Image-Text Multi-modal Sponsored Review Detection

Jiyoon Kim
Artificial Intelligence
Sungkyunkwan University
abcdef@skku.edu

Kyutae Kim
Data Science
Sungkyunkwan University
rlarbxo0324@gmail.org

Bogeun Kim
Statistics
Sungkyunkwan University
kbg927@gmail.com

Sun Huh
Culture and Technology
Sungkyunkwan University
abcxyz02@g.skku.edu

Abstract

This study aims to develop a multimodal approach to detect sponsored reviews on the e-commerce platform Coupang, based solely on review content without using user information. Using a balanced dataset of 6,238 reviews (3,119 sponsored and 3,119 unsponsored) collected across various product categories, features were extracted from text, images, and metadata. The proposed model employs an intermediate fusion approach, combining text features extracted with a Text-CNN model, visual features from a pre-trained VGG19 network, and metadata features through a fully connected layer. This model outperformed baseline approaches like VSCNN and SMPC, achieving an overall accuracy of 0.934, precision of 0.94/0.92, and recall of 0.93/0.93 for sponsored/unsponsored reviews respectively. Analysis revealed that sponsored reviews tend to have more images, higher ratings, longer text, and more structured formatting (e.g., more line breaks, presence of titles) compared to unsponsored reviews. The study demonstrates the feasibility of identifying sponsored reviews based solely on content, which has academic implications for advancing content-based detection methods and practical implications for maintaining consumer trust and brand integrity. Limitations include the inability to use multiple images per review and the need for a larger dataset, suggesting areas for future work.

1. Introduction

The e-commerce market is rapidly growing around the world. Coupang, one of South Korea's e-commerce platforms, reported annual revenue of approximately 31.83 trillion won in 2023 (Coupang, 2024). The number of Coupang's customers has grown more and more, with monthly active users becoming 21 million in 2023, a 16% increase from the previous year's 18.1 million (Coupang, 2024). One of the important factors in the growth of e-commerce is consumer reviews. Research said that the higher the number of consumer reviews, the more likely it is to stimulate purchases (Mudambi & Schuff, 2010). As a result, many companies sponsor products to consumers and encourage them to write reviews. However, sometimes, this can be misused. For example, some reviews are written to appear voluntary, although it was sponsored. Such undisclosed sponsored reviews would harm consumer trust and impact brand image negatively. According to the Fair Trade Commission, only in 2022, there were approximately 20,000 cases of fake reviews (Fair Trade Commission, 2023). Thus, detecting between sponsored and unsponsored reviews is important for enhancing consumer decision-making and the reliability of reviews. Previous attempts to identify sponsored reviews have heavily relied on user information. Jindal and Liu (2007) utilized user data to detect sponsored reviews. Also, Ott et al. (2011) combined review data with machine learning techniques to detect fake reviews. However, these methods face limitations due to user anonymity and data collection problems, underlining the need for content-based detection methods. We aim to develop a multimodal approach that combines image and text data to identify sponsored reviews without relying on user information. Also, we seek to analyze the differences between sponsored and unsponsored reviews. Our research questions are defined as follows:

- RQ1: Can sponsored reviews be identified based solely on review content without user information?
- RQ2: What are the differences between sponsored and unsponsored reviews?

Unlike most studies that utilize user information, our research focuses on identifying sponsored reviews based solely on content, demonstrating that limited information can still be effective. Our model achieves high performance in predicting sponsored reviews by employing multimodal techniques that integrate image and text features.

2. Related Works

Intermediate fusion combines features from different modalities at an intermediate stage before classification. Boulahia reported that intermediate fusion captures complex inter-modal interactions, improving feature representation and demonstrating superior classification accuracy compared to early and late fusion. Additionally, it effectively handles data synchronization issues and maintains robust inter-modal interactions.

Text-CNN models have been widely used in text classification due to their ability to capture local dependencies and hierarchical structures in text data. Kim (2014) introduced a simple yet powerful text-CNN model for sentence classification, which outperformed traditional models by leveraging convolutional filters to extract n-gram features. This model has been adapted and extended in various applications, such as spam detection (Zhang et al., 2015), highlighting its versatility and effectiveness in handling diverse text classification tasks.

The VGG network, introduced by Simonyan and Zisserman (2014), is known for its simplicity and effectiveness, utilizing small (3x3) convolution filters to increase the depth of the network. Kiela and Bottou (2014) reported that VGG19 can be used to extract image features, which were then combined with textual features from a recurrent neural network for image captioning tasks. Ngiam et al. (2015) utilized VGG19 within a multimodal fusion framework to integrate visual and auditory data, resulting in significant improvements in audiovisual speech recognition.

The inclusion of specific features like line breaks and title presence has been explored in various classification tasks to enhance model performance. Shin et al. (2016) reported the impact of structural features, including line breaks and title presence, in document classification. They found that incorporating these features, alongside traditional textual features, significantly boosted the classification accuracy of news articles.

3. Dataset

In this study, we collected review data from Coupang, which includes the reviewer, review content, date of writing, rating, images, and sponsorship status. The data was collected by Selenium in Python. We collected product data whose categories are such as home interior, beauty, and nutritional supplements. These categories were chosen because they have a high volume of sponsored reviews on Coupang and consist of items frequently used by consumers. Initially, we collected a total of 13,444 reviews. For data preprocessing, we labeled reviews containing phrases like “Get sponsored” as sponsored. Reviews with a “Coupang Experience Group” badge were also labeled as sponsored. To ensure clear learning, phrases like “Get sponsored” were removed from the review data. For review text preprocessing, we deleted very short reviews and reviews without images. After these preprocessing steps, we reduced the dataset to 6,238 reviews. The dataset was balanced with 3,119 sponsored and 3,119 unsponsored reviews, 1:1 ratio. We extracted several features for our analysis. The first image of each review was used for modeling, indexed from the review image column. The review clean column, a combination of the cleaned review text and title, was used for text features. Also, we extracted review meta features such as the number of images (img num), presence of a title (title1), number of helpful votes (helpfulness), star rating (rate), length of the review text (review num), and number of line breaks (line breaks). Reviewer meta features included whether the reviewer used their real name (realname reviewer) and whether the reviewer has a top badge (top reviewer), consolidated into a single feature. The final dataset included key columns such as label, which indicates whether the review is sponsored; review clean, which contains the cleaned review text; review image, which is the first image associated with the review; and review meta, which contains metadata related to the review such as the number of images (img num), presence of a title (title1), number of helpful votes (helpfulness), rating (rate), and review length (review num). Additionally, reviewer meta contains metadata related to the reviewer, such as whether the reviewer used their real name (realname reviewer) and has a top badge (top reviewer). For the modeling process, we used a combination of features including the cleaned review text, the first image from the review image column, and all metadata from the review meta and reviewer meta columns. To ensure the features selected had low correlation, we analyzed their relationships and visualized the correlations using a heatmap.

4. Method

An overview of our model is presented in Fig. 2. We adopted intermediate fusion for this task. This approach was chosen because it showed better performance compared to early fusion and late fusion methods. Our model consists of modules that extract features from different modalities and a part that integrates these features.

4.1. Text Feature Module

The input to this module is the list of words of the reviews. We employ convolutional neural networks (Text-CNN) to extract textual features, similar to the approach

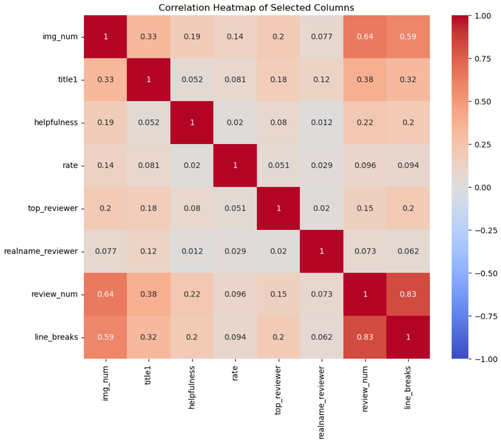

Figure 1. Heatmap of the features' correlation

Column	Description
review image	The first image associated with each review
review clean	Combination of the cleaned review text and title
img num	Number of images
title1	Presence of a title
helpfulness	Number of helpful votes
rate	Star rating
review num	Length of the review text
line breaks	Number of line breaks
date	Days elapsed since the review date
realname reviewer	Whether the reviewer used their real name
top reviewer	Whether the reviewer has a top badge
label	Indicates whether the review is sponsored

Table 1. Description of Collected Data

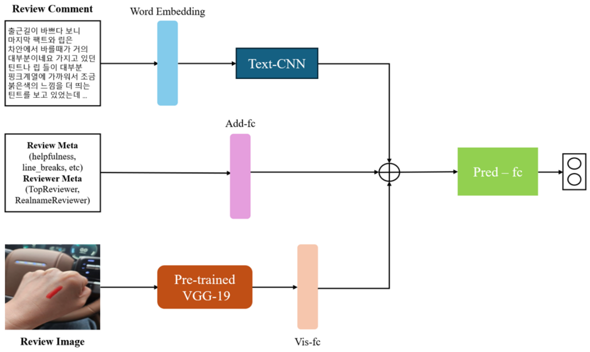

Figure 2. Overview of our model

used by Wang et al. in EANN.

First, each word in the text is represented as a word embedding vector. For the i-th word in the sentence, the corresponding k-dimensional word embedding vector is denoted as $T_i \in \mathbb{R}^k$. Thus, a sentence with n words is represented as:

$$T_{1:n} = T_1 \oplus T_2 \oplus \ldots \oplus T_n$$

where $\oplus$ is the concatenation operator. In the convolution layer, multiple filters of varying sizes are applied to capture different n-gram features in the text. Specifically, the operation of a convolutional filter with a window size h starting with the i-th word can be represented as:

$$t_i = \sigma(W_c \cdot T_{i:i+h-1})$$

where $\sigma()$ is the ReLU activation function and W_c is the weight of the filter. Then we get a feature vector $t = [t_1, t_2, \cdots, t_{n-h+1}]$ for a sentence.

After the convolutional steps, we apply a max-pooling operation for each feature vector t to capture the most important features and reduce the dimension. The results from the pooling operations are then concatenated into a single feature vector. Finally, the concatenated vector is passed through a fully connected layer with a LeakyReLU activation function, and R_T is obtained as:

$$R_T = \sigma(W_{tf} \cdot R_T^*)$$

where R_T^* is the textual features after the max-pooling, and W_{tf} is the weight matrix of the fully connected layer.

4.2. Visual Feature Module

The images are passed through a VGG-19 network that has been pre-trained on the ImageNet dataset. To match the required input dimension, images are resized to 224×224. The extracted feature vectors are then passed through a fully connected layer, referred to as Vis-fc. The visual feature representation R_V is obtained as:

$$R_V = \sigma(W_{vf} \cdot R_V^*)$$

where R_V^* is the visual feature obtained from the pre-trained VGG-19 network, and W_{vf} is the weight matrix of the Vis-fc.

4.3. Meta Feature Module

This module is designed for additional features of the review data. We performed standard scaling to normalize these features, which helps stabilize the learning process and improve the performance. The normalized vectors are then passed through a fully connected layer. With this procedure, the output of this module is denoted as:

$$R_M = FC(M_{scaled})$$

where M_{scaled} is the normalized feature and $FC(\cdot)$ is the sequence of the fully connected layer, including batch normalization, ReLU activation function, and dropout.

4.4. Model Integration

The outputs from each module (128-dimensional vectors from both text and visual modules, and a 64-dimensional vector from the meta feature module) are concatenated into a single combined feature vector denoted as $R_C = R_T \oplus R_V \oplus R_M$. This combined vector is passed through a fully connected layer with softmax to predict whether the review is sponsored or unsponsored.

To train the model, we employ the cross-entropy loss function, which is widely used for binary classification tasks. The cross-entropy loss L_d for the fake review detection is computed as:

$$L_d = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

where N is the number of samples, y_i is the true label for the i-th sample, and $\hat{y}_i$ is the predicted probability that the i-th review is fake. Therefore, the objective is to minimize the loss function L_d. Adam optimizer is used with a learning rate of 0.0005.

5. Experiment

5.1. Baseline

SMPC: boosting social media popularity prediction with caption. [12] : SMPC (Social Media Popularity Prediction with Caption) architecture processes caption information at the word, sentence, and length levels to extract three distinct caption features, which are then concatenated with multimodal features. These combined representations are subsequently trained using a CatBoost regression model to achieve optimal performance.

Multimodal deep learning framework for image popularity prediction on social media. [13] : It is a dual-CNN model, which utilizes VGG-19 for extracting visual features and PCA reduction for social features to predict post popularity within a dataset.

5.2. Model Description

VGG19 Model: VGG19 is a convolutional neural network that extracts visual features from images. It uses 19 layers with small 3x3 convolution filters to increase depth and capture complex image features.

Text-CNN Model: The Text-CNN model applies convolutional neural networks to text data, using convolutional filters of various sizes to capture n-gram features in sentences. It processes text to extract important textual features for classification.

VSCNN Model: This model combines visual features from the VGG19 network and textual features from the Text-CNN model using a multimodal approach. It integrates these features to improve classification accuracy for detecting sponsored reviews.

SMPC: SMPC(Social Media Popularity Prediction with Caption) architecture processes caption information at the word, sentence, and length levels to extract three distinct caption features, which are concatenated with multimodal features. These combined representations are trained using a CatBoost regression model.

Ours (Proposed Model): The proposed model employs intermediate fusion to combine text, image, and meta features. It uses Text-CNN for text features, VGG19 for visual features, and a fully connected layer for meta features, integrating them to predict sponsored reviews with high accuracy.

6. Result

6.1. Experiment Result

As shown in Table 3a, our model outperformed the existing models in all metrics, including VSCNN's recall score. This indicates the superior performance of our model compared to the baselines.

The results of the ablation study, presented in Table 3b, indicate that the text modality plays a more significant role in successful predictions compared to the image modality. Additionally, the review meta data appears to be more important than the reviewer meta data. However, our experimental approach, which utilizes all modalities, outperformed cases where certain modalities were excluded across all metrics.

6.2. Difference between Sponsored Review and Unsponsored Review

Sponsored reviews and unsponsored reviews exhibited differences across various features, as shown in Table 2.

Review Type	img num	rate	review num	title1	title1(X)	line break
Sponsored	6.79	4.91	730.48	2294	315	30.1
Unsponsored	4.61	4.84	353.32	825	2804	15.22

Table 2. Compare feature values by review type

According to Table 2, sponsored reviews recorded higher values in all aspects compared to unsponsored reviews. Sponsored reviews had higher numbers in terms of the number of images and ratings, and the length of the reviews was approximately twice as long as that of unsponsored reviews. Notably, in elements related to the structure of the reviews, such as the presence of a title (title1), the absence of a title (title1(X)), and the number of line breaks, sponsored reviews demonstrated significantly higher structural quality. Specifically, the differences were approximately threefold, ninefold, and twofold respectively, indicating a much more

Model	Label	Accuracy	Precision	Recall	F1-Score
image (VGG19)	Sponsored	0.78	0.81	0.78	0.80
	Unsponsored		0.75	0.79	0.77
text (Text-CNN)	Sponsored	0.88	0.87	0.92	0.89
	Unsponsored		0.89	0.84	0.86
VSCNN	Sponsored	0.91	0.90	**0.94**	0.92
	Unsponsored		0.92	0.87	0.90
SMPC	Sponsored	0.88	0.88	0.92	0.90
	Unsponsored		0.90	0.84	0.87
ours	Sponsored	**0.934**	**0.94**	**0.94**	**0.93**
	Unsponsored		**0.92**	**0.93**	**0.92**

(a) Comparison of Model Performance

Model	Label	Accuracy	Precision	Recall	F1-Score
image	Sponsored	0.78	0.81	0.78	0.80
	Unsponsored		0.75	0.79	0.77
text	Sponsored	0.88	0.87	0.92	0.89
	Unsponsored		0.89	0.84	0.86
image + text	Sponsored	0.89	0.87	0.93	0.90
	Unsponsored		0.91	0.83	0.87
image + text + reviewer meta	Sponsored	0.89	0.89	0.92	0.90
	Unsponsored		0.90	0.87	0.88
image + text + review meta	Sponsored	0.92	0.92	0.93	0.92
	Unsponsored		0.92	0.90	0.91
ours	Sponsored	**0.934**	**0.94**	**0.93**	**0.93**
	Unsponsored		**0.92**	**0.93**	**0.92**

(b) Comparison of Ablation Study Results

Figure 3. Model Performance and Ablation Study Results

organized structure in sponsored reviews. The detailed analysis of line break is presented in Fig. 4. We have confirmed the statistical significance of the differences in feature values between all sponsored and unsponsored reviews using t-tests and p-values.

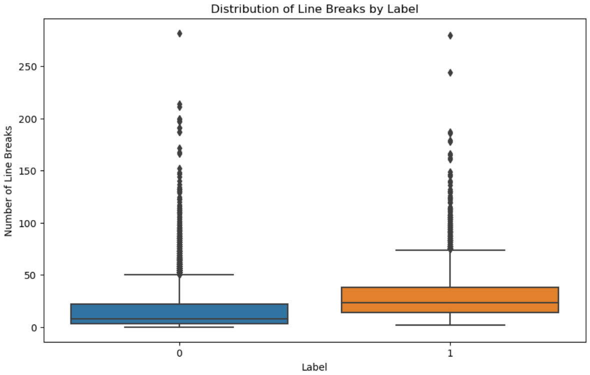

Figure 4. Difference of line break by review type

7. Discussion

7.1. Overview of the Research

In this study, we aimed to develop a multimodal approach to detect sponsored reviews based on image and text data, without relying on user information. We collected and preprocessed a dataset from Coupang, focusing on reviews in categories with a high volume of sponsored content. Our model integrates text, image, and meta features to predict whether a review is sponsored. Through our experiments, we demonstrated that our model achieves high performance in identifying sponsored reviews, highlighting the importance of multimodal techniques in this task.

7.2. Summary of Research Findings

- **RQ1: Can sponsored reviews be identified based solely on review content without user information?** - We successfully built a model with an accuracy of 0.92 using only content-based data, excluding reviewer information.
- **RQ2: What are the differences between sponsored and unsponsored reviews?** - Sponsored reviews often included titles and had, on average, about two times of line breaks than unsponsored reviews. This indicates that sponsored reviews are generally more structured compared to unsponsored reviews.

7.3. Implications

7.3.1 Academic Implication

We have demonstrated that it is possible to build an effective model for predicting sponsored reviews without using user information. This advances the field by focusing on content-based features.

7.3.2 Practical Implication

By identifying fake reviews where users pretend not to have received sponsorship, we can reduce the negative impact of undisclosed sponsored content. This can help maintain consumer trust and protect brand integrity.

7.4. Limitations and Future Works

We were unable to use multiple images per review and had to rely on just one first image for each review. For the same issue, we were not able to conduct experiments with a larger dataset.

References

[1] Coupang. (2024). ”Annual revenue and customer growth statistics.” Retrieved from Coupang.

[2] Fair Trade Commission. (2023). ”Cases of fake reviews in 2022.” Retrieved from Fair Trade Commission.

[3] Jindal, N., & Liu, B. (2007). ”Analyzing and detecting review spam.” In Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), pages 547-552. IEEE.

[4] Ott, M., Choi, Y., Cardie, C., & Hancock, J.T. (2011). ”Finding deceptive opinion spam by any stretch of the imagination.” In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 309-319.

[5] Mudambi, S. M., & Schuff, D. (2010). ”What makes a helpful review? A study of customer reviews on Amazon.com.” MIS Quarterly, 34(1), 185-200.

[6] Boulahia, T., Amiri, H., & Dornaika, F. (2019). ”Intermediate fusion for multi-modal object recognition.” Neural Processing Letters, 49(3), 1001-1018.

[7] Kim, Y. (2014). ”Convolutional neural networks for sentence classification.” arXiv preprint arXiv:1408.5882.

[8] Zhang, X., Zhao, J., & LeCun, Y. (2015). ”Character-level convolutional networks for text classification.” In Advances in Neural Information Processing Systems (pp. 649-657).

[9] Simonyan, K., & Zisserman, A. (2014). ”Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556.

[10] Kiela, D., & Bottou, L. (2014). ”Learning image embeddings using convolutional neural networks for improved multi-modal semantics.” In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 36-45).

[11] Shin, D., Park, J., & Joo, C. (2016). ”Importance of structural features in document classification: Line breaks and title presence.” Journal of Information Science, 42(6), 765-779.

[12] Liu, A. A., Wang, X., Xu, N., et al. (2023). ”SMPC: boosting social media popularity prediction with caption.” Multimedia Systems, 29, 577–586. https://doi.org/10.1007/s00530-022-01030-5 4

[13] Abousaleh, F. S., Cheng, W. H., Yu, N. H., & Tsao, Y. (2020). ”Multimodal deep learning framework for image popularity prediction on social media.” IEEE Transactions on Cognitive and Developmental Systems, 13(3), 679-692. 4