Information Processing and Management 59 (2022) 102984

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Popularity prediction for marketer-generated content: A text-guided attention neural network for multi-modal feature fusion

Yang Qian [a,b], Wang Xu [a], Xiao Liu [c], Haifeng Ling [a,b], Yuanchun Jiang [a,b,*], Yidong Chai [a,b], Yezheng Liu [a,d]

[a] School of Management, Hefei University of Technology, Hefei, Anhui 230009, China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei, Anhui 230009, China
[c] School of Information Technology, Deakin University, Melbourne 00113B, Australia
[d] National Engineering Laboratory for Big Data Distribution and Exchange Technologies, Shanghai 200436, China

ARTICLE INFO

Keywords:
Popularity prediction
Marketer-generated content
Text-guided attention neural network
Multi-modal data

ABSTRACT

In this paper, we focus on the popularity prediction for marketer-generated content (MGC), which has not been investigated by current studies. To address this problem, we propose a novel deep learning approach, namely the text-guided attention neural network (TGANN) model, to make full use of heterogeneous and multi-modal data related to MGCs (e.g., text descriptions and images). In the proposed model, we first design a filter-based topic model to filter out the noise words and extract topic features from textual descriptions. To lessen the influence of irrelevant information in images, we then propose a text-guided attention mechanism to use the text's topic features to guide the image region representations. Lastly, to determine the contribution of each topic and each image, the TGANN model introduces the attention computation for each visual modality and textual modality. Our experiments are conducted on two real-world datasets of MGC. The quantitative results show that the proposed model outperforms several state-of-the-art methods. The qualitative experiments demonstrate that our model can accurately capture topic attentions, image attentions, and image region attentions. The proposed model provides important practical implications for marketers and online platforms, e.g., estimating the success of online advertising campaigns, creating more attractive marketing contents, and improving recommendation systems.

1. Introduction

In the era of Web 2.0, online social media (e.g., Twitter and Facebook) has deeply impacted our daily life. Large amounts of textual or visual contents are generated and spread on social media platforms every day. An interesting observation is that some contents received millions of views, likes, shares, and comments, whereas others attracted much less attention and disseminated narrowly (Li & Xie, 2020; Shin et al., 2020). The possible reasons for this observation include the quality of contents (e..g, labels, texts, and images) and information of the publisher (e..g, user reputation). This observation motivates us to analyze the popularity of the content on social

* Corresponding author at: School of Management, Hefei University of Technology, Hefei, Anhui 230009, China.
E-mail address: ycjiang@hfut.edu.cn (Y. Jiang).

https://doi.org/10.1016/j.ipm.2022.102984
Received 30 October 2021; Received in revised form 13 April 2022; Accepted 16 May 2022
Available online 23 May 2022
0306-4573/© 2022 Elsevier Ltd. All rights reserved.

media. The task of popularity prediction attempts to estimate the number of interactions (e.g., number of views or number of likes) between users and given contents, or to identify the potential hot contents in advance from a large amount of information. Currently, accurate popularity prediction has gained increasing attention from both academia and industry due to its wide applications such as decision making for advertisers, content recommendation, and public opinion monitoring.

Recently, many studies have focused on popularity prediction for social media data, with most of them targeted on a conventional form of user-generated contents (UGCs). The approaches for the UGCs' popularity prediction can be divided into two categories: point process methods and feature-based methods (Mishra et al., 2016; Wu et al., 2018). Point process methods are used to exploit popularity evolution patterns, which regard the popularity cumulation of UGCs as a micro arrival point process of view events (Gao et al., 2019). The core of point process methods is to determine the intensity function of the point process. For example, Kong, Rizoiu & Xie (2020) introduced a connection between the recovery time distributions in Susceptible-Infected-Recovered (SIR) and the kernel functions in Hawkes Intensity Process (HIP), to improve the prediction of popularity. Although these point process methods can nicely describe the formation process of popularity, most of them simply consider time-series information while ignoring a series of key features (e.g., features extracted from text and image) for the popularity prediction. Moreover, in practice, it is hard to discern the real hypothesis or form of the intensity function, which limits the use of the point process model. Feature-based methods aim to extract various types of features from UGCs and use classical machine-learning models for prediction, e.g., Logistic Regression (LR) (Cheng et al., 2014), Support Vector Machine (SVM) (Trzciński & Rokita, 2017) and boosting algorithm (Kang et al., 2019). The main shortcoming of these classical methods is that they typically rely on laborious feature engineering to obtain useful features.

Compared with the popularity prediction for traditional UGCs, this paper focuses on the popularity prediction for marketer-generated content, which has not been investigated by current studies. Marketer-generated content (MGC) refers to the product- or brand-related messages posted by occupational marketers on social media platforms, and MGC plays an important role in driving customer engagement (Zhao et al., 2022). Due to the nature of MGCs, predicting the popularity of MGCs faces at least three main challenges as follows.

The first challenge lies in the heterogeneous and multi-modal data. Through our observations, we find that each MGC often consists of various types of data such as textual descriptions, images, and tags. These types of data provide complementary cues for enhancing prediction performance. For example, vivid picture expressions can attract more users to view or click, and appropriate labels may improve the effectiveness of information retrieval for users. However, the heterogeneous and multi-modal data in MGCs make it more difficult to construct models that can represent and relate the predictive signals from multiple modalities. Additionally, the information from different modalities may have varying predictive power, e.g., images, and textual descriptions in MGCs. Thus, how to assess the impact of different modal data on popularity also needs to be addressed.

The second challenge is the noise information. Compared with UGCs, according to the statistics, textual descriptions for MGCs are more lengthy and each MGC may contain multiple images. Although textual descriptions and images in MGCs are more normative than those in UGCs, they also suffer from noise information, since not all texts and images are meaningful and valuable for the prediction of their popularity. Thus, filtering out the irrelevant or unimportant parts from textual descriptions and images has become an important task for the popularity prediction of MGCs.

The third challenge is the fusion mechanism of multi-modal features. In previous studies on multi-modal data (Ding et al., 2019; Hsu et al., 2019), they simply combine the image feature with the text feature vector, since each record of the data only contains one picture and one text description. However, in MGCs, the longer text descriptions may convey richer information than the images. In addition, each image, each region of an image, and each feature of a text description can have different impacts on the popularity prediction for MGC. Therefore, it is necessary to design a new fusion mechanism for multi-modal features.

To address these challenges, we propose a novel deep learning approach, namely text-guided attention neural network (TGANN), which makes full use of heterogeneous and multi-modal data related to MGCs, including textual descriptions, images, labels, title, author and time information. More specifically, the TGANN model contains four parts: feature extraction, text-guided attention mechanism, feature fusion, and popularity prediction. For the feature extraction, we propose a filter-based topic model, an extension of latent Dirichlet allocation (LDA) (Blei et al., 2003), to filter out the noise words and then effectively learn the representation of textual descriptions. Given the success of deep learning in visual computing, we adopt the pre-trained VGGNet model (Simonyan & Zisserman, 2015) to extract visual representation for each image in MGCs. We capture the features from labels and title information using the embedding method and convert the time information to features using the periodicity property. In addition, we use the number of followers and followings to represent the author's influence. To diminish the impact of irrelevant and unimportant information in images on popularity prediction, we incorporate an attention mechanism for obtaining new textual-based visual features. Also, we provide attention to the hierarchical representation of textual features and visual features, which determines the relative importance of each information for the popularity prediction. To capture the interaction between textual and visual modalities, we use a Multimodal Compact Bilinear (MCB) model to combine visual with textual features, which has been proved to achieve satisfactory performance in many applications (Fukui et al., 2016). Other features (e.g., labels and title) which can provide additional information, are also incorporated into the model for improving the accuracy of popularity prediction. Finally, we formulate the task of popularity prediction as a classification problem and predict the popularity levels for all of the MGCs.

To validate the effectiveness of our proposed model, we conduct comprehensive experiments on two real-world datasets collected from Taobao[1] and Autohome,[2] respectively. Experimental results demonstrate that our proposed model achieves significantly better

[1] https://www.taobao.com/.

[2] https://www.autohome.com.cn/.

performance than several state-of-the-art methods. To analyze the influences of different types of features, we also perform several ablation experiments, which offer useful insights into the fusion of heterogeneous and multi-modal features for MGCs' popularity prediction.

In summary, the contribution of this paper is three-fold:

(1) To the best of our knowledge, this paper is the first attempt to address the popularity prediction problem for marketer-generated content. To effectively predict the popularity of MGCs, we make full use of the features from heterogeneous and multi-modal data related to MGCs, such as textual descriptions and images. This paper can help marketers estimate the success of online advertising campaigns and thus support marketers in making decisions.
(2) We propose a novel text-guided attention neural network that incorporates multiple types of features and explores their benefits for the popularity prediction of MGCs. In the model, we use a text-guided attention mechanism to lessen the influence of irrelevant information in images on the prediction. Moreover, advanced unsupervised methods (e.g., topic model and VGGNet) are proposed or used to extract features from visual and textual contents, so as to avoid heavy work on feature engineering.
(3) We evaluate our proposed model with two MGC datasets that are collected from Taobao and Autohome respectively. Experimental results show that our method can achieve significantly better predictive performance than all baselines including state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature and identifies the research gaps. Section 3 introduces the details of our proposed model. Section 4 presents the experimental results that evaluate our model and compare with baselines. Finally, Section 5 concludes this paper and points out the future work.

2. Related work

In this section, we provide an overview of three different streams of literature devoted to popularity prediction on social media, deep learning models, and the attention mechanism.

2.1. Popularity prediction on social media

Recently, a series of efforts have focused on popularity prediction based on social media data. The core idea is to build regression models or classification models for popularity prediction by exploiting various features or time-series patterns (Gao et al., 2019). There are two types of features mainly used to make predictions. One is the textual features and the other is visual features. Zhang et al. (2021) proposed a Bi-layered long short-term memory (LSTM) network that incorporated user reputation and historical tweets to learn latent features for tweet popularity prediction. Similarly, Yuan et al. (2020) proposed a deep information echoing model based on a bi-directional LSTM, which incorporated the market-related tweets echo with financial news to predict the market trend. These studies mainly consider the text information without exploring the image information, which restrains their performance on popularity prediction.

With the development of visual processing techniques, several studies have focused on using visual features to predict the popularity of content on social media. For example, Lin et al. (2019) introduced a layer-wise deep stacking (LDS) model for image popularity prediction. The visual features and social cue features (e.g., title length and tag count) are combined into their methods for predicting the count of views. Zhang et al. (2018) suggested that the image itself and its corresponding textual description are complementary with each other. Based on visual representation and textual representation, they proposed a co-attention network to predict social image popularity. Although these methods incorporate visual features and textual features, they mainly consider the impact of a single image corresponding to the textual content (e.g., post). In addition, relaxed for the irrelevant and unimportant parts of the image or text, these methods would draw into noise information, resulting in the deviation for the prediction.

In this paper, we explore the popularity prediction problem for MGCs. A MGC in social media includes various types of data such as the title, tags, long descriptions, multiple images, user information, and time information. And we focus on filtering out irrelevant and unimportant information from unstructured textual and visual modalities, and then integrating them with other additional features into a unified model to predict popularity for MGCs.

2.2. Deep learning models

The proposed model for popularity prediction is based on a neural network. Neural network models have achieved significant performance in tackling bottleneck problems, especially in applications of computer vision, natural language processing, and social media analysis (Behera et al., 2021; Briskilal & Subalalitha, 2022; Sawhney et al., 2021). Neural network models are efficient in extracting feature representation from different types of data (e.g., text, images, or audio), which avoids constructing features with human labeling. For instance, Ayoub et al. (2021) applied Bidirectional Encoder Representations from Transformers (BERT) and Shapley Additive explanation (SHAP) methods to construct interpretable features for detecting misinformation about COVID-19. Sawhney et al. (2021) proposed a hierarchical time-aware hyperbolic LSTM (HTLSTM) for modeling temporal sequences of online information.

In addition, deep neural networks are well-matched for solving multimodal tasks (Sawhney et al., 2020). Shraga et al. (2020) proposed a multimodal deep learning architecture for the web table retrieval task, and the experiments demonstrated that the fusion of

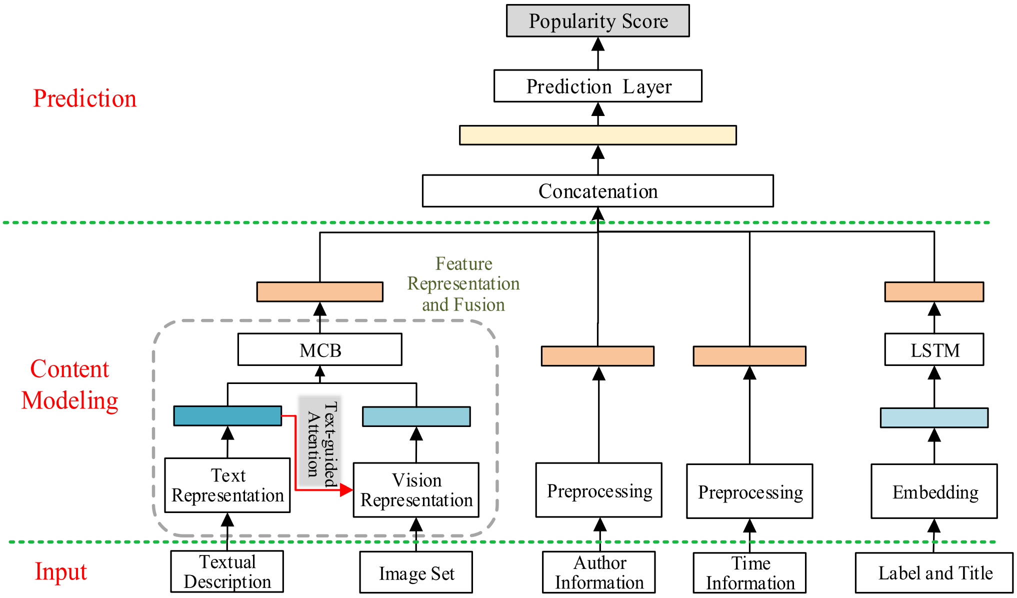


Fig. 1. The overall architecture of the TGANN model.

multimodal features in the model performs better than that using only one modality feature. In recommender systems, Ma et al. (2019) introduced a co-attention mechanism to model the multimodal microblogs and to perform the hashtag recommendation. In addition, multimodal deep learning models have been studied in popularity prediction (Hsu et al., 2020).

Inspired by their success, in this work, we propose a text-guided attention neural network to fuse multimodal data for popularity prediction. In our scenario, we consider several unique facets of MGCs for popularity prediction, e.g., textual descriptions with a longer length and multiple images in each MGC.

2.3. Attention mechanism

Our model is taken inspiration from the attention mechanism that has been widely used in neural network models. The attention mechanism is inspired by psychology, which suggests that human brains often focus on a subset of total information according to their demands (Wilterson & Graziano, 2021). Based on this idea, introducing the attention mechanism in neural network models offers the ability to avoid noise in the input, and to select the most pertinent parts of information for improving the model performance. For example, Peng et al. (2017) developed a novel framework, namely the object-part attention model (OPAM), for the fine-grained image classification. Instead of processing an entire image, OPAM can select the relevant regions of the image and ignore the irrelevant regions. Wang et al. (2020) proposed a hierarchical attention network (HAN) for code summarization. The HAN model can determine the importance of tokens and statements in a code snippet. Although these attention-based models are efficient, they mainly focus on modeling visual attention or textual attention alone.

Recently, a few efforts have explored the multi-modal attention models. For sentiment analysis, Yang et al. (2019) proposed a co-attention LSTM network that adopts a multiple-hops co-attention mechanism to learn nonlinear representations of context and target simultaneously. In visual question answering (VQA), Yu et al. (2019) presented a deep modular co-attention network (MACN) that jointly performed the self-attention of questions and images, as well as the question-guided-attention of images. To discriminative visual regions related to the sounds, Cheng et al. (2020) explored three different co-attention modules to learn the cross-modal representations of audio and visual events.

Despite producing remarkable performance, the models based on the co-attention mechanism are difficult to meet popularity prediction for MGCs. The reason is that the co-attention mechanism combines textual and visual information by capturing the common and relevant features between these two modalities, ignoring the contribution of the unique and key information in these modalities to prediction. In this work, we argue that the informative long text in MCGs plays a more important role than images. Thus, we propose a text-guided attention model for popularity prediction of MGC, which can learn visual attention from textual features to enhance the predictive performance. Based on the proposed model, we can easily filter out noisy information in images that make it more accurate to predict the popularity of MGC.

3. Model

In this section, we present a text-guided attention neural network (TGANN) model for the popularity prediction of MGC. Fig. 1 shows the overall architecture of this model. Different from conventional models, TGANN integrates the textual features learned by the

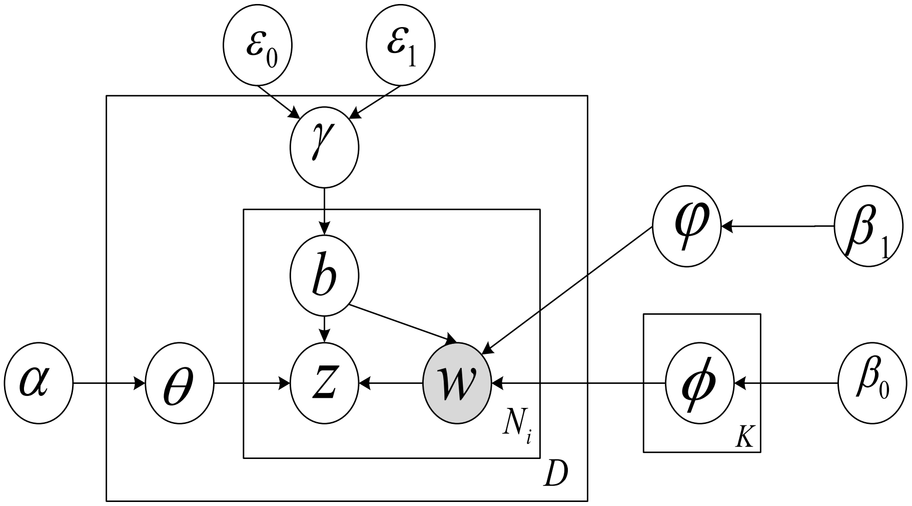
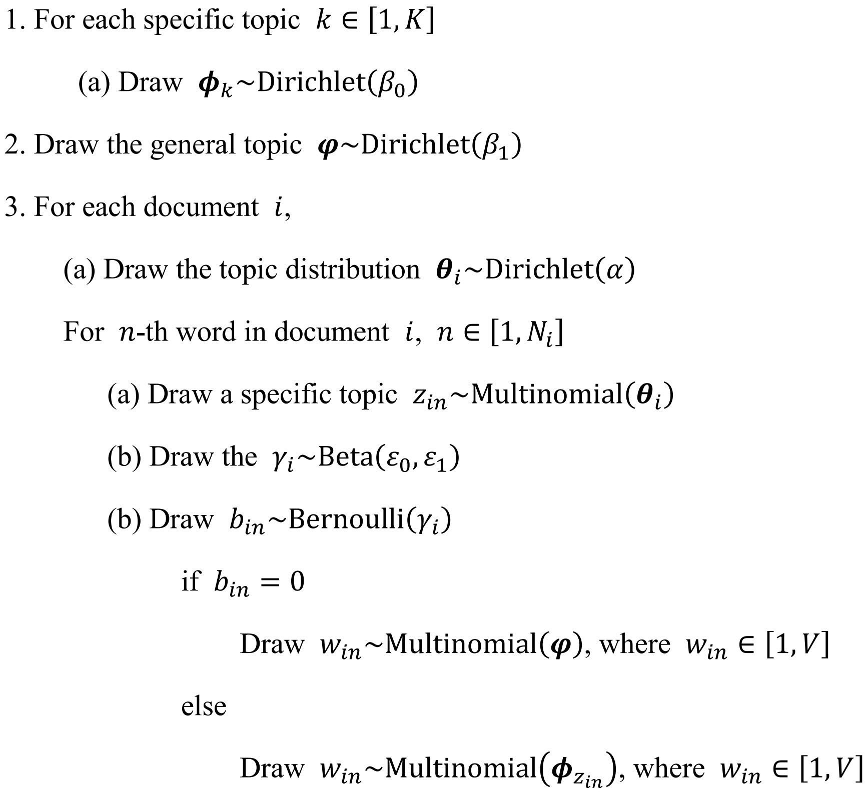


Fig. 2. Graphical structure of the FBT model.



```
1. For each specific topic $k \in [1, K]$
      (a) Draw $\boldsymbol{\phi}_k$~Dirichlet($\beta_0$)
2. Draw the general topic $\boldsymbol{\varphi}$~Dirichlet($\beta_1$)
3. For each document $i$,
      (a) Draw the topic distribution $\boldsymbol{\theta}_i$~Dirichlet($\alpha$)
      For $n$-th word in document $i$, $n \in [1, N_i]$
         (a) Draw a specific topic $z_{in}$~Multinomial($\boldsymbol{\theta}_i$)
         (b) Draw the $\gamma_i$~Beta($\varepsilon_0, \varepsilon_1$)
         (b) Draw $b_{in}$~Bernoulli($\gamma_i$)
               if $b_{in} = 0$
                     Draw $w_{in}$~Multinomial($\boldsymbol{\varphi}$), where $w_{in} \in [1, V]$
               else
                     Draw $w_{in}$~Multinomial($\boldsymbol{\phi}_{z_{in}}$), where $w_{in} \in [1, V]$
```

Fig. 3. The generative process of the FBT model.

topic model, image features guided by textual features, and other auxiliary features (e.g., label and title) to enhance the prediction performance. To ease understanding, we describe the proposed model in five ways: problem definition, feature extraction, text-guided attention mechanism, multimodal compact bilinear for feature fusion, and prediction and training.

3.1. Problem definition

We start with some notations. Suppose that there is a collection of D MGCs. Indexed by i, a MGC is represented by a textual description $\boldsymbol{d}_i = \{w_{i,1}, w_{i,2}, \cdots, w_{i,n}, \cdots, w_{i,N_i}\}$, with the corresponding image set $\boldsymbol{I}_i = \{p_{i,1}, p_{i,2}, \cdots, p_{i,m}, \cdots, p_{i,M_i}\}$, its label set $\boldsymbol{l}_i$, title information t_i, author information a_i and time information T_i. Let $w_{i,n}$ denote the n-th word in $\boldsymbol{d}_i$ and $p_{i,m}$ the m-th image in I_i. Let N_i be the number of words in $\boldsymbol{d}_i$ and M_i the number of images in $\boldsymbol{I}_i$. Using feature extraction methods, we obtain the feature representations of different modalities for MGC i, denoted by $\mathbf{F} = \{\mathbf{f}_i^{text}, \mathbf{f}_i^{image}, \mathbf{f}_i^{label}, \mathbf{f}_i^{title}, \mathbf{f}_i^{author}, \mathbf{f}_i^{time}\}$. Following prior works (Liao et al., 2019; Xiong et al., 2021), we regard the popularity prediction task as a classification. We discretize the number of total views and define $y_i \in \{0, 1\}$ to represent popularity score of MGC $\boldsymbol{i}$.

Based on the above notations, we define our problem: Given the MGC $\boldsymbol{i}$, our task is to learn a function $funciton: \{\mathbf{f}_i^{text}, \mathbf{f}_i^{image}, \mathbf{f}_i^{label}, \mathbf{f}_i^{title}, \mathbf{f}_i^{author}, \mathbf{f}_i^{time}\} \to y_i$ to predict its popularity score.

3.2. Feature extraction

3.2.1. Filter-based topic model for the textual description

Although textual descriptions in MGCs contain meaningful and valuable information, they suffer from noise problems. For example, a MGC is often mixed with both document-specific words and general words, where document-specific words can be regarded as discriminative information. In contrast, general words are frequently occurring in all MGCs, which can be regarded as noise words that are meaningless for popularity prediction. To solve the noise problem in textual descriptions of MGCs, we propose a filter-based topic (FBT) model to filter out the trivial information and to maintain the valuable information for popularity prediction. The FBT model extends the LDA algorithm (Blei et al., 2003) by introducing a background topic. Fig. 2 illustrates the plate notation of the proposed FBT model.

Furthermore, we describe the generative process of the FBT model in Fig. 3. Regarding the generation of the document (textual description), each document is viewed as a mixture of K specific topics and a general topic, similar to LDA. Specifically, we first sample each specific topic $k \in [1, K]$ using symmetric Dirichlet prior with fixed parameter β_0, and sample the general topic using symmetric Dirichlet prior with fixed parameter β_1. Then, to generate the $\boldsymbol{n}$-th word in the document, we sample the per-document topic distribution $\boldsymbol{\theta}_i$ from a Dirichlet prior with parameter α. The unique preference of document $\boldsymbol{i}$ based on specific topics or the general topic expressing the opinion is described by the Bernoulli parameter γ_i, which is sampled from a Beta distribution with parameters ε_0 and ε_1. We use γ_i to determine the value of the binary variable b_{in}, with $b_{in} = 1$ denoting that the word w_{in} is sampled from a specific topic, whereas $b_{in} = 0$ meaning that w_{in} is sampled from the general topic.

In this paper, we use collapsed Gibbs sampling to estimate parameters $\boldsymbol{\theta}_i$, ϕ_k and $\boldsymbol{\varphi}$. Due to the space limation, we omit the detailed derivation and just give the equations that we use in the estimation process.

$$\theta_{ik} = \frac{c_i^k + \alpha}{\sum_{k'=1}^{K}(c_i^{k'} + \alpha)} \tag{1}$$

$$\phi_{kv} = \frac{c_{k,b=1}^{v} + \beta_0}{\sum_{v'=1}^{V}\left(c_{k,b=1}^{v'} + \beta_0\right)} \tag{2}$$

$$\varphi_v = \frac{c_{b=0}^{v} + \beta_1}{\sum_{v'=1}^{V}(c_{b=0}^{v'} + \beta_1)} \tag{3}$$

where c_i^k denotes the number of words assigned to the specific topic k. $c_{k,b=1}^{v}$ denotes the number of word v assigned to the specific topic k when the binary variable b is equal to 1. $c_{b=0}^{v}$ denotes the number of word v assigned to the general topic when the binary variable b is equal to 0. Based on the inference, we can filter out the trivial information and maintain the valuable information for popularity prediction. Using the estimated topic-word distribution ϕ_k, we embed the topic k with a vector using the weights of top probable words. Combined with the document-topic distribution $\boldsymbol{\theta}_i$, we can denote the topic feature for textual description $\boldsymbol{i}$ as a feature matrix, i.e., $\mathbf{f}_i^{text} = [\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^K]$.

3.2.2. Image feature representation

In our context, each MGC contains multiple images. For each image, we apply a pre-trained VGGNet-16 (Simonyan & Zisserman, 2015) to obtain its feature representation. Specifically, we rescale all images in MGCs to 224 × 224 pixels. Rather than constructing the image feature representation with a global vector, we build the spatial feature representation of different regions based on the last pooling layer of the VGGNet. To this end, we split each image in MGC into N (7 × 7) regions. The dimension size of the feature vector for each region is set as 512. Thus, we can represent an image $p_{i,m}$, using $\mathbf{v}_{i,m} = [\boldsymbol{v}_{i,m,1}, \cdots, \boldsymbol{v}_{i,m,N}]$. To make the calculation more convenient, we exploit a single full connection layer to transfer each image vector into a new vector that has the same dimension as the feature vector of textual description.

3.2.3. Other feature representations

In addition to the text description and images, each MGC is also associated with other information, including label, title, author, and time information. To enhance the prediction performance, we represent such information as auxiliary feature vectors, which the processing is given as follows:

Features from label and title. For the label set $\boldsymbol{l}_i$ and title information t_i of MGC $\boldsymbol{i}$, we use the long-short term memory (LSTM)

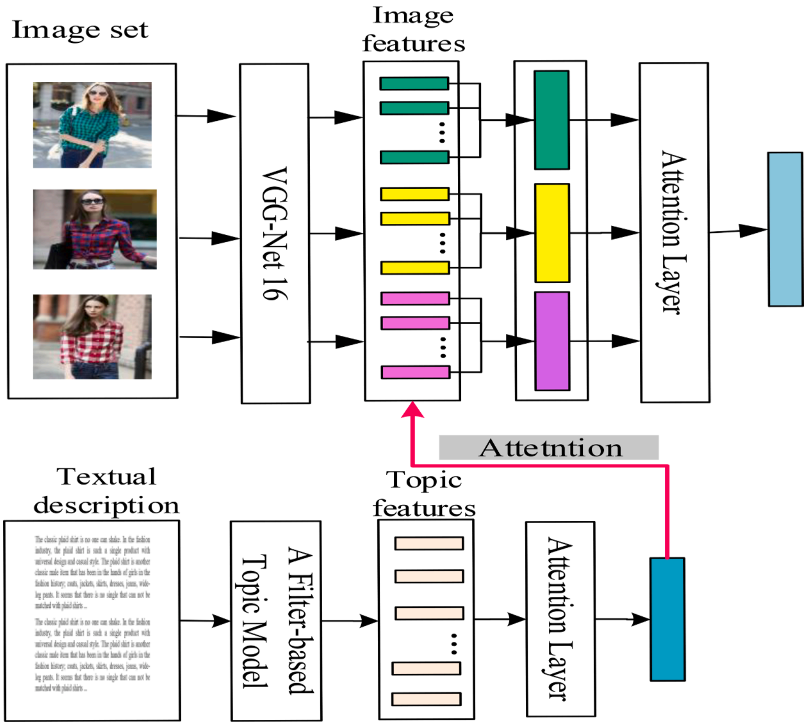


Fig. 4. The module of text-guided attention mechanism.

(Hochreiter & Schmidhuber, 1997) to extract the features from these information. As an efficient method for Natural Language Processing (NLP) tasks, the LSTM can learn high-quality vector representations from textual data. Specifically, taking the label data as an example, we first represent the label $l_{i,j}$ as a one-hot vector $\boldsymbol{x}_{i,j}$ with the size of the label vocabulary. Then, we use the one-hot embeddings of all labels as input. After the transformation of LSTM, we obtain the hidden state vector as the label feature vector. The details for processing label data through LSTM are given as follows:

$$\mathbf{i}_{i,j} = \sigma\left(\mathbf{W}_{input}\boldsymbol{x}_{i,j} + \mathbf{U}_{input}\mathbf{h}_{i,j-1} + \mathbf{b}_{input}\right) \tag{4}$$

$$\mathbf{fo}_{i,j} = \sigma\left(\mathbf{W}_{forget}\boldsymbol{x}_{i,j} + \mathbf{U}_{forget}\mathbf{h}_{i,j-1} + \mathbf{b}_{forget}\right) \tag{5}$$

$$\mathbf{o}_{i,j} = \sigma\left(\mathbf{W}_{output}\boldsymbol{x}_{i,j} + \mathbf{U}_{output}\mathbf{h}_{i,j-1} + \mathbf{b}_{output}\right) \tag{6}$$

$$\mathbf{c}_{i,j} = \mathbf{fo}_{i,j} \odot \mathbf{c}_{i,j-1} + \mathbf{i}_{i,j} \odot \tanh\left(\mathbf{W}_{cell}\boldsymbol{x}_{i,j} + \mathbf{U}_{cell}\mathbf{h}_{i,j-1} + \mathbf{b}_{cell}\right) \tag{7}$$

$$\mathbf{h}_{i,j} = \mathbf{o}_{i,j} \odot \tanh\left(\mathbf{c}_{i,j}\right) \tag{8}$$

where $\mathbf{i}_{i,j}$, $\mathbf{fo}_{i,j}$, $\mathbf{o}_{i,j}$ and $\mathbf{c}_{i,j}$ represent the input gate, forget gate, output gate, and memory cell. $[\mathbf{W}_{input}, \mathbf{W}_{forget}, \mathbf{W}_{output}, \mathbf{W}_{cell}, \ \mathbf{U}_{input}, \mathbf{U}_{forget}, \mathbf{U}_{output}, \mathbf{U}_{cell}\ ]$ are weight matrices. The $\odot$ denotes element-wise multiplication; $\sigma(\cdot)$ and $\tanh(\cdot)$ denote the sigmoid function and hyperbolic function, respectively. We first construct the label feature matrix for the label set $\boldsymbol{l}_i$, by combining the hidden states. Then, we use a mean-pooling operation to obtain a vector representation $\mathbf{f}_i^{label}$ for the label information of MGC i. Similarly, we can get the vector representation $\mathbf{f}_i^{title}$ for the title information of MGC i.

Features from author information. The author of MGC also influences its popularity. For example, an author with a high number of followers often publishes some high-quality content, which can attract more users to view and click. In this paper, we use the number of followers, followings and MGCs, to represent the author feature, $\mathbf{f}_i^{author}$.

Features from time information. The published date of the MGC would be associated with its popularity. For example, MGCs on weekends may have higher views than those on weekdays. Following the Hsu et al. (2019), we turn the published date of the MGC into five sub-features, i.e., hour, week, day, month, and year, represented as follows:

$$\mathbf{f}_i^{time} = \left[f_i^{hour}, f_i^{weak}, f_i^{day}, f_i^{month}, f_i^{year}\right] \tag{9}$$

3.3. Text-guided Attention mechanism

For the popularity prediction of MGCs, we suggest that the long text descriptions convey richer information against the images. Thus, we design a text-guided attention mechanism to use the text's topic features to guide the image representations. In addition, to infer the importance of each feature vector in both text's topic features and image features towards the popularity prediction task, we introduce the attention computation for each visual modality and textual modality. The module of the text-guided attention mechanism is illustrated in Fig. 4.

(1) *Attention computation for topics*. Based on the topic feature for textual description, we first introduce the attention mechanism to determine such topics' importance for the task of popularity prediction. Specifically, we first use a multi-layer perceptron (MLP) to obtain the hidden representation $\overline{x}_{i,k}$ for the topic feature $\boldsymbol{x}_i^k$. Then, we compute the importance of the topic k using the similarity between $\overline{x}_{i,k}$ and a topic-level context vector $\boldsymbol{W}_a$. Based on the softmax function, we calculate the importance score $a_{i,k}$ for the topic k. After that, using a weighted sum of topic features, we construct a new vector representation $\widetilde{\boldsymbol{u}}_i^{text}$ for the textual description of MGC i. The details for this operation are as follows:

$$\overline{\boldsymbol{x}}_{i,k} = \tanh\left(\boldsymbol{W}_{topic}\boldsymbol{x}_i^k + \boldsymbol{b}_{topic}\right) \tag{10}$$

$$a_{i,k} = \frac{\exp\left(\overline{\boldsymbol{x}}_{i,k}^{\mathrm{T}}\boldsymbol{W}_a\right)}{\sum_{k'=1}^{K}\exp\left(\overline{\boldsymbol{x}}_{i,k'}^{\mathrm{T}}\boldsymbol{W}_a\right)} \tag{11}$$

$$\widetilde{\boldsymbol{u}}_i^{text} = \sum\nolimits_{k=1}^{K} a_{i,k}\boldsymbol{x}_i^k \tag{12}$$

where $\boldsymbol{W}_{topic}$, $\boldsymbol{W}_a$, and $\boldsymbol{b}_{topic}$ are the parameters.

(1) ***Text-guided attention computation.*** In most cases, a text description of MGC is just associated with few regions in its corresponding images. To filter out the irrelevant and unimportant regions in images for popularity prediction, we design a text-guided attention mechanism to obtain new textual-based visual features. This module can adaptively attune visual features according to the constraint of the corresponding textual input. Specifically, based on the text vector representation $\widetilde{\boldsymbol{u}}_i^{text}$, we first calculate the new region vector representation $\overline{\boldsymbol{v}}_{i,m,n}$ for region feature $\boldsymbol{v}_{i,m,n}$.For the convenience of computation, we apply a single full connection layer to align the dimension of each region's vector to the same dimension as the vector representation $\widetilde{\boldsymbol{u}}_i^{text}$. This step uses textual features to single-directionally guide the region features through the generation of filters. Then, for each region, we compute the attention score $a_{i,m,n}$ using the similarity between $\overline{\boldsymbol{v}}_{i,m,n}$ and the context vector $\boldsymbol{W}_v$. After that, we can generate weight scores for region-level visual features in each image. Next, with a weighted sum of the region features, we construct a new vector representation $\widetilde{\boldsymbol{v}}_{i,m}^{image}$ for the image m in MGC i. By doing this, we aggregate a set of region features together by reducing the impact of irrelevant visual features. The details for this operation are given as follows:

$$\overline{\boldsymbol{v}}_{i,m,n} = \tanh\left(\boldsymbol{W}_{region}\boldsymbol{v}_{i,m,n}\right) \odot \tanh\left(\boldsymbol{W}_{text}\widetilde{\boldsymbol{u}}_i^{text}\right) \tag{13}$$

$$a_{i,m,n} = \frac{\exp\left(\overline{\boldsymbol{v}}_{i,m,n}^{\mathrm{T}}\boldsymbol{W}_v\right)}{\sum_{n'=1}^{N}\exp\left(\overline{\boldsymbol{v}}_{i,m,n'}^{\mathrm{T}}\boldsymbol{W}_v\right)} \tag{14}$$

$$\widetilde{\boldsymbol{v}}_{i,m}^{image} = \sum\nolimits_{n=1}^{N} a_{i,m,n}\boldsymbol{v}_{i,m,n} \tag{15}$$

where $\boldsymbol{W}_{region}$, $\boldsymbol{W}_{text}$ and $\boldsymbol{W}_v$ are parameters. Based on the text-guided attention mechanism, we can construct a model with a focus on those important regions in each image, thus enhancing the prediction performance.

(1) ***Attention computation for multiple images in each MGC.*** Typically, each MGC contains multiple images. In addition, not all images in each MGC are equivalently relevant to predicting popularity. To determine the contribution of each image to the task of prediction, we apply the attention mechanism to fuse features of multiple images. The details for this operation are given as follows:

$$\boldsymbol{h}_{i,m} = \tanh\left(\boldsymbol{W}_{image}\widetilde{\boldsymbol{v}}_{i,m}^{image} + \boldsymbol{b}_{image}\right) \tag{16}$$

$$a_{i,m} = \frac{\exp\left(\boldsymbol{h}_{i,m}^{\mathrm{T}} \boldsymbol{W}_h\right)}{\sum_{m'=1}^{M_i} \exp\left(\boldsymbol{h}_{i,m'}^{\mathrm{T}} \boldsymbol{W}_h\right)} \tag{17}$$

$$\widetilde{\boldsymbol{v}}_i^{image} = \sum_{m=1}^{M_i} a_{i,m} \widetilde{\boldsymbol{v}}_{i,m}^{image} \tag{18}$$

where $\boldsymbol{h}_{i,m}$ denotes the hidden representation for the image feature $\widetilde{\boldsymbol{v}}_{i,m}^{image}$; $a_{i,m}$ is attention score for image m in MGC i; $\widetilde{\boldsymbol{v}}_i^{image}$ is the new vector representation by fusing visual images in MGC i. $\boldsymbol{W}_{image}$, $\boldsymbol{W}_h$ and $\boldsymbol{b}_{image}$ are parameters.

3.4. Multimodal compact bilinear for visual and textual feature

Prior studies mainly apply the element-wise product, sum, or concatenation to combine representations of different modalities. Although these operations construct a joint representation, they might not fully reflect the complicated interaction between two different modalities. To efficiently and expressively fuse our visual and textual representations, we use a Multimodal Compact Bilinear (MCB) pooling (Fukui et al., 2016). With the interaction between all elements of these two vectors, MCB pooling has been widely used in many tasks, e.g., visual question answering (Do et al., 2019) and event classification (Abavisani et al., 2020). In our context, MCB pooling take the outer product of our textual vector $\widetilde{\boldsymbol{u}}_i^{text}$ and visual vector $\widetilde{\boldsymbol{v}}_i^{image}$, then learning a linear model $\mathscr{L}$ as follows:

$$\mathscr{C}_i = \mathscr{L}\left[\widetilde{\boldsymbol{u}}_i^{text} \otimes \widetilde{\boldsymbol{v}}_i^{image}\right] \tag{19}$$

where $\mathscr{C}_i$ denotes the fused representation; $\otimes$ is the outer product; and [] represents linearizing the matrix in a vector. From Eq. (19), we note that the vectors' interaction in bilinear pooling is based on the multiplicative way, which generates the high dimensional representation and thus makes it difficult to learn massive parameters in the model $\mathscr{L}$. Inspired by Fukui et al. (2016), we apply the Count Sketch projection function $\Psi(\cdot)$ to project the outer product to a lower-dimensional space, which reduces the number of parameters in the model $\mathscr{L}$. Pham and Pagh (2013) suggested that the count sketch of the outer product of two vectors can be decomposed into the convolution of both count sketches, which is given as follows:

$$\Psi\left(\widetilde{\boldsymbol{u}}_i^{text} \otimes \widetilde{\boldsymbol{v}}_i^{image}, p, q\right) = \Psi\left(\widetilde{\boldsymbol{u}}_i^{text}, p, q\right) * \Psi\left(\widetilde{\boldsymbol{v}}_i^{image}, p, q\right) \tag{20}$$

where p and q denote vectors that are randomly initialized from a uniform distribution; $*$ denote the convolution operator. Also, the convolution theorem argues that convolution in the time domain equals the element-wise product in the frequency domain. To improve the efficiency, we use the element-wise product in Fast Fourier Transform (FFT) space to operate the convolution.

$$\boldsymbol{u}^{'} * \boldsymbol{v}^{'} = \mathrm{FFT}^{-1}(\mathrm{FFT}(\boldsymbol{u}^{'}) \odot \mathrm{FFT}(\boldsymbol{v}^{'})) \tag{21}$$

where $\odot$ denote element-wise product; $\boldsymbol{u}^{'} = \Psi(\widetilde{\boldsymbol{u}}_i^{text}, p, q)$ and $\boldsymbol{v}^{'} = \Psi(\otimes \widetilde{\boldsymbol{v}}_i^{image}, p, q)$.

3.5. Prediction and training

Based on the final presentations obtained from the above process, we concatenate the features of $[\widetilde{\boldsymbol{u}}_i^{text}, \widetilde{\boldsymbol{v}}_i^{image}, \mathbf{f}_i^{title}, \mathbf{f}_i^{label}, \mathbf{f}_i^{author}, \mathbf{f}_i^{time}]$ and obtain the global feature $\mathbf{f}_{global}$. Then, we feed the global feature into the sigmoid function for predicting the popularity score:

$$\widehat{y} = \mathrm{sigmoid}\left(\boldsymbol{W}_g \mathbf{f}_{global} + \boldsymbol{b}_g\right) \tag{22}$$

where $\boldsymbol{W}_g$ and $\boldsymbol{b}_g$ are weights that need to be trained. The larger $\widehat{y}$ indicates the higher probability of the MGC is to be popular.

For the optimization of our model, we apply the cross-entropy as loss function, which is defined as follows:

$$J = -\sum_{i=1}^{S} (y_i \log \widehat{y}_i + (1 - y_i) \log(1 - \widehat{y}_i)) \tag{23}$$

where S is the total number of elements in the training set. y_i and $\widehat{y}_i$ are the true popularity score and the predictive result of the i-th MGC, respectively. We use Adam Optimizer (DP & Ba, 2015) with backpropagation to minimize the objective loss function. In addition, we use the dropout for the regularization.

4. Experiments

In this section, we assess the performance of the proposed method using two real-world datasets. We first present the experimental details, including the datasets, baselines, and evaluation metrics. Then, we report the results, i.e., overall performance and ablation study. Next, we give the qualitative analysis by visualization. Finally, we present several important practical implications of the

Table 1
Summary statistics.

Dataset	Hot samples	Cold samples
Taobao.com	7036	7036
Autohome.com	9597	9597

proposed model.

4.1. Datasets and setting

To test the performance of our proposed model, we crawl two datasets from Taobao.com and Autohome.com. We count the number of clicks of MGCs in Autohome.com and Taobao.com. And we find that these distributions of MGCs' clicks obey the power-law distributions, which are highly skewed. However, the clicks of MGCs in Taobao.com are far below that in Autohome.com. The main reason is that users on Taobao.com mainly use its search function and pay less attention to MGCs comparatively. For MGCs in Taobao.com, we divide the overall popularity into "hot" and "cold" with the cutoff point set as 300 clicks. And for MGCs in Autohome.com, we divide the overall popularity into "hot" and "cold" with the cutoff point set as 10,000 clicks. In addition, we find out that the number of "cold" MGCs is much larger than that of "hot" MGCs, thus resulting in unbalanced effects. To avoid the predictive bias towards the "cold" category, we construct the balanced datasets using an undersampling technique for selecting samples (Chawla et al., 2004).

We preprocess and standardize our raw datasets for our proposed model. Specifically, we apply the Jieba tool to segment the textual description and title into meaningful Chinese words. Then, for textual description and title, we remove all of the stop words. We also remove low-frequency words, whose frequency of occurrence is no more than 2% in the textual description, ensuring that the topic results are not influenced by outlier words. In addition, we select the MGCs that also contain both images and labels. Table 1 summarizes the detailed statistics of our final datasets. To train our proposed model, we randomly split 80% of the datasets as the train set, 10% as the validation set, and 10% as the test set.

Our related experiments are carried out on an Ubuntu 16.04 (GNU/Linux) server with a 2.10 GHz Intel(R) Xeon(R) E5–2620 CPU, 128 GB of memory, and four TITAN X GPU. We implement the FBT model by Java, and run this model with hyperparameters $\alpha = \frac{50}{K}$, $\beta_0 = \beta_1 = 0.01$, $\varepsilon_0 = \varepsilon_1 = 0.01$. In addition, we set the number of iterations as 4000, and use a statistical approach, namely perplexity score (Blei et al., 2003), to optimize the number of topics for Taobao and Autohome datasets. The proposed TGANN model is implemented and trained based on TensorFlow that is a popular open-source library for deep learning. The hidden states in the LSTM is set to 500. The Adam optimizer with learning rate $\varepsilon = 0.001$, $\beta = (0.9, 0.999)$ is adopted to train the TGANN model. The batch size is 256. In addition, the rate of dropout is set to 0.4 in our training procedure.

4.2. Baselines

To analyze the effectiveness of the proposed TGANN model, we compare its prediction performance with four state-of-the-art baselines, including some traditional feature-based methods and a deep learning model based on attention mechanisms. We select these methods as baselines because they are widely used in popularity prediction (Chen et al., 2019; Zhang et al., 2018). Brief descriptions of baselines are listed as follows:

Logistic Regression (LR): LR is the simplest model for classification, which offers benchmark performance. To train this model, we first use the pre-trained VGGNet and obtain each image feature based on the output of the full connection layer in VGGNet. Then, we use the average of all image features for each MGC as the input for the LR model. In addition, we directly apply document topic distribution learned by the FBT model, to represent the textual description of each MGC.

Support Vector Machine (SVM): SVM (Chang & Lin, 2011) is a widely used model for classification. To train the model, we consider multimodal features similar to those included in LR. In addition, we use the grid search method to optimize parameters for SVM.

LightGBM: LightGBM (Ke et al., 2017) is an efficient boosting model in ensemble learning, which can build a good balance between keeping the accuracy for learned decision trees and reducing the number of data instances. Following the previous work (He et al., 2019), we feed the multiple features into LightGBM to predict the popularity scores for MGCs.

Co-Attention Network (CAN): The last strong baseline method is taken inspiration from the co-attention network for multi-modal learning. The CAN is the state-of-the-art model for visual question answering and recommendation (Lu et al., 2016; Ma et al., 2019). However, this model can simply combine both the textual and visual information by the co-attention network. To adapt to our datasets, we extend the model by introducing additional information such as labels, and titles.

4.3. Evaluation metrics

Since we regard the popularity prediction task as a classification, we use three standard metrics, including precision, recall, and F1-score, to evaluate the model performance. Specifically, precision denotes the ratio of correct positive predictions to all MGCs predicted as positive samples. Recall calculates the ratio of positive MGCs' predictions that are successfully identified out of all MGCs; F1-score is calculated by a harmonic mean of precision and recall. The precision, recall, and F1-score are defined as follows:

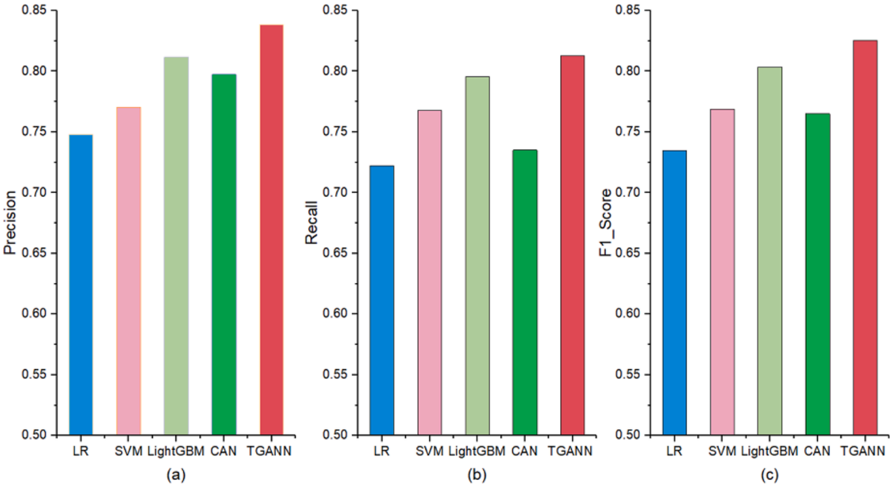


Fig. 5. Performance of our proposed model and the baselines on the Taobao dataset.

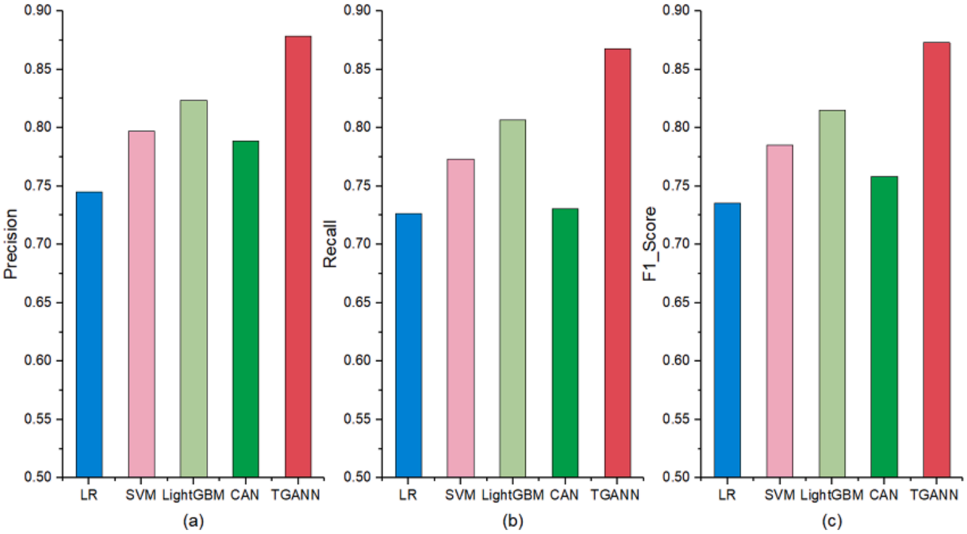


Fig. 6. Performance of our proposed model and the baselines on the Autohome dataset.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{24}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{25}$$

$$\text{F1}_{\text{Score}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

where *TP* is the number of MGCs successfully predicted as “hot” MGCs; *FP* is the number of “cold” MGCs that are wrongly predicted as “hot” MGCs; *FN* is the number of “hot” MGCs that are predicted as “cold” MGCs. Better predictive performance is revealed by a higher precision, recall, and F1-score over the test set.

4.4. *Overall performance*

In this section, we give the prediction performances by comparing the proposed TGANN model with the four baselines. For the robust comparison, we run the proposed model and each baseline five times and calculate their average performance. Fig. 5 displays the comparison results on the Taobao dataset. From this figure, we observe that the proposed model outperforms all the baselines across all metrics such as precision (0.8385), recall (0.8127), and F1-score (0.8254). Fig. 6 shows the performance of the Autohome

Table 2
Improvements between our proposed model and the baseline.

Model	Taobao dataset			Autohome dataset		
	Precision	Recall	F1-score	Precision	Recall	F1-score
LR	10.79%***	11.16%***	10.97%***	15.21%***	16.24%***	15.73%***
SVM	8.14%***	5.53%***	6.83%***	9.26%***	10.87%***	10.08%***
LightGBM	3.21%***	2.08%**	2.64%***	6.24%***	7.03%***	6.64%***
CAN	4.90%***	9.54%***	7.31%***	10.23%***	15.82%***	13.13%***

Note: * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$. All differences are statistically significant at $p < 0.01$ on paired t-tests.

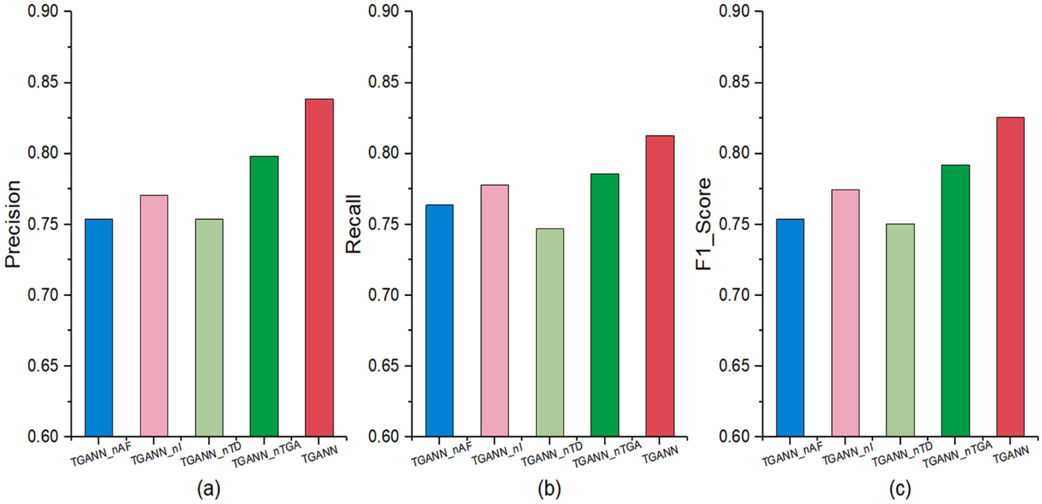


Fig. 7. Results of ablation studies on the Taobao dataset.

dataset. Also for the prediction task on the Autohome dataset, we find out that our proposed model achieves the best performance in terms of precision (0.8384), recall (0.8676), and F1-score (0.8730).

To facilitate the analysis, we also compute the improvements between our proposed model and the baseline. For example, for the LR model on the Taobao dataset, the precision is 0.7480, with the improvement being (0.8385–0.7480)/0.8385 = 10.79%. Table 2 shows the improved results on our two datasets. We can find that LR does not perform well, and SVM performs better than LR. In addition, LightGBM achieves the second-best performance, indicating that it is also an efficient method to predict the popularity of MGCs. Compared with the LightGBM method, the proposed TGANN model exhibits performance improvement, about 3.21% in precision, 2.08% in recall and 2.64% in F1-score on the Taobao dataset, respectively; and about 6.24% in precision, 7.03% in recall and 6.64% in F1-score on the Autohome dataset, respectively. Compared with the co-attention model CAN, our proposed model shows absolute improvement in terms of 4.90% in precision, 9.54% in recall and 7.31% in F1-score on the Taobao dataset, respectively; and 10.23% in precision, 15.82% in recall and 13.13% in F1-score on the Autohome dataset, respectively. To help ensure model robustness, we apply precision, recall, and F1-score evaluated over different runs of the proposed model and a baseline, and further measure the statistical significance of the experimental results by conducting the paired *t*-test (Kokkodis & Ipeirotis, 2021). The null hypothesis supposes that the competing model performs worse than the proposed model. As shown in Table 2, we note that the improvements of our model over the baselines, are statistically significant at $p < 0.01$ on paired t-tests.

From this result, we can conclude that the text-guided attention mechanism is more suitable for the task of popularity prediction for MGC than using the co-attention mechanism. The co-attention mechanism generates text and image attention simultaneously, which aims to find consistent information between two modalities (Liu et al., 2021; Tay et al., 2018). However, for the popularity prediction of MGCs, we need to consider not only consistent information but also unique information in each modality. Moreover, we observe that the improvement of TGANN in all metrics on the Autohome dataset is better than that on the Taobao dataset. The main reason is that the text length of the Autohome dataset is longer than that of the Taobao dataset. And the longer the text length is, the more information the TGANN model exploits. In summary, the proposed TGANN model has revealed its superior performance, validating the effectiveness of the proposed multi-modal feature fusion method.

4.5. Ablation study

To further illustrate the advantages of the proposed TGANN model structure, we also conduct ablation studies to test the contribution of each used feature. To this end, we simplify our TGANN model with four versions: TGANN_nAF, TGANN_nI, TGANN_nTD, and

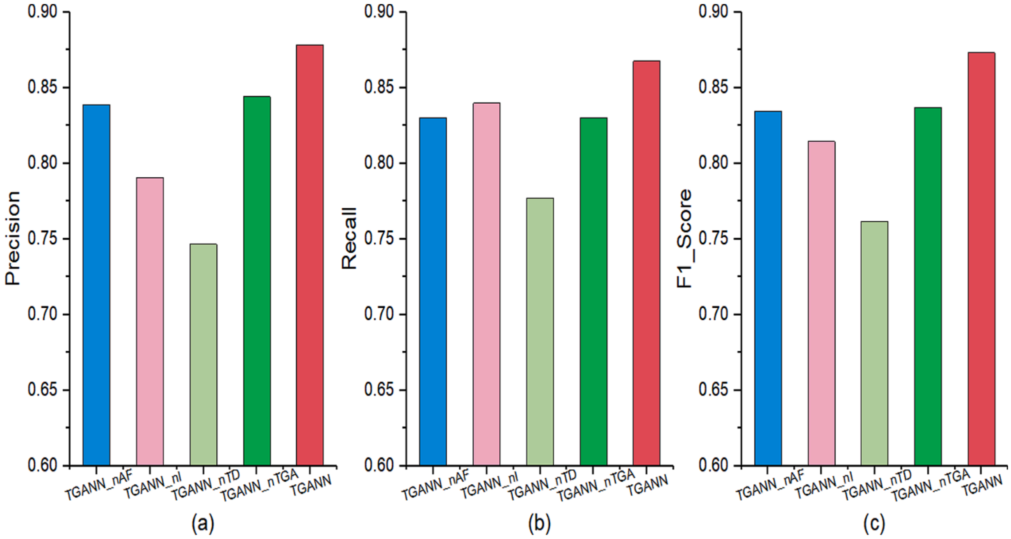


Fig. 8. Results of ablation studies on the Autohome dataset.

Table 3
Improvements between our proposed model and the variant method.

Model	Taobao dataset			Autohome dataset		
	Precision	Recall	F1-score	Precision	Recall	F1-score
TGANN_nAF	10.15%***	6.04%***	8.11%***	4.48%***	4.33%***	4.40%**
TGANN_nI	8.09%***	4.25%**	6.18%***	9.98%***	3.22%**	6.70%***
TGANN_nTD	10.09%***	8.05%***	9.07%***	15.01%***	10.41%***	12.76%***
TGANN_nTGA	4.80%***	3.31%**	4.05%**	3.90%**	4.34%***	4.12%***

Note: * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$. All differences are statistically significant at $p < 0.01$ on paired t-tests.

Table 4
Performance comparisons for different fusion strategies.

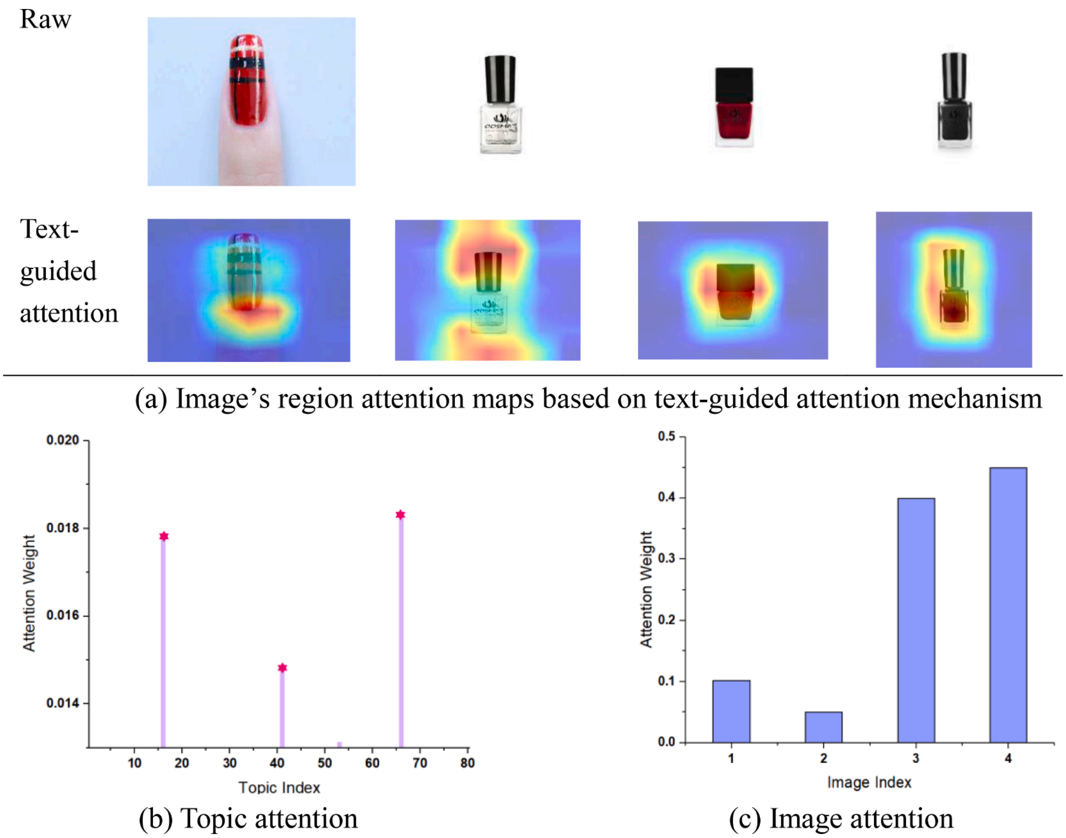
Fusion Strategy	Taobao dataset			Autohome dataset		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Element-wise sum	81.98***	80.11***	81.04***	85.92***	84.59**	85.25***
Element-wise product	82.44**	80.79**	81.61***	86.05***	84.98**	85.51***
MCB pooling	83.85	81.27	82.54	87.84	86.76	87.30

Note: * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$. All differences are statistically significant at $p < 0.01$ on paired t-tests.

TGANN_nTGA. The TGANN_nAF model only integrates the textual features learned by the topic model, and image features guided by textual features, without considering other **a**uxiliary **f**eatures, i.e., label, title, author, and time information. For the TGANN_nI model, we do **n**ot consider the **i**mage features and remove MCB pooling from the structure of the proposed model. For the TGANN_nTD model, we do **n**ot consider the features of **t**extual **d**escriptions and remove MCB pooling from the structure of the proposed model. Compared with our model, the TGANN_nTGA model does **n**ot consider the **t**ext-**g**uided **a**ttention mechanism.

Figs. 7 and 8 show the results of ablation studies on the Taobao dataset and Autohome dataset, respectively. Table 3 presents the improvements between our proposed model and the variant method. Based on the paired t-tests in Table 3, our model outperforms all baseline methods for both datasets in terms of precision, recall, and F1-score with statistically significant margins ($p < 0.01$). On both datasets, we can see that the proposed TGANN model performs better than TGANN_nTGA, which shows that introducing the text-guided attention mechanism can significantly enhance the predictive power for MGCs' popularity. This is intuitive because our model can extract more useful information from both textual features and visual features with the help of the attention-guided mechanism. The TGANN_nTGA performs better than TGANN_nAF, TGANN_nI, and TGANN_nTD. This demonstrates that ignoring important features (e.g., image features) would significantly affect prediction performance.

In addition, we note that TGANN_nI performs better than TGANN_nTD, which indicates that using features extracted from the textual description of MGCs can give better popularity prediction than that of using images. Based on the performance of the TGANN_nAF model, we suggest that the auxiliary features have a significant impact on the popularity prediction of both MGC datasets. Interestingly, the auxiliary features on the Autohome dataset have a greater impact on the popularity prediction than that on the



(a) Image’s region attention maps based on text-guided attention mechanism

(b) Topic attention

(c) Image attention

Fig. 9. Case study 1 for attention generation.

Taobao dataset. This is intuitive because online users pay less attention to auxiliary information (e.g., author and time) of automotive products. For apparel products in the Taobao dataset, MGCs’ popularity is more vulnerable to factors such as timeliness and professional author.

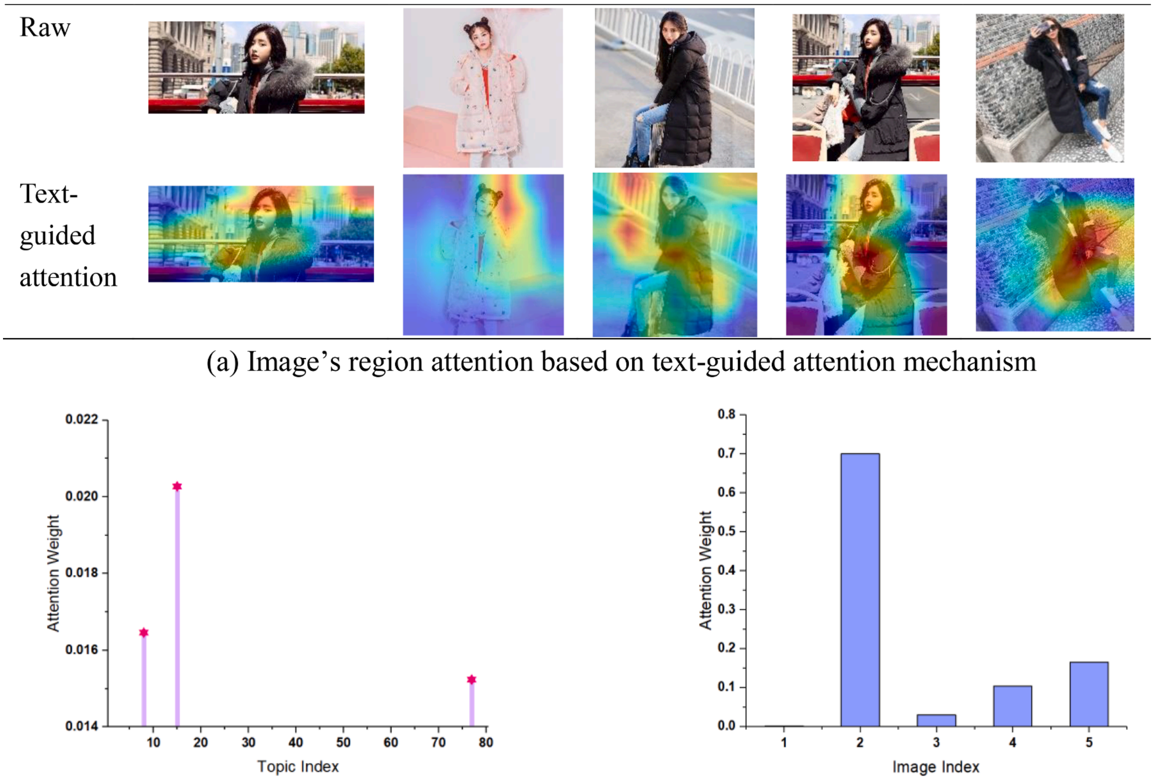
Moreover, compared with the TGANN_nTD model, our proposed model presents significant improvement in terms of 10.09% in precision, 8.05% in recall, and 9.07% in F1-score on the Taobao dataset; in terms of 15.01% in precision, 10.41% in recall, and 12.76% in F1-score on the Autohome dataset. Also, we note that the improvement of the TGANN_nTD model on the Autohome dataset is more obvious than that on the Taobao dataset.

To verify the effectiveness of MCB pooling in our model, we also compare the performance of our pooling method with the non-bilinear methods, i.e., the element-wise sum and the element-wise product. The comparison results are shown in Table 4. We find that MCB pooling outperforms the element-wise sum and the element-wise product methods, on both two datasets. The main reason is that MCB pooling can make full use of the interactions between the visual and text representations, which is consistent with previous studies (Fukui et al., 2016; Gao et al., 2016).

4.6. Qualitative analysis

Apart from the above quantitative results, we give the qualitative analysis by visualizing three attentions (region attentions, topic attentions, and image attentions) generated by the proposed model. Specifically, Figs. 9 and 10 give representations of two MGC cases in our Taobao dataset. Table 5 gives the 10 most likely words of six topics that appear in Figs. 9 and 10. From case study 1 in Fig. 9(a), we find that our model can obtain fine image’s region attention maps since the attended regions are the important and meaningful elements. And most of the background information is filtered out. However, the second image in Fig. 9(a), suggests that the proposed model may pick some irrelevant regions. We suggest that two factors led to this result. One is that the high similarities between the object regions and the background regions in an image pose a bias in learning the representations of the image’s regions (Fan et al., 2020; Ning et al., 2010). Another is that the VGGNet-16 we use is pre-trained on an ILSVRC dataset, which is different from our datasets. Thus, crudely using the pre-trained weights may result in deviations for the feature representations (Zhang et al., 2017). And the feature representations are critical to the performance of the image’s region attention.

Based on Fig. 9(b), we can easily determine the contribution of each topic to the popularity of this MGC. For this case, we note that topic 66, topic 16, and topic 41 are useful information on its popularity prediction. From Table 5, we can see that topic 66 is mainly about graffiti of fingers, with representative words such as “graffiti,” “finger,” “vase” and “dye.” Topic 16 corresponds to the hygiene



(a) Image's region attention based on text-guided attention mechanism

(b) Topic attention

(c) Image attention

Fig. 10. Case study 2 for attention generation.

Table 5
Representative words for each topic.

Topic Index	Representative words
8	suit, autumn, winter, suede, down coat, innovate, high-quality, fashion, city
15	clothing, double-deck, Korea style, cool, counterattack, city, life, cotton-coat, famous-brand, hairy
16	nail, hygiene, health, armor, delicate, colorful, neat, season, ornament, beautiful
41	water, tasteless, collocation, pink, square, headset, grapefruit, sexy, life, purple
66	graffiti, finger, vase, dye, color, polish, lipstick-pen, claret, light, lovely
77	British style, plus-velvet, fashion, lovely, leisure, Western, cloth shoes, Martin boots, mild, nimble

and health of nails, with representative words such as "nail," "hygiene," "health" and "armor." Topic 41 mainly focuses on color matching or shapes for nails, which are described by words including "water," "tasteless", "collocation" and "pink". These three topics can well reveal the characteristics of the product in Fig. 9(a). In addition, from Fig. 9(c), we note that images 3 and 4 are more vivid, contributing more to the prediction result.

From case study 2 in Fig. 10(a), we find that the fashionable clothes in these images are clearly marked as important elements. Surprisingly, part of background knowledge about the architectural style is also captured, which influences the popularity prediction of this MGC. As shown in Fig. 10(b), we can see that the topic attention focuses on topic 15, topic 8, and topic 77. From Table 5, we find that the word "city" appears in both topics 8 and 15, and this is why part of the background information in images is marked as important elements. From Fig. 10(c), we observe that image 2 has a higher impact on the prediction result than other images in this MGC. To sum up, the qualitative analysis reveals that the proposed TGANN model has the advantages of estimating the topic importance, image importance, and image's region importance towards the popularity of the MCC. For marketers, our model can discover more explanatory results, which serve as informative signals for them to adjust their advertising strategies and to design purposeful marketing activities.

4.7. Practical implications

In online markets, MGC is a very important type of information that has the ability to influence consumers' decision-making

Algorithm 1
Filter-based topic (FBT) model.

```
1. For each specific topic $k \in [1, K]$
    (a) Draw $\phi_k \sim \text{Dirichlet}(\beta_0)$
2. Draw the general topic $\boldsymbol{\varphi} \sim \text{Dirichlet}(\beta_1)$
3. For each document $i$,
    (a) Draw the topic distribution $\boldsymbol{\theta}_i \sim \text{Dirichlet}(\alpha)$
    For $n$-th word in document $i$, $n \in [1, N_i]$
        (a) Draw a specific topic $z_{in} \sim \text{Multinomial}(\boldsymbol{\theta}_i)$
        (b) Draw the $\gamma_i \sim \text{Beta}(\varepsilon_0, \varepsilon_1)$
        (b) Draw $b_{in} \sim \text{Bernoulli}(\gamma_i)$
        if $b_{in} = 0$
            Draw $w_{in} \sim \text{Multinomial}(\boldsymbol{\varphi})$, where $w_{in} \in [1, V]$
        else
            Draw $w_{in} \sim \text{Multinomial}(\phi_{z_{in}})$, where $w_{in} \in [1, V]$
```

process. The MGC usually contains various types of multimedia such as texts, images, labels, or other display formats. This paper contributes to the popularity prediction of MGC by fusing its multi-modal features. Our results provide several important practical implications for marketers and online platforms.

First, for marketers, our model can gain the key insight about "What kind of MGC is more attractive to consumers?" We easily determine the specific and important parts of images and text, with the help of the attention mechanism. Thus, our results can help marketers improve writing quality and create more attractive MGCs. In addition, we can readily capture the topic structure in the whole market, which enhances marketers' ability to identify potential explosion points to improve the marketing effect. Last but not least, our results reveal that it would be beneficial to combine texts with images to attract consumers.

Second, the large-scale MGCs generated and disseminated on online platforms every day, not only bring great challenges to the management of the online platforms but also easily cause the trouble of information overload to consumers. Predicting the popularity of MGC would help online platforms optimize content quality and upgrade search engines. In addition, our results can be used to improve recommendation systems that allocate more attractive MGCs to target consumers.

5. Conclusion and future work

In this study, we attempted to address the problem of the popularity prediction for marketer-generated content (MGC). To this end, we proposed a text-guided attention neural network (TGANN) model that fuses the multi-modal features and explores their benefits for the popularity prediction of MGCs. Specifically, the TGANN integrates the textual features learned by the topic model, image features guided by textual features, and other auxiliary features (e.g., label and title) to enhance the prediction performance. In addition, we successfully addressed three major challenges in the task of the popularity prediction of MGCs including the heterogeneous and multi-modal data, noise problems in text and image information, and the fusion mechanism of multi-modal features. To evaluate the performance of TGANN and compare it with other state-of-the-art methods, we constructed two MGC datasets collected from Taobao and Autohome respectively. Based on the quantitative comparison and ablation studies, the experimental results showed that our proposed model can achieve better performance than other baseline methods.

There are some interesting aspects for future work. First, the MGCs' popularity prediction problem can also be regarded as a regression task. Future research may select appropriate regression (e.g., linear regression and Poisson regression) for our model framework. Second, though multi-modal features are highly informative to the MGCs' popularity prediction, some external uncertainties (e..g, online celebrities, and other social media) also impact the popularity of MGCs. Thus, another aspect of future work would be to design a new method that considers the uncertainty in popularity prediction. Third, in this study, we focus on only popularity prediction based on multi-modal features of MGCs. It would be interesting to test the impact of different features (e.g., image qualities, topics in textual contents) of the MGCs on user engagement (e.g., number of clicks) via empirical study.

Algorithm 1

CRediT authorship contribution statement

Yang Qian: Conceptualization, Methodology, Writing – original draft. **Wang Xu:** Conceptualization, Methodology. **Xiao Liu:** Writing – original draft, Validation. **Haifeng Ling:** Formal analysis, Validation. **Yuanchun Jiang:** Conceptualization, Validation. **Yidong Chai:** Data curation. **Yezheng Liu:** Visualization.

Declaration of Competing Interest

The authors certify that there is no conflict of interest in the subject matter discussed in the manuscript.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (72101072, 91846201, 72171071, 72101079), the

Postdoctoral Research Foundation of China (2021M690852), the Fundamental Research Funds for the Central Universities (JZ2021HGQB0272), and the National Engineering Laboratory for Big Data Distribution and Exchange Technologies.

References

Abavisani, M., Wu, L., Hu, S., Tetreault, J., & Jaimes, A. (2020). Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14679–14689).

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management, 58*(4), Article 102569.

Behera, R. K., Jena, M., Rath, S. K., & Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing & Management, 58*(1), Article 102435.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Briskilal, J., & Subalalitha, C. (2022). An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. *Information Processing & Management, 59*(1), Article 102756.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3), 1–27.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter, 6*(1), 1–6.

Chen, J., Liang, D., Zhu, Z., Zhou, X., Ye, Z., & Mo, X. (2019). Social media popularity prediction based on visual-textual features with xgboost. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 2692–2696).

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014). Can cascades be predicted?. In *Proceedings of the 23rd international conference on world wide web* (pp. 925–936).

Cheng, Y., Wang, R., Pan, Z., Feng, R., & Zhang, Y. (2020). Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 3884–3892).

Ding, K., Wang, R., & Wang, S. (2019). Social media popularity prediction: A multiple feature fusion approach with deep neural networks. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 2682–2686).

Do, T., Do, T. T., Tran, H., Tjiputra, E., & Tran, Q. D. (2019). Compact trilinear interaction for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 392–401).

DP, K., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd international conference for learning representations (ICLR)*.

Fan, D. P., Ji, G. P., Sun, G., Cheng, M. M., Shen, J., & Shao, L. (2020). Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2777–2787).

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the EMNLP*.

Gao, X., Cao, Z., Li, S., Yao, B., Chen, G., & Tang, S. (2019). Taxonomy and evaluation for microblog popularity prediction. *ACM Transactions on Knowledge Discovery from Data (TKDD), 13*(2), 1–40.

Gao, Y., Beijbom, O., Zhang, N., & Darrell, T. (2016). Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 317–326).

He, Z., He, Z., Wu, J., & Yang, Z. (2019). Feature construction for posts and users combined with lightgbm for social media popularity prediction. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 2672–2676).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hsu, C. C., Kang, L. W., Lee, C. Y., Lee, J. Y., Zhang, Z. X., & Wu, S. M. (2019). Popularity prediction of social media based on multi-modal feature mining. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 2687–2691).

Hsu, C. C., Tseng, W. H., Yang, H. T., Lin, C. H., & Kao, C. H. (2020). Rethinking relation between model stacking and recurrent neural networks for social media prediction. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 4585–4589).

Kang, P., Lin, Z., Teng, S., Zhang, G., Guo, L., & Zhang, W. (2019). Catboost-based framework with additional user information for social media popularity prediction. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 2677–2681).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.

Kokkodis, M., & Ipeirotis, P. G. (2021). Demand-aware career path recommendations: A reinforcement learning approach. *Management Science, 67*(7), 4362–4383.

Kong, Q., Rizoiu, M. A., & Xie, L. (2020). Modeling information cascades with self-exciting processes via generalized epidemic models. In *Proceedings of the 13th international conference on web search and data mining* (pp. 286–294).

Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research, 57*(1), 1–19.

Liao, D., Xu, J., Li, G., Huang, W., Liu, W., & Li, J. (2019). Popularity prediction on online articles with deep fusion of temporal process and content features. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 200–207).

Lin, Z., Huang, F., Li, Y., Yang, Z., & Liu, W. (2019). A layer-wise deep stacking model for social image popularity prediction. *World Wide Web, 22*(4), 1639–1655.

Liu, Y., Zhang, X., Zhang, Q., Li, C., Huang, F., Tang, X., et al. (2021). Dual self-attention with co-attention networks for visual question answering. *Pattern Recognition, 117*, Article 107956.

Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems, 29*, 289–297.

Ma, R., Qiu, X., Zhang, Q., Hu, X., Jiang, Y. G., & Huang, X. (2019). Co-attention memory network for multimodal microblog's hashtag recommendation. *IEEE Transactions on Knowledge and Data Engineering, 33*(2), 388–400.

Mishra, S., Rizoiu, M. A., & Xie, L. (2016). Feature driven and point process approaches for popularity prediction. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 1069–1078).

Ning, J., Zhang, L., Zhang, D., & Wu, C. (2010). Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition, 43*(2), 445–456.

Peng, Y., He, X., & Zhao, J. (2017). Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing, 27*(3), 1487–1500.

Pham, N., & Pagh, R. (2013). Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 239–247).

Sawhney, R., Agarwal, S., Thakkar, M., Wadhwa, A., & Shah, R. R. (2021a). Hyperbolic online time stream modeling. In *Proceedings of the 44th International ACM SIGIR conference on research and development in information retrieval* (pp. 1682–1686).

Sawhney, R., Agarwal, S., Wadhwa, A., & Shah, R. R. (2020). Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (pp. 8415–8426).

Sawhney, R., Wadhwa, A., Agarwal, S., & Shah, R. (2021b). FAST: Financial news and tweet based time aware network for stock trading. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume* (pp. 2164–2175).

Shin, D., He, S., Lee, G. M., Whinston, A. B., Cetintas, S., & Lee, K. C. (2020). Enhancing social media analysis with visual data analytics: A deep learning approach. *MIS Quarterly, 44*(4), 1459–1492.

Shraga, R., Roitman, H., Feigenblat, G., & Cannim, M. (2020). Web table retrieval using multimodal deep learning. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1399–1408).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International conference on learning representations.

Tay, Y., Luu, A. T., & Hui, S. C. (2018). Multi-pointer co-attention networks for recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2309–2318).

Trzciński, T., & Rokita, P. (2017). Predicting popularity of online videos using support vector regression. *IEEE Transactions on Multimedia, 19*(11), 2561–2570.

Wang, W., Zhang, Y., Sui, Y., Wan, Y., Zhao, Z., Wu, J., et al. (2020). Reinforcement-learning-guided source code summarization via hierarchical attention. *IEEE Transactions on software Engineering*.

Wilterson, A. I., & Graziano, M. S. (2021). The attention schema theory in a neural network agent: Controlling visuospatial attention using a descriptive model of attention. *Proceedings of the National Academy of Sciences, 118*(33).

Wu, Q., Yang, C., Zhang, H., Gao, X., Weng, P., & Chen, G. (2018). Adversarial training model unifying feature driven and point process perspectives for event popularity prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 517–526).

Xiong, J., Yu, L., Zhang, D., & Leng, Y. (2021). DNCP: An attention-based deep learning approach enhanced with attractiveness and timeliness of news for online news click prediction. *Information & Management, 58*(2), Article 103428.

Yang, C., Zhang, H., Jiang, B., & Li, K. (2019). Aspect-based sentiment analysis with alternating coattention networks. *Information Processing & Management, 56*(3), 463–478.

Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6281–6290).

Yuan, K., Liu, G., Wu, J., & Xiong, H. (2020). Dancing with Trump in the stock market: A deep information echoing model. *ACM Transactions on Intelligent Systems and Technology (TIST), 11*(5), 1–22.

Zhang, Q., Wang, J., Huang, H., Huang, X., & Gong, Y. (2017). Hashtag recommendation for multimodal microblog using co-attention network. In *Proceedings of the IJCAI* (pp. 3420–3426).

Zhang, W., Wang, W., Wang, J., & Zha, H. (2018). User-guided hierarchical attention network for multi-modal social image popularity prediction. In *Proceedings of the world wide web conference* (pp. 1277–1286).

Zhang, Z., Yin, Z., Wen, J., Sun, L., Su, S., & Yu, P. (2021). DeepBlue: Bi-layered LSTM for tweet popUlarity Estimation. *IEEE Transactions on Knowledge and Data Engineering*.

Zhao, K., Zhang, P., & Lee, H. M. (2022). Understanding the impacts of user-and marketer-generated content on free digital content consumption. *Decision Support Systems, 154*, Article 113684.