Expert Systems With Applications 210 (2022) 118375

Contents lists available at ScienceDirect

Expert Systems With Applications

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

DemoHash: Hashtag recommendation based on user demographic information

Dahye Jeong[1], Soyoung Oh[1], Eunil Park *

Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul 03063, Republic of Korea

ARTICLE INFO

Keywords:
Hashtag recommendation
Multi-modal model
Demographic information

ABSTRACT

Social network services have become widely used, and hashtags, which are implicitly involved in delivering specific information, have shown to greatly improve user engagement. A number of prior studies have attempted to recommend appropriate hashtags for each social media user considering his/her posts by consequently extracting the important features from text and images. To develop this multi-dimensionality with hashtag recommendation, user demographic information also plays a significant role in the manner of personalized hashtag recommendation. Thus, this paper proposes the demographic hashtag recommendation (*DemoHash*) model to utilize users' demographic information extracted from their selfie images, in addition to textual and visual information. The experimental results with the datasets from *Instagram* show that our proposed model achieves a greater performance with F_1-score, Precision, and Recall than the existing hashtag recommendation methods by average of 4.19%, 18.45%, and 3.91%, respectively. Our approach effectively combined the content-based as well as user-oriented modeling for personalized hashtag recommendation.

1. Introduction

Among several social network services (SNSs), *Instagram*, which is photo-oriented, allows users to upload and share their images or short videos. When users share their images or videos on social network services, a hashtag, generally used with the "#" prefix, presents the abstracted meaning of the images or videos. Further, a hashtag has become prevalent and efficient communication means widely adopted by SNS users to share information and express their contents. Therefore, recommending specific hashtags, which are examined by content-based filtering procedures with image features, is one of the primary areas for capturing semantic understandings of the uploaded images or videos (Gong, Ke, Isard, & Lazebnik, 2014; Wei et al., 2014).

The hashtags are sometimes employed to present humorous meanings, or to express specific opinions, it is not only meaningful to understand users' tagging behavior, but also its abstracted/inherent associations with the uploaded images or videos. Thus, to allow users to efficiently present specific hashtags in SNSs, several applications for recommending personalized hashtags based on users' information have been developed. The majority of prior studies in this regard employed users' images, texts, locations, tagging habits and social groups as primary features for recommending specific hashtags (Rawat & Kankanhalli, 2016; Wei et al., 2019; Zhang et al., 2019; Zhu, Aloufi, & El Saddik, 2015).

However, because the majority of these approaches have mainly employed images and texts in users' posts as the main features, these approaches have notable limitations in fully capturing the essence of hashtag recommendations. As users' socio-demographic information is significantly related to their hashtags usage behavior (Denton, Weston, Paluri, Bourdev, & Fergus, 2015), utilizing their socio-demographic information as one of the main features can be effective for recommending appropriate hashtags. From a broader perspective, user behavior on social networks can be differentiated by culture (Sheldon, Rauschnabel, Antony, & Car, 2017), gender (Tomorn & Bao, 2020), age (Jang, Han, Shih, & Lee, 2015), the number of followers/followings (Neal, 2017), and so on.

Therefore, the main focus of this paper is to employ users' demographic information as one of the primary features for recommending user-personalized hashtags. In addition, we consider their demographic information including age, gender, race, emotion, the number of followers, followings, posts, and introductory texts. Based on such demographic information, we develop a personalized hashtag recommendation model. Fig. 1 presents the overview of the proposed model. The proposed model is publicly available[2]. Our main contributions are summarized as follows:

* Correspondence to: 310 International Hall, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Republic of Korea.
E-mail addresses: gwg03391@skku.edu (D. Jeong), sori424@g.skku.edu (S. Oh), eunilpark@skku.edu (E. Park).

[1] Equally contributed first authors.

[2] https://github.com/dxlabskku/DemoHash

https://doi.org/10.1016/j.eswa.2022.118375
Received 4 December 2021; Received in revised form 13 July 2022; Accepted 1 August 2022
Available online 9 August 2022

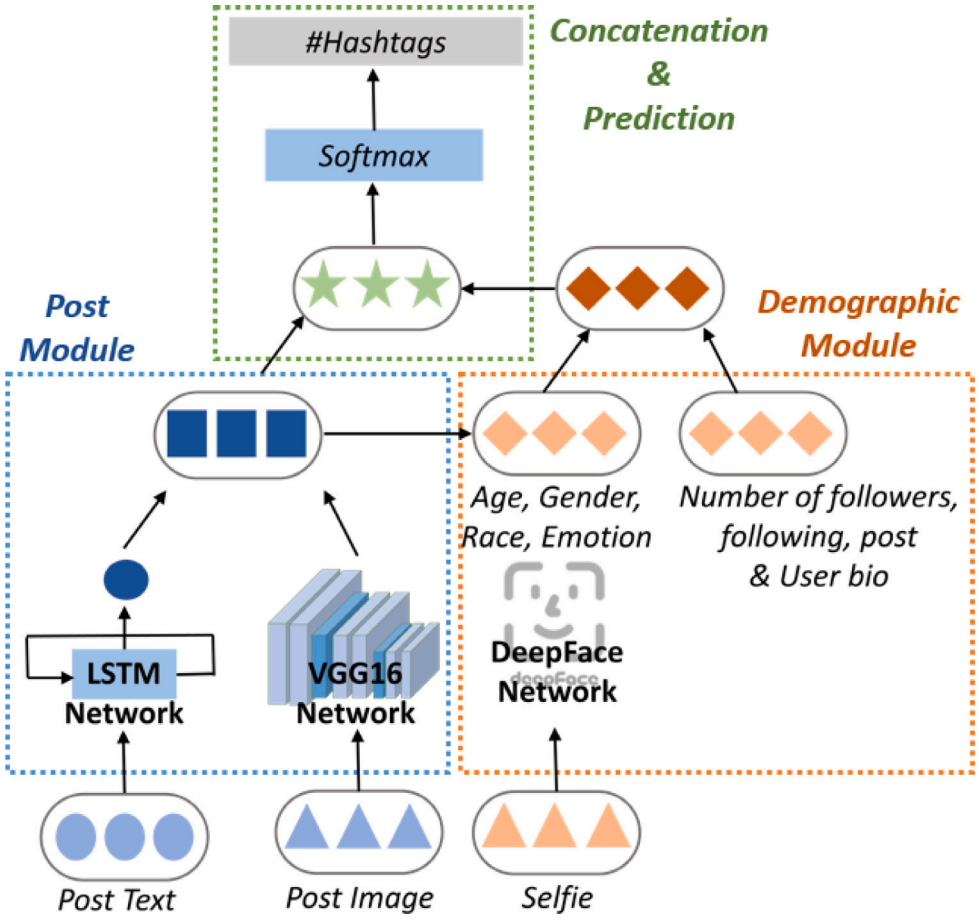

Fig. 1. The overview of DemoHash; (1) Blue box: post module represents the process of extracting post feature from images and text. (2) Orange box: demographic module represents the process extracting of demographic information and user information feature. (3) Green box: final concatenation of two modules recommends hashtags to users using overall features.

- We proposed a DemoHash model, that uses image, text, and demographic information for user personalized hashtag recommendation.
- To incorporate the multi-modal features from each data type channel, we proposed an attention-based neural network model for consideration of both post and user demographic features.
- Experimental results using a dataset, which includes demographic information, demonstrate that our model outperformed than other baseline hashtag recommendation methods.

2. Related works

2.1. Hashtag recommendation by text

Several prior studies have proposed hashtag recommendation systems with various feature variables. Textual representation, one of the most classical natural language processing solutions, has been mainly employed for hashtag recommendation. Godin, Slavkovikj, De Neve, Schrauwen, and Van de Walle (2013) used a topic modeling method, Latent Dirichlet Allocation (LDA), to assign general hashtags of specific tweets and achieved an accuracy of 91%.

Also, Gong and Zhang (2016) employed an attention-based neural network model including trigger words and achieved an F_1-score of approximately 9.4% in microblog hashtag recommendation tasks. In addition, Liu, He, and Huang (2018) employed a hashtag embedding approach by associating tweets to resolve the sparsity, polysemy, and synonym issues with hashtag recommendations. Their clustering evaluation metric achieved the highest score, at 64.8%, compared to the other baseline methods. Moreover, Zhang et al. (2019) proposed the term frequency–inverse document frequency (TF-IDF) of tweets to recommend customized hashtags and demonstrated a 12.8–23.6% improvement in precision compared to the co-attention model.

Recent research on hashtag recommendation employed BERT embedding and outperformed LDA, SVM, and TTM with the recall measure (Kaviani & Rahmani, 2020). In addition, Cantini, Marozzo, Bruno, and Trunfio (2021) proposed sentence-hashtag embedding translation, which has two latent spaces to embed the text content of a post and hashtags. They obtained a 15% improvement in F-score comparing the relevant techniques.

2.2. Hashtag recommendation by multimodal feature

Moreover, prior studies employed both textual and visual information to recommend hashtags (Table 1). The majority of such studies focused on the suggested images, when it was too difficult to identify certain meanings from textual representations. Hwang and Grauman (2012) employed the image object detection approach to capture implicit information of objects and proposed new tag recommendation solutions. With more than 5,000 images, the proposed approach achieved 81% precision among the top seven retrievals. Wang et al. (2016) proposed the unified framework of an end-to-end model to learn the semantics of images and multiple tags. In addition, Zhang, Wang, Huang, Huang, and Gong (2017) used a co-attention network model incorporating users' visual and textual contents to suggest specific hashtags for multimodal posts. Considering more than 402,000 tweets, the proposed model achieved a 31.1% F_1-score, which outperformed all the

Table 1
Summary of prior research on hashtag recommendation.

Category	Sources	Method	Feature	Datasets
Text-based	Kaviani and Rahmani (2020)	Neural network based on BERT embedding	text	100,000 tweets
	Cantini et al. (2021)	Semantic mapping	text	2.5 million tweets
Multi-modal	Yang, Wu, et al. (2020)	Attention neural network with GRU	image and text	56,861 posts
	Kang et al. (2020)	Semantic generation based on BERT	image, text, location, and time	181,620 posts
	Yang, Wang, and Jiang (2020)	Space learning model	image, text, and sentiment	40,049 micro-videos
	Sun et al. (2021)	Hierarchical attention model	item, user information, and contents	CiteULike/MovieLens

baseline models. Similarly, Kang, Kim, Shin, and Myaeng (2020) proposed attention-based multimodal neural network model using Gated Recurrent Unit (GRU) mechanism. They concatenated image features and text features on *Instagram* and transferred them into the sequence to sequence model. The experimental result showed that the model achieved 65.49% of accuracy.

Extracting additional information from textual and visual representations with multiple labels has been proposed and employed by several researchers. Rawat and Kankanhalli (2016) utilized image context, spatial location, and time information with a trained end-to-end model and attempted effectively to solve a multi-label classification problem. Based on their model and the YFCC100M dataset, their implementation using additional information increased the accuracy, precision, and recall by 71%, 25%, and 20%, respectively. In addition, Zhang et al. (2019) employed users' tagging preferences for automated image tagging tasks. The concept of an external memory component is proposed to characterize the users' historical tagging behaviors in the habit modeling module. Considering 624,520 *Instagram* posts, the habit model proposed by Zhang et al. (2019) achieved an F_1-score improvement of up to 13.4%. Moreover, Yang, Wang, and Jiang (2020) considered content and sentiment features for micro-video hashtag recommendation. To analyze sentiment tags, they developed common space learning techniques, which can decrease fusion and disagreement problems. Moreover, to recommend personalized tag for users, a hierarchical attention model using information of user-item pairs was proposed (Sun, Zhu, Jiang, Liu, & Wu, 2021).

Because hashtags can be treated as the self-expression of users, personalized hashtag recommendation has attracted considerable attention. However, to conceptualize user representation issues, the majority of prior research focused on content features, which is directly displayed on users' post (Kang et al., 2020; Sun et al., 2021). Our study examines users' demographic information (e.g. age, gender, race, and emotion), from users' selfie images of posts, and employs the information as a potential feature to provide more user-specific hashtags. In addition, to efficiently compute the features with different modalities, we propose the co-attention mechanism model. Compared to the findings of prior research on attention mechanism with image and text (Kang et al., 2020; Zhang et al., 2017), we develop unique demographic module, which calculates a correlation level between content features and demographic information.

3. Approach

Considering a set of text, image, and corresponding user demographic information, the task of our approach is to effectively generate appropriate hashtags for each post. To examine this task for each post, multi-class classification is performed.

The overview of our proposed model is presented in Fig. 1. The image, and text as well as user demographic information, which is extracted using a selfie image and each user's account information, are employed as the input. For a better understanding, we present the details of our suggested *Demographic Hashtag Recommendation* (DemoHash) model in the following sections. Section 3.1 elucidates the employed feature extraction procedures, and the baseline model and our DemoHash are discussed in Sections 3.2 and 3.3, respectively.

3.1. Feature selection

Text. Each text was embedded into sequences with an embedding size of 128, where all words in the text are represented as vectors x_i. The text can be indicated as $[x_1, x_2, \dots, x_N]$, where N denotes the maximum text length. Then, we adopted a long short-term memory (LSTM) network with 300 units to maintain the sequential nature of the text to generate text feature efficiently (Gers, Schmidhuber, & Cummins, 2000). In each step, the current word embedding sequence x_i and the output of the previous unit h_{i-1} from the LSTM units were concatenated to obtain the output word representation of h_i. From the LSTM networks, we obtain output u as the text feature matrix, and u can be represented as $u = [h_1, h_2, \dots, h_N]$ with $h_i \in \mathbb{R}, i = 1, 2, \dots, N$. Identical text feature extraction process was employed to create inputs for both the baseline and our proposed models (Oh, Ji, Kim, Park, & del Pobil, 2022).

Image. The images were resized to 224 × 224 to be used as an input to the VGG-16 network pre-trained model with the Imagenet dataset (Simonyan & Zisserman, 2014). The outputs of the last pooling layer of the VGG network were extracted as the image features. Specifically, tensors for the 7 × 7 regions which is represented via a 512 dimensional vector are kept by the multiple feature vectors for each image. That is, denoting v^* as an image feature matrix, v^* consists of $v^* = [v_1^*, v_2^*, \dots, v_M^*]$ with $v_i^* \in \mathbb{R}^{\mathbb{D}}, i = 1, 2, \dots, M$ with $D = 512$. For the computational convenience, we added fully connected layer after VGG network to convert each D-dimensional regional feature vector into a new vector with the same dimension of the text feature vectors. Therefore, the image feature matrix is as $v = [v_1, v_2, \dots, v_m]$ with $v_i \in \mathbb{R}, i = 1, 2, \dots, M$.

Demographic information. We extracted the users' age, gender, race, and emotion from the collected selfie posts by utilizing *Deepface*[3] (Serengil & Ozpinar, 2021), where *Deepface* is a lightweight face recognition and facial attribute analysis framework. *Deepface* was examined with the VGG-Face model, while the output layers were customized for demographic prediction. The output layer for age prediction has 101 nodes, and predicts age levels from 0 to 100 years as a continuous variable. The race model has six outputs: Asian, Black, Latino Hispanic, Middle Eastern, and White. As an exception, the emotion model is not examined with the VGG Face structure, because of the limited channel numbers in the datasets. The emotion model has 5 convolution layers and 3 fully connected layers, allowing users to have one of the following labels: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral (Serengil & Ozpinar, 2021). The categorical values of gender, race, and emotion were presented as integer values.

Moreover, we additionally collected the user information which is organized by the number of each user's posts, the number of followings/followers, and his/her bio. The bio indicates the text of the introduction from the user's profile. Each word of the user bio is embedded as a vector, and the embedded words are constructed to be represented in matrix form. We concatenated age a_i, gender g_i, race r_i, emotion e_i, and collected user information ui_i, thereby employing the combined demographic information as demographic information features. Demographic information w_i is indicated as $w_i = [a_i, g_i, r_i, e_i, ui_i]$,

[3] https://github.com/serengil/deepface

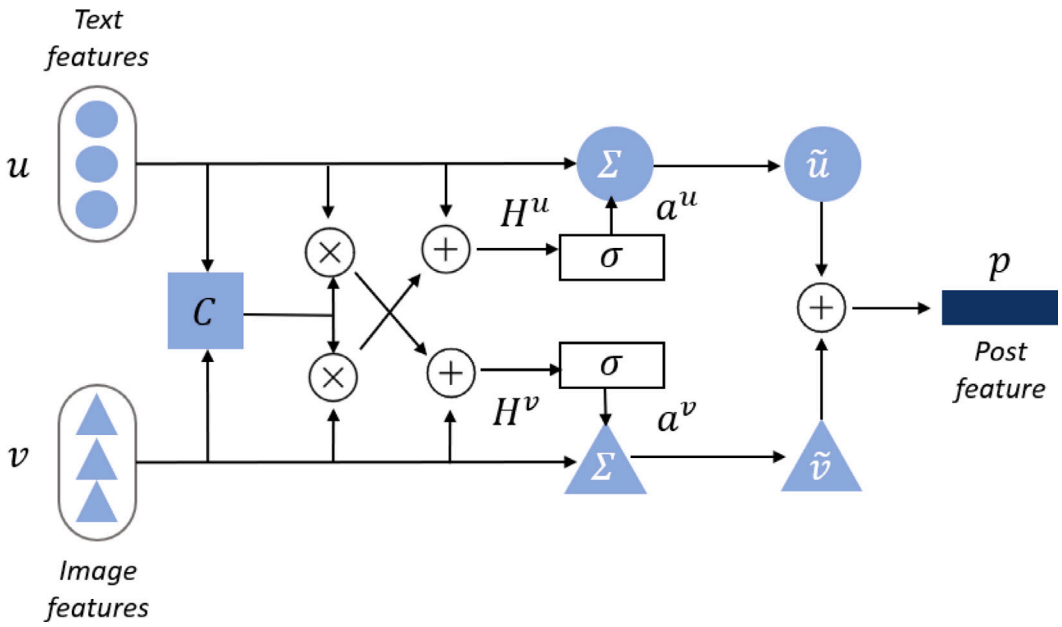

Fig. 2. The post modeling module of our proposed model; coherent feature for post feature (p) was extracted by co-attention mechanism from text (u) and image (v) features. H^u and H^v mean new text and image features, respectively. The attention weights are denoted as a^u (text) and a^v (image).

with $w_i \in \mathbb{R}, i = 1, 2, \ldots, K$, where K represents the number of users. Then, we finally obtain w as the demographic information feature matrix, which can be represented as $d = [w_1, w_2, \ldots, w_K]$.

3.2. Our proposed DemoHash model

In this study, we implemented two modules for learning post features and demographic features. To process the causal relationship between images and texts, we employed the parallel co-attention mechanism (Lu, Yang, Batra, & Parikh, 2016). For the demographic features, we calculated the similarities between the post feature and the demographic information. Then, we measured the weighted sum of the demographic information where weights are determined by the similarities. Finally, post feature p and demographic feature d are concatenated to make the hashtag recommendations. We employed a softmax layer to compute the probability of each hashtag, and a ranked list of top-K hashtags are returned for a given post and demographic information.

Post module. For the content modeling of the given post, the overall procedure is as in Fig. 2. Considering hashtags allow users to indicate the topics of their posts, it is important to extract the significant features from the posts to effectively recommend the hashtags. By using parallel co-attention network, we suggest the post module as follows. The details of the post module and demographic module are as follows.

An affinity matrix $C \in \mathbb{R}^{N \times M}$ in the parallel co-attention network is defined and based on the similarity between the corresponding feature vector pairs of the text vector matrix u and image vector matrix v. In addition, N and M mean the maximum length of text and image features, respectively. C is represented as follows:

$$C = tanh(u^T W_b v), \tag{1}$$

where $W_b \in \mathbb{R}^{d \times d}$ indicates the correlation parameter to be updated through the training procedures, when d means its representation dimension. Then, we obtain a new representation of the text and image matrices based on the affinity matrix C. As a new text feature matrix, $H^u \in \mathbb{R}^{d \times M}$ is represented by the integration of text features, where $d \times M$ is representation dimension rows and image feature size columns. The computed affinity matrix C as follows:

$$H^u = tanh(W_u u + (W_v v) \cdot C^T). \tag{2}$$

In the new text feature representation of H^u, the image feature matrix v is multiplied by C^T and added to the text features, and W_u and $W_v \in \mathbb{R}^{d \times d}$ are the parameters of the text and image, respectively. The new image feature matrix $H^v \in \mathbb{R}^{d \times M}$ is represented as follows:

$$H^v = tanh(W_v v + (W_u u) \cdot C). \tag{3}$$

With the new text and image feature matrices, H^u and H^v, the attention weights of the text and image are calculated as follows:

$$a^u = softmax(W_{hu}^T H^u + b_{hu}), \qquad a^v = softmax(W_{hv}^T H^v + b_{hv}), \tag{4}$$

where the parameters are $W_{hu}, W_{hv} \in \mathbb{R}^d$ and $b_{hu}, b_{hv} \in \mathbb{R}$. We then obtain the global feature vectors for the text and image using the attention weights. The global feature vectors are represented by the weighted sum of the partial text and image feature vectors, as follows:

$$\tilde{u} = \sum_{i=1}^{N} a_i^u u_i, \tilde{v} = \sum_{i=1}^{M} a_i^v v_i, \tag{5}$$

In $\tilde{u}$ and $\tilde{v}$, a_i^u and a_i^v indicate the attention weights of a certain word and image region, respectively. Finally, added the two global feature vectors to construct the multimodal post feature p as below.

$$p = \tilde{u} + \tilde{v}. \tag{6}$$

p and demographic information feature d are transferred to a softmax layer for predicting specific hashtags. The softmax layer is employed to compute the probability of each hashtag, and a ranked list of top-K hashtags is returned for a given post.

Demographic module. As demographic information of the users are related to content of the posts, we calculate the similarities between them. To find out the most significant demographic information to recommend hashtags, we measure the weighted sum from demographic information where the weights are determined by the similarities. The overall modeling procedure is as in Fig. 3. The details are as follows.

We measured the similarity between the post feature and each demographic information as follows,

$$s_i = tanh(p \odot di_i), \tag{7}$$

where $\odot$ means element-wise multiplication and s_i represents the correlation vector between the post and the ith demographic information. By combining all the vectors, we extract the similarity matrix

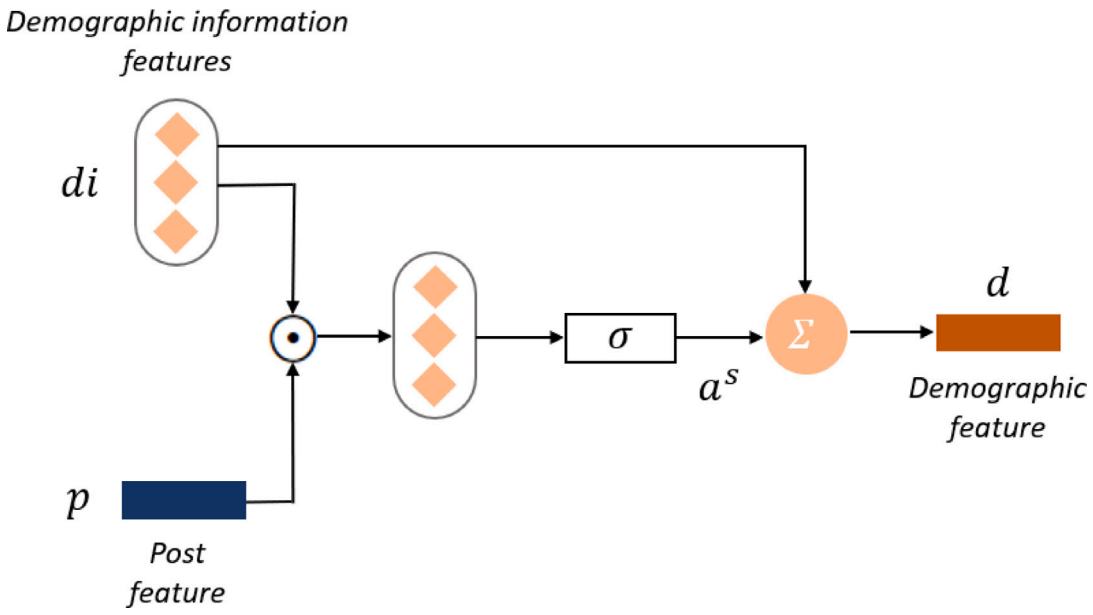

Fig. 3. The demographic modeling module of our proposed model; demographic feature (d) is extracted by combined representation of each demographic information weighted by the similarities between post (p) features and demographic information (di) features.

$s = [s_1, s_2, \ldots, s_K]$, where K is the number of demographic information. Based on correlation vector s, we compute the weights of each demographic information as

$$a^s = softmax(W_s^T s + b_s), \tag{8}$$

where $W_s \in \mathbb{R}^d$, $b_s \in \mathbb{R}^K$ is a vector containing the weights of demographic information.

At last, the demographic feature d can be computed as follows,

$$d = \sum_{i=1}^{K} a_i^s di_i. \tag{9}$$

which represents the combined features of the corresponding demographic information weighted by the similarities between the post and demographic information.

3.3. Baseline models

We selected the following machine learning methods and state-of-the-art models for our baselines:

- **Tweet2vec**: It encodes the text representation in vector-space by learning complex, non-local dependencies in character sequences to recommend hashtag (Dhingra, Zhou, Fitzpatrick, Muehl, & Cohen, 2016).
- **CNN-Att**: It is a text-based model to incorporate trigger words into hashtag recommendation by using convolutional neural network and attention mechanism (Gong & Zhang, 2016).
- **NB**: As a form of multi-class classification task, a Naive Bayes classifier is employed to present the posterior probability of each hashtag given the textual, visual and demographic information (Rish et al., 2001).
- **SVM**: We adopted the Support Vector Machine for the classification task to recommend the hashtags with text, image and demographic features (Noble, 2006).
- **Co-Att**: It is one of the state-of-the-art hashtag recommendation methods incorporating textual and visual information with co-attention mechanism to recommend hashtags (Zhang et al., 2017).
- **Habit**: It employs the parallel co-attention mechanism to address content modeling module and external memory unit to characterize the historical tagging habit of each user (Zhang et al., 2019).
- **AMNN**: It employs self-attention mechanisms to image and text, respectively. And, merged features are trained by GRU networks (Yang, Wu, et al., 2020).

4. Experiment

4.1. Data collection and construction

We found *Instagram* posts via the *#selfie* query to extract users' demographic information, and the searched posts were filtered by photos that included a single person. As the users use *#selfie* hashtag regardless of their native language, we collected the user data with different languages. Moreover, we collected the number of followers, followings, and user bio for the same user ID. To evaluate the baseline model, which requires user tagging habits, we crawled two more posts for each user. The example of a crawled data for the experiment is presented as in Fig. 4.

With these procedures, we collected 8,965 posts. Then, we excluded posts that did not contain all the necessary information, images, text, or at least one hashtag. Among the 8,965 posts, 5,125 posts (57.2%) were excluded, while we used 3,840 posts (42.8%), which included all the required information. We then extracted image, text, hashtag, user demographic information (age, gender, race, and emotion) and user information (the number of posts, followers, followings, and user bio). The final dataset in this study contained 3,840 posts (of 1,280 users), 80,781 unique hashtags, and 49,549 distinct words. The distributions of the user demographic information are presented as in Fig. 5. Then, we randomly split two-thirds of posts as a training set and the other as a test set. Table 2 presents an overview of the collected data sets.

We can use pseudonymous information on Instagram according to three data-related laws and enforcement decrees in Korea. The laws and enforcement decrees enable the government, companies, and institutions to use personal information, which is fully anonymized. It means that they can use the information, after eliminating user identities (The Korea Times, 2020). To convert the collected data to pseudonymous data, we conducted a data anonymization procedure.

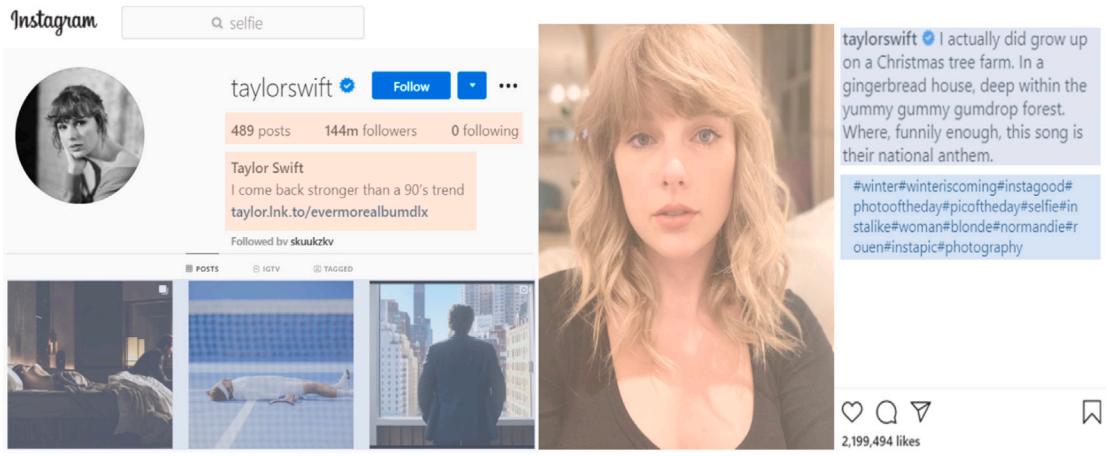

Fig. 4. The example of crawled data by "selfie" search query of user where highlighted parts are collected by each user; post information and demographic and user information.

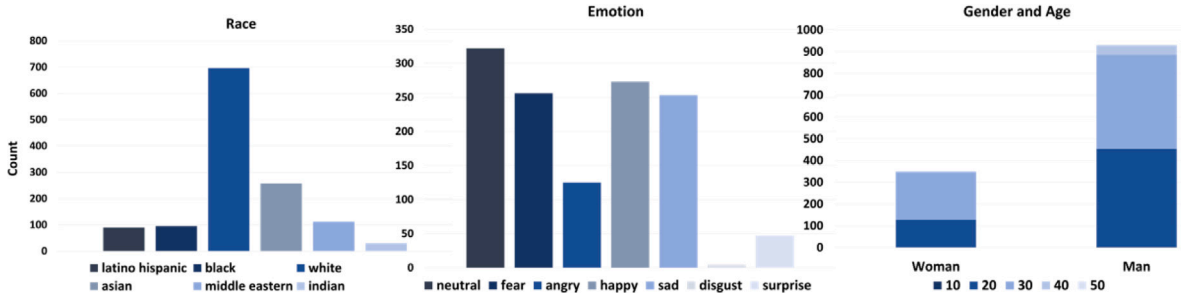

Fig. 5. Distribution of demographic information from the collected dataset.

Table 2
Statistics of the *Instagram* dataset.

Posts		Users	Hashtag	Words
Train	Test			
2,560	1,280	1,280	80,781	49,549

In this procedure, we converted all user information into encoded values, and directly used both extracted images and text-based numeric features for our proposed model (Kim, Lee, Park, & Han, 2020). Moreover, instead of sharing our original dataset, we shared our embedded features examined by feature extraction models after this procedure in agreement with Instagram's Terms & Conditions (Meta, 2022)[4]. We mainly followed the social media guidelines of the European Data Protection Board in our data anonymization procedure[5], with the Institutional Review Board (IRB) approval of Sungkyunkwan University, Seoul, Korea.

4.2. Implementation details

The output of the image feature layer from VGG-16 is $7 \times 7 \times 512$. Then, we used dense layer including *tanh* activation function for 300 embedding size for feature concatenation. The text feature is set to having maximum text-sequence length of 328. The LSTM layer extracted text feature of 300 embedding size. The 133 dimension of demographic feature, which includes age, gender, race, emotion, number of each user's posts, the number of followings/followers, and his/her bio, was deployed as another input to the model. For the feature concatenation, demographic feature is also resized to having 300 embedding size by dense layer with *tanh* activation.

In addition, we used a dropout rate of 0.75. The batch size of training procedures is set to 5, where the number of epochs is computationally obtained by dividing the number of train data by the batch size (Lee, Jeong, & Park, 2022). Finally, the number of epoch is 10 and the parameters are trained by ADAM optimizer.

4.3. Evaluation metrics

Following the evaluation guidelines of prior studies on hashtag recommendations (Denton et al., 2015; Zhang et al., 2019), three evaluation metrics, *precision* (P), *recall* (R), and F_1-score (F_1), were employed. We varied the number of recommended hashtags for each post, K, where the results of each evaluation metric, P, R, and F_1 at K are denoted as P_k, R_k, and F_{1k}, respectively. Also, the rank (p, u, k) and GroundTruth (p, u) are denoted the set of top k-ranked hashtags presented by the model and the set of hashtags actually tagged by the user u for the post p, respectively. The final evaluation metrics are defined as follows:

$$P@K = \frac{1}{N}\sum_{i=1}^{N}\frac{|Rank(p_i,u_i,k_i) \cap GroundTruth(p_i,u_i)|}{|Rank(p_i,u_i,k_i)|},$$

where N is the number of test sets. $P@K$ is defined as the precision level of the probability of whether a specific GroundTruth hashtag is presented in the ranked list of top-K predicted hashtags. We measured $R@K$, which represents a fraction of the relevant presented hashtags for each post in the actual hashtags as follows:

$$R@K = \frac{1}{N}\sum_{i=1}^{N}\frac{|Rank(p_i,u_i,k_i) \cap GroundTruth(p_i,u_i)|}{|GroundTruth(p_i,u_i)|}.$$

Moreover, based on $P@K$ and $R@K$, the F_1-score is computed as follows:

[4] https://github.com/dxlabskku/DemoHash

[5] https://edpb.europa.eu/our-work-tools/documents/public-consultations/2020/guidelines-082020-targeting-social-media-users_en

Table 3
Evaluation results of models (K=7)

Model	Precision	Recall	F_1	Time
Tweet2Vec	0.264	0.081	0.124	2486 s
CNN-Att	0.274	0.094	0.127	391 s
NB	0.174	0.056	0.079	54 s
SVM	0.098	0.040	0.049	800 s
Co-Att	0.333	0.111	0.153	791 s
Habit	0.356	0.119	0.164	1300 s
AMNN	0.262	0.091	0.122	23993 s
DemoHash	**0.426**	**0.166**	**0.208**	1705s

Table 4
Results of 3-Fold validation (K=7); Originally, we randomly split each user's 3 posts (2 posts for the train and 1 post for the test sets). The denoted numbers (0, 1, 2) mean the order of splited posts.

DemoHash	Precision	Recall	F_1	Time
$train_{1,2}$ + $test_0$	0.402	0.146	0.194	3289 s
$train_{0,1}$ + $test_2$	0.391	0.142	0.187	3303 s
$train_{0,2}$ + $test_1$	0.464	0.186	0.233	2214 s

$$F_1@K = \frac{2 \times P@K \times R@K}{P@K + R@K}.$$

5. Results

Comparison analysis. We first compared DemoHash model with the baseline models. Table 3 summarizes the results of each evaluation metric with the number of 7 hashtag recommendation (half of the average number of hashtags in a post). DemoHash model as our proposed model outperformed the other models (7.0%–32.8%, 4.7%–12.6%, 4.4%–15.9% of precision, recall, and F_1-score). The remarkable improvements demonstrate the effectiveness of our approach. We also conducted the 3-Fold validation procedures for DemoHash model. Table 4 shows the summary of the results with 3-fold validation approaches.

Next, we analyzed the performance of DemoHash model with the baseline models with varying K as shown in Table 5. The K means the number of predicted hashtags. Considering the recommendation of a single hashtag, DemoHash model demonstrated the greatest precision (DemoHash: 0.614). Compared to other competitors, when the recommended number of hashtags, K, was varied from 1 to 5, DemoHash model achieved 0.3%–40.1% (Mean: 18.45%), 0.31%–9.3% (Mean: 3.91%), and 0.5%–10.7% (Mean: 4.19%) improvements in precision, recall, and F_1-score, respectively. As the users use several hashtags containing demographic information at the same time, the most significant improvement captured with the increase of recommendation numbers as in Table 5.

Performance gain analysis. Then, we analyzed the performance gain of the proposed DemoHash model. To analyze the effectiveness of the demographic feature, we varied the models with the following four components: (1) $DemoHash_{t+i}$, the text of users' posts and the image of users' posts (Text+Image); (2) $DemoHash_{i+d}$, image and user demographic information (Image+Demo) which is names as ; (3) $DemoHash_{t+d}$, text and demographic information (Text+Demo); (4) DemoHash which is our proposed model, text, image, and the demographic information (Text+Image+Demo). Then, we varied the number of recommended hashtags from 1 to 5 to evaluate each model under identical conditions. Table 6 shows the summary of the evaluation metrics from all the variants. Compared to the variant, which removed the demographic information from the input feature, the $DemoHash_{i+t}$ model, all the models used demographic features outperformed from 0.1% to 3.5% in terms of F_1-score. With the number of one hashtag recommendation, our proposed model outperformed all the other variants by average of 1.7%, 0.3%, and 0.4% in terms of precision, recall, and F_1-score, respectively. Also, for the number of five hashtag recommendation, the DemoHash outperformed all the others by average of 7.7%, 4.2%, and 3.1% with each evaluation metric.

Table 5
Precision, recall, and F_1-Score with different numbers of recommended hashtags; K = 1 to 5.

Top-K		1	2	3	4	5
Tweet2Vec	$P@K$	0.541	0.483	0.322	0.286	0.257
	$R@K$	**0.041**	0.049	0.058	0.066	0.075
	$F_1@K$	**0.072**	0.089	0.098	0.107	0.116
CNN-Att	$P@K$	**0.636**	0.460	0.381	0.344	0.312
	$R@K$	**0.041**	0.053	0.061	0.072	0.080
	$F_1@K$	0.070	0.085	0.095	0.107	0.114
NB	$P@K$	0.213	0.222	0.199	0.189	0.185
	$R@K$	0.010	0.021	0.029	0.035	0.043
	$F_1@K$	0.036	0.046	0.055	0.062	0.065
SVM	$P@K$	0.469	0.341	0.228	0.171	0.137
	$R@K$	0.028	0.040	0.040	0.040	0.040
	$F_1@K$	0.050	0.065	0.060	0.057	0.054
Co-Att	$P@K$	0.606	0.469	0.400	0.379	0.347
	$R@K$	0.038	0.053	0.064	0.078	0.087
	$F_1@K$	0.066	0.086	0.099	0.117	0.126
Habit	$P@K$	0.571	**0.490**	0.434	0.397	0.358
	$R@K$	0.037	**0.056**	0.069	0.080	0.088
	$F_1@K$	0.063	0.091	0.108	0.122	0.129
AMNN	$P@K$	**0.636**	0.455	0.381	0.332	0.302
	$R@K$	**0.041**	0.052	0.062	0.070	0.077
	$F_1@K$	0.070	0.084	0.096	0.104	0.110
DemoHash	$P@K$	0.614	0.469	**0.454**	**0.400**	**0.429**
	$R@K$	**0.041**	0.053	**0.079**	**0.087**	**0.133**
	$F_1@K$	0.070	0.087	**0.117**	**0.127**	**0.161**

Error analysis. To better understand the errors our proposed DemoHash model makes, we analyzed the posts with a high error rate among the test sets with seven hashtag recommendations. We extracted 319 test cases from the bottom 25% of F_1-score. To figure out the reason for the low performance of the given test cases, we assumed two possibilities for these errors as follows: (i) As the demographic module is one of our key module for the model performance, the irrelevance post characteristic with demographic information might affect the recommendation performance. (ii) As one of the key module we used within the model is co-attention network between post text and image, the short length of the post text might incur the errors. Based on these two assumptions, we further scrutinized two possibilities within this section.

First, we analyzed the assumption by manually classifying the extracted erroneous posts into three categories; advertisement, only emoticon, and multilingual. The distribution of each class is as in Fig. 6 where advertisement posts form 5.64% (18), only emoticon posts as 10.03% (32), multilingual posts as 10.97% (35), and the others 73.66% (233) of the erroneous cases. The post with advertisement might shares lack of similarities between demographic information which results in low performance. For the same reason, the post with using only emoticons might lack of features to share the similarities with demographic information. The multilingual posts use post texts and hashtags with languages that cannot be captured from the demographic information. As the demographic information of race does not contain the nationality of the users, this lack of information might result in recommending hashtags with most frequently used ones with given race information. For example, in Fig. 7, the users are tagged with the race as white which results in the proposed model recommending the most frequently used hashtags with white races (e.g., blonde, selfie). Furthermore, Table 7 presented the languages of the multilingual posts.

To figure out the plausibility of the second assumption, we employed word count of error rate top 25% posts as shown in Fig. 8. The posts with the number of words under 25 takes the 69.90% of all the error cases (223 out of 319). Moreover, the average of the word

Table 6
The evaluation metrics of performance gain analysis.

top-K	DemoHash			DemoHash_{t+i}			DemoHash_{t+d}			DemoHash_{i+d}		
	P@K	R@K	$F_1@K$	P@K	R@K	$F_1@K$	P@K	R@K	$F_1@K$	P@K	R@K	$F_1@K$
1	**0.614**	**0.041**	**0.070**	0.606	0.038	0.066	0.611	0.039	0.067	0.574	0.037	0.063
2	0.469	0.053	0.087	0.469	0.053	0.086	**0.677**	**0.108**	**0.155**	0.507	0.065	0.100
3	0.454	0.079	0.117	0.400	0.064	0.099	**0.462**	**0.080**	**0.119**	0.429	0.077	0.114
4	0.400	0.087	0.127	0.379	0.078	0.117	0.327	0.069	0.102	**0.422**	**0.094**	**0.135**
5	**0.429**	**0.133**	**0.161**	0.347	0.087	0.126	0.299	0.077	0.110	0.410	0.109	0.154

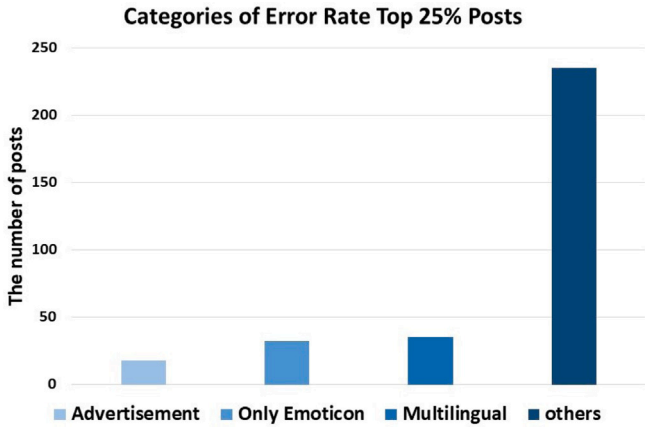

Fig. 6. The distribution of the categories of error rate top 25% posts.

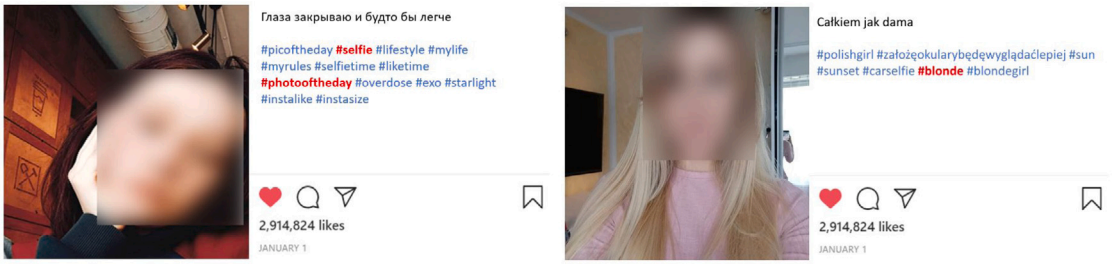

Fig. 7. Examples of the DemoHash model with Russian and Polish; Red colored hashtags are recommended hashtags; Only the frequently used hashtags were recommended: '#selfie', '#photooftheday'; Unable to recommend hashtags with Polish: '#zalozeokularybedewygladaclepiej'.

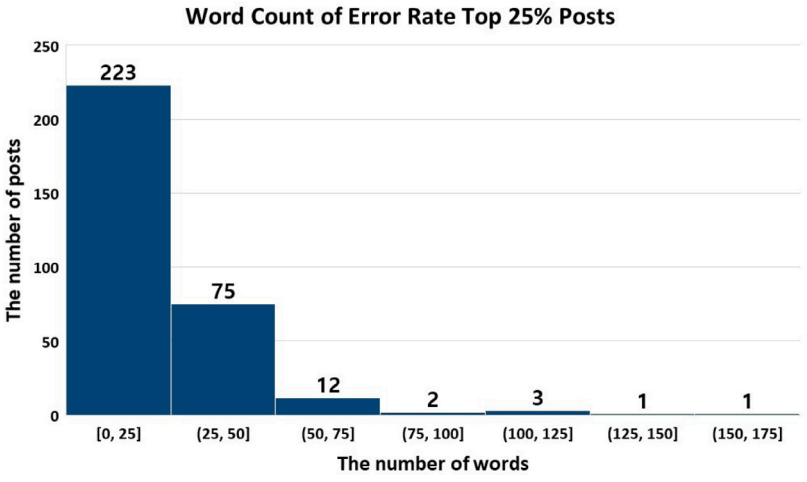

Fig. 8. The distribution of the word count of text from error rate top 25% posts; x-axis indicates the word count of text and y-axis indicates the number of posts.

Table 7
The number of multilingual posts among the error rate top 25% posts in a descending order.

Language	Number of posts (Ratio)
Spanish	9 (25.71%)
Hindi	5 (14.29%)
German	4 (11.43%)
Russian	3 (8.57%)
Persian	3 (8.57%)
Portuguese	3 (8.57%)
Japanese	2 (5.71%)
Indonesian	1 (2.86%)
Thai	1 (2.86%)
Polish	1 (2.86%)
French	1 (2.86%)
Hungarian	1 (2.86%)
Dutch	1 (2.86%)

count with the class of others defined within the first assumption is 14. In addition, 63.52% of the (148 out of 233) class of others have text length under 10. Therefore, these results imply that short length of the text might deteriorate the performance of co-attention network which results in low performance with the proposed model.

6. Conclusion

This paper proposes a new approach for recommending personalized hashtags to users. Our developed DemoHash model employs user demographic information as one of its key components in an attention-based neural network model. Two experimental evaluations with the *Instagram* dataset demonstrate that the DemoHash model achieves higher performance over previous models. This means that DemoHash, a multimodal recommendation approach based on users' demographic information, provides a better understanding and characterization of recommending personalized hashtags. It also indicates that a user's demographic information should be emphasized more in recommending specific hashtags.

We plan to extend the findings of the current study and the structure of the employed model to address multi-lingual hashtag recommendations. Moreover, cultural differences can be one of the potential research areas for recommending user-oriented hashtags. For instance, because several Korean hashtags tend to be employed as a meaning of certain sentences by positioning sequential presentations, future research can employ the results of the current study as an academic foundation.

CRediT authorship contribution statement

Dahye Jeong: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Soyoung Oh:** Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Eunil Park:** Conceptualization, Methodology, Validation, Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

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

Data availability

We share our code and data via https://github.com/dxlabskku/DemoHash.

Acknowledgments

This research was supported by a grant (2022-01) from Gyeonggi-do researcher-centered R&D support project funded by Gyeonggi Province (Gyeonggido Business & Science Accelerator). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (NRF-2020R1C1C1004324).

References

Cantini, R., Marozzo, F., Bruno, G., & Trunfio, P. (2021). Learning sentence-to-hashtags semantic mapping for hashtag recommendation on microblogs. *ACM Transactions on Knowledge Discovery from Data*, *16*(2), 1–26.

Denton, E., Weston, J., Paluri, M., Bourdev, L., & Fergus, R. (2015). User conditional hashtag prediction for images. In *Proc. of KDD '15* (pp. 1731–1740).

Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., & Cohen, W. (2016). Tweet2Vec: Character-based distributed representations for social media. In *Proc. of ACL '16* (pp. 269–274).

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471.

Godin, F., Slavkovikj, V., De Neve, W., Schrauwen, B., & Van de Walle, R. (2013). Using topic models for twitter hashtag recommendation. In *Proc. of WWW '13* (pp. 593–596).

Gong, Y., Ke, Q., Isard, M., & Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, *106*(2), 210–233.

Gong, Y., & Zhang, Q. (2016). Hashtag recommendation using attention-based convolutional neural network. In *Proc. of IJCAI '16* (pp. 2782–2788).

Hwang, S. J., & Grauman, K. (2012). Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International Journal of Computer Vision*, *100*(2), 134–153.

Jang, J. Y., Han, K., Shih, P. C., & Lee, D. (2015). Generation like: Comparative characteristics in instagram. In *Proc. of CHI '15* (pp. 4039–4042).

Kang, J., Kim, J., Shin, S., & Myaeng, S.-H. (2020). A sequence-oblivious generation method for context-aware hashtag recommendation. Retrieved from https://ui.adsabs.harvard.edu/abs/2020arXiv201202957K. (Accessed 13 July 2022).

Kaviani, M., & Rahmani, H. (2020). Emhash: Hashtag recommendation using neural network based on bert embedding. In *Proc. of ICWR '20* (pp. 113–118). IEEE.

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, *10*(1), 1–6.

Lee, S., Jeong, D., & Park, E. (2022). MultiEmo: Multi-task framework for emoji prediction. *Knowledge-Based Systems*, *242*, Article 108437.

Liu, J., He, Z., & Huang, Y. (2018). Hashtag2Vec: Learning hashtag representation with relational hierarchical embedding model. In *Proc. of IJCAI '18* (pp. 3456–3462).

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.

Meta (2022). Instagram data policy. Retrieved from https://help.instagram.com/519522125107875/?maybe_redirect_pol=0.(Accessed 13 July 2022).

Neal, M. (2017). Instagram influencers: The effects of sponsorship on follower engagement with fitness instagram celebrities. Retrieved from https://scholarworks.rit.edu/theses/9654/. (Accessed 13 July 2022).

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*(12), 1565–1567.

Oh, S., Ji, H., Kim, J., Park, E., & del Pobil, A. P. (2022). Deep learning model based on expectation-confirmation theory to predict customer satisfaction in hospitality service. *Information Technology & Tourism*, *24*(1), 109–126.

Rawat, Y. S., & Kankanhalli, M. S. (2016). ConTagNet: Exploiting user context for image tag recommendation. In *Proc. of ACM MM '16* (pp. 1102–1106).

Rish, I., et al. (2001). An empirical study of the naive Bayes classifier. In *Proc. of IJCAI '01 workshop on empirical methods in artificial intelligence* (pp. 41–46).

Serengil, S. I., & Ozpinar, A. (2021). Hyperextended LightFace: A facial attribute analysis framework. In *Proc. of ICEET '21* (pp. 1–4).

Sheldon, P., Rauschnabel, P. A., Antony, M. G., & Car, S. (2017). A cross-cultural comparison of Croatian and American social network sites: Exploring cultural differences in motives for instagram use. *Computers in Human Behavior*, *75*, 643–651.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Retrieved from https://arxiv.org/abs/1409.1556. (Accessed 13 July 2022).

Sun, J., Zhu, M., Jiang, Y., Liu, Y., & Wu, L. (2021). Hierarchical attention model for personalized tag recommendation. *Journal of the Association for Information Science and Technology*, *72*(2), 173–189.

The Korea Times (2020). Passage of data bills. Retrieved from https://www.koreatimes.co.kr/www/opinion/2020/01/202_281848.html. (Accessed 13 July 2022).

Tomorn, R., & Bao, Y. (2020). Thai's fascination WITH food: Race and gender differences in instagram hashtag use. *International Journal of Organizational Innovation*, *12*(3), 190–206.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). Cnn-rnn: A unified framework for multi-label image classification. In *Proc. of CVPR '16* (pp. 2285–2294).

Wei, Y., Cheng, Z., Yu, X., Zhao, Z., Zhu, L., & Nie, L. (2019). Personalized hashtag recommendation for micro-videos. In *Proc. of ACM MM '19* (pp. 1446–1454).

Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., et al. (2014). CNN: Single-label to multi-label. Retrieved from https://arxiv.org/abs/1406.5726. (Accessed 13 July 2022).

Yang, C., Wang, X., & Jiang, B. (2020). Sentiment enhanced multi-modal hashtag recommendation for micro-videos. *IEEE Access*, *8*, 78252–78264.

Yang, Q., Wu, G., Li, Y., Li, R., Gu, X., Deng, H., et al. (2020). Amnn: Attention-based multimodal neural network model for hashtag recommendation. *IEEE Transactions on Computational Social Systems*, *7*(3), 768–779.

Zhang, Q., Wang, J., Huang, H., Huang, X., & Gong, Y. (2017). Hashtag recommendation for multimodal microblog using co-attention network. In *Proc. of IJCAI '17* (pp. 3420–3426).

Zhang, S., Yao, Y., Xu, F., Tong, H., Yan, X., & Lu, J. (2019). Hashtag recommendation for photo sharing services. In *Proc. of AAAI '19* (pp. 5805–5812).

Zhu, S., Aloufi, S., & El Saddik, A. (2015). Utilizing image social clues for automated image tagging. In *Proc. of ICME '15* (pp. 1–6). IEEE.