COOPNet: Multi-modal Cooperation for Improving Gender Prediction in Social Media User Profiling

LIN LI*, Wuhan University of Technology, China
KAIXI HU, Wuhan University of Technology, China
YUNPEI ZHENG, Wuhan University of Technology, China
JIANQUAN LIU, NEC Corporation, Japan
KONG AIK LEE, Institute for Infocomm Research, A*STAR, Singapore

Gender prediction in user profiling is of great significance for pervasive user-based services (e.g. recommendation and search). The principal way of performing user profiling is to investigate accumulated social media data. However, the problem of information asymmetry generally exists in user generated content since users can post various modal contents in social media freely. It is a challenging task to narrow information differences on different modal contents and work together to improve the accuracy of prediction models. To better balance information and facilitate information exchange between different modalities, it is a kind feasible means of establishing a bridge to enhance their cooperation. In this paper, we propose a novel text-image cooperation framework (COOPNet), a bridge connection network architecture that exchanges information between texts and images in social networks. The cooperation of COOPNet is mainly reflected on cooperative representation, cooperative learning and cooperative decision. First, we map the representations of both visual and sentiment enriched textual modalities into a cooperative semantic space to derive a cooperative representation. Second, the representations of texts and images joined with their cooperative representation are all added into the learning process of framework. Finally, a multi-modal regression is leveraged to make a cooperative decision. Extensive experiments on the public PAN-2018 dataset demonstrate the superiority of our framework over the state-of-the-art methods on the premise of automatic feature learning.

CCS Concepts: • **Social and professional topics → User characteristics**; • **Information systems → Web searching and information discovery**; • **Computing methodologies → Supervised learning**.

Additional Key Words and Phrases: Cooperative Representation, Gender Prediction, Multi-modal, User Profiling

1 INTRODUCTION

In social media, user profiling is to digitize users and offers reliable assistance to ubiquitous web services. The key work of user profiling is to analyze user-related data and extract general and representative labels [49]. As one of the basic statistical attribute labels in user profile, gender has long been the top factor considered in various web applications such as recommendation [2, 10, 42], advertisements [13, 30], search [17, 18]. Therefore, to provide more personalized and precise services, accurate and reliable prediction of gender is a necessity for helping enterprises to effectively target their core user groups. The earliest user profile consisted of nothing more than questionnaires [15, 34], telephone interviews [6], face-to-face communications [44]. Nowadays, the abundance of social media activities has attracted lots of users and accumulated amounts of user data, including content data, behavior data and social data. The manual survey method is no longer able to serve such a large number of users. Machine learning methods are coming onto the stage.

*Lin Li is the corresponding author.

Authors' addresses: Lin Li, Wuhan University of Technology, China, cathylilin@whut.edu.cn; Kaixi Hu, Wuhan University of Technology, China, issac_hkx@whut.edu.cn; Yunpei Zheng, Wuhan University of Technology, China, PPgirl87@foxmail.com; Jianquan Liu, NEC Corporation, Japan, jqliu@nec.com; Kong Aik Lee, Institute for Infocomm Research, A*STAR, Singapore, Lee_Kong_Aik@i2r.a-star.edu.sg.

However, it is still very challenging for service providers to acquire high-accurate gender prediction in a big social network due to the asymmetric modality information in user generated content. This is mainly reflected in the following two aspects:

Semantic Complementarity Between Different Modalities. The original methods based on single-modal data [9, 30] have much limited ability to access enough available information. It is a necessity to add other data modalities to enrich semantic information, because they share a high degree of similarity in deep semantic space [38, 48]. However, on account of the diversity of user expression, different data modalities may present different information. As the user A in Figure 1, there is no explicit gender-specific clues can be found from the fourth micro-blog post. We get nothing more than the user A gets a ticket for a match of Ireland versus France, which shows positive sentiment. Similarly, we only know there is a football match from the images of user A and cannot acquire any context information. By taking both texts and images from the user A into consideration, it can be inferred that the user A watched a football match on the scene. However, it is doubly intractable to establish a bridge from two totally different data modalities. Simply concatenating two representations [1, 28] or voting on the results [27] are all not good solutions, which cannot learn the association relationship between texts and images.

Insufficient Information in Unilateral Modality. Despite the general phenomenon of symbiosis or co-occurrence of multi-modal data, there are still some situation that user only sent posts or images. It is an interesting thing that a model is not only capable of making prediction on multi-modal data but also on multiple single-modal data. However, texts are a product of cognitive intelligence which own abundant semantic information. Images are a product of perception intelligence which are usually used to supplement the expression of texts. As the user C in Figure 1, we can acquire enough information to decide the user's gender. If other users who are like user B only send an image, we cannot make any decision. By cooperative learning from different users, image representation component can memory a prior knowledge that users who send an image like that of user C are more likely to be male. Therefore, texts which are regarded as prior knowledge are introduced into images during the learning process. The performance of single-modal image model with poor information can be improved.

Gender prediction in multimodal user profiling heavily focuses on text-image matching relationship. In addition, the semantic understanding of gender such as sentiment preference [49] is indispensable as well, which is different from other general binary classification tasks. In this paper, we aim to balance the semantic information and capture the relevance between two different modalities in user generated content to improve the accuracy of gender prediction. To address the aforementioned challenges, we design a text-image cooperation framework which has the capacity for deep semantic information exchange. *First*, we design a bridge connection network between text presentation embedding sentiment polarity and image presentation component. A mapping matrix will be learned as the major carrier for information migration, which eventually generates a cooperative representation. *Second*, the derived cooperative representation joins with single-modal representations to transfer the knowledge and enhance the representation ability. *Finally*, we employ a secondary classifier to make a multi-modal regression decision.

The main contributions of this work are summarized as follows:

- We propose a text-image cooperation framework COOPNet that is capable of capturing the association relationship and transfer knowledge between visual and textual modalities. In particular, the cooperations of COOPNet mainly consist of cooperative representation, cooperative learning and cooperative decision.

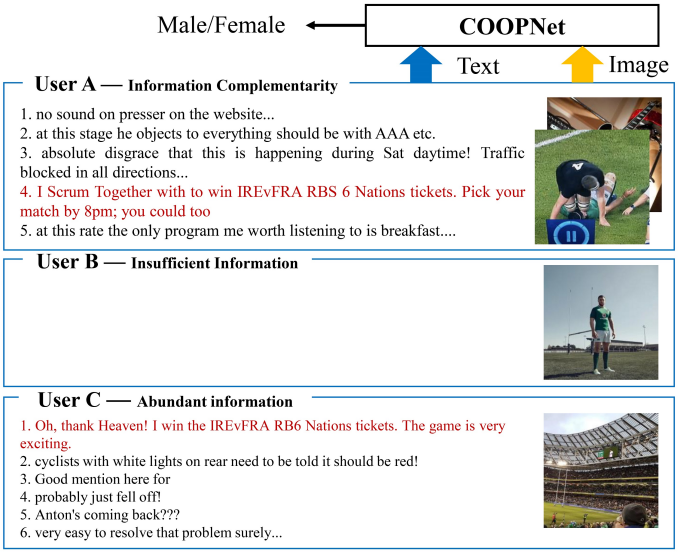


Fig. 1. Some micro-blog posts and corresponding images from users. User A presents the situation of information complementarity. Use B provides insufficient information. User C provides abundant information.

- In the aspect of cooperative representation learning, we design a bridge connection network between different modal representation components that aligns the image semantic information to the text semantic information and jointly maps their aligned semantic representation into the cooperative semantic space.
- We perform extensive experiments on PAN-2018 dataset. Evaluation results demonstrate that the COOPNet outperforms state-of-the-art baselines in terms of accuracy without expert knowledege involved feature engineering.

The rest of paper is organized as follows. We first discuss the related work in Section 2. We introduce the overview of our work in Section 3. The details of our gender prediction framework COOPNet will be described in Section 4. The evaluation results are presented in Section 5. We conclude this paper in Section 6.

2 RELATED WORK

Numerous prior studies are in the perspective of single-modal data, mainly in textual data. Recently, researchers observe that images often co-occur with texts. This mode of expression can present more abundant information. Therefore, the methods combined with texts and images are gradually coming into researchers' view.

Single-modal user profile has attracted a large number of researchers for a long time. This can be reflected in the following three aspects. *Firstly*, various types of data are exploited. Burger et al. [4] investigate a variety of textual data, such as nicknames, full names, self-descriptions, blog posts, etc. They further explore several different classifier types on the dataset. Preotiuc-Pietro et al. [26] conducte on a new annotated corpus of Twitter users. They extract some word clusters and embeddings, and then propose a gaussian process model to predict a user's occupational class. *Secondly*, plentiful features are explored. Volkova et al. [37] learn a log-linear model using lexical features to infer various traits from user communications in social media. Li et al. [19] explore characters and word n-grams on Chinese posts and leverage their embeddings to predict gender. Zhao et al. [47] separate a single user's interest profile into several behavioral features and construct better user profiles. *Thirdly*, diverse model architectures are

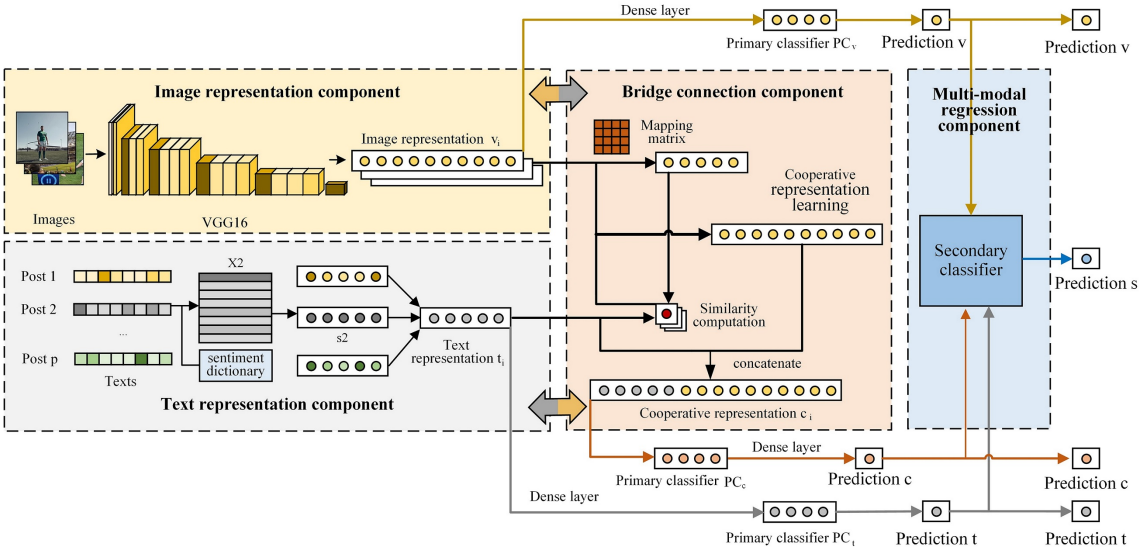


Fig. 2. The text-image cooperation framework (COOPNet)

proposed. Wu et al. [41] develop a general neural framework to jointly capture the user preferences and item attributes. Kang et al. [16] propose a deep-neural-network-based approach for predicting user interests by combining Bi-GRU and CNN models. Wang et al. [39] attempt to make the relevant attribute classification tasks and employ the auxiliary representation to integrate the learning process in a muti-task model. Zhang et al. [45] incorporate multiple textual perspectives to classify users by a ensemble LSTM model.

Multi-modal user profile is a hot topic in both academia and industry. Pardo et al. [24] report some classical fusion method in PAN competition. The problem of multi-modal fusion can be unfolded from the following three aspects. *Firstly*, texts and images are merged in early phase, which captures the correlation between features. Patra et al. [25] convert images to texts by using image caption system. Some researchers also utilize concatenation method but make too little improvement [1, 28]. *Secondly*, results from different modalities are combined in late phase without dependency among data sources. Schaetti et al. [27] apply CNN to each post and RestNet to each image respectively, then calculate the average class probabilities over all instances. Some researchers also exploit the method of average probabilities or voting, but make no more progress [31]. Martinc et al. [21] only use images when several conditions are met. Namely, they process texts and images independently and then integrate them in a relatively simple way. *Thirdly*, the learning process is a joint participation mode. Farnadi et al. [11] merge multimodal user data, such as profile pictures, status updates and page likes. They incorporate correlations among data sources and outcomes from other tasks, such as age, gender and extrovert, which is a multiple targets task.

In summary, we observe that there are deep semantic patterns between texts and their attached images. Some available information can be inferred by jointly learning and information alignment from a higher semantic level. The simple operation of concatenation or vote which most methods adopts is insufficient to capture these patterns. Our work focuses more on the intermediate semantic layer which jointly maps two modalities into a cooperative semantic space to balance the information gap by semantic alignment.

3 OVERVIEW OF OUR WORK

3.1 Problem Description

In this section, we begin with some mathematical notations and then formally present the problem formulation of gender prediction. Assume that we have a set of texts $\boldsymbol{T} = \{T^1, T^2, ..., T^n\}$ and a set of images $\boldsymbol{V} = \{V^1, V^2, ..., V^n\}$ from n users. Each user has a gender label $y_i \in \boldsymbol{Y} = \{y_1, y_2, ..., y_n\}$, which is a binary value. Besides, we consider some extra data, e.g., sentiment dictionary D, pre-trained word embeddings as side information. We refer to an individual text from one user as T_j^i and an individual image as V_j^i, where i, j are defined as the index for the user, text or image respectively.

Gender Prediction. With the aforementioned notations and definitions, the problem of gender prediction can be formally stated as follows: given the text T^i, the image V^i from the same user and the extra sentiment dictionary D which is introduced as a knowledge to enhance the representation of texts. The object of our task is to learn a predictive framework $\boldsymbol{F}$ which infers the user's gender in the real world. Briefly, we attempt to learn a function $\boldsymbol{F}$, which maps a user's texts and images into a scalar $\hat{y}$ in Eq.(1).

$$\hat{y}_i = \boldsymbol{F}(V^i, T^i) \tag{1}$$

In this paper, we adopt the binary cross entropy $\boldsymbol{e}$ to calculate the error rate in Eq.(2).

$$E = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{e}(y_i - \hat{y}_i) \tag{2}$$

3.2 Proposed Framework

The COOPNet framework is a bridge neural network architecture which solves the gender prediction problem formulated above. We present the model architecture in Figure 2, where four components work together. Before presenting the details of our framework, we elaborate the motivations of the model design that attempt to address the challenges mentioned in Section 1.

- To acquire a high-quality and low-cost text representation and image representation, we employ the transfer learning method. We transfer the VGG16 model to generate image representation, the sentiment and self-attention enhanced model proposed in this paper to generate text representation and then they are all fed into the bridge connection component. Note that we introduce a sentiment dictionary and a self-attention mechanism to the hierarchical text representation model which learns the relationship between posts from one user.
- To eliminate the semantic gap between texts and images and stimulate the exchange of information, we design a bridge connection network that employs the weight vector as a medium to generate the cooperative representation. Additionally, COOPNet employs three different modal representations to generate three different prediction results and then add them into the loss function to provide power for information exchange.
- To further enhance information exchange, we exploit a multi-modal regression classifier to jointly decide the final gender. Meanwhile, the primary classifier is not affected which still has the ability for prediction in single-modal data.

4 TECHNICAL DETAILS

In this section, we present the details of our framework. COOPNet consists of four major components: text representation, image representation, bridge connection and multi-modal regression. We explain these four components in the following subsections.

4.1 Text Representation

In a large social network, a user usually posts more than one blog posts. A general method is to integrate them into a big virtual document. As shown in the lower left gray block of Figure 2, our text representation component regards all posts from one user as a input and encode them into a text representation vector. Inspired by HAN (Hierarchical Attention Networks) model [43], we propose to adopt a sentiment and self-attention [35] enhanced text representation model to solve the one-to-many situation. In particular, we embed sentiment into the word representation, since female's sentiment is more subjective and obvious than male [14, 46]. Various approaches are developed to achieve sentiment embedding, e.g. extra corpus [40, 49], sentiment dictionary [5, 20]. As the sentiment representation largely depends on the quality of extra corpus, we concatenate the word representations with its corresponding sentiment polarity in sentiment dictionary to relax our model from heavy corpus learning and focus on the effect of sentiment.

The details of our text representation component are shown in Figure 3, which presents the double encoding process of posts. In the word encoder, a pre-trained word embedding is introduced to derive a word embedding matrix $\boldsymbol{X}$ of each post. We use the word embedding matrix $\boldsymbol{X}$ to calculate the query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$ and the value matrix $\boldsymbol{V}$ in Eq.(3). Additionally, we concatenated the hidden state of each word with its sentiment polarity as the input of GRU (Gate Recurrent Unit). Similar to the word encoder, we regard a whole post as one sentence and fed into the post encoder after a Dense layer. Finally, we derive the text single-modal representation t^i, which is the input of primary classifier PC_t and bridge connection component.

$$\begin{aligned} &Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \\ &Q = XW^Q,\ K = XW^K,\ V = XW^V \end{aligned} \tag{3}$$

where d_k represents the dimension of key queries in the key matrix, W^* represents the parameter matrix of Q, K and V respectively.

4.2 Image Representation

The image representation from one user is also a one-to-many relationship. As shown in the upper left yellow block of Figure 2, our image representation component encodes all images from one user and transform them into image representation space. This work is different from the image caption system [36]. Concretely, we extract the hidden state in the fifth block of VGG16 model [29] and employ a full connection layer to transform the dimension of hidden state into that of image semantic space. We formally define the process in Eq.(4). The representation of each image from one user is further fed into the bridge connection component in turn to align to the text representation. In the process of back propagation, the parameters of VGG16 and full connection layer will be fine-tuned. The mean v^i_{pri} of all images is computed as input of the primary classifier PC_v in Eq.(5).

$$v^i_j = Dense_v(VGG16(V^i_j)) \tag{4}$$

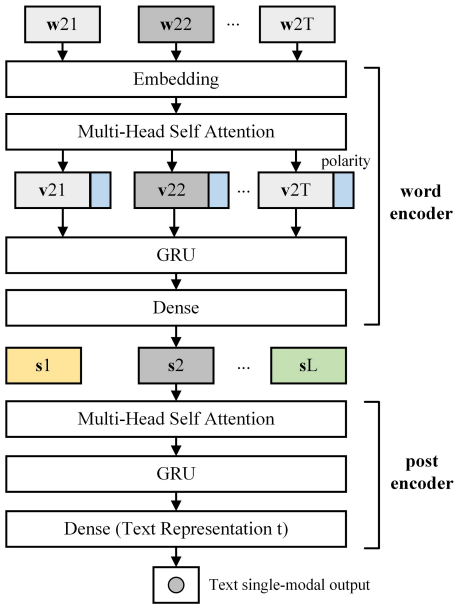


Fig. 3. Text representation component.

$$v_{pri}^{i} = \frac{1}{m}\sum_{j=1}^{m} v_{j}^{i} \tag{5}$$

where m represents the number of images from user i, $VGG16(\cdot)$ represents the VGG16 model, $Dense(\cdot)$ represents the full connection layer.

4.3 Bridge Connection

As shown in the middle red block of Figure 2, the representations of text and images are all fed into the bridge connection component and generate a cooperative representation after semantic alignment. One limitation of capturing the correlation relationship between different data modalities lies in that the information differences impedes the exchange of deep semantic information. In fact, textual information is more pure than visual information which contains other irrevant information. To overcome this limitation, we develop a bridge connection network to filter irrelevant image information and allow the different representation component modelling the semantic relationshop. In particular, representations of two modalities are jointly mapped into the cooperative semantic space via semantic alignment, where the semantic patterns between texts and images are captured and the semantic relationship can be smoothly transferred to the modality with insufficient information .

The visualization of semantic capturing is present in Figure 4. Firstly, each image representation is mapped into the text representation space by a mapping matrix W_m as shown in Eq.(6) and then the semantic relevance between them is calculated in Eq.(7). This is the first step of semantic alignment where the text freely selects the most semantically relevant images in text representation space and some irrelevant information is filtered. Secondly, the weight vector is used to weight the multiple image representations and an intermediate image representation v_c^i is generated as shown in Eq.(8). Finally, the intermediate image representation v_c^i is concatenated with the text representation t^i and the

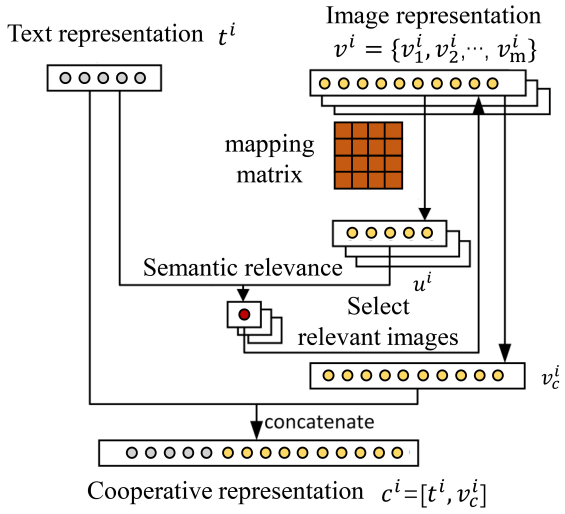


Fig. 4. The bridge connection component.

cooperative representation c^i which conains the semantic relationship between texts and images is derived. Formally, the cooperative representation process is presented as follows:

$$u_j^i = \tanh(W_m v_j^i + b_m) \tag{6}$$

$$\alpha_j^i = \frac{\exp({t^i}^T u_j^i)}{\sum_{j=1}^{n} \exp({t^i}^T u_j^i)} \tag{7}$$

$$v_c^i = \sum_{j=1}^{m} \alpha_j^i v_j^i \tag{8}$$

$$c^i = [t^i, v_c^i] \tag{9}$$

In above equations, W_m is a mapping matrix which maps each image representation into text space and its dimension depends on the representation dimension of texts and images. Specially, $t^i \in \mathbb{R}^{d_t}$ and d_t is the dimension of text representation space. v^i is a set of image representations from user i. $v_j^i \in \mathbb{R}^{d_v}$ and d_v is the dimension of image representation space. Thus, $W_m \in \mathbb{R}^{d_t \times d_v}$ and b_m is the bias. u_i is the mapped image representation. α_i is the softmax similarity between t^i and u_i. $\alpha_i \in (0, 1)$ and $\sum_n \alpha_i = 1$.

In this component,we leverage the text representation as a learning guidance to select a set of images, which is also applicable in reverse. Finally, we feed the cooperative representation as input into a primary classifier to predict user's gender. It is worth noting that we may only have either texts or images in many situations. Apart from the cooperative prediction result, outputs of single-modal representation also can be obtained as shown in Figure 2.

4.4 Multi-modal Regression

While the bridge connection component exchanges information between text and image representation component, we can acquire three predictive probabilities from their own primary classifiers each of which is a three-layers perception

Algorithm 1 The Cooperative Learning Process for COOPNet

Input: texts: $T = \{T^1, T^2, ..., T^n\}$;
images: $V = \{V^1, V^2, ..., V^n\}$;
true gender label: $y = \{y^1, y^2, ..., y^n\}$;
sentiment dictionary: D;
Output: text prediction label: $\hat{y_t} = \{\hat{y}_t^1, \hat{y}_t^2, ..., \hat{y}_t^n\}$;
image prediction label: $\hat{y_v} = \{\hat{y}_v^1, \hat{y}_v^2, ..., \hat{y}_v^n\}$;
cooperative prediction label: $\hat{y_c} = \{\hat{y}_c^1, \hat{y}_c^2, ..., \hat{y}_c^n\}$;
final prediction label: $\hat{y} = \{\hat{y}^1, \hat{y}^2, ..., \hat{y}^n\}$;

Load the parameters of text representation component $\boldsymbol{f_t}$ and image representation component $\boldsymbol{f_v}$;
repeat
  for i = 1 to n **do** ▷ n is the number of users
    $t^i = \boldsymbol{f_t}(T^i)$; ▷ text representation
    $\hat{y}_t^i = \boldsymbol{PC_t}(t^i)$; ▷ primary text prediction
    $v^i = \boldsymbol{f_v}(V^i)$; ▷ image representation
    $v_{pri}^i = \frac{1}{m}\sum v^i$; ▷ m is the number of images
    $\hat{y}_v^i = \boldsymbol{PC_v}(v_{pri}^i)$; ▷ primary image prediction
    $c^i = \boldsymbol{f_c}(t^i, v^i)$; ▷ cooperative representation
    $\hat{y}_c^i = \boldsymbol{PC_c}(c^i)$; ▷ primary cooperative prediction
  end for
  $Loss1 = \frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{e}(y_t^i, \hat{y}_t^i) + \boldsymbol{e}(y_v^i, \hat{y}_v^i) + \boldsymbol{e}(y_c^i, \hat{y}_c^i)]$
  Loss1.backward()
until minimize Loss1
repeat
  for i = 1 to n **do**
    $\hat{y}^i = \boldsymbol{SC}(\hat{y}_t^i, \hat{y}_v^i, \hat{y}_c^i)$ ▷ multi-modal regression
  end for
  $Loss2 = \frac{1}{n}\sum \boldsymbol{e}(y^i, \hat{y}^i)$
  Loss2.backward()
until minimize Loss2

(MLP). A straightforward idea is that we can enhance the ability of information exchange by combining the results from different modalities. As shown in the rightmost blue block of Figure 2, the prediction results of primary classifiers are all fed into the multi-modal regression component and generate the final result. In this work, a logarithmic probability regression [7] is employed as the secondary classifier to achieve multi-modal regression and derive the final result. The multi-modal regression can be represented as Eq.(10).

$$\hat{y}^i = \frac{1}{1 + e^{-(w_t^T \hat{y}_t^i + w_v^T \hat{y}_v^i + w_c^T \hat{y}_c^i + b_c)}} \tag{10}$$

where $\hat{y}_t$, $\hat{y}_v$ and $\hat{y}_c$ are the gender prediction results of user i from text representation component, image representation component and bridge connection component respectively. w_t, w_v and w_c are the weight metrics. b_c is the bias.

4.5 Cooperative Learning Process

In this subsection, we describe the learning process of our COOPNet in Algorithm 1. According to the transfer learning theory [23], we first save the optimal parameters of single-modal model and load them into the multi-modal model

Table 1. Technical details in different multi-modal baselines.

Model	Text	Image	Fusion(Early/Late)	Feature Engineering
Ciccone et al. [8]	N-grams, TF-IDF	6 classifiers	LinearSVC(Late)	Y
Nieuwenhuis et al. [22]	N-grams, Glolve	13 features	LR(Early)	Y
Tellz et al. [33]	N-grams	Image caption	Weighted average(Late)	Y
Aragon et al. [1]	N-grams	VGG16	SVM(Early)	Y
Sierra et al. [28]	Bag-of-Words	ResNet50, VGG16	Concat(Early)	Y
Kerner et al. [12]	N-grams, Stylistic features	SIFT, Color, VGG	Weighted average(Late)	Y
Martinc et al. [21]	CNN, TF-IDF	Face detection	three conditions(Late)	Y
Stout et al. [31]	TF-IDF, N-gram, RNN	CNN, Pool	Weighted average(Late)	Y
Patra et al. [25]	Word2vec, TF-IDF, LSA, LDA	Image caption	SVM(Early)	Y
*Takahashi et al. [32]	RNN	VGG16	Direct-product(Early)	N
Schaetti et al. [27]	Character based CNN	ResNet18	Average probabilities(Late)	N
COOPNet(ours)	Self-attention, GRU	VGG16	Bridge(Early), LR(Late)	N

as the initial parameter (line 1). On the basis of initialized single-modal model, we further fine-tune the single modal representation (lines 4-7) and train the bridge connection network (lines 8-10). Finally, we use the prediction results of primary classifiers as training data to optimize the secondary classifier (lines 15-21). In this work, we use Stochastic Gradient Descent (SGD) [3] to learn the parameters of COOPNet. We define our loss function of the first training phase in Eq.(11) and the fusion training phase in Eq.(12).

$$Loss1 = \frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{e}(y_t^i, \hat{y}_t^i) + \boldsymbol{e}(y_v^i, \hat{y}_v^i) + \boldsymbol{e}(y_c^i, \hat{y}_c^i)] \tag{11}$$

$$Loss2 = \frac{1}{n}\sum \boldsymbol{e}(y^i, \hat{y}^i) \tag{12}$$

5 EVALUATION

In this section,we conduct experiments to evaluate the performance of COOPNet on PAN-2018 dataset. In particular, we aim to answer the following questions:

- **Q1:** How does our COOPNet framework perform compared with the state-of-the-art multi-modal gender prediction models?
- **Q2:** Are there any information difference between texts and images?
- **Q3:** How significantly bridge connection network contributed to the exchange of information from different modal data?
- **Q4:** What is the role of the multi-modal regression decision component?
- **Q5:** Does the sentiment have great impact on text representation?

5.1 Experimental Setup

5.1.1 Data. We use the dataset from PAN-2018 competition [1]. Users are grouped by the language of their tweets: English, Arabic and Spanish. As baselines keep consistent performance in different languages, we choose the universal English dataset for our experiments. There are 4900 users in total. For each user, a total of 100 tweets and 10 images are

[1]https://pan.webis.de/clef18/pan18-web/author-profiling.html

Table 2. Comprehensive accuracy comparison between COOPNet and Ciccone in prediction phase. Ciccone's model is based on traditional feature engineering and ours is based on automatic feature learning.

Model	Image input	Text input	Multi-modal input
COOPNet(ours)	80.63	80.73	81.05
Ciccone et al.	69.63	80.74	81.32

provided. The training set contains 3000 samples and test set contains 1900 samples. The male-to-female ratio is 1:1 in both of them.

According to preliminary statistics, most every user owns 100 posts and 10 images in PAN-2018. Small number of users have less than 10 images or even none. In this work, we use the zero matrix to fill up the missing image and resize all images into the size of (150,150,3). Besides, we limit the maximum length of a post to 50 and the word embedding dimension to 300. Similar to Takahashi [32], we average the RGB value of all images to normalize the images further. In our work, the word embeddings are derived by the model itself instead of pre-trained word embeddings.

5.1.2 Parameter Settings. We implement our framework based on TensorFlow and use SGD as our optimizer to learn the model parameters. In this paper, we set learning rate as 0.01, the head number of self-attention as 2, the dimension of text representation as 64, image representation as 1000 and mapping matrix as 1000×64.

5.1.3 Baselines. To objectively evaluate the effectiveness of our model, we compare COOPNet with the state-of-the-art baselines. Pardo al et. [24] summarize some top methods and results in the PAN-2018 competition. We list all of them and describe their technical details in Table 1. These methods can be boiled down to early fusion and late fusion. The majority of them depend on traditional feature engineering which introduces expert knowledege. Takahashi et al. [32] is the champion in the PAN-2018 competition whose method is also an end-to-end model. To make a fair comparison, we reproduce their original neural network model and strip out extra streaming tweets used for pre-trained word embeddings.

Additionly, we also add some variants of COOPNet to get a better understanding of the proposed framework and evaluate the key components. COOPNet-T, COOPNet-V, COOPNet-C represents the whole multi-modal COOPNet model without multi-modal regression component. These simplified models are trained in multi-modal way by minimizing Eq. (11) and capable of providing single-modal predictive results. In particular, COOPNet-T, COOPNet-V, COOPNet-C generates the gender prediction result from the primary three-layers MLP classifier added to the text representation, image representation and cooperative representation respectively. Additionally, COOPNet-OT, COOPNet-OV represents the single-modal model of text and image representation component which are trained by minimizing only their corresponding error terms. COOPNet-OTS denotes the text representation component without sentiment.

5.1.4 Evaluation Metrics. In this paper, we adopt accuracy to evaluate performance between COOPNet and the baselines. AUC (Area Under The Curve) and F1-Measure are also utilized to analyze various variants of COOPNet.

5.2 Performance Comparison (Q1)

To investigate the performance of all compared baselines on multi-modal prediction, we use different colors to distinguish whether expert knowledege is introduced and show their accuracy in Figure 5. In addition, we compare COOPNet with Ciccone et al.'s model to analyze the potential of our model further. We have the following key observations.

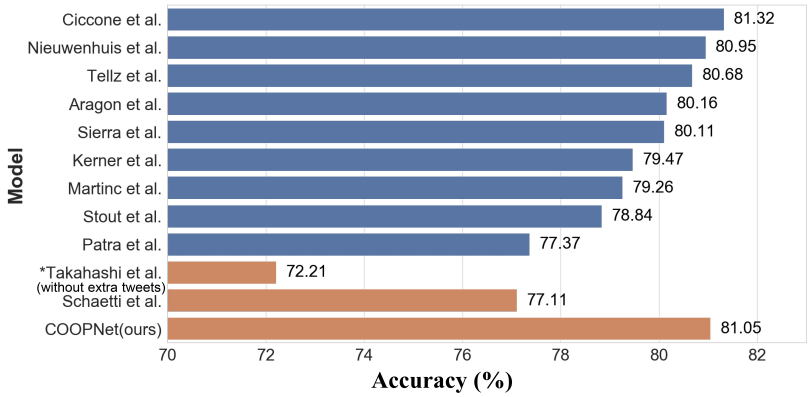


Fig. 5. Accuracy comparison between our model and the baselines in multi-modal prediction. The blue bars present traditional feature engineering based methods and the orange bars present the automatic feature learning based methods which reduce human effort and are convenient for engineering application.

COOPNet shows promissing performance. In Figure 5, we observe that our COOPNet far outperforms other baselines by comparing orange bars of automatic feature learning based methods. For example, COOPNet achieves relatively 12.24% and 5.11% improvements over Takahashi et al.'s and Schaetti et al.'s model respectively. Particularly, such automatic feature learning based methods is easier to achieve a good result by end-to-end training instead of manual human effort, which reduces the complexity of the engineering. COOPNet is significantly superior to Takahashi et al.'s model which ranks first in the competition if the extra tweets is stripped out. It proves that COOPNet reduces dependence on extra data through a better modalities exchange and make up the lack of extra knowledge on some degree. Compared to traditional feature engineering based methods, COOPNet outperforms most of them even if they introduce lots of expert knowledge.

COOPNet achieves great improvement in visual modality. To further analyze the difference between COOPNet and Ciccone et al.'s model, we compare the performance of different modalities in Table 2. We observe that the multi-modal accuracy of COOPNet is close to the single-modal results. Compared to Ciccone et al.'s result, COOPNet has the absolute predominance of image-based result. The gap between text-based results is so tiny that we can ignore it. The observation demonstrates that the information difference between text representation and image representation is much small in COOPNet as the bridge connection network is capable of transferring knowledge. However, Ciccone et al.'s text representation owns more distinctive information on account of the introduction of expert knowledge. Their multi-modal result achieves an obvious improvement. Therefore, our COOPNet is capable of improving the performance of single modal input and making it close to that of multi-modal input, which is not present in other previous approaches.

5.3 Asymmetric Modality Information (Q2)

To analyze the information difference between texts and images, we show the performance of different modal models in Table 3. We observe that multi-modal method outperforms single-modal methods. Text-based model is better than image-based model in various evaluation metrics. Therefore, the fusion of different modalities can enhance the information and perform semantic complementarity. In line with the view from Martinc el al. [21], it is harder to predict user's gender from images due to some haphazard contents in images such as landscape and animals, which causes insufficient

Table 3. Different evaluation metrics on the variants of COOPNet.

Variants	Accuracy(%)	AUC	F1_male	F1_female
COOPNet-OT[1]	79.31	0.8301	0.7800	0.8047
COOPNet-OV[1]	75.57	0.7757	0.7775	0.7292
COOPNet-OTS[2]	77.57	0.8112	0.7857	0.7649
COOPNet-T	80.73	0.8696	0.8000	0.8142
COOPNet-V	80.63	0.8486	0.8015	0.8108
COOPNet-C	80.57	0.8718	0.7964	0.8142
COOPNet	**81.05**	**0.8730**	**0.8058**	**0.8150**

[1] The single-modal model of only text(OT) or image(OV).
[2] The COOPNet-OT without sentiment.

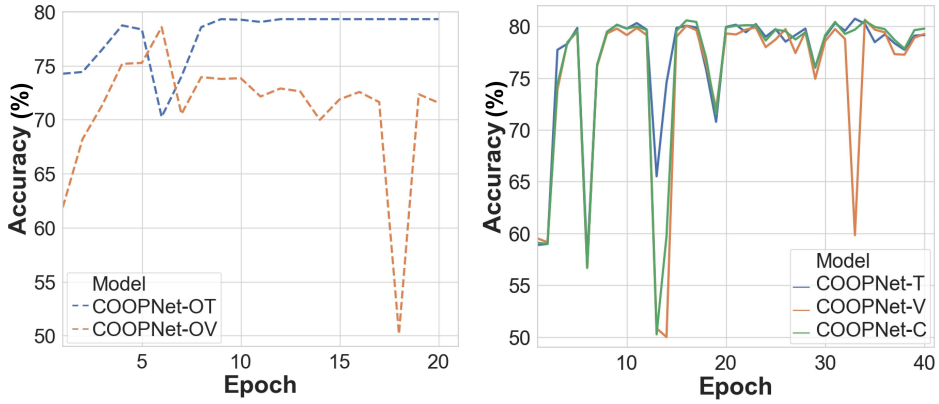
information. Therefore, an appropriate image selection method has a large effect on the gender prediction result which is capable of filtering irrelevant information. Our bridge connection network has the ability of information selection by computing the relevance.

5.4 Evaluations on Variants of COOPNet (Q3 and Q4)

In this subsection, we attempt to conduct some ablation studies to get a better understanding of the key components of COOPNet.

The bridge connection network works well. To analyze the affection of bridge connection network, we compare the results of primary classifiers in multi-modal model to that of single-modal models. The results among primary classifiers are also discussed. In Table 3, we observe that COOPNet-T, COOPNet-V achieves relatively 1.8% and 6.7% accuracy improvements over COOPNet-OT, COOPNet-OV. The improvement of image prediction is quite obvious. The prediction accuracy of COOPNet-T, COOPNet-V and COOPNet-C are almost the same. Meanwhile, in Figure 6(a), we observe that the blue dash line derives a great accuracy and remains relatively steady in the late phase. The orange dash line of COOPNet-OV is clearly below the blue of COOPNet-OT and falls into the valley in the late epoch phase. In Figure 6(b), the orange solid line of COOPNet-V has almost the same behaviors with the blue of COOPNet-T. The green solid line of COOPNet-C also keeps pace with the primary classifiers of texts and images. The behaviors of three solid lines tend to be homogeneous. They are easier to fall into the valley in the early phase and relatively steady in the late phase. The above observations reveal that texts and images become complementary knowledge of each other. Through our bridge connection network, lots of text knowledges are transferred to image representation component and relatively few image knowledges are transferred to text representation. The performance of image representation component is a little unsteady during the learning phase. When exchanging information at early cooperative learning phase, the representation components of text and bridge connection are easy to be affected by images and fall into the valley. This phenomenon disappears at the late phase and the fluctuation range of images also decreases. Hence, the ability of information transfer of our bridge connection network is effective and the comparison results are in line with our prior knowledge which texts have more knowledge about gender.

The multi-modal regression enhances the information exchange. In Table 3, we observe that COOPNet outperforms three primary classifiers in different evaluation metrics. The three primary classifiers from different backgrounds still keep their own characteristics even if the bridge connection network makes them learn from each



(a) Single-modal model (b) Multi-modal model

Fig. 6. Accuracy in different epochs.

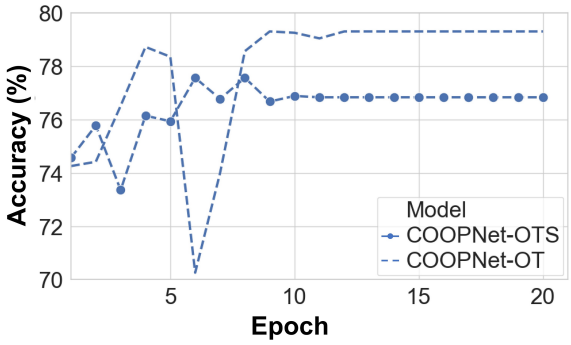


Fig. 7. The study of sentiment in different epochs.

other. Therefore, the secondary classifier in multi-modal regression component can further stimulate the exchange of knowledge of user gender prediction.

5.5 Impact of Sentiment (Q5)

In addition to the above analysis of components, we further discuss the effect of sentiment in texts and focus on the text representation component instead of the whole COOPNet to eliminate the influence of other components. In Figure 7, we observe that the blue dash line without circlse is above the one with circles in the stable phase. And the accuracy of COOPNet-OT achieves 2.24% gain over COOPNet-OTS. When we switch male to female as the positive class to compute the F1 value, the performance of two models arises a small shock in Table 3. This phenomenon denotes that male are neutral in sentiment, which is in line with Zhang et al. [46].

5.6 Runtime Analysis

We conduct our experiments in Centos 7.6 with TITAN Xp. In the training phase, COOPNet spends 45 minutes in each epoch on average and steps into the stable phase after 16 epochs. COOPNet-OT spends about 3 minutes in each epoch

and steps into the stable phase after 12 epochs. On account of the complex contents and insufficient information in images, COOPNet-OV spends about 30 minutes in each epoch and steps into the stable phase after 20 epochs. Overall, COOPNet spends more than 1.5 hour for multi-modal cooperation than single-modal component. The observation demonstrates that COOPNet does not take much time to align images to texts in semantics and achieves a remarkable trade off between performance and time cost.

6 CONCLUSION AND FUTURE WORK

Gender prediction is a challenging and important task in multi-modal user profiling. Knowledge exchanging and semantic association relationship capturing are two hard and vital problems for multi-modal prediction model. In this paper, a novel bridge connection architecture is explored to explicitly model the semantic relationship between different data modalities and a gender prediction framework COOPNet is further proposed to stimulate the exchange of knowledge. The framework is evaluated on PAN-2018 dataset. The results demonstrate that our framework outperforms the state-of-art methods without expert knowledege involved feature engineering. COOPNet is a general framework to facilitate the exchange of knowledge and capture the semantic association relationship. We will apply the framework to a much broader set of user profile, such as age, location and preference.

REFERENCES

[1] Mario Ezra Aragón and Adrián Pastor López-Monroy. 2018. A Straightforward Multimodal Approach for Author Profiling: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[2] Ricardo Mitollo Bertani, Reinaldo A. C. Bianchi, and Anna Helena Reali Costa. 2020. Combining novelty and popularity on personalised recommendations via user profile learning. *Expert Syst. Appl.* 146 (2020), 113149.

[3] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.

[4] John D. Burger, John C. Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *EMNLP*. ACL, 1301–1309.

[5] Yi Cai, Kai Yang, Dongping Huang, Zikai Zhou, Xue Lei, Haoran Xie, and Tak-Lam Wong. 2019. A hybrid model for opinion mining based on domain sentiment dictionary. *Int. J. Machine Learning & Cybernetics* 10, 8 (2019), 2131–2142.

[6] Thomas TW Chiu and Arran SL Leung. 2006. Neck pain in Hong Kong: a telephone survey on prevalence, consequences, and risk groups. *Spine* 31, 16 (2006), E540–E544.

[7] Ronald Christensen. 2006. *Log-linear models and logistic regression*. Springer Science & Business Media.

[8] Giovanni Ciccone, Arthur Sultan, Léa Laporte, Elöd Egyed-Zsigmond, Alaa Alhamzeh, and Michael Granitzer. 2018. Stacked Gender Prediction from Tweet Texts and Images: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[9] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Luis Redondo-Expósito. 2020. Automatic construction of multi-faceted user profiles using text clustering and its application to expert recommendation and filtering problems. *Knowl. Based Syst.* 190 (2020), 105337.

[10] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *RecSys*. ACM, 242–250.

[11] Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User Profiling through Deep Multimodal Fusion. In *WSDM*. ACM, 171–179.

[12] Yaakov HaCohen-Kerner, Yair Yigal, Elyashiv Shayovitz, Daniel Miller, and Toby P. Breckon. 2018. Author Profiling: Gender Prediction from Tweets and Images: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[13] Sean F. Higgins, Maurice D. Mulvenna, Raymond Bond, Aodheen McCartan, Stephen Gallagher, and Darren Quinn. 2018. Multivariate Testing Confirms the Effect of Age-Gender Congruence on Click-Through Rates from Online Social Network Digital Advertisements. *Cyberpsychology Behav. Soc. Netw.* 21, 10 (2018), 646–654.

[14] Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, and Teruo Higashino. 2013. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems* 51 (2013), 35–47.

[15] Ion Juvina and Herre Van Oostendorp. 2004. Predicting user preferences: from semantic to pragmatic metrics of Web navigation behavior. In *Proceedings of the conference on Dutch directions in HCI*. 10.

[16] Jaeyong Kang, Hongseok Choi, and Hyunju Lee. 2019. Deep recurrent convolutional networks for inferring user interests from social media. *J. Intell. Inf. Syst.* 52, 1 (2019), 191–209.

[17] Gaurav Khatwani and Praveen Ranjan Srivastava. 2017. Modeling Gender based Customer Preferences of Information Search Channels. *JGIM* 25, 2 (2017), 52–67.

[18] Ritesh Kumar, Bhanodai Guggilla, and Rajendra Pamula. 2019. Book search using social information, user profiles and query expansion with Pseudo Relevance Feedback. *Appl. Intell.* 49, 6 (2019), 2178–2200.

[19] Wen Li and Markus Dickinson. 2017. Gender Prediction for Chinese Social Media Data. In *RANLP*. INCOMA Ltd., 438–445.

[20] Kui Lu and Jiesheng Wu. 2019. Sentiment analysis of film review texts based on sentiment dictionary and SVM. *ACM International Conference Proceeding Series* Part F148152, 73 – 77.

[21] Matej Martinc, Blaz Skrlj, and Senja Pollak. 2018. Multilingual Gender Classification with Multi-view Deep Learning: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[22] Moniek Nieuwenhuis and Jeroen Wilkens. 2018. Twitter Text and Image Gender Classification with a Logistic Regression N-Gram Model: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[23] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. 2011. Domain Adaptation via Transfer Component Analysis. *IEEE Trans. Neural Networks* 22, 2 (2011), 199–210.

[24] Francisco M. Rangel Pardo, Paolo Rosso, Manuel Montes-y-Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[25] Braja Gopal Patra, Kumar Gourav Das, and Dipankar Das. 2018. Multimodal Author Profiling for Twitter: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[26] Daniel Preotiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *ACL (1)*. The Association for Computer Linguistics, 1754–1764.

[27] Nils Schaetti. 2018. Character-based Convolutional Neural Network and ResNet18 for Twitter Author Profiling: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[28] Sebastián Sierra and Fabio A. González. 2018. Combining Textual and Visual Representations for Multimodal Author Profiling: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[29] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[30] Atakan Simsek and Pinar Karagoz. 2020. Wikipedia enriched advertisement recommendation for microblogs by using sentiment enhanced user profiles. *J. Intell. Inf. Syst.* 54, 2 (2020), 245–269.

[31] Luka Stout, Robert Musters, and Chris Pool. 2018. Author Profiling based on Text and Images: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[32] Takumi Takahashi, Takuji Tahara, Koki Nagatani, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. 2018. Text and Image Synergy with Feature Cross Technique for Gender Identification: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[33] Eric Sadit Tellez, Sabino Miranda-Jiménez, Daniela Moctezuma, Mario Graff, Vladimir Salgado, and José Ortiz-Bejar. 2018. Gender Identification through Multi-modal Tweet Analysis using MicroTC and Bag of Visual Words: Notebook for PAN at CLEF 2018. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 2125. CEUR-WS.org.

[34] Gholamreza Torkzadeh and Jungwoo Lee. 2003. Measures of perceived end-user computing skills. *Information & Management* 40, 7 (2003), 607–615.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. IEEE Computer Society, 3156–3164.

[37] Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring Latent User Properties from Texts Published in Social Media. In *AAAI*. AAAI Press, 4296–4297.

[38] Cheng Wang, Haojin Yang, and Christoph Meinel. 2016. A deep semantic framework for multimodal representation learning. *Multimedia Tools Appl.* 75, 15 (2016), 9255–9276.

[39] Jingjing Wang, Shoushan Li, and Guodong Zhou. 2017. Joint Learning on Relevant User Attributes in Micro-blog. In *IJCAI*. ijcai.org, 4130–4136.

[40] Lu Wang and Claire Cardie. 2014. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. In *ACL (2)*. The Association for Computer Linguistics, 693–699.

[41] Peizhi Wu, Yi Tu, Zhenglu Yang, Adam Jatowt, and Masato Odagaki. 2018. Deep Modeling of the Evolution of User Preferences and Item Attributes in Dynamic Social Networks. In *WWW (Companion Volume)*. ACM, 115–116.

[42] Zhenxing Xu, Ling Chen, Haodong Guo, Mingqi Lv, and Gencai Chen. 2018. User similarity-based gender-aware travel location recommendation by mining geotagged photos. *IJES* 10, 5 (2018), 356–365.

[43] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *HLT-NAACL*. The Association for Computational Linguistics, 1480–1489.

[44] Sevgi Yilmaz, Murat Zengin, and Nalan Demircioglu Yildiz. 2007. Determination of user profile at city parks: A sample from Turkey. *Building and Environment* 42, 6 (2007), 2325–2332.

[45] Dong Zhang, Shoushan Li, Hongling Wang, and Guodong Zhou. 2016. User Classification with Multiple Textual Perspectives. In *COLING*. ACL, 2112–2121.

[46] Yulei Zhang, Yan Dang, and Hsinchun Chen. 2013. Research note: Examining gender emotional differences in Web forum communication. *Decis. Support Syst.* 55, 3 (2013), 851–860.

[47] Zhe Zhao, Zhiyuan Cheng, Lichan Hong, and Ed Huai-hsin Chi. 2015. Improving User Topic Interest Profiles by Behavior Factorization. In *WWW*. ACM, 1406–1416.

[48] Ziyuan Zhao, Huiying Zhu, Zehao Xue, Zhao Liu, Jing Tian, Matthew Chin Heng Chua, and Maofu Liu. 2019. An image-text consistency driven multimodal sentiment analysis approach for social media. *Inf. Process. Manag.* 56, 6 (2019).

[49] Yunpei Zheng, Lin Li, Jianwei Zhang, Qing Xie, and Luo Zhong. 2019. Using Sentiment Representation Learning to Enhance Gender Classification for User Profiling. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 3–11.