International Journal of Computational Intelligence Systems (2024) 17:97
https://doi.org/10.1007/s44196-024-00495-8

RESEARCH ARTICLE

# VSEM-SAMMI: An Explainable Multimodal Learning Approach to Predict User-Generated Image Helpfulness and Product Sales

**Chengwen Sun**[1] · **Feng Liu**[2]

Received: 30 November 2023 / Accepted: 30 March 2024


**Abstract**
Using user-generated content (UGC) is of utmost importance for e-commerce platforms to extract valuable commercial information. In this paper, we propose an explainable multimodal learning approach named the visual–semantic embedding model with a self-attention mechanism for multimodal interaction (VSEM-SAMMI) to predict user-generated image (UGI) helpfulness and product sales. Focusing on SHEIN (i.e. a fast-fashion retailer), we collect the images posted by consumers, along with product and portrait characteristics. Moreover, we use VSEM-SAMMI, which adopts a self-attention mechanism to enforce attention weights between image and text, to extract features from UGI then use machine learning algorithms to predict UGI helpfulness and product sales. We explain features using a caption generation model and test the predictive power of embeddings and portrait characteristics. The results indicate that when predicting commercial information, embeddings are more informative than product and portrait characteristics. Combining VSEM-SAMMI with light gradient boosting (LightGBM) yields a mean squared error (MSE) of 0.208 for UGI helpfulness prediction and 0.184 for product sales prediction. Our study offers valuable insights for e-commerce platforms, enhances feature extraction from UGI through image–text joint embeddings for UGI helpfulness and product sales prediction, and pioneers a caption generation model for interpreting image embeddings in the e-commerce domain.

**Keywords** User-generated content · User-generated image · Deep learning · Image analysis · E-commerce · Multimodal learning

## 1 Introduction

The development of the Internet and mobile devices has significantly propelled advancements in e-commerce. User-generated content (UGC) published on e-commerce platforms helps consumers gain product awareness. To extract valuable commercial information such as product sales [1, 2], purchase intention [3–5], and brand perception [6], previous literature has made valuable contributions to utilizing text content in UGC. The full potential of user-generated image (UGI) may yet be unexploited, however, given the sole reliance upon pre-trained image recognition techniques or text mining methodologies to retrieve labels from images. Previous multimodal fusion studies emphasized the challenges that come from interpreting such embeddings [6, 7]. For example, while these models can predict product sales, they struggle to identify how each feature affects product sales. Further exploration is urgently needed to leverage the features extracted from UGI, which are essential for predicting commercial information. Significant research questions arise from these circumstances: (1) How can we extract more information from UGI, and (2) How can we achieve the interpretability of multimodal learning model?

To fill this gap, we present the VSEM-SAMMI: visual–semantic embedding model with a self-attention mechanism for multimodal interaction, a novel approach for explainable multimodal learning. Specifically, VSEM-SAMMI includes three parts, i.e., visual embedding generation with pre-trained ResNet, semantic embedding generation with Word2Vec, and multimodal interaction with

✉ Feng Liu
liufeng@sdu.edu.cn

Chengwen Sun
scw0424@mail.sdu.edu.cn

[1] SDU-ANU Joint Science College, Shandong University, Weihai, China

[2] Business School, Shandong University, Weihai, China

Published online: 18 April 2024

a self-attention mechanism. This model extracts information from UGI. We then leverage machine learning algorithms for two prediction studies: UGI helpfulness prediction and product sales prediction based on SHEIN datasets—SHEIN is a B2C e-commerce site specializing in fast fashion. For the multimodal learning model to achieve interpretable results, we determine the best-performing machine learning model for each prediction study and use it to examine the importance of each feature through SHapley Additive exPlanations (SHAP). A caption generation model is used to describe the embeddings in natural language.

This research has several objectives. First, the data are divided into four types: multimodal, discrete, dimensional continuous, and non-dimensional continuous. We then employ VSEM-SAMMI to obtain multimodal data. Second, we use 16 machine learning approaches and obtain the best-performing model for each prediction task to predict UGI's helpfulness and product sales. Third, to interpret the image embeddings, SHAP is implemented with the best-performing model and a caption generation model based on long short-term memory (LSTM). Finally, an ablation study is conducted to evaluate whether image–text joint embeddings are more informative than feature-engineered features.

The first contribution that this research makes to the literature on multimodal learning and e-commerce is the development of VSEM-SAMMI to embed and extract information from images. This work also demonstrates how VSEM-SAMMI can expand to address other challenges related to UGI utilization since employing joint embeddings enhanced the prediction accuracy of both UGI helpfulness and product sales, yielding mean squared error (MSE) values of 0.208 and 0.184, respectively. We further present a novel caption generation model for interpreting image embeddings to generate natural language. Overall, this study offers insight for e-commerce platform managers to evaluate the quality of UGI and provide decision support related to production planning, marketing strategies, inventory control, and supply chain management.

## 2 Related Works

### 2.1 User-Generated Content

UGC refers to all forms of social media output created and shared by non-professional users, including text, image, video, and audio [8–10]. Understanding how to utilize UGC to extract information is an important topic for scholars studying purchase intention mining [3–5], disinformation detection [11–13], and brand perceptions [9, 14, 15]. Gupta [16] pioneered the extraction of seven features from UGC text to predict purchase intent, and Smith [17] presented a preliminary set of dimensions to compare brand-related UGC by examining the differences in UGC across Twitter, Facebook, and YouTube. Roma and Aloini [9] used empirical analysis to further expand the comparative framework of UGC utilization dimensions. The main focus of these studies is the text of the UGC and the use of natural language processing (NLP) technology to extract labeled features.

The literature also includes a methodological exploration of extracting embeddings from the text content of UGC. Wei [18] established CAND, a new framework that went beyond conventional text-based analysis to detect disinformation from tweets by aggregating the extracted judgments using an unsupervised Bayesian aggregation model. Different from Wie's [19] method of extracting a singular type of text embedding, Zhang [5] utilized two distinct deep learning models, convolutional neural network (CNN) and LSTM, to forecast product adoption intentions through distinct text representations. Image representations, however, are overlooked, especially UGI in e-commerce. While UGC covers a broad range of content from text and audio to video and image, UGI focuses on the particular images or photos generated by users that carry unique information not captured by text or other media formats. Given the wealth of information in the images, it is crucial to understand how to use UGI to predict commercial information.

### 2.2 Deep Learning-Driven User-Generated Image

E-commerce studies frequently employ deep learning to extract information from UGI [6, 19, 20]. To forecast business survival, Zhang and Luo [19] utilized UGI to extract photographic attributes, photo contents, and image captions. Overgoor [20] expanded the visual complexity framework by extracting visual complexity attribute information from visual content. Unlike these studies, Liu [6] collected labeled images from the Flickr website and proposed a CNN to measure brand perceptions based on UGI. By leveraging the unique information in UGI and the efficiency of deep learning, these studies performed well in the e-commerce field, but they only labeled features when extracting information from images. The literature is limited, though, regarding how to explore the representative images and text to obtain image-text joint embeddings. Our research contributes to the existing knowledge by filling this void, using VSEM-SAMMI to retrieve image-text joint embeddings and machine learning algorithms to predict UGI helpfulness and product sales. The ablation study also helps evaluate whether image–text joint embeddings are more informative than labeled features.

# 3 Proposed Approach

## 3.1 Framework of VSEM-SAMMI

VSEM-SAMMI comprises three modules: visual embedding generation with pre-trained ResNet, semantic embedding generation with Word2Vec, and multimodal interaction with a self-attention mechanism. The framework of VSEM-SAMMI is demonstrated in Fig. 1. First, we extracted visual embeddings from images to obtain hidden information in images. Second, we used Word2Vec to map word sequences to semantic embeddings and match the embedding dimension of the image with the text. Third, we used a self-attention mechanism to integrate image and text embeddings to accomplish the multimodal interaction.

## 3.2 Visual Embedding Generation with Pre-trained ResNet

The semantic embedding generation involves the utilization of a pre-trained ResNet50 model, enhanced by the incorporation of a self-attention mechanism. The ResNet architecture, proposed by He et al. [21], is a sophisticated deep learning framework renowned for its residual blocks. Since its establishment, ResNet and residual blocks have garnered significant popularity in visual embedding generation [22–25]. The architecture employs skip connections and shortcut connections to effectively mitigate the vanishing gradient issue, commonly observed in neural networks with a significant depth. We removed the last fully connected layer in ResNet50 to maintain the model's capability for feature extraction. As a result, the image was presented as a 1-of-2048 dimensional embedding, $I^{(1\times2048)}$.

## 3.3 Semantic Embedding Generation with Word2Vec

We embed product descriptions using Word2Vec, following the deep learning literature [26–28]. Word2Vec is an embedding technique that can be implemented using either of its two primary models: Continuous Bag of Words (CBOW) or Skip-Gram [29]. We selected the Skip-Gram model to finish our task, as it is more efficient with less training data compared with CBOW. We implement text embeddings through the following three steps: first, we divided multi-word entities into separate words; second, we trained the Word2Vec model to vectorize each word; finally, we utilized the average pooling operation to extract the feature for the entire description.

Assuming a particular product is $K_{description}$, we can abstract the product description as: $K_{description} = \{D_1, D_2, D_3 \ldots, D_n\}$, where $D_i$ is a descriptive word or phrase for a particular aspect of the product. We

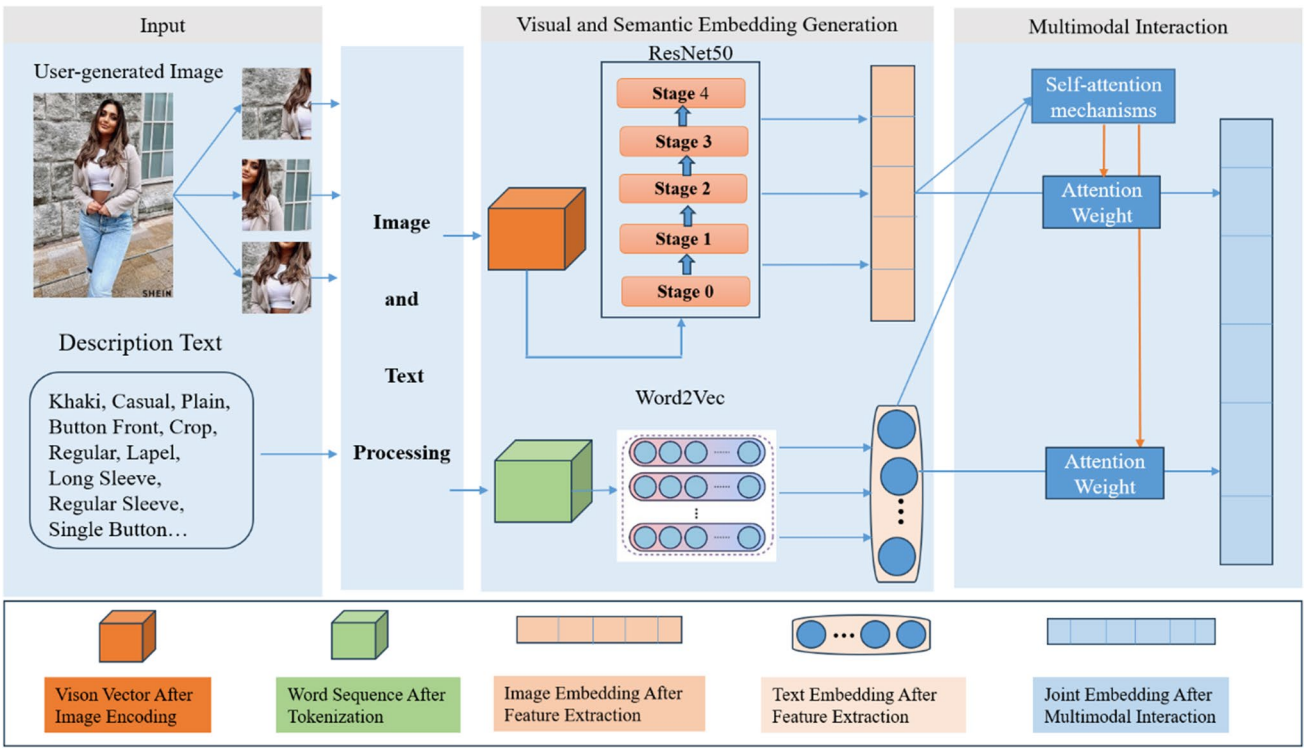


**Fig. 1** The overall architecture of VSEM-SAMMI

tokenize each product description since product descriptions are word-based and do not entail contextual relationships. For instance, the original description $D_i$ is "Medium Wash;" after tokenization, it was {"*Medium*", "*Wash*"}. We obtained the tokenization results by adding all tokenized words to a list. The Skip-Gram model encodes the words as 100-dimensional tensors. After encoding, an average pooling operation was implemented for the product description words to create a 1-of-100-dimensional tensor, appending the words to the description tensor list. Finally, we used the technique of average pooling to convert a list of description tensors into a descriptive vector with a dimensionality of 100 $\boldsymbol{T}^{(1\times100)}$.

### 3.4 Multimodal Interaction with Self-Attention Mechanism

We utilized a self-attention-based joint embedding approach to obtain image–text joint embeddings to fulfill multimodal interactions. Self-attention is a method of attention that establishes correlations between distinct positions within a singular sequence, producing a cohesive representation of said sequence [30]. By employing this approach, we derive weights for text and image embeddings, obtaining joint embeddings of images and text.

Our self-attention mechanism involves two layers: the fully connected layers and the activation layers. This configuration enables us to derive weights for image and text embeddings. We posit a tensor $\boldsymbol{I}^{1\times2048}$ expresses the image embedding and a tensor $T^{1\times100}$ respresents the text embeddings. Initially, the tensor of the image and text passes through a fully connected layer, aligning both to the same tensor dimension.

$$\boldsymbol{T}^{(1\times100)}_{\text{embedded}} = \boldsymbol{T}^{(1\times100)} \times \boldsymbol{W}^{(100\times100)}_{\text{text}}, \tag{1}$$

$$\boldsymbol{I}^{(1\times100)}_{\text{embedded}} = \boldsymbol{I}^{(1\times2048)} \times \boldsymbol{W}^{(2048\times100)}_{\text{image}}. \tag{2}$$

Next, after summing the image and text tensors, they pass through a tanh activation layer $\sigma$. Subsequently, they undergo another fully connected layer followed by a Softmax layer, denoted as $\delta$, resulting in the weight distribution for the two embeddings.

$$\boldsymbol{A}^{(1\times100)}_{\text{input}} = \sigma\left(\boldsymbol{T}^{(100)}_{\text{embedded}} + \boldsymbol{I}^{(1\times100)}_{\text{embedded}}\right), \tag{3}$$

$$\boldsymbol{A}^{(1\times1)}_{\text{weights}} = \boldsymbol{\delta}\left(\text{attention linear}\left(\boldsymbol{A}^{(1\times100)}_{\text{input}}\right)\right). \tag{4}$$

Finally, we employ a fully connected layer $\boldsymbol{W}^{(100\times2048)}_{\text{text\_proj}}$ to project the text tensor back to the dimension of the image tensor. Utilizing attention weights, we amalgamate the embeddings of the text and image with weighted fusion, resulting in the final embedding vector $\boldsymbol{F}^{(1\times2048)}$.

$$\boldsymbol{T}^{(1\times2048)}_{\text{projected}} = \boldsymbol{T}^{(1\times100)}_{\text{embedded}} \times \boldsymbol{W}^{(100\times2048)}_{\text{text\_proj}}, \tag{5}$$

$$\boldsymbol{F}^{(1\times2048)} = \boldsymbol{A}^{(1\times1)}_{\text{weights}} \odot \boldsymbol{T}^{(1\times2048}_{\text{projected}} + \left(1 - \boldsymbol{A}^{(1\times1)}_{\text{weights}}\right) \odot \boldsymbol{I}^{(1\times2048)}. \tag{6}$$

### 3.5 Caption Generation Model

To achieve the interpretability of VSEM-SAMMI, we present a caption generation model to generate description for image embeddings [31]. We extracted image features through pre-trained ResNet50. We used Word2Vec to contract text features with a fixed dimension. Specifically, our model takes the image $I^{3\times224\times224}$ and its description $S^{1\times37}$ encoded as a sequence of 1-of-100 embeddings, where 100 is the size of description words, and 37 is the longest image's description word numbers. In addition, we padded the <pad> tag pad descriptions with less than 37 words to reach the required length $S = (S_0, \ldots, S_N)$. For images, we employed a pre-trained ResNet50 model to extract feature vectors $x_{-1} = \text{ResNet}(I^{3\times224\times224})$.

We used a LSTM network, a recurrent neural system [32], to generate captions. For new input for LSTM $x_t$, the probability $p_{t+1}$ of a word was defined as

$$x_t = W_e S_t \in \{0, \ldots N-1\}, \tag{7}$$

$$p_{t+1} = \text{LSTM}(x_t),\ t \in \{0 \ldots N-1\}. \tag{8}$$

Additionally, the image embeddings $I^{3\times224\times224}$ were inputted a single time, at $t = -1$, to convey the image's contents to the LSTM.

## 4 Data Collection and Data Processing

### 4.1 Data Collection

SHEIN, a B2C e-commerce site specializing in fast fashion, has gained substantial international recognition. This company effectively incorporates the advantages of China's supply chain to cater to the high demand for quick fashion among young women in Europe and America. This platform can effectively represent the product preferences of individuals belonging to the younger demographic. We focused on UGI and implemented a crawler that collected 7974 images from 27 categories. As shown in Fig. 2, the style gallery section of the SHEIN platform features a buyer show, exemplifying a category of UGI, including


**Fig. 2** Examples of buyer show

types such as summertime, spring and summer, girls' outings, etc. In addition to the style gallery images, we obtained the number of likes images received, listed in Table 1. We also considered the product attributes: the mean star rating, product descriptions, review number, discounts, availability, and stock status (whether sold out). Especially given the variance in the product launch dates, the reviews gathered by our crawler span from January 1 to July 1, 2023.

Furthermore, we gathered portrait data for every photograph to label images to assess the predictive power of features achieved by VSEM-SAMMI and labeled features. We used the face attribute analysis and human body attribute analysis API in Baidu AI Cloud to extract the portrait data of the image [33]. The detailed data definitions are in Table 1.

## 4.2 Data Processing

After gathering data from three categories, we utilized diverse approaches to extract pertinent information. We can categorize our collected data into four distinct types: multimodal data (comprising descriptions and images), discrete, non-dimensional continuous, and dimensional continuous. These types are in Table 2. First, we implemented VSEM-SAMMI for multimodal data to obtain image–text joint embeddings. The ResNet50 model was pre-trained on a collection of one million images in an ImageNet dataset [34], and the Word2Vec model was pre-trained on a collection of 10 billion words in a Google News dataset [35].

**Table 1** Data descriptions for Study 1

| Category | Data | Description |
|---|---|---|
| UGI characteristics | Image | Buyer show images |
| | Image likes | Number of likes received by viewers |
| Product characteristics | Description | Description of each product |
| | Stars | Average score of the product |
| | Review number | The number of product reviews between January 1 and July 1, 2023 |
| | Price | Price of the product |
| | Discount | Indicator coded as 1 if discount applied, and 0 otherwise |
| | Sold out | Indicator coded as 1 if the product sells out and 0 otherwise |
| Portrait characteristics | Age | Age of the user in the image |
| | Expression | Expression of the user in the image (i.e., none, smile, laugh) |
| | Face shape | Face shape of the user in the image (i.e., heart, oval, round) |
| | Glass type | Glass type of the user in the image (i.e., common, none, sun) |
| | Emotion | Emotion of the user in the image (i.e., surprise, fear, sad) |
| | Lower color | Color of the lower clothing (i.e., unsure, red, black) |
| | Bag type | Type of bag (i.e., none, shoulder bag, backpack) |
| | Headwear type | Indicator coded as 1 if wearing a hat, and 0 otherwise |
| | Upper wear | Type of upper wear (i.e., t-shirt, jacket, shirt) |
| | Cellphone | Status of using a cellphone (i.e., viewing phone, not using, calling) |
| | Lower wear | Lower body clothing (i.e., shorts, skirt, trousers) |
| | Upper color | Color of the upper clothing (i.e., unsure, red, black) |
| | Lower cut | Indicator coded as 1 if there is a lower truncation and 0 otherwise |
| | Carrying item | Indicator coded as 1 if there is an item in hand and 0 otherwise |
| | Angle roll | Face angle of tilt in three-dimensional rotation |
| | Angle yaw | Face angle of rotation within a plane |
| | Angle pitch | Face angle of left–right rotation in three-dimensional space |

**Table 2** Types of data and separate processing

| Type | Data | Processing |
|---|---|---|
| Multimodal data | Image, description | VSEM-SAMMI |
| Discrete data | Upper wear, headwear type, expression, face shape, bag type, glass type, emotion, lower color, cellphone, lower wear, upper color | One-hot encoding |
| Dimensional continuous data | Age, angle roll, angle yaw, angle pitch, stars, price | Min–max normalization |
| Non-dimensional continuous data | Image likes, discount, sold out, lower cut, carrying item | |

The parameters of the two pretrained models were frozen. In order to maintain a consistent size for image input, we cropped the image to a $224 \times 224$ pixel square. Second, we used the one-hot encoding technique to convert discrete data such as glass kind, cell phone, and emotion into non-dimensional continuous data. One-hot encoding involves utilizing Euclidean space to expand discrete data [36]. Considering a dataset consisting of $K$ discrete observations, converting each observation into a binary value, represented by either 0 or 1, is possible. Therefore, we can represent the discrete data as a vector with $K$ dimensions. Finally, we employed normalization to transform dimensional continuous data into non-dimensional continuous data.

## 5 The Study Framework

To predict UGI helpfulness and product sales, we proposed VSEM-SAMMI to obtain image–text embedding based on datasets collected from the SHEIN platform and implement prediction tasks by machine learning algorithms. In Study 1, we make precise predictions about the helpfulness of UGI and evaluate the potential of using embeddings as features in UGI helpfulness prediction. Furthermore, we demonstrate that embeddings are more informative in predicting UGI helpfulness than product and portrait characteristics. In Study 2, we predict product sales to assess the reproducibility of our methodology in delivering accurate outcomes for product sales forecast across various e-commerce commercial information predictions. Finally, to showcase the efficacy of VSEM-SAMMI, we use the identical predictor in conjunction with different ways of UGI utilization, finding that our approach can obtain the highest accuracy for product sales predictions. Figure 3 illustrates the diagrammatic portrayal of the framework employed in each investigation and shows a detailed illustration of the conceptual structure underlying each study.

## 6 Study 1

### 6.1 UGI Helpfulness Prediction

Our dependent variable for Study 1 is UGI helpfulness, defined as the total number of likes received by reviews according to past literature [37, 38]. This variable is conducive to reducing consumers' confusion when making purchase decisions [38]. It is worth noting that the distribution of UGI helpfulness has a left-skewed pattern. As a result, we employed the logarithm of

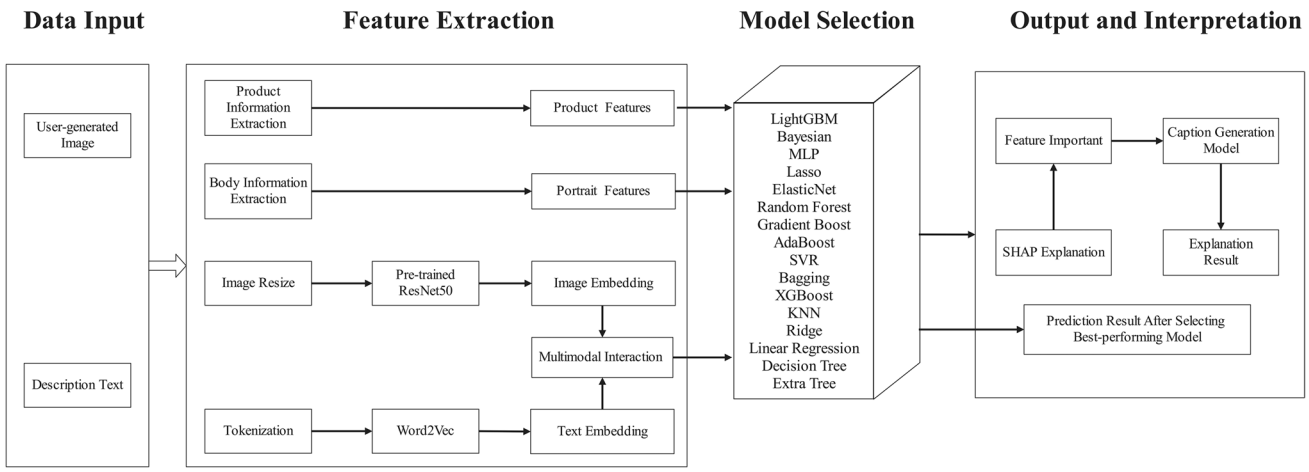


**Fig. 3** Flowchart of research framework

UGI helpfulness in our analysis. Online Appendix A elucidates the descriptive statistics of the dependent and independent variables. We used 16 distinct machine learning algorithms to forecast UGI helpfulness: linear regression (LR), k-nearest neighbors (KNN), support vector regression (SVR), lasso regression (Lasso), ridge regression (Ridge), elastic net regression (ElasticNet), Bayesian ridge (Bayesian), multilayer perceptron (MLP), decision tree (DT), extra tree (ET), extreme gradient boosting (XGBoost), random forest (RF), adaptive boosting (AdaBoost), gradient boost (GB), bagging, and light gradient boosting (LightGBM). Machine learning techniques have been pivotal in sales prediction endeavors [1, 39]. In choosing the most appropriate model for enhanced project performance prediction and data interpretation, it is crucial to assess multiple algorithms and determine which one has greater accuracy for UGI helpfulness prediction.

We deployed the MSE as our metric to compare the performance of different models:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2, \tag{9}$$

where $y_i$ is the actual output variable, $\widehat{y_i}$ represents the predicted output variable, and $n$ is the total number of samples. The MSE is a widely accepted approach in comparing the performance of machine learning algorithms; the smaller the MSE, the higher the model's accuracy [40, 41]. Since average processing time is a commonly used measure to compare computational complexity [42], we used it to evaluate the machine learning algorithms. We split the dataset into three sections: training, validation, and testing sets. The training set is used to train the model, the validating set is utilized to optimize the model parameters, and the testing set is employed to evaluate the performance of the models. Given that the optimal splitting configuration varies for each study, we analyzed 16 different splitting schemes for Study 1, as displayed in the first row of Table 3. The value in bold indicates the best performance for UGI helpfulness prediction. For instance, a 70%:15%:15% split indicates that we used 70% of the data for training, 15% for validation, and 15% for testing. Table 3 compares the machine learning techniques we utilized regarding their MSE and processing time. The results indicate that modifying the distribution of the training dataset by decreasing its share and allocating a larger proportion to validation and testing positively impacts the performance of the ExtraTree and DecisionTree models. However, this adjustment negatively affects the performance of LinearRegression and BayesianRidge models. Also, the results indicate that in minimizing MSE, the best method is LightGBM with a 60%:20%:20% split, demonstrating the best performing for our data. Therefore, considering model's accuracy and the computational complexity, we used LightGBM with a 60%:20%:20% split when conducting SHAP interpretation as our model.

### 6.2 Model Interpretability

To assess the significance of each independent variable's impact on UGI helpfulness, we conducted an analysis using an interpretable model that integrates LightGBM and SHAP. SHAP is a game-theoretic approach that we can use to explain the output of any machine learning model [43, 44]. Positioned as a post hoc evaluation tool, SHAP draws upon the foundational Shapley values from game theory. By individually altering each feature within the model's input and monitoring the resultant changes in the model's output, one can discern the incremental contribution of each feature [45, 46]. The corresponding mathematical expression for this computation is below:

$$\mathrm{SHAPvalue}_i = \sum S \subseteq N\backslash(i)\frac{|S|!(M-|S|-1)!}{M!}\left[f_x(S\cup\{i\}) - f_x(S)\right], \tag{10}$$

where $i$ is one input feature, $N$ denotes the complete collection of input features, $M$ is the total count of all input features, $S$ is a subset of $N$, of which the number of input features is $|S|$. Furthermore, $\mathrm{f}_x(S\cup\{i\})$ is the forecasted value using the collection of features $S\cup\{i\}$, while $\mathrm{f}_x(S)$ is the forecasted value using the collection of features $S$. In addition, it should be noted that the embeddings mentioned in Sects. 6.2 and 7.2 refer to image–text joint embeddings. For example, the 27th dimension of the image–text joint embeddings is the 27th embedding.

Table 4 shows that stars, 551st embeddings, 2027th embeddings, 1008th embeddings, 1373rd embeddings, and upper_wear_short-sleeved_shirt negatively impact UGI helpfulness (the total SHAP values of the variables are in Online Appendix C). In contrast, 473rd, 657th, 1962nd, 1477th, and 188th embeddings positively affect UGI helpfulness.

To elucidate the embeddings via captions, we also utilized the caption-generating model described in Sect. 3.5. The 473rd embeddings were used to clarify how our caption generation model described the embeddings in natural language. We extracted the 473rd embedding of all images, keeping the other embeddings at zero value. A trained autoencoder then decoded this 1–2048 dimensional embedding to vector $V^{3\times224\times224}$, which was finally input at a single time, $t=-1$, to convey the image contents to the LSTM and acquire a caption to describe the embedding. During caption collection, we excluded the <pad> tags, rearranged the word order for clarity, and selected the top ten most frequently occurring words for each embedding. Table 4 shows that the 473rd embedding connects with

**Table 3** MSE prediction performance and average processing time for Study 1

| Splitting | 80%:10%:10% | 78%:11%:11% | 76%:12%:12% | 74%:13%:13% | 72%:14%:14% | 70%:15%:15% | 68%:16%:16% | 66%:17%:17% | 64%:18%:18% |
|---|---|---|---|---|---|---|---|---|---|
| LightGBM | 0.214 | 0.212 | 0.216 | 0.214 | 0.212 | 0.211 | 0.216 | 0.212 | 0.209 |
| Bayesian | 0.213 | 0.213 | 0.216 | 0.213 | 0.211 | 0.210 | 0.215 | 0.212 | 0.209 |
| MLP | 0.216 | 0.214 | 0.217 | 0.215 | 0.213 | 0.211 | 0.216 | 0.214 | 0.210 |
| Lasso | 0.216 | 0.215 | 0.217 | 0.215 | 0.213 | 0.212 | 0.217 | 0.213 | 0.210 |
| ElasticNet | 0.216 | 0.215 | 0.217 | 0.215 | 0.213 | 0.212 | 0.217 | 0.213 | 0.210 |
| RF | 0.216 | 0.212 | 0.217 | 0.218 | 0.211 | 0.211 | 0.217 | 0.214 | 0.212 |
| GB | 0.216 | 0.214 | 0.218 | 0.216 | 0.213 | 0.214 | 0.217 | 0.216 | 0.217 |
| AdaBoost | 0.220 | 0.226 | 0.227 | 0.223 | 0.221 | 0.219 | 0.218 | 0.219 | 0.217 |
| SVR | 0.224 | 0.224 | 0.226 | 0.220 | 0.222 | 0.220 | 0.224 | 0.22 | 0.219 |
| Bagging | 0.241 | 0.235 | 0.240 | 0.229 | 0.230 | 0.235 | 0.234 | 0.238 | 0.231 |
| XGBoost | 0.257 | 0.259 | 0.253 | 0.247 | 0.240 | 0.239 | 0.242 | 0.253 | 0.248 |
| KNN | 0.256 | 0.258 | 0.258 | 0.253 | 0.252 | 0.261 | 0.245 | 0.238 | 0.256 |
| Ridge | 0.296 | 0.313 | 0.309 | 0.307 | 0.311 | 0.319 | 0.334 | 0.336 | 0.335 |
| LR | 0.302 | 0.321 | 0.321 | 0.312 | 0.318 | 0.338 | 0.340 | 0.344 | 0.347 |
| DT | 0.414 | 0.431 | 0.447 | 0.427 | 0.401 | 0.416 | 0.414 | 0.426 | 0.410 |
| ET | 0.442 | 0.468 | 0.392 | 0.443 | 0.408 | 0.431 | 0.422 | 0.422 | 0.438 |
| Splitting | 62%:19%:19% | 60%:20%:20% | 58%:21%:21% | 56%:22%:22% | 54%:23%:23% | 52%:24%:24% | 50%:25%:25% | Lowest MSE | Average processing time |
| LightGBM | 0.210 | **0.208** | 0.209 | 0.210 | 0.209 | 0.209 | 0.208 | 0.208 | 1.634 |
| Bayesian | 0.210 | 0.209 | 0.209 | 0.210 | 0.21 | 0.210 | 0.209 | 0.209 | 11.642 |
| MLP | 0.210 | 0.210 | 0.210 | 0.213 | 0.211 | 0.214 | 0.209 | 0.209 | 86.147 |
| Lasso | 0.211 | 0.210 | 0.211 | 0.212 | 0.211 | 0.211 | 0.210 | 0.210 | 0.296 |
| ElasticNet | 0.211 | 0.210 | 0.211 | 0.212 | 0.211 | 0.211 | 0.210 | 0.210 | 0.374 |
| RF | 0.212 | 0.210 | 0.213 | 0.213 | 0.211 | 0.211 | 0.210 | 0.210 | 1232.212 |
| GB | 0.212 | 0.215 | 0.212 | 0.214 | 0.212 | 0.213 | 0.213 | 0.212 | 400.784 |
| AdaBoost | 0.215 | 0.217 | 0.217 | 0.216 | 0.217 | 0.219 | 0.216 | 0.215 | 127.707 |
| SVR | 0.220 | 0.22 | 0.219 | 0.221 | 0.217 | 0.220 | 0.217 | 0.217 | 40.252 |
| Bagging | 0.233 | 0.229 | 0.233 | 0.235 | 0.235 | 0.233 | 0.235 | 0.229 | 138.562 |
| XGBoost | 0.238 | 0.241 | 0.247 | 0.250 | 0.246 | 0.244 | 0.245 | 0.238 | 85.938 |
| KNN | 0.236 | 0.268 | 0.266 | 0.251 | 0.242 | 0.245 | 0.245 | 0.236 | 0.227 |
| Ridge | 0.345 | 0.368 | 0.373 | 0.388 | 0.377 | 0.390 | 0.408 | 0.296 | 1.212 |
| LR | 0.356 | 0.380 | 0.396 | 0.405 | 0.391 | 0.446 | 0.431 | 0.302 | 4. 486 |
| DT | 0.407 | 0.423 | 0.405 | 0.455 | 0.440 | 0.397 | 0.432 | 0.397 | 24.444 |
| ET | 0.429 | 0.419 | 0.430 | 0.438 | 0.417 | 0.431 | 0.417 | 0.392 | 6.505 |

**Table 4** The Top 10 SHAP values of the determinants of UGI Helpfulness

| Variable | Description | Absolute SHAP value | Effects | Sign |
|---|---|---|---|---|
| Stars | Average score of the product | 0.015 | −0.465 | Negative |
| 473rd | Sculptural square polyurethane cap paisley, polyester preppy strap, leg push | 0.008 | 0.622 | Positive |
| 551st | 55% breasted buckle case bust wrap iron non-stretch red small butterfly | 0.002 | −0.462 | Negative |
| 2027th | Cinch collar, metal rose medium materials burgundy 5% elegant brown | 0.002 | −0.590 | Negative |
| 1008th | Clean, wash chunky track grey 21.3% 69% shirt breasted knot | 0.002 | −0.560 | Negative |
| 657th | Shawl appliques, metalized cut 94% fur, cool pleated polyamide out bra | 0.002 | 0.594 | Positive |
| 1373rd | Flare 47% out, camisole heart bodysuit puffer vacation blouse | 0.002 | −0.564 | Negative |
| Upper_wear_short-sleeved_shirt | Indicator coded as 1 if wearing a short-sleeved shirt and 0 otherwise | 0.001 | −0.592 | Negative |
| 1962nd | Low-top fringe pocket, leg 3-mesh, 76% butterfly, tight 43% | 0.001 | 0.398 | Positive |
| 188th | 34% 10% pc dry bishop preppy 0.5% soft stand maroon beaded | 0.001 | 0.564 | Positive |

The captions for embeddings had a similar structure to the product description, which was made up of words and phrases

the clothing's material composition. Online Appendix E includes the total captions for every embedding.

The above results indicate that image–text joint embeddings significantly impact the model, as evidenced by eight types of embeddings ranking among the top ten, and show that star ratings negatively affect UGI helpfulness and that the reviewer perceives the image as less helpful when the user wears a short-sleeved shirt.

In summary, we devised a deep learning-driven framework to improve the predictive precision of UGI helpfulness. The seamless integration of an attention mechanism has successfully established a combined embedding for image and text data. By employing these embeddings in conjunction with other variables to forecast the helpfulness of UGI, we have determined that LightGBM, utilizing a data split ratio of 60%:20%:20%, achieves an MSE of 0.208. Furthermore, we implemented a caption generation model that utilizes embedding techniques to enhance the interpretability of these encodings. Two pivotal inquiries arise, which we explore in Study 2: (1) Is the low MSE of VSEM-SAMMI replicable for different problems, and (2) Is there a way to confirm that embeddings are more informative in making predictions?

## 7 Study 2

### 7.1 Product Sales Prediction

In Study 2, the dependent variable is product sales, defined as the sales number within a given sales cycle. We measure this variable by the total number of reviews between January 1 and July 1, 2023 [47, 48]. We excluded products with zero sales, given the uncertainty of whether these products are past their sales cycle or within the processing phase [37]. Since the sales cycle and processing phase are not the focal points of our discussion, we restricted our analysis to products that recorded at least one sale between January 1 and July 1, 2023. In addition, the distribution of product sales is left-skewed, so we used the logarithm of product sales. It should be noted that we removed products with multiple buyer show images, as our study does not concentrate on the multifaceted interactions of UGI.

The dataset illustrates a notably imbalanced distribution in product sales, a situation that has the potential to significantly compromise the effectiveness of predictive models [49]. We addressed this concern using an over-sampling resampling regression technique developed by Torgo [49]. More specifically, we generated new data points within the sparse numerical product sales ranges by replicating existing ones to mitigate predictive bias. Consequently, we centered our research on a dataset of 1850 images. The descriptive statistics of these variables are in Online Appendix B.

We kept other configurations the same in Study 1 and used the machine learning algorithms mentioned in Study 1 to forecast product sales. Table 5 demonstrates different algorithms with their MSE and average processing time. Table 5 shows that our model consistently achieves a low MSE in forecasting product sales. The value in bold indicates the best performance for product sales prediction. Slightly different from Study 1, the MSE for each model initially decreases and then gradually increases while the percentage of validation and test datasets increases. The best-performing model remains LightGBM, and the data split is almost unchanged at 62%:19%:19%. Therefore, considering model's accuracy and the computational complexity, we used LightGBM with 62%:19%:19% data split for SHAP interpretation.

**Table 5** MSE prediction performance and average processing time for Study 2

| Splitting | 80%:10%:10% | 78%:11%:11% | 76%:12%:12% | 74%:13%:13% | 72%:14%:14% | 70%:15%:15% | 68%:16%:16% | 66%:17%:17% | 64%:18%:18% |
|---|---|---|---|---|---|---|---|---|---|
| LightGBM | 0.226 | 0.226 | 0.213 | 0.226 | 0.230 | 0.214 | 0.215 | 0.215 | 0.197 |
| Bayesian | 0.291 | 0.249 | 0.255 | 0.280 | 0.216 | 0.240 | 0.300 | 0.258 | 0.248 |
| MLP | 0.253 | 0.266 | 0.261 | 0.248 | 0.272 | 0.266 | 0.258 | 0.271 | 0.260 |
| Lasso | 0.276 | 0.283 | 0.351 | 0.256 | 0.311 | 0.337 | 0.286 | 0.289 | 0.303 |
| ElasticNet | 0.278 | 0.313 | 0.310 | 0.288 | 0.305 | 0.301 | 0.291 | 0.303 | 0.293 |
| RF | 0.298 | 0.306 | 0.326 | 0.294 | 0.327 | 0.303 | 0.306 | 0.284 | 0.287 |
| GB | 0.311 | 0.302 | 0.330 | 0.288 | 0.347 | 0.316 | 0.332 | 0.329 | 0.321 |
| AdaBoost | 0.596 | 0.413 | 0.385 | 0.312 | 0.395 | 0.290 | 0.533 | 0.442 | 0.543 |
| SVR | 0.317 | 0.304 | 0.335 | 0.291 | 0.353 | 0.321 | 0.340 | 0.335 | 0.328 |
| Bagging | 0.360 | 0.312 | 0.403 | 0.301 | 0.345 | 0.310 | 0.363 | 0.376 | 0.331 |
| XGBoost | 0.445 | 0.396 | 0.321 | 0.440 | 0.347 | 0.386 | 0.428 | 0.457 | 0.490 |
| KNN | 0.438 | 0.431 | 0.460 | 0.416 | 0.423 | 0.426 | 0.430 | 0.423 | 0.400 |
| Ridge | 0.453 | 0.520 | 0.755 | 0.428 | 0.456 | 0.445 | 0.467 | 0.594 | 0.399 |
| LR | 0.707 | 0.784 | 0.653 | 0.799 | 0.820 | 0.830 | 0.835 | 0.803 | 0.763 |
| DT | 0.790 | 0.810 | 0.808 | 0.816 | 0.810 | 0.797 | 0.784 | 0.783 | 0.772 |
| ET | 0.790 | 0.810 | 0.808 | 0.816 | 0.810 | 0.797 | 0.784 | 0.783 | 0.772 |

| Splitting | 62%:19%:19% | 60%:20%:20% | 58%:21%:21% | 56%:22%:22% | 54%:23%:23% | 52%:24%:24% | 50%:25%:25% | Lowest MSE | Average processing time |
|---|---|---|---|---|---|---|---|---|---|
| LightGBM | 0.188 | 0.194 | 0.195 | **0.184** | 0.203 | 0.218 | 0.205 | 0.184 | 1.402 |
| Bayesian | 0.238 | 0.275 | 0.248 | 0.259 | 0.288 | 0.349 | 0.378 | 0.216 | 2.188 |
| MLP | 0.262 | 0.270 | 0.275 | 0.255 | 0.283 | 0.301 | 0.343 | 0.248 | 13.426 |
| Lasso | 0.280 | 0.305 | 0.322 | 0.294 | 0.345 | 0.352 | 0.360 | 0.256 | 0.119 |
| ElasticNet | 0.301 | 0.309 | 0.319 | 0.295 | 0.336 | 0.355 | 0.373 | 0.278 | 0.112 |
| RF | 0.287 | 0.297 | 0.292 | 0.279 | 0.314 | 0.342 | 0.329 | 0.279 | 118.385 |
| GB | 0.342 | 0.362 | 0.345 | 0.329 | 0.370 | 0.400 | 0.458 | 0.288 | 60.169 |
| AdaBoost | 0.492 | 0.445 | 0.460 | 0.455 | 0.545 | 0.623 | 0.551 | 0.290 | 23.884 |
| SVR | 0.350 | 0.369 | 0.350 | 0.335 | 0.379 | 0.407 | 0.466 | 0.291 | 2.327 |
| Bagging | 0.377 | 0.393 | 0.400 | 0.339 | 0.393 | 0.463 | 0.503 | 0.301 | 13.005 |
| XGBoost | 0.543 | 0.451 | 0.502 | 0.494 | 0.601 | 0.536 | 0.604 | 0.321 | 10.937 |
| KNN | 0.399 | 0.436 | 0.416 | 0.363 | 0.373 | 0.442 | 0.437 | 0.363 | 0.053 |
| Ridge | 0.406 | 0.423 | 0.443 | 0.474 | 0.416 | 0.432 | 0.446 | 0.399 | 0.215 |
| LR | 0.857 | 0.815 | 0.861 | 0.791 | 0.804 | 0.876 | 0.834 | 0.653 | 1.578 |
| DT | 0.768 | 0.771 | 0.756 | 0.752 | 0.764 | 0.752 | 0.743 | 0.743 | 2.273 |
| ET | 0.768 | 0.771 | 0.756 | 0.752 | 0.764 | 0.752 | 0.743 | 0.743 | 0.746 |

**Table 6** The Top 10 SHAP values of the determinants of product sales

| Variable | Description | Absolute SHAP value | Effects | Sign |
|---|---|---|---|---|
| IS_SOLD_OUT | Indicator coded as 1 if the product sells out and 0 otherwise | 0.374 | −0.968 | Negative |
| Stars | Average score of the product | 0.046 | 0.692 | Positive |
| 51st embedding | Deep work wool-like clean, wash 46% bust buckle, modest, ripped | 0.020 | 0.650 | Positive |
| IS_DISCOUNT | Indicator coded as 1 if there is a set discount and 0 otherwise | 0.019 | −0.849 | Negative |
| 346th embedding | Hat bucket boys notched zipper, nylon floral cap 21% plants | 0.019 | −0.646 | Negative |
| 888th embedding | Rhinestone, embroidery 64% low-top mocha 1pc 62.8% Cotton glitter green | 0.019 | 0.569 | Positive |
| 1849th embedding | Wayfarer lace cross front zipper, ripped 96% 73% tie skinny | 0.017 | −0.560 | Negative |
| 60th embedding | Sporty yoga fly, polyurethane materials pinafore notched 1% molded support | 0.015 | −0.730 | Negative |
| 1862nd embedding | Waist top shorts materials chevron natural raglan 51% 54% figure | 0.015 | 0.609 | Positive |
| 744th embedding | 68% ultra-polyamide, polyester, rusty patched, sexy Sequin, 26% cut | 0.014 | 0.657 | Positive |

## 7.2 Model Interpretability

Similar to Study 1, SHAP is utilized to assess the significance of each independent variable and its impact on product sales while also employing a caption generating model to elucidate the embeddings via captions. Table 6 shows the SHAP value of the top ten product sales attributes (the total SHAP values of the variables are in Online Appendix C). First, seven embeddings are in the top ten and dominant, confirming the results of Study 1. In addition, the SHAP value of IS_SOLD_OUT is 0.37, indicating that product sales depend more on whether products were sold out than other factors. Furthermore, stars ($b = 0.046$) increase product sales, suggesting that sellers on e-commerce platforms should focus on word-of-mouth [50]. Specifically, IS_DISCOUNT ($b = -0.849$) leads to decreased product sales, as discounting techniques may decrease the perceived quality of these products, thereby reducing sales [51].

To evaluate the superiority of VSEM-SAMMI and compare the predictive power of characteristics and embeddings, we conducted an ablation study [20]. Specifically, we considered the following model specifications: (1) product (incorporating all product characteristics variables), (2) product + portrait (incorporating product and portrait characteristics), (3) VSEM-SAMMI, (4) Bag of Words (BoW) + VGG16, and (5) Bidirectional Encoder Representations from Transformers (BERT) + Vision Transformers (ViT). We used concatenation for multimodal interaction in the experiments of "BoW + VGG16" and "BERT + ViT" [52–54]. The five models depicted in Fig. 4 were evaluated using MSE as the performance metric. The figure demonstrates that our proposed VSEM-SAMMI outperforms "BoW + VGG16"

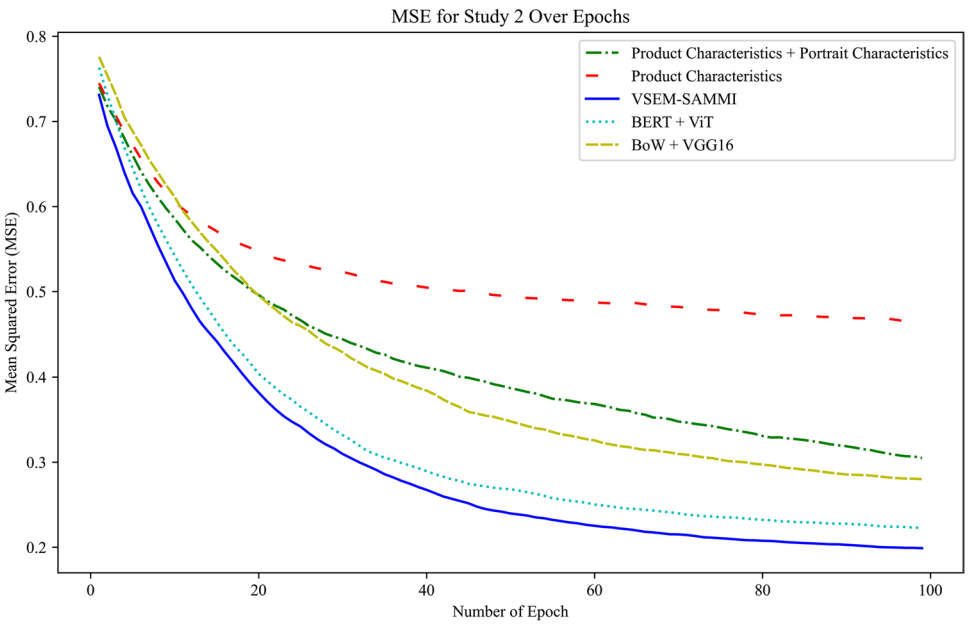


**Fig. 4** MSE performance for product sales prediction with different models

and "BERT + ViT" with concatenation. Embeddings are more informative than portrait characteristics, indicating that the primary source of predictive power stems from the embedding. In brief, our findings highlight the superiority of VSEM-SAMMI and the significant impact of embeddings when accurately predicting product sales.

Study 2's findings illustrate the considerable precision in forecasting product sales using VSEM-SAMMI. The predictor produces high accuracy and successfully generates embedding captions in different tasks. Moreover, the embeddings are the primary source of predictive power in predicting product sales.

## 8 Conclusion

This paper focuses on implementing deep learning in the context of UGI and proposes an explainable multimodal learning approach—VSEM-SAMMI—to predict UGI helpfulness and product sales. After obtaining multimodal data from the SHEIN e-commerce platform, we incorporated and merged it using VSEM-SAMMI, employing 16 machine-learning algorithms to predict UGI helpfulness and product sales. Additional features were analyzed using SHAP and a caption generation model, making valuable contributions to the general prediction of commercial information and specific forecasts of UGI helpfulness and product sales. This paper provides valuable contributions to predicting commercial information, especially UGI helpfulness and product sales prediction.

First, this paper introduces a new approach called VSEM-SAMMI, which representing images and text to get image-text joint embeddings. It leverages UGI to generate image–text joint embeddings, yielding MSE values of 0.208 for UGI helpfulness prediction and 0.184 for product sales prediction. Second, the ablation study in Sect. 7.2 demonstrates that using joint embeddings, instead of non-UGI data or feature-engineered UGI data, enhances the accuracy of predicting UGI helpfulness, product sales, and other commercial information. Third, to comprehend the meaning of image embeddings, our research presents a LSTM-based caption generation model to articulate the image embeddings. Overall, our model improves the accuracy of UGI helpfulness and product sales forecasting, which is crucial for e-commerce platform operation.

Our study has certain limitations and proposes avenues for future research. First, this paper focuses on one platform; future research might collect UGI from multiple platforms (e.g., Amazon, Zalando) to evaluate the robustness of our VSEM-SAMMI. Second, the absence of timestamps indicating when images receive helpful votes prevents us from quantifying each photograph's annual distribution of helpful votes [20]. Subsequent research could benefit from utilizing data that include timestamps for each vote, thereby enhancing the measurement of temporally varying information related to helpful votes. Third, because the product descriptions comprise words or phrases, we overlooked the contextual relationship when developing the caption generation model—this is something researchers should consider in the future.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s44196-024-00495-8.

**Acknowledgements** We gratefully acknowledge insightful suggestions from the editors and the anonymous reviewers, which substantively improved this article. We would also like to thank the members of Star-lights Research Team for their comments on earlier versions of the manuscript.

**Author Contributions** Conceptualization: CS and FL; data curation: CS and FL; formal analysis: CS and FL; funding acquisition: FL; investigation: CS and FL; methodology: CS and FL; project administration: FL; resources: FL; software: CS and FL; supervision: FL; validation: CS; visualization: CS; writing: CS and FL.

**Funding** This work was supported by the Humanities and Social Sciences Foundation of the Ministry of Education of China [Grant No. 21YJC630076].

**Data availability** The data that support the findings of this study are available from the corresponding author, upon reasonable request.

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.



## References

1. Bi, X., Adomavicius, G., Li, W., Qu, A.: Improving sales forecasting accuracy: a tensor factorization approach with demand awareness. INFORMS J. Comput. **34**(3), 1644–1660 (2022)
2. Chen, G., Huang, L., Xiao, S., Zhang, C., Zhao, H.: Attending to customer attention: a novel deep learning method for leveraging

multimodal online reviews to enhance sales prediction. Inf. Syst. Res. (2023). https://doi.org/10.1287/isre.2021.0292
3. Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., Mora, H.: A framework for big data analytics in commercial social networks: a case study on sentiment analysis and fake review detection for marketing decision-making. Ind. Mark. Manag. **90**, 523–537 (2019)
4. Nilashi, M., Abumalloh, R.A., Samad, S., Alrizq, M., Alyami, S., Alghamdi, A.: Analysis of customers' satisfaction with baby products: the moderating role of brand image. J. Retail. Consum. Serv. **73**, 103334 (2023)
5. Zhang, Z., Wei, X., Zheng, X., Li, Q., Zeng, D.D.: Detecting product adoption intentions via multiview deep learning. INFORMS J. Comput. **34**(1), 541–556 (2022)
6. Liu, L., Dzyabura, D., Mizik, N.: Visual listening in: extracting brand image portrayed on social media. Mark. Sci. **39**(4), 669–686 (2020)
7. Chen, J., Wu, Z., Yang, Z., Xie, H., Wang, F.L., Liu, W.: Multimodal fusion network with contrary latent topic memory for rumor detection. IEEE Multimedia **29**(1), 104–113 (2022)
8. Santos, M.L.B.D.: The "so-called" UGC: an updated definition of user-generated content in the age of social media. Online Inf. Rev. **46**(1), 95–113 (2022)
9. Roma, P., Aloini, D.: How does brand-related user-generated content differ across social media? Evidence reloaded. J. Bus. Res. **96**, 322–339 (2019)
10. Song, T., Huang, J., Tan, Y., Yu, Y.: Using user-and marketer-generated content for box office revenue prediction: differences between microblogging and third-party platforms. Inf. Syst. Res. **30**(1), 191–203 (2019)
11. Alturayeif, N., Luqman, H., Ahmed, M.: A systematic review of machine learning techniques for stance detection and its applications. Neural Comput. Appl. **35**(7), 5113–5144 (2023)
12. Bonet-Jover, A., Sepúlveda-Torres, R., Saquete, E., Martínez-Barco, P.: A semi-automatic annotation methodology that combines Summarization and Human-In-The-Loop to create disinformation detection resources. Knowl. Based Syst. **275**(5), 110723 (2023)
13. Papadopoulou, O., Zampoglou, M., Papadopoulos, S., Kompatsiaris, I.: A corpus of debunked and verified user-generated videos. Online Inf. Rev. **43**(1), 72–88 (2019)
14. Hartmann, J., Heitmann, M., Schamp, C., Netzer, O.: The power of brand selfies. J. Mark. Res. **58**(6), 1159–1177 (2021)
15. Zhang, M., Fan, B., Zhang, N., Wang, W., Fan, W.: Mining product innovation ideas from online reviews. Inf. Process. Manag. **58**(1), 102389 (2021)
16. Gupta, V., Varshney, D., Jhamtani, H., Kedia, D., Karwa, S.: Identifying purchase intent from social posts. Proc. Int. AAAI Conf. Web Soc. Media **8**(1), 180–186 (2014)
17. Smith, A.N., Fischer, E., Yongjian, C.: How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? J. Interact. Mark. **26**(2), 102–113 (2012)
18. Wei, X., Zhang, Z., Zhang, M., Chen, W., Zeng, D.D.: Combining crowd and machine intelligence to detect false news on social media. MIS Q. **46**(2), 977–1008 (2022)
19. Zhang, M., Luo, L.: Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp. Manag. Sci. **69**(1), 25–50 (2023)
20. Overgoor, G., Rand, W., van Dolen, W., Mazloom, M.: Simplicity is not key: understanding firm-generated social media images and consumer liking. Int. J. Res. Mark. **39**(3), 639–655 (2022)
21. He, K., Zhang, X., Ren, S., & Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Bian, P., Zheng, Z., & Zhang, D.: Light-weight multi-channel aggregation network for image super-resolution. In: Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III, vol. 4, pp. 287–297 (2021)
23. Zhang, D., Zheng, Z., Li, M., He, X., Wang, T., Chen, L., Lin, F.: Reinforced similarity learning: Siamese relation networks for robust object tracking. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 294–303(2020)
24. Ma, W., Zhou, T., Qin, J., Xiang, X., Tan, Y., Cai, Z.: Adaptive multi-feature fusion via cross-entropy normalization for effective image retrieval. Inf. Process. Manag. **60**(1), 103119 (2023)
25. Xiong, Q., Zhang, X., He, S., Shen, J.: Data augmentation for small sample iris image based on a modified sparrow search algorithm. Int. J. Comput. Intell. Syst. **15**(1), 110 (2022)
26. Bonner, M.F., Epstein, R.A.: Object representations in the human brain reflect the co-occurrence statistics of vision and language. Nat. Commun. **12**(1), 4081 (2021)
27. Feng, J., Cui, J., Wei, Q., Zhou, Z., Wang, Y.: A classification model of legal consulting questions based on multi-attention prototypical networks. Int. J. Comput. Intell. Syst. **14**(1), 204 (2021)
28. Wu, J., Liu, C., Wu, Y., Cao, M., Liu, Y.: A novel hotel selection decision support model based on the online reviews from opinion leaders by best worst method. Int. J. Comput. Intell. Syst. **15**(1), 19 (2022)
29. Mikolov T., Chen K., Corrado G., Dean J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings (2013)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V., Garnett, R. (eds.) Adv. Neural Inform. Processing Systems, pp. 5998–6008. Neural Information Processing Systems Foundation, Inc., La Jolla (2017)
31. Bai, S., An, S.: A survey on automatic image caption generation. Neurocomputing **311**, 291–304 (2018)
32. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
33. Wang, J., Zhu, S.: A novel stock index direction prediction based on dual classifier coupling and investor sentiment analysis. Cogn. Comput. **15**(3), 1023–1041 (2023)
34. Mihaltz, M.: Word2vec google news model. https://github.com/mmihaltz/word2vec-GoogleNews-vectors. Accessed 16 Sept 2022
35. Kaiming, H.: ResNet50. https://download.pytorch.org/models/resnet50-19c8e357.pth. Accessed 23 Nov 2022
36. Wu, Z., Jing, L., Wu, B., Jin, L.: A PCA-AdaBoost model for E-commerce customer churn prediction. Ann. Oper. Res. (2022). https://doi.org/10.1007/s10479-022-04526-5
37. Yang, Y., Wang, Y., Zhao, J.: Effect of user-generated image on review helpfulness: perspectives from object detection. Electron. Commer. Res. Appl. **57**, 101232 (2023)
38. Zhuang, W., Zeng, Q., Zhang, Y., Liu, C., Fan, W.: What makes user-generated content more helpful on social media platforms? Insights from creator interactivity perspective. Inf. Process. Manag. **60**(2), 103201 (2023)
39. Ferreira, K.J., Lee, B.H.A., Simchi-Levi, D.: Analytics for an online retailer: demand forecasting and price optimization. Manuf. Serv. Oper. Manag. **18**, 69–88 (2015)

40. Zhang, S., Luo, J., Wang, S., Liu, F.: Oil price forecasting: a hybrid GRU neural network based on decomposition–reconstruction methods. Expert Syst. Appl. **218**, 119617 (2023)
41. Liu, C., Li, Y., Fang, M., Liu, F.: Using machine learning to explore the determinants of service satisfaction with online healthcare platforms during the COVID-19 pandemic. Serv. Bus. **17**, 449–476 (2023)
42. Erkan, U.: A precise and stable machine learning algorithm: eigenvalue classification (EigenClass). Neural Comput. Appl. **33**(10), 5381–5392 (2021)
43. Liu, F., Wang, R., Fang, M.: Mapping green innovation with machine learning: evidence from China. Technol. Forecast. Soc. Change **200**, 123107 (2024)
44. Liu, F., Huang, W., Zhang, J., Fang, M.: Corporate social responsibility in family business: using machine learning to uncover who is doing good. Technol. Soc. **76**, 102453 (2024)
45. Wang, M., Yu, Y., Liu, F.: Does digital transformation Curb the formation of Zombie firms? A machine learning approach. Technol. Anal. Strateg. Manag. (2023). https://doi.org/10.1080/09537325.2023.2296007
46. Zhang, J., Zhu, M., Liu, F.: Find who is doing social good: using machine learning to predict corporate social responsibility performance. Oper. Manag. Res. **2023**, 1–14 (2023)
47. Ye, Q., Law, R., Gu, B.: The impact of online user reviews on hotel room sales. Int. J. Hosp. Manag. **28**(1), 180–182 (2009)
48. Ye, Q., Law, R., Gu, B., Chen, W.: The influence of user-generated content on traveler behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Comput. Hum. Behav. **27**(2), 634–639 (2011)
49. Torgo, L., Branco, P., Ribeiro, R.P., Pfahringer, B.: Resampling strategies for regression. Expert. Syst. **32**(3), 465–476 (2015)
50. Feng, J., Li, X., Zhang, X.: Online product reviews-triggered dynamic pricing: theory and evidence. Inf. Syst. Res. **30**(4), 1107–1123 (2019)
51. DelVecchio, D., Puligadda, S.: The effects of lower prices on perceptions of brand quality: a choice task perspective. J. Prod. Brand. Manag. **21**, 465–474 (2012)
52. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings (2015)
53. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019—2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proceedings of the Conference, vol. 1, pp. 4171–4186 (2019)
54. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations, pp. 1–21 (2020)