Macro-F1 metric, indicating enhanced prediction stability and generalization performance under class-imbalanced conditions. These findings suggest that generative Synthetic Reviews can distort the training data distribution of sentiment analysis models, thereby degrading predictive accuracy.

Furthermore, category-wise analysis revealed that the effects of Synthetic Review filtering were not uniform across domains. Categories characterized by functional and information-oriented reviews, such as home appliances and IT devices, exhibited larger performance gains after Synthetic Review removal. In contrast, categories dominated by subjective and affective expressions, such as fashion, showed relatively limited improvement. This result indicates that linguistic characteristics and sentiment expression patterns inherent to each review domain act as important moderating factors in the effectiveness of Synthetic Review filtering.

Taken together, these findings support Research Hypothesis H1, which posits that the impact of Synthetic Review filtering on sentiment analysis performance differs significantly depending on the review category. This study contributes by positioning data refinement strategies not merely as a preliminary preprocessing step, but as a core design consideration that must account for domain-specific characteristics in sentiment analysis systems.

Nevertheless, this study has several limitations. First, the Synthetic Reviews considered in this work were restricted to reviews generated by a single generative AI model through stylistic and expressive transformations of real user reviews, and thus do not encompass a broader range of generative models or generation strategies. Second, although the generation model and prompt conditions were fixed throughout the experiments to ensure fair comparison, the effects of different generation mechanisms on sentiment analysis performance were not explored. Third, the analysis was limited to the shopping review domain; therefore, future research should extend the scope to diverse text domains to further validate the generalizability of Synthetic Review filtering effects.

Overall, this study highlights the importance of managing Synthetic Reviews in various application contexts, including e-commerce platforms, user opinion analysis systems, and review-based decision support services.

# References

**[1].** Hu Xu, Bing Liu, Lei Shu, Philip S. Yu, “BERT Post-Training for Review Reading Comprehension and Aspect-Based Sentiment Analysis”, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/N19-1242/

**[2].** Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang, “How to Fine-Tune BERT for Text Classification?”, in *Proceedings of the 18th China National Conference on Chinese Computational Linguistics (CCL)*, 2019. DOI: 10.1007/978-3-030-32381-3_16

**[3].** Jie Yang, Yue Zhang, “NCRF++: An Open-source Neural Sequence Labeling Toolkit”, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/P18-4013/

 

**[4].** Sebastian Gehrmann, Hendrik Strobelt, Alexander M. Rush, "GLTR: Statistical Detection and Visualization of Generated Text", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. [Online]. Available: https://aclanthology.org/P19-3019/

**[5].** Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, Chelsea Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature", in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR, 2023. [Online] Available from: https://proceedings.mlr.press/v202/mitchell23a.html

**[6].** Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators", in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=r1xMH1BtvB

**[7].** AI Hub, "Aspect-Based Sentiment Analysis Dataset for Online Shopping Reviews", Ministry of Science and ICT (MSIT), National Information Society Agency (NIA), Republic of Korea. [Online]. Available: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71603

**[8].** Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *Proceedings of NAACL-HLT*, Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/N19-1423/

 