Sentiment Enhanced Multi-Modal Hashtag Recommendation for Micro-Videos

CHAO YANG, (Member, IEEE), XIAOCHAN WANG, AND BIN JIANG, (Member, IEEE)

College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

Corresponding author: Chao Yang (yangchaoedu@hnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702176, and in part by the Natural Science Foundation of Hunan Province under Grant 2017JJ3038.

ABSTRACT

Recommending hashtags for micro-videos is a challenging task due to the following two reasons: 1) micro-video is a unity of multi-modalities, including the visual, acoustic, and textual modalities. Therefore, how to effectively extract features from multi-modalities and utilize them to express the micro-video is of great significance; 2) micro-videos usually include moods and feelings, which may provide crucial cues for recommending proper hashtags. However, most of the existing works have not considered the sentiment of media data for hashtag recommendation. In this paper, the senTiment enhanced multi-mOdal Attentive haShtag recommenDation (TOAST) model is proposed for micro-video hashtag recommendation. Different from previous hashtag recommendation models, which merely consider content features, sentiment features of modalities are further incorporated in TOAST to improve the recommendation performance of the sentiment hashtags (e.g., #funny, #sad). Specifically, the multi-modal content features and the multi-modal sentiment features are modeled by a content common space learning branch based on self-attention and a sentiment common space learning branch, respectively. Furthermore, the varying importance of the multi-modal sentiment and content features are dynamically captured via an attention neural network according to their consistency with the hashtag semantic embedding by an attention neural network. Extensive experiments on a real-world dataset have demonstrated the effectiveness of the proposed method compared with the baseline methods. Meanwhile, the findings from the experiments may provide new insight for future developments of micro-video hashtag recommendation.

INDEX TERMS

Hashtag recommendation, micro-videos, multiple modalities, self-attention mechanism, sentiment features.
important to design a specific model to recommend hashtags for micro-videos.

However, recommending hashtags for micro-videos is a challenging task. On the one hand, modalities of micro-videos usually have distinct distributions and representations, but share a certain of semantic consistency. Therefore, how to effectively leverage the valuable information in the multiple modalities and capture the semantic consistency is a significant yet challenging problem. On the other hand, hashtags are not just labels but have semantic information intrinsically, as they are mainly used to mark topics and emphasize the main content of the videos, as well as share moods and feelings. Therefore, it is significant to effectively explore and leverage the semantic information of hashtags in the hashtag recommendation task. Hashtags can be briefly divided into sentiment hashtags (e.g., #funny, #sad) and content hashtags (e.g., #kid, #dinner and #piecemeal) according to their semantics. The former expresses the sentiment of the users or the people in micro-videos, and the latter reveal entities in micro-videos. Due to the fact that the content features mainly focus on identifying objects in videos, we argue that it is insufficient to predict sentiment hashtags with only content features of micro-videos. For instance, if only considering content features of the two micro-videos in Fig. 1, we could predict the content hashtags (e.g., #dinner, #meal, #eating, #kid), but ignore their sentiment hashtags (e.g., #funny in the first micro-video and #crying in the second micro-video), let alone distinguish the two opposite sentiments. For hashtag recommendation, existing works have studied various kinds of information (e.g., user information [8]–[10], time effect [11] and hashtag co-occurrence [12]), but they have not taken sentiment features into consideration. In the light of this, we propose to additionally incorporate sentiment features of each modality to improve the hashtag recommendation performance.

As manually labeled tags for marking topics and emphasizing target content, hashtags have proven to be useful in many applications, including sentiment analysis [13], content recommendation [14], [15], and even been used as the manual supervision and annotation for training vision models [16]. Hashtag recommendation has gained great attention in

as TOAST, to address the two challenges mentioned above. The TOAST model consists of two branches as the sentiment common space learning branch and the content common space learning branch. Both of the branches take three modalities (i.e., visual, acoustic, and textual modalities) as inputs and learn a common representation subspace to bridge the modality gap. The sentiment common space learning branch is to model the sentiment common space features of the three modalities, and the content common space learning branch takes advantages of self-attention mechanism on the basis of Bi-directional LSTM (Bi-LSTM) to capture the information that are related to the content of the micro-videos. Thereafter, the varying importance of the multi-modal sentiment and content features are dynamically captured by an attention neural network according to their consistency with the semantics of candidate hashtags. Finally, we employ a multi-layer perceptron (MLP) network to predict the interactions between hashtags and micro-videos. Experimental results demonstrate that our model significantly outperforms the baselines.

Our main contributions are summarized as follows:

- We present an integrated framework to perform micro-video hashtag recommendation, which jointly considers the multi-modal sentiment and content information of micro-videos and hashtag embeddings. Our model can effectively capture the correlations between micro-videos and hashtags for hashtag recommendation.
- To the best of our knowledge, this is the first work which attempts to simultaneously leverage sentiment features and content features from multiple modalities to tackle the micro-video hashtag recommendation task.
- Self-attention mechanism with Bi-directional LSTM is introduced in the multi-modal content feature learning to filter out noises and capture information that are most relevant to the corresponding hashtags.
- We evaluate our proposed model on the real-world dataset collected from Vine. Extensive experiments have demonstrated the effectiveness of our proposed model. As a side contribution, the data and the codes are released.

The rest of the paper is organized as follows. In section II, we briefly review the related work. Section III and IV detail the feature extraction of each modality and our proposed TOAST model, respectively. The experiment results are presented in section V. Finally, section VI concludes the work and points out the future directions.

**II. RELATED WORK**

**A. HASHTAG RECOMMENDATION**

As manually labeled tags for marking topics and emphasizing target content, hashtags have proven to be useful in many applications, including sentiment analysis [13], content recommendation [14], [15], and even been used as the manual supervision and annotation for training vision models [16]. Hashtag recommendation has gained great attention in

6https://microvh.wixsite.com/toast
the field of text and image from different perspectives. Ding et al. [3] propose a topical translation model from content to hashtags for microblog hashtag suggestion. Lu and Lee [17] propose a hashtag recommendation method that captures the temporal clustering effect of latent topics in tweets. Kowald et al. [11] incorporate the time effect on hashtag reusing to build a hashtag recommendation algorithm called \(\text{BLiS}\).

Meanwhile, the neural networks have also shown their superiorities in hashtag recommendation tasks. For example, Denton et al. [8] adopt convolutional neural network to automatically extract features from images and then perform hashtag recommendation with user metadata. Wang et al. [6] make the first endeavour to annotate hashtags with a novel sequence generation framework via viewing hashtags as short sequences of words. To recommend hashtags for multi-modal microblogs, a topical translation model incorporating textual and visual information is proposed in [4]. Since it is hard to predict some hashtags when merely leveraging textual information, Zhang et al. [5] propose a co-attention network to recommend hashtags for multimodal tweets with both textual and visual information. Recently, some other works [7], [10] based upon co-attention mechanism have also make great progress in this task. For micro-video hashtag recommendation, [9] firstly proposes a Graph Convolution Network (GCN) based Personalized Hashtag Recommendation model, which can comprehensively model the interactions between users, hashtags, and micro-videos. And [18] incorporates the GCN and LSTMs to address the hashtag long-tail phenomenon. However, most of existing works overlook the hashtag category information and sentiment analysis when recommending hashtags. In the light of this, we propose to incorporate sentiment features of modalities and hashtag semantics to enhance the hashtag recommendation performance.

**B. MULTI-MODALITY REPRESENTATION LEARNING**

By virtue of the capability to learn more expressive information, utilizing multiple modalities has attracted a surge of interest and achieved good performance in various fields such as question answering [19]–[21], information retrieval [22], [23], location prediction [24], [25], sentiment analysis [26] and video description [27]. Generally, the prior efforts on multi-modality representation learning can be briefly divided into three categories, namely, collaborative training, multiple kernel learning and subspace learning. With the success of deep neural network in single-modal scenarios, efforts on multi-modality representation learning incorporating deep neural network have also made great progress. In [22], the deep Bolzmann machine is used to learn a joint representation over multimodal inputs. To explore the correlation between modalities, Gao et al. [21] propose a motion-appearance co-memory network to jointly model the motion and appearance information. Recently, RNN and CNN are widely used in multi-modal tasks. Gao et al. [28] propose an end-to-end framework which integrates attention mechanism with LSTM to capture salient structures of video, and explores the correlation between multimodal representations. Liu et al. [29] present a learning model with three parallel LSTMs to capture the common information of visual, acoustic and textual modality in a common subspace for micro-video categorization. Based on GRU and CNN, [30] simultaneously utilizes multi-modal features by a fusion strategy for cross-modal video-text retrieval. Furthermore, generative adversarial networks have also been applied to perform cross-modal common representation learning by generating one modality to another through adversarial training process [31], [32]. Inspired by the success of multi-modality fusion, our proposed TOAST model performs hashtag recommendation by jointly considering visual, acoustic and textual modalities of micro-videos. Furthermore, both sentiment and content information of these three modalities are exploited simultaneously.

**C. MULTI-MODAL SENTIMENT ANALYSIS**

Sentiment analysis is the task of automatically classifying the sentiment polarity of a given text, phrase or sentence [33]. Due to the surge of online social media, sentiment analysis has been introduced as a tool for automatically extracting sentiment from the user-generated data. Since the work [34] in 2011, multi-modal sentiment analysis (MSA) has emerged as an important area of research which goes beyond the traditional text-based analysis. MSA integrates two or more modalities to improve the performance of the user sentiment detection. And fuzzy logic is also considered to deal with the mixed and complex emotions [35], [36]. A detailed introduction of MSA is presented in [37].

With the success of deep leaning, neural networks have been widely used as the modality feature extractors to replace manual feature engineering in various tasks, including MSA. In [38], visual information extracted by VGG-16 is utilized as alignment for pointing out the important sentences of a document. Poria et al. [39] utilize text-CNN to extract features from the textual modality and 3D-CNN to obtain visual features from the video, and then employ LSTM to capture contextual information. In [40], Object-VGG and Scene-VGG Model are employed to detect visual semantic features, and GloVe is employed as the textual feature detector. The features mentioned above can be generally summarized as content information. However, the content information of modalities is somewhat obscure for understanding user’s sentiment and thus needs complicated subsequent model for MSA task. Directly utilizing sentiment features could be a easier way to capture the simplified sentiment polarity and has shown promising performance in other tasks. In [41], sentiment features of images are extracted and fed into Logistic Regression to classify the sentiment of Microblogs. [42] utilizes sentiment features of text and a linear SVM classifier for sentiment polarity analysis. In addition to the classical computer vision features, sentiment features are exploited to classify the induced sentiment of a video for affective video analysis in the work [43],
And Chen et al. [51] extract sentiment features for micro-video popularity prediction, as the sentiment of UGCs has been proven to be strongly correlated with their popularity. To improve the recommendation performance of the sentiment hashtags, sentiment features of multi-modalities are incorporated in our model for exploring the sentiment of micro-videos, and further combined with the content features to perform hashtag recommendation.

III. FEATURE EXTRACTION

For each modality, we respectively extract two kinds of features of each sequential unit (i.e., frame, audio clip and word), namely, content features and sentiment features.

A. FEATURES IN VISUAL MODALITY

The maximum length of micro-videos in our dataset crawled from Vine is set as 6 seconds [29]. The short-length characteristic enables us to employ a few key frames to represent the micro-video. Therefore, we extract 12 key frames of each micro-video by FFmpeg\(^7\) with equal intervals.

ResNet-152, which is pre-trained on ImageNet,\(^8\) is utilized to extract visual content features. Finally, each frame is represented as a 2048 dimensional feature vector.

We extract additional sentiment features of each key frame through the deep CNN model trained on the SentiBank dataset [44]. As the ANPs with the same adjective reveal almost the same emotion, we summarize the output probabilities of the ANPs according to their adjective to reduce the dimensionality. As a result, we obtain a 231 dimensional sentiment feature vector for each frame.

B. FEATURES IN ACOUSTIC MODALITY

To extract features from acoustic modality, we first use FFmpeg to detach the audio channel from video and then segment it into 6 equal-length clips.

We utilize the SoundNet CNN [45] to obtain a 1024 dimensional content feature vector from each audio clip. The content features of acoustic modality are used to model the content common space.

We then employ Librosa\(^9\) to extract widely-used acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCC) [46], Zero Crossing Rate [47], etc. Finally, we obtain a 512 dimensional feature vector from each audio clip. Since the features obtained by Librosa are generally used in many sentiment-related tasks [39], [48], we employ this kind of features to learn the sentiment common space in this work.

C. FEATURES IN TEXTUAL MODALITY

Textual modality has proved its usefulness for hashtag recommendation in previous works [4], [49]. For textual descriptions, we first eliminate non-English characters, followed by removing the stop words. We then extract 300 dimensional features for each word via the pre-trained Glove model.\(^10\)

With the help of the Sentiment Analysis tool in Stanford CoreNLP,\(^11\) we embed each text into a 5 dimensional vector, which corresponds to the probability of being very negative, negative, neutral, positive and very positive, respectively.

All the features of the three modalities that we extract are summarized in Table 1.

| Modalities     | Extracted features                                      |
|----------------|---------------------------------------------------------|
| Visual         | 231-D Sentibank, 2,048-D CNN                           |
| Acoustic       | 512 Librosa, 1,024-D SoundNet                          |
| Textual        | 5-D CoreNLP, 300-D GloVe                               |

D. HASHTAG EMBEDDING

Embedding layer is the general way to automatically learn the hashtag semantic embedding in deep-learning models, but we argue that the information learned in this way is limited. Therefore, we propose to transfer the knowledge learned in the pre-trained GloVe model to compensate the hashtag id embedding, so as to provide stronger semantics. Inspired by [50], hashtag segmentation is conducted in our model. Firstly, we devise an algorithm to automatically segment each unspaced hashtag into isolated words which are already properly embedded in GloVe model. Afterwards, the average pooling is employed on all these embeddings of the segmented words. Taking the hashtag “#britishvinefamily” as an example, we segment it into “british”, “vine”, and “family”, and take an average of embeddings of the three words as the semantic representation for “#britishvinefamily”. Consequently, each hashtag is represented by a 600 dimensional vector, which is the combination of the average pooling vector and the hashtag id embedding vector by concatenation operation.

IV. OUR PROPOSED MODEL

The overall architecture of our proposed TOAST model is shown in Fig. 2. As illustrated, the model comprises of three components: 1) the sentiment common space learning network, which maps the sentiment features extracted from visual, acoustic and textual modalities into a shared common space with the same length; 2) self-attentive content common space learning network, which maps the sentiment features extracted from visual, acoustic and textual modalities into a shared common space with the same length; and 3) hashtag prediction network which estimates the relevance scores between the hashtags and the attention weighted sum representations of micro-videos.

A. NOTATIONS AND PROBLEM FORMULATION

For notations, we use bold capital letters (e.g., \(X\)) and bold lowercase letters (e.g., \(x\)) to represent matrices and vectors, respectively. We employ non-bold letters (e.g., \(x\)) to represent scalars, squiggly letters (e.g., \(\chi\)) to denote sets, and Greek letters (e.g., \(\lambda\)) to represent parameters.

\(^1\)https://www.ffmpeg.org/
\(^2\)http://www.image-net.org/
\(^3\)https://github.com/librosa/librosa
\(^4\)https://nlp.stanford.edu/projects/glove/

\(^7\)https://www.ffmpeg.org/
\(^8\)http://www.image-net.org/
\(^9\)https://github.com/librosa/librosa
\(^10\)https://nlp.stanford.edu/projects/glove/
\(^11\)https://stanfordnlp.github.io/CoreNLP/
Formally, suppose we have $I$ micro-videos and $J$ hashtags, denoted as $\mathcal{X} = \{x_i\}_{i=1}^I$, $\mathcal{Y} = \{y_j\}_{j=1}^J$. The task of our research is to automatically recommend a hashtag list of $\mathcal{Y}$ for a given micro-video that the video may be relevant to. Towards this end, for each micro-video $x_i \in \mathcal{X}$, we pre-segment it into three modalities (i.e., $x_i = \{x_i^v, x_i^a, x_i^t\}$), where the superscripts $v$, $a$, and $t$ represent the visual, acoustic, and textual modalities. Thereafter, for each modality $m \in \mathcal{M} = \{v, a, t\}$, we extract two kinds of features, namely, sentiment features and content features. In this work, we use $e_{m(s)}^{i} \in \mathbb{R}^{D_{m(s)}}$ to denote the $D_{m(s)}$ dimensional sentiment feature vector and $e_{m(c)}^{i} \in \mathbb{R}^{D_{m(c)}}$ to denote the $D_{m(c)}$ dimensional content feature vector of $i$-th micro-video from the $m$-th modality, respectively.

B. SENTIMENT COMMON SPACE LEARNING
Common space learning has been proved to be an effective way in dealing with data containing multiple modalities [24], [29] due to its ability in alleviating the fusion and disagreement problems. In order to better predict those sentiment hashtags (e.g., #funny, #bored, and some other phrase-like hashtags which contains sentiment words, we briefly categorize the rest hashtags as content hashtags. Naturally, content hashtags are much more complicated than sentiment hashtags and thus need a more comprehensive and accurate understanding of the key information of the video. Therefore, we propose a self-attentive sequential feature learning method to model the crucial information of videos and then improve the recommendation performance over content hashtags.

The framework of our self-attentive sequential feature learning is illustrated in Fig. 3. It can be briefly divided into two learning stages: in the first stage, the content features

\begin{align}
\tilde{h}_i^{v(s)} &= f_{\text{map}}^v(e_{i}^{v(s)}) \\
\tilde{h}_i^{a(s)} &= f_{\text{map}}^a(e_{i}^{a(s)}) \\
\tilde{h}_i^{t(s)} &= f_{\text{map}}^t(e_{i}^{t(s)}),
\end{align}

where $f_{\text{map}}^v$, $f_{\text{map}}^a$, $f_{\text{map}}^t$ are three mapping functions which can be learned by the MLPs, and $\tilde{h}_i^{v(s)}$, $\tilde{h}_i^{a(s)}$, $\tilde{h}_i^{t(s)} \in \mathbb{R}^s$ are the embedded feature vectors in the common sentiment space of visual, acoustic, textual modalities. $\mathbb{R}^s$ is the dimension of the sentiment common space.

C. CONTENT COMMON SPACE LEARNING
Apart from single-word sentiment hashtags, such as "#funny", "#bored", and some other phrase-like hashtags which contains sentiment words, we briefly categorize the rest hashtags as content hashtags. Naturally, content hashtags are much more complicated than sentiment hashtags and thus need a more comprehensive and accurate understanding of the key information of the video. Therefore, we propose a self-attentive sequential feature learning method to model the crucial information of videos and then improve the recommendation performance over content hashtags.

The framework of our self-attentive sequential feature learning is illustrated in Fig. 3. It can be briefly divided into two learning stages: in the first stage, the content features
extracted from each modality are sequentially fed into a Bi-LSTM neural network to capture the sequential cues. It is worth noting that the outputs of the three parallel Bi-LSTMs are set as equal dimension to directly map the content feature vectors of the three modalities into a common space at the same time. In the second learning stage, self-attention mechanism is employed to produce a weight vector for the embeddings of all units.

1) PARALLEL BI-LSTM

For a given multi-modal micro-video \( x_i = \{x_i^v, x_i^a, x_i^t\} \), each modality is pre-processed into a sequence of units \( \{x_i^{v(1)}, x_i^{v(2)}, x_i^{v(3)}, \ldots, x_i^{v(T)}\} \), \( T \in \{T_v, T_a, T_t\} \), where \( T_v, T_a, T_t \) respectively represent the number of units in visual, acoustic and textual modality. We then obtain the corresponding content feature vector \( e_i^{m(c)} \) for each unit \( x_i^m \), where \( x_i^m \) denotes the \( t \)-th unit of the \( m \)-th modality obtained from the \( t \)-th micro-video. The content features extracted from units of each modality can be treated as sequential data. Inspired by [29], three parallel Bi-LSTMs are employed to capture the sequential structure in the three modalities. Each Bi-LSTM takes a sequence of feature vectors \( \{e_i^{v(1)}, e_i^{v(2)}, e_i^{v(3)}, \ldots, e_i^{v(T)}\} \) as input and outputs a forward hidden state \( h_i^{c(v)} \) and a backward hidden state \( h_i^{c(b)} \) for the \( t \)-th unit of the \( m \)-th modality. We then concatenate \( h_i^{c(v)} \) and \( h_i^{c(b)} \) to obtain a hidden state \( h_i^{c(m)} \), which contains the global information from the \( t \)-th unit and the context unit before and after \( t \), to benefit the comprehensive understanding of the micro-video.

For the \( t \)-th unit, the aforementioned process can be formulated as follows:

\[
\begin{align*}
\hat{h}_t^{c(v)} &= f_{\text{extr}}(x_t^v) \\
\hat{h}_t^{c(b)} &= \text{LSTM}_b(e_t^{c(v)}, h_{t-1}^{c(b)}) \\
\hat{h}_t^{c(m)} &= \text{LSTM}_m(e_t^{c(v)}, h_{t-1}^{c(m)}) \\
\end{align*}
\]

where “\( f_{\text{extr}} \)” means the different content feature extracting process of each modality.

2) SELF-ATTENTIVE SEQUENTIAL FEATURE LEARNING

After encoding by a Bi-LSTM, we obtain a sequence of feature vectors \( \{h_i^{c(v)}, h_i^{c(a)}, h_i^{c(t)}\} \) from the hidden states of each modality. Hashtags are usually used to describe topics or highlight key information of videos, that is, hashtags pay more attention to the crucial information rather than all scenes in the micro-videos. Therefore, how to identify the importance of each sequential unit, and eliminate the useless even noisy features is of great significance in the hashtag recommendation task.

Motivated by the success of attention mechanism [53]–[55], a self-attentive pooling is adopted to explicitly capture the varying importance of each sequential unit by assigning it an attention weight. For each modality, the self-attentive pooling takes the whole output sequence \( h_i^{m(c)}, h_i^{m(c)}, h_i^{m(c)}, \ldots, h_i^{m(c)} \) of the Bi-LSTM as input and uses a neural attention network to produce a weight vector. Formally, the weight of the \( t \)-th unit of the \( m \)-th modality for the \( i \)-th micro-video is calculated as:

\[
\begin{align*}
\tilde{p}_{i(t)}^{m(c)} &= \frac{\exp(\alpha_t^{m(c)})}{\sum_{t=1}^{T_m} \exp(\alpha_t^{m(c)})} \\
\end{align*}
\]

where \( \tilde{p}_{i(t)}^{m(c)} \) is the parameter matrix, \( w_m \) and \( b_m \) is parameter vectors. We employ ReLU as the activation function for the hidden layer and softmax to normalize \( \tilde{p}_{i(t)}^{m(c)} \) into an attention score \( p_{i(t)}^{m(c)} \) (i.e., \( \sum_{t=1}^{T_m} p_{i(t)}^{m(c)} = 1 \)).

Then the \( m \)-th modality is represented by a weighted sum of the vector representations of all time steps:

\[
\hat{h}_i^{c(m)} = \frac{1}{T_m} \sum_{t=1}^{T_m} p_{i(t)}^{m(c)} h_{i(t)}^{c(m)}. 
\]

D. HASHTAG PREDICTION

Thus far, for a given video-hashtag pair \((x_i, y_j)\), we have obtained the sentiment common space feature vectors (i.e., \( \tilde{h}_i^{c(v)}, \tilde{h}_i^{c(a)}, \tilde{h}_i^{c(t)} \) and the content common space feature vectors (i.e., \( \tilde{h}_i^{c(c)}, \tilde{h}_i^{c(a)}, \tilde{h}_i^{c(t)} \)), as well as the hashtag semantic representation \( h_i \). We then respectively concatenate the sentiment and content feature vectors of each modality and obtain the multi-modal sentiment features \( x_i^{s} = \tilde{h}_i^{c(v)} \oplus \tilde{h}_i^{c(a)} \oplus \tilde{h}_i^{c(t)} \) and the multi-modal content features \( x_i^{c} = \tilde{h}_i^{c(c)} \oplus \tilde{h}_i^{c(a)} \oplus \tilde{h}_i^{c(t)} \). \oplus denotes the vector concatenation operation.

As mentioned earlier, hashtags can be semantically divided into two categories as sentiment hashtags and content hashtags. The semantic category determines that the candidate hashtags have different attention for sentiment features and content features. Therefore, we propose to assign attentive weights to the two kinds of multi-modal features (i.e., the multi-modal sentiment features \( h_i^{s} \) and the multi-modal content features \( h_i^{c} \) ) of the given micro-video \( x_i \), according to their consistencies with the semantic embedding \( h \) of the candidate hashtag \( y_j \).

\[
\begin{align*}
\alpha_t^{s} &= \frac{1}{\sum_{t=1}^{T_m} \exp(\alpha_t^{s})} \exp(\alpha_t^{s}) \\
\alpha_t^{c} &= \frac{1}{\sum_{t=1}^{T_m} \exp(\alpha_t^{c})} \exp(\alpha_t^{c}) \\
\end{align*}
\]

where \( \mathbf{W}, \mathbf{w} \) and \( \mathbf{b} \) are learnable parameters. \( \alpha_t^{s} \) and \( \alpha_t^{c} \) are the obtained weights of \( x_i^{s} \) and \( x_i^{c} \), respectively. To compute the attention scores, \( x_i^{s} \) is concatenated by zero vector to \( \tilde{x}_i^{s} \), which is of the same size as \( x_i^{c} \).
Thereafter, we derive a cross-modal representation \( \mathbf{x}_{i,j} \) for the current video-hashtag pair by employing the weighted concatenation as follows:

\[
\mathbf{x}_{i,j} = \alpha^i_i \mathbf{x}^i_i \odot \alpha^j_j \mathbf{x}^j_j \ominus \mathbf{h},
\]  

(6)

To estimate the relevance score of the given micro-video and hashtag, we then feed \( \mathbf{x}_{i,j} \) into a MLP, where the complicated interactions and correlations among the embedding of the micro-video and hashtag can be well captured via the strong representation power of non-linear hidden layers. The hidden layers of MLP are defined as:

\[
\begin{align*}
\mathbf{o}_1 &= \text{ReLU}(\mathbf{W}_1 \mathbf{x}_{i,j} + \mathbf{b}_1) \\
\mathbf{o}_2 &= \text{ReLU}(\mathbf{W}_2 \mathbf{o}_1 + \mathbf{b}_2) \\
&\vdots \\
\hat{s}_{i,j} &= \text{sigmoid}(\mathbf{W}_{l+1} \mathbf{o}_l),
\end{align*}
\]  

(7)

where \( \mathbf{W}_1, \mathbf{b}_1, \mathbf{o}_1 \), and \( \text{ReLU} / \hat{s}_{i,j} \) denote the weight matrix, bias vector, output vector of \( l \)-th hidden layer and the adopted activation function, respectively. \( \hat{s}_{i,j} \) is the relevance score of the given video-hashtag pair \( (x_i, y_j) \), and \( \mathbf{W}_{l+1} \) is the weight matrix for transforming the output vector of \( l \)-th hidden layer into a number. Finally, for a micro-video, we can recommend the top-n hashtags according to the score ranking of its candidate hashtag set, which can be obtained by random sampling or other sampling strategies.

In our research, the hashtag recommendation task is specified as a binary classification problem as [18]. In detail, if the given video-hashtag pair \( (x_i, y_j) \) is observed in our dataset, the target score \( s_{i,j} \) will be assigned the value 1, and otherwise 0. We optimize the cross entropy loss to force the prediction score \( \hat{s}_{i,j} \) to be close to \( s_{i,j} \) as follows:

\[
L = - \sum_{(i,j) \in S} s_{i,j} \log \hat{s}_{i,j} + (1 - s_{i,j}) \log(1 - \hat{s}_{i,j}),
\]  

(8)

where \( S \) denotes the set of all the positive and negative instances.

V. EXPERIMENTS

In this section, to thoroughly verify the effectiveness of our proposed TOAST model, we carry out extensive experiments to answer the following four research questions:

- **RQ1** Can our proposed TOAST approach outperform other state-of-the-art competitors? Do the proposed self-attention mechanism and hashtag embedding as well as the multi-modal sentiment and content feature importance capture contribute to our model performance?

- **RQ2** Is each modality equally important? And how does TOAST perform under different modality combinations?

- **RQ3** Do the additional sentiment features of each modality help to improve the hashtag recommendation performance?

- **RQ4** Do the sentiment features and content features have their specialties in hashtag recommendation?

A. EXPERIMENTAL SETTINGS

1) DATASET

In experiment, we rearrange data from a public micro-video dataset released by Chen et al. [51] to perform the hashtag recommendation. This original dataset, which contains the URL of user-generated micro-videos in Vine with corresponding text and social information, is used for micro-video popularity prediction. We process the dataset by retaining micro-videos with 3 modalities (i.e., the visual, acoustic, and textual modalities) and hashtag information after data cleaning. Furthermore, we eliminate hashtags which occurred less than 10 times in our dataset. Finally, we obtain 40049 available micro-videos and 1935 different hashtags.

We divide our dataset into three parts, with around 80% (i.e., 32000 videos), 10% (i.e., 4000 videos), and 10% (i.e., 4049 videos) randomly select micro-videos and their corresponding hashtags for training, validation and testing, respectively. We use the training set to adjust the hyper parameters and then leverage validation set to verify the performance of our model. The testing set is only used for testing the final solution to confirm the actual predictive power of our model with optimal parameters.

2) EVALUATION PROTOCOLS

Following the evaluation method adopted in [52], we randomly pair a positive instance (i.e., an observed video-hashtag pair) in the testing with 100 negative hashtags that have never been marked to the micro-video before as its candidate hashtag set. Then each method outputs prediction scores for these 101 hashtags. The widely used metrics — Recall and NDCG are employed to evaluate the performance of the Top-N hashtag recommendation list. The Recall measures whether the testing item is presented in the Top-N list, while NDCG accounts for the position of the hit by assigning higher scores to hits at top positions.

3) IMPLEMENTATION DETAILS

We implement our method based on PyTorch. Adam optimizer is used for all gradient-based methods, where the mini-batch size and learning rate are searched in \([64, 128, 256, 512]\) and \([0.0001, 0.0005, 0.001, 0.005, 0.01]\), respectively. The number (i.e., \( \lambda \)) of randomly sampled negative hashtags for each micro-video and the dimension of sentiment common space and content common space are all hyper parameters in our work. We conduct experiments with \( \lambda \) and the dimension of sentiment common space and content common space searched in \([1, 2, 4, 6, 8], [30, 50, 100, 150]\) and \([100, 200, 300, 400]\), respectively. The experiments results reveal that the model obtains best results when \( \lambda = 6 \) and the dimensions of sentiment common space and content common space equal 50 and 300. To avoid overfitting and gradient vanishing, dropout and Batch Normalization are employed. We repeat each setting for 5 times and report the average results.

---

12 https://pytorch.org/
4) BASELINES

To demonstrate the effectiveness of our framework, we compare it with the following methods.

- **RSDAE** [56]: This is a textual modality based hashtag recommendation algorithm, which uses stacked denoising autoencoders to perform deep representation learning and an EM-style algorithm for relational learning under a probabilistic framework. We adopt the released implementation\(^\text{13}\) and modify its evaluation scheme to adapt to our testing scenario.

- **Co-Attention** [5]: This is the state of the art hashtag recommendation algorithm which incorporates textual and visual information to recommend hashtags for multimodal tweets. Finally, the hashtags are predicted by a single-layer softmax classifier. We introduce co-attention mechanism in learning the sequential content features of the micro-video textual and visual modalities with Bi-LSTMs, and modify its evaluation strategy.

- **TMALL** [51]: This is a transductive multi-modal learning model, which is designed to find the optimal latent common space to predict the popularity of the micro-videos. We adopt the core idea of the model by directly mapping the sentiment and content feature vectors of the three modalities into the optimal common space. We then cast the representations of multiple modalities into a softmax classifier to perform the hashtag recommendation.

- **EASTERN** [29]: This work employs three parallel LSTMs to capture the sequential structure and adopts the last state of each LSTM (i.e., last pooling) as the final embedding for each modality. Finally, a convolutional neural network is employed to learn the sparse concept level representations of micro-videos to perform micro-video categorization. We replace the last convolutional neural network with a softmax classifier and utilize it as one of our baselines. For a fair comparison, both sentiment and content features are used in TMALL and EASTERN.

- **TOAST-L, TOAST-A**: They are two variants of our TOAST method by replacing the self-attention mechanism with last pooling (TOAST-L) and average pooling (TOAST-A). They are implemented to demonstrate the effect of our proposed self-attentive sequential feature pooling.

- **TOAST-H**: Instead of using hashtag segmentation to pre-obtain the semantic embedding for hashtags as the additional knowledge, we just employ the embedding layer to automatically learn the hashtag id embedding from scratch.

- **TOAST-D**: In this model, the multi-modal sentiment and content features are directly concatenated with the hashtag semantic embedding into the cross-modal representation, while their varying importance are ignored.

\(^{13}\)http://www.wanghao.in/publication.html

| Methods   | K=5  | NDCG | K=10 | NDCG@ |
|-----------|------|------|------|-------|
| RSDAE     | 0.4029 | 0.3128 | 0.4836 | 0.3389 |
| Co-Attention | 0.5100 | 0.4062 | 0.6093 | 0.4382 |
| TMALL     | 0.5946 | 0.4872 | 0.6826 | 0.5167 |
| EASTERN   | 0.5981 | 0.4887 | 0.6970 | 0.5212 |
| TOAST-L   | 0.6289 | 0.5194 | 0.7298 | 0.5473 |
| TOAST-A   | 0.6218 | 0.5175 | 0.7275 | 0.5486 |
| TOAST-H   | 0.6432 | 0.5264 | 0.7329 | 0.5568 |
| TOAST-D   | 0.6449 | 0.5286 | 0.7324 | 0.5581 |
| TOAST     | 0.6481* | 0.5317* | 0.7360* | 0.5618* |

B. OVERALL PERFORMANCE COMPARISON (RQ1)

The performance comparison among all the methods are summarized in Table 2. It can be seen that:

- Our proposed TOAST model achieves the best result as shown in Table 2, substantially surpassing all the baselines by a significant margin. As compared with the state-of-the-art hashtag recommendation model (Co-Attention), TOAST achieves increases of 12.67% in Recall (k=10) and 12.36% in NDCG (k=10). This is mainly because TOAST considers information of all the three modalities and hashtag semantic simultaneously.

- RSDAE achieves worst performance since it only considers textual modality for hashtag recommendation but ignores the information from visual and acoustic modalities, which are two significant aspects to convey the main idea of micro-videos.

- When simultaneously utilizing three modalities with sentiment and content features, TOAST achieves much better performance than both TMALL and EASTERN. This verifies the effectiveness of the structure of TOAST in exploring the correlation between micro-videos and hashtags.

- While TMALL considers all the three modalities, it directly averages the feature vectors of all the units and then employs MLPs to map the averaged feature vectors of each modality into a common space to conduct modality fusion. It hence fails to capture sequential cues, which are exploited in EASTERN via the LSTM neural network. As a result, TMALL performs poorly than EASTERN. The improvement of EASTERN verifies the necessity of the sequential feature learning of micro-videos.

- It can be seen that TOAST shows consistent improvements over TOAST-L, TOAST-A, TOAST-H and TOAST-D. This observation demonstrates the effectiveness of our proposed self-attentive sequential feature learning and the hashtag semantic learning, as well as the necessity of capturing the varying significance of the multi-modal sentiment features and content features. This is because changing self-attentive pooling into last pooling (i.e., TOAST-L) or average pooling
(i.e., TOAST-A) overlooks the different importance of each modality unit, and is unable to eliminate the irrelevant features.

**TABLE 3. Performance of various modality combinations.** * indicates that the improvements over all other methods are statistically significant for \( p < 0.05.\)

| Methods | \(K=5\) | \(K=10\) |
|---------|---------|---------|
|         | Recall  | NDCG    | Recall  | NDCG    |
| Visual-(both) | 0.5487  | 0.4294  | 0.6475  | 0.4789  |
| Textual-(both) | 0.4397  | 0.4098  | 0.6488  | 0.4672  |
| Acoustic-(both) | 0.4286  | 0.3823  | 0.6339  | 0.4590  |
| Visual+Acoustic-(both) | 0.5665  | 0.4688  | 0.6762  | 0.4827  |
| Visual+Textual-(both) | 0.6013  | 0.4871  | 0.6985  | 0.5171  |
| Visual+Acoustic+Textual-(both) | 0.6296  | 0.5048  | 0.7260  | 0.5357  |
| All-(sentiment) | 0.6088  | 0.5014  | 0.7029  | 0.5279  |
| All-(content) | 0.6315  | 0.5216  | 0.7153  | 0.5487  |
| All-(both)  | 0.6481* | 0.5317* | 0.7360* | 0.5618* |

**FIGURE 4.** Performance comparison of models with (i.e, content+sentiment) or without (i.e, content only) sentiment features.

**D. EFFECTIVENESS OF MULTI-MODAL SENTIMENT FEATURES (RQ3)**

We conduct an empirical study to investigate whether the sentiment features do benefit our hashtag recommendation. The comparison results are shown in Fig. 4. It can be seen that:

- From Fig. 4, we can obviously observe that all the models show consistent and substantial improvements after incorporating sentiment features. It is consistent with our expectation and further demonstrates that the sentiment features are effective and essential for hashtag recommendation.

- Compared with other variants of TOAST, TMALL and EASTERN observe greater improvements after taking sentiment features into consideration. And all the variants of TOAST performs better than TMALL and EASTERN when only content features are incorporated. This indicates that incorporating sentiment features is particularly beneficial to strengthen the performance, especially when the model performs poorly in understanding the content features of the given micro-videos.

- When combining both sentiment and content features, TOAST-L, TOAST-A, TOAST-H and TOAST-D all achieve better performance than merely utilizing content features. Note that, TOAST and TOAST-D will become the same model in the single content features senario, so we give the same results in Fig. 4 when only content features are considered. Meanwhile, we can find that TOAST-H obtains the largest improvement among all the variant models. It is probably because the automatically learned hashtag semantics by the embedding layer is insufficient. Therefore, it is hard for the model to comprehend the hashtags and then distinguish the interactions between hashtags and micro-videos without enough information (i.e., ignoring the sentiment features).
FIGURE 5. Examples of 6 testing micro-videos in our real-world dataset and the top 10 recommended hashtags when only using sentiment features (i.e., TOAST-sentiment), content features (i.e., TOAST-content), and both types of features (i.e., TOAST-both). Ground Truth(GT) shows the user-generated hashtags. The different colors are just for a easy understanding of the analysis of the qualitative results.
E. QUALITATIVE RESULTS (RQ4)

To understand the specialities of the sentiment features and the content features, as well as their combination, we further display the top 10 recommended hashtags for several micro-videos in Fig. 5. Specifically, each micro-video is displayed with four key frames and its corresponding user-generated text. The acoustic modality is omitted for simplicity. In addition, we also report the ground-truth hashtags (GT) of each micro-video. According to the examples, we have the following observations:

- In most of the cases (e.g., micro-video 1, micro-video 2 and micro-video 4), the sentiment features and content features recommend significantly different hashtags. For micro-video 4, TOAST-content mainly recommend some object-related hashtags (e.g., #sports, #basketball, #football) while TOAST-sentiment and TOAST-both provide more correct sentiment-related hashtags (e.g., #depression, #depressed, #anxiety). Although some sentiment hashtags (e.g., #comedy, #funny, #lol) are given by TOAST-content for micro-video 4, these hashtags are completely contrary to its real emotion (i.e., very sad and negative). This shows that only utilizing content features is insufficient to predict reasonable sentiment hashtag sometimes. For micro-video 1, micro-video 2 and micro-video 6, utilizing both sentiment features and content features (i.e., TOAST-both) achieves better performance, indicating the necessity to integrate sentiment features and content features for micro-video hashtag recommendation.

- The acoustic modality in micro-video 1 and micro-video 2 are a bit similar to each other (i.e., both contain the sound of the guitar and singing), so it is reasonable for TOAST-both to recommend the same hashtag “#music” for both of them. However, the hashtags recommended by TOAST-content are distinct. This is mainly because the content features in visual modality of the two videos are quite different. For micro-video 1, TOAST-both mainly focuses on the sound of singing and the moving mouth (i.e., recommending #lipsyning, #lipsyncing), while paying more attention to the animal headgear in micro-video 2 (i.e., recommending #furry, #fursuiter, and #spongebobsquarepants).

- For micro-video 3, TOAST-sentiment and TOAST-content recommend very similar content hashtags (e.g., #soccer, #football, #sports). It is probably because micro-video 3 is almost silent, and its available visual and textual modalities do not convey obvious emotions. Meanwhile, apart from predicting simple hashtags (e.g., #soccer, #football, #sports), TOAST-both is capable of recommending complicated abstract hashtags (e.g., #realmadrid, #messi and #beastmode).

- For micro-video 4, TOAST-content regards the detected dot and empty image as a ball and sports field. Therefore, TOAST-content recommends many sport-related hashtags (e.g., #sports, #basketball, #thezone). This is consistent with micro-video 3.

- Micro-video 5 and micro-video 6 show two of our failure cases, where TOAST-both achieves unsatisfactory results as compared to GT. The ground-truth hashtags of micro-video 5 are too complicated so that the information of multi-modalities is insufficient to capture the correlation of these hashtags. For micro-video 6, it is probably because its information is mainly transmitted by visual modality, which is too abstract for the model to understand.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present a novel attentive multi-modal model (TOAST) to perform hashtag recommendation for micro-videos and creatively aggregate sentiment features and content features to enhance the performance. In particular, TOAST works by learning separate optimal common space for sentiment and content features from multi-modalities, which is devised to exploit the latent correlation among heterogeneous modalities. We employ three MLPs to directly learn the sentiment common space of the three modalities (i.e., visual, acoustic and textual modalities). While learning the content common space, a Bi-LSTM is employed to capture the sequential information and self-attention mechanism is adopted to adaptively identify the crucial units in each modality. Finally, the sentiment features and content features as well as the semantic embeddings of hashtags are integrated by a weighted concatenation to figure out the correlation between hashtags and the micro-videos. Extensive experiments on our rearranged Vine dataset have validated the effectiveness of our proposed model.

In the future, we plan to extend our work from the following directions: 1) We will capture more precise sentiment information of each modality, such as take the emoji in textual modality into consideration to obtain more accurate sentiment information for micro-videos; 2) We will strive to explore more effective interactions between micro-video sentiment information, content information, and hashtag semantic information, along with user information.

REFERENCES

[1] P. Heymann, D. Ramage, and H. Garcia-Molina, “Social tag prediction,” in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR), Singapore, Jul. 2008, pp. 531–538.
[2] S. M. Kywe, T.-A. Hoang, E.-P. Lim, and F. Zhu, “On recommending hashtags in Twitter networks,” in Proc. SocInfo, Lausanne, Switzerland, Dec. 2012, pp. 337–350.
[3] Z. Ding, X. Qiu, Q. Zhang, and X. Huang, “Learning topical translation model for microblog hashtag suggestion,” in Proc. IJCAI, Beijing, China, Aug. 2013, pp. 2078–2084.
[4] Y. Gong, Q. Zhang, and X. Huang, “Hashtag recommendation for multimodal microblog posts,” Neurocomputing, vol. 272, pp. 170–177, Jan. 2018.
[5] Q. Zhang, J. Wang, H. Huang, X. Huang, and Y. Gong, “Hashtag recommendation for multimodal microblog using co-attention network,” in Proc. 26th Int. Joint Conf. Artif. Intell., Melbourne, VIC, Australia, Aug. 2017, pp. 3420–3426.
[6] Y. Wang, J. Li, I. King, M. R. Lyu, and S. Shi, “Microblog hashtag generation via encoding conversation contexts,” in Proc. Conf. North, Minneapolis, MN, USA, Jun. 2019, pp. 1624–1633.
Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, “Co-attention network recommendation for micro-videos,” in Proc. 27th ACM Int. Conf. Multimedia, Nice, France, Oct. 2019, pp. 1446–1454.

S. Zhang, Y. Yao, F. Xu, H. Tong, X. Yan, and J. Lu, “Hashtag recommendation for photo sharing services,” in Proc. AAAI Conf. Artif. Intell., Honolulu, HI, USA, vol. 33, Jul. 2019, pp. 938–942.

D. Kowald, S. C. Pujari, and E. Lec., “Temporal effects on hashtag reuse in Twitter: A cognitive-inspired hashtag recommendation approach,” in Proc. 26th Int. Conf. World Wide Web, Perth, WA, Australia, Apr. 2017, pp. 1401–1410.

N. Garg and I. Weber, “Personalized, interactive tag recommendation for flickr,” in Proc. ACM Conf. Recommender Syst. (RecSys), Lausanne, Switzerland, Sep. 2010, pp. 49–56.

X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, “Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach,” in Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), Glasgow, U.K., Oct. 2011, pp. 1031–1040.

K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu, “Collaborative personalized tweet recommendation,” in Proc. 35th ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR), Portland, OR, USA, Aug. 2012, pp. 661–670.

I. Ronen, I. Guy, E. Kravi, and M. Barnea, “Recommending social media content to community owners,” in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR), Gold Coast, QLD, Australia, Jul. 2014, pp. 243–252.

A. Veit, M. Nickel, S. Belongie, and L. van der Maaten, “Separating self-expression and visual content in hashtag supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 5919–5927.

H.-M. Lu and C.-H. Lee, “A Twitter hashtag recommendation model that accommodates for temporal clustering effects,” IEEE Intell. Syst., vol. 30, no. 3, pp. 18–25, May 2015.

M. Li, T. Gan, M. Liu, Z. Cheng, J. Yin, and L. Nie, “Long-tail hashtag recommendation for micro-videos with graph convolutional network,” in Proc. 28th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), Beijing, China, Nov. 2019, pp. 509–518.

C. Yang, M. Jiang, B. Jiang, W. Zhou, and K. Li, “Towards micro-video understanding,” IEEE Access, vol. 7, pp. 40771–40781, 2019.

M. Lao, Y. Guo, H. Wang, and X. Zhang, “Multimodal local perception bilinear pooling for visual question answering,” IEEE Access, vol. 6, pp. 57923–57932, 2018.

J. Gao, R. Ge, K. Chen, and R. Nevatia, “Motion-appearance co-memory networks for video question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 6576–6585.

P. Zhang, S. Zhu, and C.-W. Ngo, “Deep multimodal learning for affective analysis and retrieval,” IEEE Trans. Multimedia, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

X. Li, S. Jiang, Y. Heng, H. Wang, X. Liu, and L. Herranz, “Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration,” IEEE Trans. Multimedia, vol. 19, no. 5, pp. 1100–1113, May 2017.

J. Zhang, L. Nie, X. Wang, X. He, X. Huang, and T. S. Chua, “Shorter-is-better: Venue category estimation from micro-video,” in Proc. ACM Multimedia Conf. (MM), Amsterdam, The Netherlands, Oct. 2016, pp. 1415–1424.

Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, “Neural multimodal cooperative learning toward micro-video understanding,” IEEE Trans. Image Process., vol. 29, pp. 1–14, Jul. 2020.

A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in Proc. Conf. Empirical Methods Natural Lang. Process., Copenhagen, Denmark, Sep. 2017, pp. 1103–1114.
[49] Y. Gong and Q. Zhang, “Hashtag recommendation using attention-based convolutional neural network,” in Proc. IJCAI, New York, NY, USA, Jul. 2016, pp. 2782–2788.

[50] P. Bansal, R. Bansal, and V. Varma, “Towards deep semantic analysis of hashtags,” in Proc. ECIR, Vienna, Austria, Mar./Apr. 2015, pp. 453–464.

[51] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, and T.-S. Chua, “Micro tells macro: Predicting the popularity of micro-videos via a transductive model,” in Proc. ACM Multimedia Conf. (MM), Amsterdam, The Netherlands, Oct. 2016, pp. 898–907.

[52] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural collaborative filtering,” in Proc. 26th Int. Conf. World Wide Web, Perth, WA, Australia, Apr. 2017, pp. 173–182

[53] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” in Proc. ICLR, Toulon, France, Apr. 2017, pp. 1–15.

[54] C. Yang, H. Zhang, B. Jiang, and K. Li, “Aspect-based sentiment analysis with alternating coattention networks,” Inf. Process. Manage., vol. 56, no. 3, pp. 463–478, May 2019.

[55] Z. Zhao, J. Lin, X. Jiang, D. Cai, X. He, and Y. Zhuang, “Video question answering via hierarchical dual-level attention network learning,” in Proc. ACM Multimedia Conf. (MM), Mountain View, CA, USA, Oct. 2017, pp. 1050–1058.

[56] H. Wang, X. Shi, and D.-Y. Yeung, “Relational stacked denoising autoencoder for tag recommendation,” in Proc. AAAI, Austin, TX, USA, Jan. 2015, pp. 3052–3058.

XIAOCHAN WANG received the bachelor’s degree in computer science and technology from Southwest University, in 2017. She is currently pursuing the master’s degree with the College of Computer Science and Electronic Engineering, Hunan University, China. Her research interests include multi modality representation learning and machine learning.

CHAO YANG (Member, IEEE) received the B.E. and M.E. degrees in computer science from Hunan University, Changsha, China, in 1999 and 2005, respectively, and the Ph.D. degree in computational intelligence and systems science from the Tokyo Institute of Technology, Tokyo, Japan, in 2010. She has ever worked as a Postdoctoral Fellow in the Tokyo Institute of Technology, Tokyo, Japan. Since 2016, she has been an Associate Professor with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. Her research interests include data mining, machine learning, intelligent systems, recommender systems, and deep learning. She is a member of ACM and CCF.

BIN JIANG (Member, IEEE) received the B.A. degree in mathematics and the M.E. degree in soft engineering from Hunan University, Changsha, China, in 1993 and 2006, respectively, and the Ph.D. degree in computational intelligence and systems science from the Tokyo Institute of Technology, Tokyo, Japan, in 2015. He is currently an Associate Professor with the College of Computer Science and Electronic Engineering, Hunan University, Changsha. He is a member of ACM and CCF. His research interests include big data technology, machine learning, data mining, intelligent computing, recommender systems, and social computing.

* * *