Knowledge-Based Systems 178 (2019) 61–73

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Visual-textual sentiment classification with bi-directional multi-level attention networks[☆]

Jie Xu[a], Feiran Huang[b,c], Xiaoming Zhang[d,*], Senzhang Wang[e], Chaozhuo Li[a], Zhoujun Li[a], Yueying He[f]

[a] *State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China*
[b] *College of Cyber Security/College of Information Science and Technology, Jinan University, Guangzhou, 510632, China*
[c] *Guangdong Key Laboratory of Data Security and Privacy Preserving, Guangzhou, 510632, China*
[d] *School of Cyber Science and Technology, Beihang University, Beijing, 100191, China*
[e] *School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, 210016, China*
[f] *National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, 100029, China*

HIGHLIGHTS

- Bi-directional attention to highlight the emotional regions and words.
- Multiple levels to excavate the emotional correlations between image and text.
- The experimental results demonstrate the superiority of the proposed model.

ARTICLE INFO

Article history:
Received 28 November 2018
Received in revised form 22 April 2019
Accepted 23 April 2019
Available online 9 May 2019

Keywords:
Multi-modal
Social image
Attention model
Sentiment analysis

ABSTRACT

Social network has become an inseparable part of our daily lives and thus the automatic sentiment analysis on social media content is of great significance to identify people's viewpoints, attitudes, and emotions on the social websites. Most existing works have concentrated on the sentiment analysis of single modality such as image or text, which cannot handle the social media content with multiple modalities including both image and text. Although some works tried to conduct multi-modal sentiment analysis, the complicated correlations between the two modalities have not been fully explored. In this paper, we propose a novel Bi-Directional Multi-Level Attention (BDMLA) model to exploit the complementary and comprehensive information between the image modality and text modality for joint visual-textual sentiment classification. Specifically, to highlight the emotional regions and words in the image–text pair, visual attention network and semantic attention network are proposed respectively. The visual attention network makes region features of the image interact with multiple semantic levels of text (word, phrase, and sentence) to obtain the attended visual features. The semantic attention network makes semantic features of the text interact with multiple visual levels of image (global and local) to obtain the attended semantic features. Then, the attended visual and semantic features from the two attention networks are unified into a holistic framework to conduct visual-textual sentiment classification. Proof-of-concept experiments conducted on three real-world datasets verify the effectiveness of our model.

1. Introduction

With billions of messages posted online every day, social networks have become one of the most important sources for people to acquire information. Meanwhile, people are accustomed to sharing their emotional experiences through these platforms. Sentiment analysis of social media content can be applied to many real-word applications, such as the prediction of stock market [1,2], political elections [3,4], and even the healthcare [5,6]. Meanwhile, the rapid development of social media prompts the emergence of multimodal content. For example, Twitter users often post images together with their tweets to make the tweet more expressive, and Flickr users often give the descriptions

[☆] No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.04.018.

[*] Corresponding author.
E-mail address: yolixs@buaa.edu.cn (X. Zhang).

https://doi.org/10.1016/j.knosys.2019.04.018
0950-7051/© 2019 Elsevier B.V. All rights reserved.

along with their images as clarification and explanation. Therefore, automatic sentiment detection of visual-textual content is of great significance in practice.

Due to the close relationship between the user emotion and their behavior, great effort has been devoted to sentiment analysis. However, most current works detect sentiment from only one modality of image or text [7–10]. These methods ignore the complementary information between visual content and the text descriptions, which is an important factor of sentiment analysis. To accommodate the diverse patterns of social media, multimodal sentiment analysis has attracted increasing research attention recently, which can be categorized into the following two types. The first type treats the features of different sources individually. Some works [11–13] concatenate the separate features of different modalities and input them into the classifier for sentiment classification. Some other works [14,15] predict the sentiment of separate modalities and combine the results as the multimodal sentiment label. Although these works have considered visual image and textual description both, they ignore their inner relations between the two modalities. The second type treats the features of different modalities jointly [16–18]. For example, You et al. [16] propose a multi-modal regression model to force the sentiment labels predicted by the visual and textual features separately to be consistent. However, it is still difficult for these methods to capture the complicated correlation between the two modalities.

Though visual-textual sentiment analysis has attracted considerable research interests, this task remains challenging due to the following two reasons. First, the correlations between image and text are bi-directional. Taking Fig. 1 as an example, the regions of the crying girl and the beautiful flower are more sentiment-related to the text description, which can be considered as the text–regions correlation. Going the other way, the words like "*flower*", "*excited*", and "*tears*" are more emotion-related to the image, which corresponds to the image–words correlation. Therefore, if the bi-directional correlations between the image–text pair can be parsed precisely, the sentiment analysis will be more accurate. Second, the correlations between images and texts are multi-level. The content in an image may correspond to a word, a phrase, and even the whole document of the text description. As illustrated in Fig. 1, the crying girl in the image corresponds to the word "*girl*" and phrase '*excited tears*". To get the comprehensive sentiment of this region, the whole sentence should also be taken into consideration. Likewise, the words in an text description may correspond to an object or the whole content of the image. As shown in Fig. 1, the word "*girl*" in the text corresponds to the region of the crying girl in the image, which may lead to a negative impression. However, if we want to know the real sentiment of the girl, the whole image should be considered, too.

Targeting at the above challenges, we propose a Bi-Directional Multi-Level Attention (BDMLA) model to excavate the complicated correlations between an image and its corresponding text description, based on which the complementary and comprehensive information can be captured for visual-textual sentiment classification. Specifically, a visual attention network is proposed first to focus on emotional image regions related to the corresponding text description, which is denoted as the text-to-image attention. To excavate the multi-level correlations, our visual attention model makes the region features interact with the text of multiple levels including words, phrases, and sentences. Then a semantic attention network is also proposed to highlight emotional words related to the corresponding image, which is denoted as the image-to-text attention. To excavate the multi-level correlations, our semantic attention model makes the semantic features interact with the visual features of both the global and local level. After that, the attended visual and semantic features obtained from the two attention networks are unified into a holistic framework with a Multi-Layer Perceptron (MLP) to conduct image–text sentiment classification. The main contributions of our work are summarized as follows:

- We for the first time explore the bi-directional attention between an image and its corresponding text description to highlight the emotional regions and words for sentiment analysis.
- We parse the image and text from multiple levels to excavate the complicated correlations between the visual and textual features for sentiment analysis.
- Extensive experiments are conducted on Flickr and GettyImage datasets. The experimental results demonstrate the superiority of our approach against the state-of-the-art baselines on the two datasets.

The remainder of this paper is organized as follows. Section 2 summarizes the related works about sentiment analysis. Section 3 introduces the framework of our proposed model BDMLA. Then, BDMLA is detailed in Section 4. In Section 4, we conduct extensive experiments and present the experimental results. Finally, we conclude our work in Section 5.

2. Related works

2.1. Single-modal sentiment analysis

Textual sentiment analysis. Textual sentiment analysis has been well studied in recent years and various approaches have been proposed [19–23]. Generally, these approaches can be broad ly categorized as lexicon-based methods and machine learning-based methods. In lexicon-based methods [24–26], the sentiment of the text is decided by the sentiment polarities of the individual words. For example, Kanayama et al. [24] propose a lexicon building method with context coherency to detect the polar clauses. In learning-based methods [20,27,28], the sentiment is predicted by the supervised statistical models. For example, Maas et al. [20] propose a vector space model to capture the sentiment information with the combination of unsupervised and supervised techniques. Cambria et al. [29] exploit computer and social sciences with a multi-disciplinary approach and transform them into a multi-dimensional space, which can be utilized for sentiment analysis. Besides, many works have applied the deep learning methods to the sentiment analysis with the increasingly popular of neural network [30–32]. For example, Dragoni et al. [32] propose a deep learning architecture to exploit the linguistic overlap among domains for inferring the sentiment polarity.

Visual sentiment analysis. Visual sentiment analysis is a grow ing field of research interest and currently the visual features extracted for emotion deduction can be categorized into the following three types. The first category is to extract the low-level features [33–35]. For example, Siersdorfer et al. [33] extract the color histograms and SIFT features of the image to deduce its sentiment. The second category is to extract the middle-level features [36,37]. For example, Borth et al. [37] detect 1200 adjective-noun pairs strongly related to sentiment and thus a novel mid-level representation is established with the presented concept detector library. The third category is to extract the high-level features [9,38,39]. For example, You et al. [9] first design a CNN architecture and then fine-tune it with a progressive strategy to reduce the effect of the noise on the sentiment analysis.

Although the sentiment analysis on single modality data has gained great success over the past few years, it cannot effectively handle the social media data with the diversity of information. Therefore, multi-modal sentiment analysis emerged at the right moment.

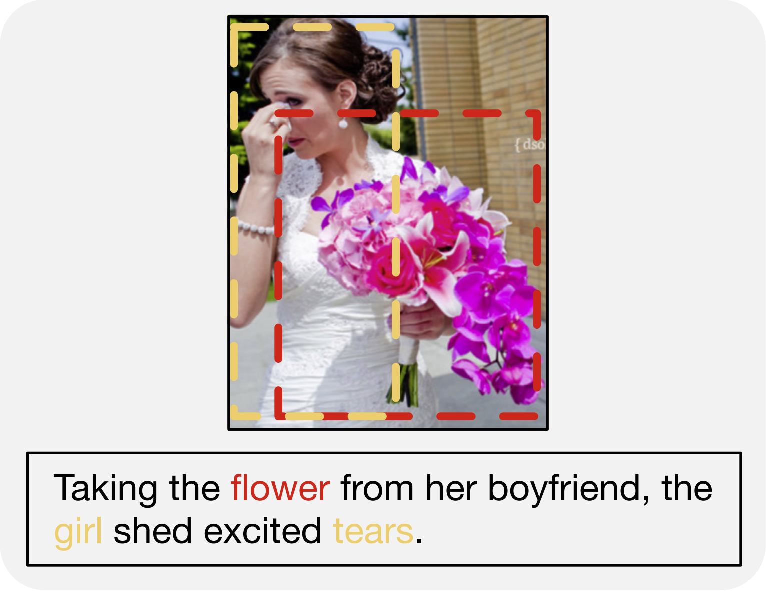

Taking the flower from her boyfriend, the girl shed excited tears.

Fig. 1. An illustration of the image–text pair. The emotional regions in the image are closely related to the emotional words in the text description, and excavating the correlations between the two modalities is helpful for visual-sentiment classification.

2.2. Multi-modal sentiment analysis

Due to the increasingly popular of the social media, multimodal sentiment analysis has recently attracted more and more research interest as a challenging task [16,40–43]. The current works can be categorized into the following two types.

The first type of works treat the features of different sources separately. Some works concatenate different features into a whole feature vector and then learn the sentiment classifier with the concatenated vector [11–13]. Poria et al. [11] combine the visual and textual features extracted from deep CNN and send them to a multiple kernel learning classifier to analyze sentiment. Wang et al. [40] embed the visual and textual content into a unified Bag-of-words representation. Then Logistic Regression is utilized to recognize the sentiment of a Weibo tweet. Morency et al. [12] take utterance-level features of textual, visual, and audio as input and utilize Hidden Markov Models to conduct sentiment classification. Besides, some other works also combine the sentiment labels predicted by single modality [14,15]. For example, Cao et al. [14] combine the image sentiment predicted with mid-level visual features and the text sentiment predicted with n-gram textual features through linear interpolation to get the multimodal sentiment. Although these methods utilize the multi-modal content, they ignore the inner relations among different modalities.

The second type of works model the features of different sources jointly [16–18,44]. You et al. [16] combine the attention mechanism with tree-structured LSTM to capture the correspondence between image and text for sentiment analysis. Zadeh et al. [17] develop a tensor fusion network to model intra-modality and inter-modality dynamics for multimodal sentiment analysis of language, visual, and acoustic content. Pang et al. [18] utilize the Deep Boltzmann Machine to learn a joint density model for sentiment classification of textual, auditory, and visual inputs. You et al. [44] introduce a multi-modal consistent regression model to force the sentiment results predicted by visual features and textual features respectively to reach a consensus. Xu et al. [45] propose an attention LSTM guided by visual feature to extract sentiment-related words to conduct sentiment analysis. Xu et al. [46] propose a Merged Neural Network (MNN) model to extract image and text features and then utilize the Early/Late Residual RMNN to fuse multimodal features for sentiment classification.

Although the existing multi-modal sentiment analysis methods have considered the inner correlations among different modal ities, most of them only parse the content from single level and capture the attention from one direction, which cannot fully exploit the comprehensive and complementary information for effective sentiment analysis. Therefore, we aim to propose a bi-directional multi-level attention network to conduct visual-textual sentiment classification.

3. Bi-directional multi-level attention networks

In this section, we introduce the architecture of the proposed model BDMLA in detail. First, we briefly present the framework. Then, we put forward visual attention network and semantic attention network to focus on emotional image regions and emotional words respectively. Finally, we integrate the two attention networks to conduct sentiment classification.

3.1. Overview

We first define the notations used in this paper. Let $V = \{V_1, \ldots, V_i, \ldots, V_n\}$ denote the set of images and $T = \{T_1, \ldots, T_i, \ldots, T_n\}$ denote the set of text descriptions, our target is to classify the sentiment (positive, negative) of the image–text pair (V_i, T_i) by considering the two modalities simultaneously.

Fig. 2 illustrates the framework of BDMLA, which consists of two attention components to learn the bi-directional correlation for image–text sentiment analysis. The visual attention network is proposed to correlate the emotional image regions with the corresponding text description, which is denoted as the text-to-image attention. To excavate the multi-level correlations, the visual attention network makes region features interact with the texts of multiple levels including words, phrases, and sentences. Similarly, the semantic attention network is proposed to select emotional words related to the corresponding image, which is denoted as the image-to-text attention. To excavate the multi-level correlations, the semantic attention network makes semantic features interact with visual features of both the global

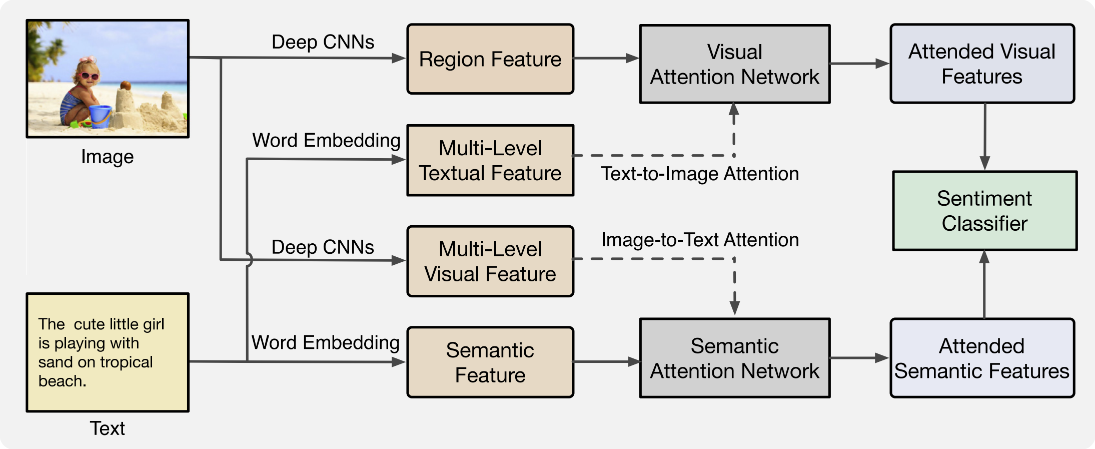


Fig. 2. The framework of BDMLA.

and local level. Then, the attended visual and semantic features output from the two attention networks are fed into a Multi-Layer Perceptron (MLP) to conduct joint sentiment classification.

3.2. Visual attention network

Rather than considering all regions in the image equally, visual attention mechanism can relate the text description to the specific meaningful image regions. It has gained great success in many computer vision related tasks, such as visual question answering [47–49], image captioning [50,51], and representation learning [52]. However, different regions may correspond to different levels of text, such as words, phrases, or even broader scopes, while the multi-level correlations between visual and textual content are ignored by most of the current methods. Therefore, we propose a multi-level visual attention network to exploit the correlations between the image and different semantic levels of the text description with a multiplicative embedding scheme. Fig. 3 illustrates the structure of the visual attention network.

Given an image–text pair (V_i, T_i), the text description with L words is denoted as $T_i = \{t_{i,1}, \ldots, t_{i,k}, \ldots, t_{i,L}\}$. The region maps obtained with the Convolution Neural Networks (CNNs) is denoted as $V_i = \{\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,j}, \ldots, \mathbf{v}_{i,R}\} \in \mathbb{R}^{R\times D}$, where $\mathbf{v}_{i,j} \in \mathbb{R}^D$ is the D-dimensional region feature and R is the number of regions. First, we embed each word t_k to a pre-trained word vector $\mathbf{t}_k^w \in \mathbb{R}^{d_w}$. Then, max-pooling is employed across each time step to obtain the word-level embedding as follows:

$$\mathbf{T}_i^w = max(\mathbf{t}_{i,1}^w, \ldots, \mathbf{t}_{i,k}^w, \ldots, \mathbf{t}_{i,L}^w) \tag{1}$$

Compared with one single word, the phrase can provide richer semantic information. Different from word-level embedding, the phrase-level embedding works on phrase composed from three words. Specifically, 1-D convolution with window size of three is employed on the word embedding to obtain the phrase features. To maintain the original length of the description, the word-level embedding $\mathbf{T}_i^w$ is 0-padded before convolution. After that, max-pooling is employed across each time step to obtain the phrase-level embedding as follows:

$$\mathbf{t}_{i,k}^p = \tanh(\mathbf{W}_c * \mathbf{t}_{k:k+2}^w) \tag{2}$$

$$\mathbf{T}_i^p = max(\mathbf{t}_{i,1}^p, \ldots, \mathbf{t}_{i,k}^p, \ldots, \mathbf{t}_{i,L}^p) \tag{3}$$

where $\mathbf{t}_k^p \in \mathbb{R}^{d_p}$ is the convolution output of the kth word, $*$ is the convolution process and $\mathbf{W}_c$ is the weight parameters.

In order to have a comprehensive understanding of the text description, LSTM is utilized to encode the whole sequence $\{\mathbf{t}_{i,1}^w, \ldots, \mathbf{t}_{i,k}^w, \ldots, \mathbf{t}_{i,L}^w\}$ at the document-level. The LSTM hidden vector at time L is considered as the document-level embedding:

$$\mathbf{T}_i^d = LSTM(\{\mathbf{t}_{i,1}^w, \ldots, \mathbf{t}_{i,k}^w, \ldots, \mathbf{t}_{i,L}^w\}; \theta_d) \tag{4}$$

To excavate the correlation between image regions and multiple semantic levels of the text description, we need to acquire the joint textual feature of the three levels first. Therefore, we embed the word embedding $\mathbf{T}_i^w \in \mathbb{R}^{d_w}$, phrase embedding $\mathbf{T}_i^p \in \mathbb{R}^{d_p}$, and document embedding $\mathbf{T}_i^d \in \mathbb{R}^{d_d}$ into a S-dimensional common space:

$$\widehat{\mathbf{T}}_i^w = \tanh(\mathbf{W}^w\mathbf{T}_i^w + \mathbf{b}_w) \tag{5}$$

$$\widehat{\mathbf{T}}_i^p = \tanh(\mathbf{W}^p\mathbf{T}_i^p + \mathbf{b}_p) \tag{6}$$

$$\widehat{\mathbf{T}}_i^d = \tanh(\mathbf{W}^d\mathbf{T}_i^d + \mathbf{b}_d) \tag{7}$$

where $\mathbf{W}^w \in \mathbb{R}^{S\times d_w}$, $\mathbf{W}^p \in \mathbb{R}^{S\times d_p}$, $\mathbf{W}^d \in \mathbb{R}^{S\times d_d}$, and $\mathbf{b}_w$, $\mathbf{b}_p$, $\mathbf{b}_d \in \mathbb{R}^S$ are parameters. The joint textual feature $\mathcal{T}_i \in \mathbb{R}^S$ can be obtained with the element-wise multiplication of $\widehat{\mathbf{T}}_i^w$, $\widehat{\mathbf{T}}_i^p$, and $\widehat{\mathbf{T}}_i^d$. To constrain the magnitude of the joint feature, the L_2 normalization is employed:

$$\mathcal{T}_i = Norm_2(\widehat{\mathbf{T}}_i^w \odot \widehat{\mathbf{T}}_i^p \odot \widehat{\mathbf{T}}_i^d) \tag{8}$$

where $\odot$ is the element-wise multiplication.

Based on the relevance with the joint textual feature $\mathcal{T}_i$, a score $\alpha_{i,j} \in [0, 1]$ is assigned to each image region $\mathbf{v}_{i,j}$ through a softmax function:

$$\alpha_{i,j} = \frac{exp(e_{i,j})}{\sum_{j=1}^R exp(e_{i,j})} \tag{9}$$

where

$$exp(e_{i,j}) = \varphi((\mathcal{T}_i)^T\mathbf{W}_1\mathbf{v}_{i,j} + \mathbf{b}_1) \tag{10}$$

calculates how well the region $\mathbf{v}_{i,j}$ is related to the joint textual feature $\mathcal{T}_i$. $\varphi(\cdot)$ is the smooth function, and tanh is usually adopted. $\mathbf{W}_1$ is the weight matrix and $\mathbf{b}_1$ is the bias term, which are learned in the training procedure. With the attention scores obtained, the strength of attention over different regions are modulated. Thus, the attended visual features can be calculated through a weighted sum of all regions:

$$\mathbf{v}^i = \sum_{j=1}^R \alpha_{i,j}\cdot\mathbf{v}_{i,j} \tag{11}$$

In such a way, we can obtain the attended visual features $\mathbf{v}^i \in \mathbb{R}^D$. By making region features interact with multiple semantic

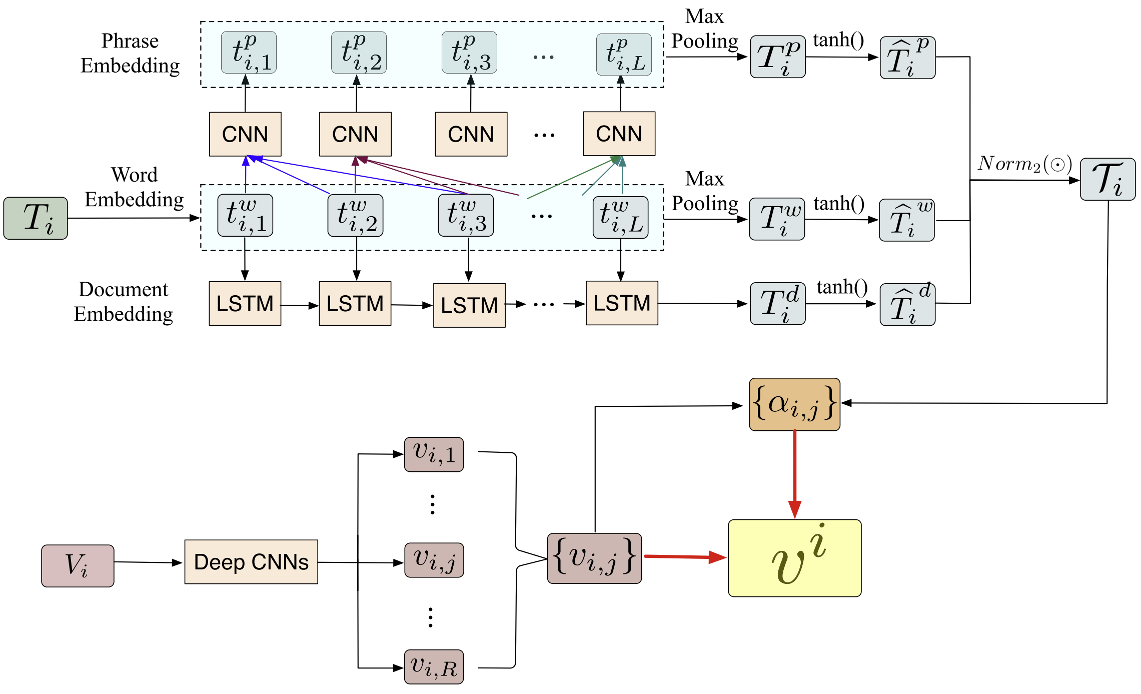


Fig. 3. Visual attention network.

levels of text, the multi-level correlations between image and text are excavated. And on this basis, emotional image regions related to the textual content are focused on.

3.3. Semantic attention network

Similar to visual attention that focuses on emotional image regions related to the text description, the sentimental words related to the image should also be highlighted, which can help understand the sentiment of the image–text pair better [53]. However, this type of attention is ignored by most of the current multi-modal sentiment analysis methods. Besides, most works represent the image with global features at a coarse level. This may make some objects in the image cannot be detected, and thus the words cannot be related to the image well. To address this issue, we propose a multi-level semantic attention network to exploit the correlations between the text and different visual levels (global level and local level) of the image. Fig. 4 illustrates the structure of the semantic attention network.

To obtain the global feature of image V_i, the VGG-19 network [54] pre-trained on the ImageNet dataset is employed to conduct the feature construction. The d_v-dimensional vector from *fc*8 layer (the last fully connected layer) is chosen as the global feature, which is denoted as $\mathbf{V}_i^g \in \mathbb{R}^{d_v}$.

To obtain the local feature of image V_i, the objects in the image should be detected first. With the confidence scores calculated from Faster R-CNN [55], we select the top-P important local objects. Thus, each object is represented as a d_v-dimensional vector from *fc*8 layer of VGG-19. The local feature of image V_i is denoted as $\mathbf{V}_i^l = \{\mathbf{V}_{i,1}^l, \ldots, \mathbf{V}_{i,j}^l, \ldots, \mathbf{V}_{i,P}^l\} \in \mathbb{R}^{P \times d_v}$. The joint visual feature $\mathcal{V}_i \in \mathbb{R}^{d_v}$ can be obtained with the element-wise multiplication of global feature $\mathbf{V}_i^g$ and local feature $\mathbf{V}_i^l$, and L_2 normalization is utilized to constrain the magnitude of the joint feature:

$$\mathcal{V}_i = Norm_2(\mathbf{V}_i^g \odot \mathbf{V}_{i,1}^l \odot \ldots \odot \mathbf{V}_{i,P}^l) \tag{12}$$

To obtain the high level text features, LSTM is utilized to optimize the initial word embedding. The output of LSTM $\mathbf{s}_{i,k} \in \mathbb{R}^h$ is considered as the high level feature at time step k. To measure how well the word is related to the image content, a score $\beta_{i,j} \in [0, 1]$ is assigned to each word $\mathbf{s}_{i,k}$. Similar to the visual attention network, $\beta_{i,k}$ is calculated through a softmax function empirically:

$$\beta_{i,k} = \frac{exp(e_{i,k})}{\sum_{k=1}^{R} exp(e_{i,k})} \tag{13}$$

where

$$exp(e_{i,k}) = \varphi((\mathcal{V}_i)^T \mathbf{W}_2 \mathbf{s}_{i,k} + \mathbf{b}_2) \tag{14}$$

The strength of attention over different words are modulated according to the attention scores obtained. With the weighted sum of all words, the attended semantic features can be calculated as follows:

$$\mathbf{t}^i = \sum_{k=1}^{L} \beta_{i,k}(\mathbf{s}_{i,k}) \tag{15}$$

In such a way, we can obtain the attended textual features $\mathbf{t}^i \in \mathbb{R}^{d_v}$. By making semantic features interact with multiple visual levels of the image, the multi-level correlations between image and text are excavated. And on this basis, emotional words related to the visual content are focused on.

3.4. Bi-directional joint learning for sentiment analysis

As discussed above, two independent networks are proposed to learn the complex non-linear bi-directional relation for the image–text pair (T_i, V_i). The emotional visual regions related to the text description are highlighted in the visual attention network, and the emotional words related to the image content are highlighted in the semantic attention network. In order to take full use of the bi-directional correlation between image and text, we utilize Multi-Layer Perceptron (MLP) to fuse the two features for sentiment analysis [56,57].

From Eq. (11) we can get the attended visual feature $\mathbf{v}^i$, and from Eq. (15) we can get the attended semantic feature $\mathbf{t}^i$. Based

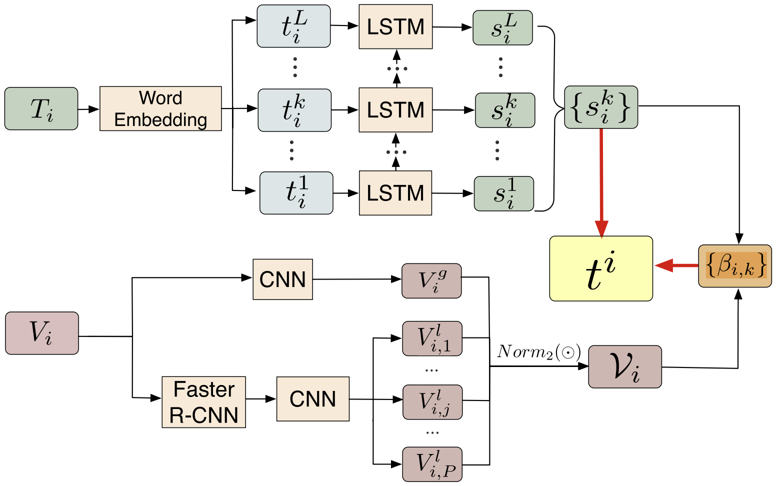


Fig. 4. Semantic attention network.

on them, the joint feature of image and text is learned through MLP as follows:

$$\mathcal{F}_i = mlp(\mathbf{v}^i, \mathbf{t}^i; \mathbf{W}_{mlp}) \tag{16}$$

where $\mathbf{W}_{mlp}$ denotes the parameters of MLP and the non-linear activation function adopted here is tanh.

After that, the joint feature $\mathcal{F}_i$ of the image–text pair(T_i, V_i) is input to a softmax classifier to conduct sentiment classification. The loss between the predicted result $p(\mathcal{F}_i)$ and the ground-truth sentiment label y_i is calculated through the negative log-likelihood(NLL) :

$$p(\mathcal{F}_i) = softmax(\mathbf{W}_s \mathcal{F}_i) \tag{17}$$

$$\mathcal{L} = -log(p(\mathcal{F}_i), y_i) \tag{18}$$

where $\mathbf{W}_s$ is the parameters of softmax function. The whole model is trained over the training set with back propagation.

4. Experiments

In this section, we first introduce the experiment setup. Then by comparing the proposed BDMLA with several state-of-the-art baselines, the effectiveness of our model is demonstrated with abundant analysis and discussions. Next, we analyze the importance of different components with ablation study. Finally, a qualitative case study is presented to further verify the effectiveness of the visual and semantic attention models of BDMLA on sentiment analysis.

4.1. Datasets

We perform our experiments on three real-world social image datasets collected from Flickr and Getty Image. The details of the datasets are as follows:

- Flickr. The SentiBank of Visual Sentiment Ontology [37] contains 1200 adjective-noun pairs (ANPs) which have strong sentiment correlations. Inspired by this, we retrieve images from Flickr by querying the 1200 ANPs through Flickr API.[1] To conduct visual-textual sentiment classification, only images with English descriptions are collected. The sentiment of the image–text pair is labeled in accordance with the corresponding ANP. To make the description more appropriate to our model, we only keep those images whose descriptions have more than 5 words and less than 100 words. The Flickr dataset consists of 298,473 weakly labeled image–text pairs and 154,931 of them have positive labels.
- Flickr-ML. To obtain exact labels for more accurate sentiment classification, crowd intelligence is employed to annotate the sentiment labels for a part of weakly labeled data in the Flickr dataset described above. Specifically, we randomly select 30,000 image–text pairs from the Flickr dataset, 15,000 of which are positive and 15,000 are negative. Then each of the candidate image–text pair is allocated to 5 annotators to get the accurate sentiment label. To guarantee the accuracy of the annotation, the pairs with less than 4 agreements are excluded. In this way, we construct the high quality Flickr-ML dataset with reliant sentiment labels. The Flickr-ML dataset consists of 21,037 manually labeled image–text pairs and 10,892 of them have positive labels.
- Getty Images. Getty Images, Inc. is a stock photo agency with an archive of 80 million images. The images in Getty usually have relatively formal descriptions and can be queried with searching system powerfully and conveniently. According to the Balanced Affective Word List Project,[2] we query Getty Images with 101 sentimental keywords to obtain sentiment-related image–text pairs. The sentiment labels of the chosen targets are consistent with the corresponding keywords. With this method, we construct the Getty Images dataset with weakly sentiment labels. The Getty Images dataset contains 401,502 image–text pairs and 198,564 have positive labels.
- Flickr-IML. Katsurai et al. [58] published two publicly available datasets for sentiment classification. This data source includes the ids of images along with the manually labeled annotations.[3] As the Instagram policies has been changed recently, we cannot download the images and the corresponding texts from the Instagram platform. Hence, we can only evaluate the proposed model on the Flickr dataset. Some images cannot be achieved currently as users may remove or delete them. Following the previous work [58], the labels of images are constructed according to the majority votes of polarity strategy. Finally 47,018 samples can be obtained, in which 23,392 images are labeled as positive and the rest are negative.

[1] https://www.flickr.com/services/developer/api/.

[2] http://www.sci.sdsu.edu/CAL/wordlist/origwordlist.html.

[3] http://mm.doshisha.ac.jp/en/senti/CrossSentiment.html.

Table 1
Statistics of the Datasets.

Dataset	Positive	Negative	Total	Labeling
Flickr	154,931	143,542	298,473	weak
Flickr-ML	10,892	10,145	21,037	strong
Getty Images	198,564	202,938	401,502	weak
Flickr-IML	23,392	23,626	47,018	strong

Table 2
Pairwise Kappa scores of Flickr-ML dataset.

Annotator	A	B	C	D	E
A	1	0.825	0.853	0.811	0.812
B	0.825	1	0.826	0.773	0.839
C	0.853	0.826	1	0.786	0.806
D	0.811	0.773	0.786	1	0.821
E	0.812	0.839	0.806	0.821	1

For the manually annotated Flick-ML dataset, the Kappa measurement is introduced to evaluate the reliability of the agreement between pairwise annotators [59]. Kappa calculates the agreement coefficient between two annotators by excluding the possibility of label matching by chance. The Kappa score usually ranges from 0 to 1, and larger numeric value means better agreement. Here the free-marginal kappa variant [60] is selected to perform the annotation quality analysis. This is because it is still effective when the annotators are not forced to assign a certain number of image–text pairs to each sentiment category. Table 2 presents the pairwise Kappa scores between five annotators named from annotator A to annotator E. From Table 2, one can see that most Kappa scores are higher than 0.8, which indicates the five annotators achieve almost perfect agreements on most samples [61]. Thus the quality of the manually labeled Flick-ML dataset should be trusted.

Although Flickr and Getty Images are labeled according to the corresponding keywords and the sentiment labels may not be reliant, the noise is fair to all methods. Thus, the weak sentiment label has little impact on the comparison between the proposed methods and the baselines. The statistics of the four datasets are shown in Table 1.

4.2. Experiment settings

To extract the visual features, images in the datasets are resized to 224 × 224 first. Then the resized images are input to VGG-19 networks [54] to obtain different visual features. Specifically, features of the last convolutional layer of VGG-19 are regarded as the region features. Thus each image can be represented as 196 region features with the dimensionality of 512 in the visual attention network. In the semantic attention network, features of the last fully connected layer are regarded as the global visual features. To extract the local visual feature, the top-10 important local objects are detected with Faster R-CNN first. Then each object is represented as a 1000-dimension vector from the last fully connected layer of VGG-19.

To extract the textual features, descriptions in the datasets are preprocessed first. The images in the social websites are usually accompanied with text descriptions provided by the image owners. However, some words (e.g., "50–200 mm", "Leica", "nikon", etc.) appearing frequently are sentiment unrelated, which should be removed before the feature extraction. Besides, there are some misspellings and rarely-used characters which should not be considered, either. Thus, only the words appearing more than 5 times are kept. Then the pre-trained 300-dimensional Glove [62] features are employed to obtain the word embedding.

The details of the BDMLA architecture are presented as follows. For the visual attention network shown in Fig. 3, each image is represented by 196 region feature vectors with the dimensionality of 512. The phrase embedding is generated by the 1D-CNN model. The LSTM model used in document embedding generation is a one-layer LSTM with output dimension of 300. For the semantic attention network shown in Fig. 4, LSTM part is the one-layer structure with 256-dimension outputs. Faster R-CNN is the pre-learned model provided in [55]. The hyper-parameters of the proposed BDMLA model are defined as: the dimension of visual region features $D = 512$, the number of visual regions $R = 196$, the dimension of global visual features $d_v = 1000$, the dimension of word embeddings $d_w = 300$, and the dimension of document embedding $d_d = 300$. To reduce overfitting, dropout is utilized with probability 0.5.

The implementation is performed on 2 × NVIDIA GeForce GTX 1080. To optimize the objective function in the training procedure, Stochastic Gradient Descent (SGD) is employed with the learning rate 0.01 and momentum 0.9. For each dataset, 60% samples are randomly selected as the training set, 20% samples are randomly picked as the validation set and the rest 20% samples are viewed as the test set. All the methods are fully trained on the training set and the hyper-parameters are tuned on the validation set with random search strategy. The checkpoint model with the best validation performance is evaluated on the test set to achieve the quantitative results. This evaluation process is repeated three times and the average scores are reported.

4.3. Baselines

In this section, we compare BDMLA with the following state-of-the-art sentiment analysis methods:

- **Single visual model** [63]. This method only uses deep visual features extracted from VGG-19 to conduct visual sentiment classification through logistic regression classifier.
- **Single textual model** [63]. This method only uses paragraph features of text description to conduct textual sentiment classification through logistic regression classifier.
- **Early fusion**. In this method, the visual features in Single Visual Model and textual features in Single Textual Model are concatenated together to conduct visual-textual sentiment classification through logistic regression classifier.
- **Late fusion** [63]. In this method, the sentiment score is the average of the visual sentiment score in Single Visual Model and the textual sentiment score in Single Textual Model.
- **CCR** [16]. This method develops a cross-modality consistent regression model that tries to make the two sentiment labels predicted by visual features and textual features respectively reach a consensus.
- **T-LSTM embedding** [44]. This method builds a tree-structured LSTM with attention mechanism to align image regions and descriptive words. Besides, visual-textual semantic embedding is introduced as an auxiliary task to learn joint visual-textual features for sentiment classification.
- **TFN** [17]. This methods develops a tensor fusion network to model intra-modality and inter-modality dynamics for multimodal sentiment analysis.

4.4. Results and analysis

In this section, we compare our model with baselines and analyze the experimental results. Following the previous work [44], we select four metrics to evaluate our proposal: Precision, Recall, F1, and Accuracy.

Table 3
Results on Flickr dataset.

Model	Precision	Recall	F1-score	Accuracy
Textual	0.721	0.728	0.724	0.738
Visual	0.735	0.741	0.738	0.742
Early fusion	0.749	0.746	0.747	0.758
Late fusion	0.755	0.749	0.752	0.766
CCR	0.809	0.815	0.812	0.811
T-LSTM embedding	0.827	0.820	0.823	0.828
TFN	0.836	0.818	0.827	0.831
BDMLA	**0.847**	**0.850**	**0.848**	**0.849**

Table 4
Results on Getty Image dataset.

Model	Precision	Recall	F1-score	Accuracy
Textual	0.745	0.729	0.737	0.741
Visual	0.738	0.708	0.723	0.734
Early fusion	0.779	0.752	0.765	0.771
Late fusion	0.781	0.798	0.789	0.793
CCR	0.812	0.828	0.820	0.809
T-LSTM embedding	0.834	0.831	0.832	0.826
TFN	0.845	0.839	0.842	0.841
BDMLA	**0.871**	**0.854**	**0.862**	**0.865**

Table 5
Results on Flickr-ML dataset.

Model	Precision	Recall	F1-score	Accuracy
Textual	0.738	0.742	0.740	0.742
Visual	0.765	0.762	0.763	0.761
Early fusion	0.787	0.801	0.794	0.798
Late fusion	0.796	0.803	0.799	0.801
CCR	0.817	0.821	0.819	0.818
T-LSTM embedding	0.839	0.843	0.841	0.836
TFN	0.848	0.845	0.846	0.845
BDMLA	**0.880**	**0.869**	**0.874**	**0.878**

Table 6
Results on Flickr-IML dataset.

Model	Precision	Recall	F1-score	Accuracy
Textual	0.691	0.697	0.694	0.698
Visual	0.705	0.721	0.713	0.707
Early fusion	0.715	0.729	0.722	0.728
Late fusion	0.732	0.743	0.737	0.741
CCR	0.779	0.796	0.787	0.783
T-LSTM embedding	0.807	0.801	0.804	0.799
TFN	0.819	0.803	0.811	0.812
BDMLA	**0.825**	**0.833**	**0.829**	**0.831**

4.4.1. Results on Flickr

Table 3 presents results on the Flickr dataset for sentiment classification. By observing the table, one can observe that:

- The single textual modality and single visual modality perform the worst. The reason may be that the two methods only consider unimodal content and ignore the complementary information between the two modalities, which hurts their performance. By comparing them with the results of Early Fusion and Late Fusion, one can verify that the multimodal content is more effective on sentiment classification than unimodal content.
- The improvement of CCR over the simple fusion strategies is significant, by nearly 5%. The main contributor is its consistent constraints imposed across the two modalities. Compared with CCR, T-LSTM Embedding explores the deep semantic features between the visual-textual content by modeling the two modalities jointly, which helps it make an improvement of nearly 2%. By modeling intra-modality and inter-modality dynamics in an end-to-end way, TFN provides competitive results compared with other baselines.
- BDMLA consistently outperforms all the baseline methods in terms of both $F1$-*score* and *Accuracy*. This demonstrates that our method can conduct multi-modal sentiment classification effectively.

4.4.2. Results on Getty Image

Table 4 presents the results on the Getty Image dataset for sentiment classification of our BDMLA and the baselines. Compared with Flickr, one can see that these methods perform better on the Getty Image dataset. This is probably because the size of the Getty Image dataset is much larger than that of the Flickr dataset. More data means the model can be trained more robust to reduce overfitting. Besides, one can see that the performance of BDMLA is the best among all the methods on all metrics, which verifies the effectiveness of our model on sentiment classification.

For fine-grained analysis, we want to know the impact of training data size on the *Accuracy* of different multi-modal sentiment classification methods. To achieve this goal, we construct new training dataset by sampling different sizes of data from the initial training data. The size interval is between 10% and 100%. Fig. 5 depicts the evolution of *Accuracy* with the increase of training data. From the result, one can see that the increase trend of *accuracy* slows down when the new training data reaches about 60% of the original data. Our conjecture is that this amount of data have trained the model approximately. Besides, it is clearly to see that BDMLA performs better than the compared methods consistently, which verifies the robustness of our model.

4.5. Results on Flickr-ML

The Flickr dataset and Getty Image dataset are both large datasets and only have weak sentiment labels. To reduce the bias introduced by weakly labeling, we employ the manually labeled dataset Flick-ML to conduct more accurate sentiment analysis. Flick-ML dataset only has 21,037 image–text pairs, which are not enough to train a new model. Therefore, Flickr dataset is employed to pre-train the model first. Then we fine-tune the model with the Flick-ML dataset.

Table 5 evaluates the performance of different methods with a 5-fold cross-validation. We find that the difference between the Single Textual Model and Sing Visual Model on the Flick-ML dataset is more significant than on the weakly labeled datasets of Flickr and Getty Image. The reason may be that the sentiment of image–text pairs in the weakly labeled datasets is labeled according to the corresponding keywords, which puts more emphasis on the textual content than visual content. Besides, the performance on Flick-ML shows comparable improvement than on Flickr. This verifies the effectiveness of the strong label.

4.6. Results on Flickr-IML

To further verify the effectiveness of our model, we compare our model against the state-of-the-art baselines on an open image sentiment classification dataset: Flickr-IML [58]. Table 6 presents the experimental results. It can be clearly seen that our BDMLA outperforms the baselines consistently, which demonstrates our method can learn better discriminative representations to improve the performance of sentiment classification. Compared with the single view based methods (Textual and Visual), BDMLA improves the F1-score by more than 10%, which proves the multi-view information fusion contributes to better capture the sentiment-related contents. The proposed BDMLA model outperforms the best baseline method (TFN) by nearly 2% in terms of the classification accuracy, which verifies the effectiveness of our method.

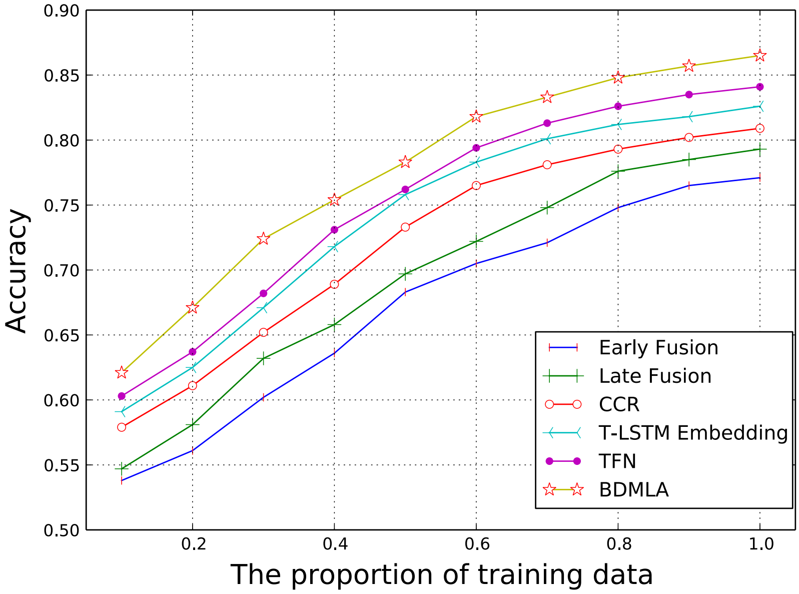


Fig. 5. Accuracy evolution with the increase of training data.

4.7. Ablation study

In this section, ablation studies are performed to quantify the effectiveness of multi-level features and bi-directional attention introduced in our work. Our model is re-trained by ablating the followings components respectively on the Flickr-ML dataset:

- Visual Attention Network, where only the text–region correspondence is considered. To study the effect caused by textual feature of each level further, we make the following four ablations:
  - Word-Level alone (W), where only word-level feature is considered as the textual feature.
  - Phrase-Level alone (P), where only phrase-level feature is considered as the textual feature.
  - Document-Level alone (D), where only document-level feature is considered as the textual feature.
  - Word-Level + Phrase-Level + Document-Level (W + P + D), where the three levels of features are modeled jointly as the textual feature, which is the same as our model. The difference lies in the lacking of semantic attention network.
- Semantic Attention Network, where only the image–words correspondence is considered. To study the effect caused by visual feature of each level further, we make the following three ablations:
  - Global alone (G), where only global-level feature is considered as the visual feature.
  - Local alone (L), where only local-level feature is considered as the visual feature.
  - Global + Local (G + L), where the two levels of features are modeled jointly as the visual feature, which is the same as our model. The difference lies in the lacking of visual attention network.

Table 7 presents the ablation results. Compared with the single attention network, the bi-directional attention obtains an obvious improvement, which verifies the effectiveness of the bi-directional attention on exploiting the deep cross-relation between the visual and textual content for sentiment analysis. Comparing the multi-level textual feature with single-level textual feature, we find that the combination of the three-levels textual feature can indeed help improve the performance of sentiment analysis. And by the same logic, the combination of global and local visual features has the same effect. Therefor, it can be concluded that there exist bi-directional multi-level correlations between the visual-textual content, and the exploitation of these correlations is conducive to more effective multi-modal sentiment classification.

Table 7
Ablation study on the Flickr-ML dataset.

Attention network	Feature level	Precision	Recall	F1-score	Accuracy
Visual attention	W	0.838	0.829	0.833	0.831
	P	0.845	0.841	0.843	0.839
	S	0.848	0.842	0.845	0.844
	W + P + S	0.866	0.855	0.860	0.862
Semantic attention	G	0.828	0.831	0.829	0.823
	L	0.831	0.839	0.835	0.834
	G + L	0.852	0.846	0.849	0.853
Bi-directional attention	BDMLA	**0.880**	**0.869**	**0.874**	**0.878**

4.8. Visualization of attention

To better interpret the attention mechanism, we visualize the attention weights with several examples to show what the model “sees” from the image and the text. Three positive examples and three negative examples are chosen from the Flickr-ML dataset respectively for illustration in consideration of their manual annotations.

For visual attention, we employ the visualization method employed in [51] with Upsampling and Gaussian filtering. Besides, for the up-sampled attention scores, we further draw the heat map with brighter colors. Therefore, the higher the attention scores, the redder the image regions. As for the semantic attention, the attention score is reflected by the background color of the words. The higher the attention score, the deeper the background color.

Fig. 6 shows the visualization results of the chosen image–text pairs. It can be seen that the attentions are almost drawn to the right regions and words generally. Take “Pos_1” as an example,

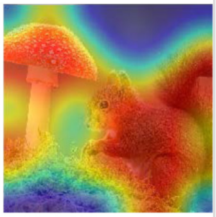
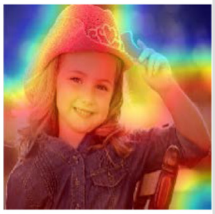
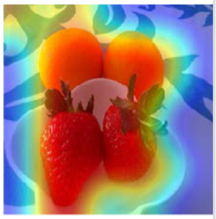
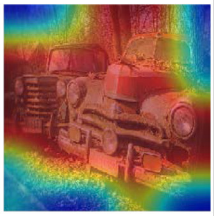
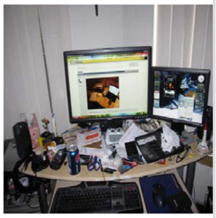
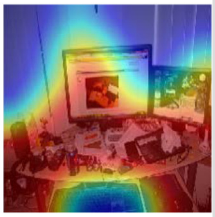
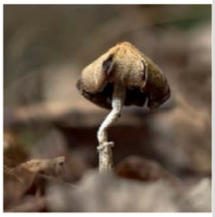
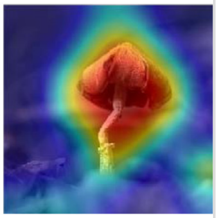
No.	Original image-text pair		Attended image	Attended text
Pos_1		A red squirrel is standing in front of a mushroom.		A red squirrel is standing in front of a mushroom.
Pos_2		A cute girl with a pink hat.		A cute girl with a pink hat.
Pos_3		This year, possibly due to the warmer than usual summer, fresh fruits such as mikan (mandarin oranges) and strawberries are plentiful and sweet.		This year, possibly due to the warmer than usual summer, fresh fruits such as mikan (mandarin oranges) and strawberries are plentiful and sweet.
Neg_1		An eclectic mix of rusty classic cars parked in a German Woodland.		An eclectic mix of rusty classic cars parked in a German Woodland.
Neg_2		My dirty desk, it's a dumping ground.		My dirty desk, it's a dumping ground.
Neg_3		This tiny mushroom was at the end of its life and looking very haggard and frail with its broken stem and split cap.		This tiny mushroom was at the end of its life and looking very haggard and frail with its broken stem and split cap.

Fig. 6. The visualization of attention.

a cute squirrel and a beautiful mushroom are highlighted in the image, which shows obvious positive sentiment. Though important words like "red", "squirrel", and "mushroom" are captured from the text, it is hard to tell the sentiment only from the text. However, combining the two attention together, it is easy to infer this pair is positive. Another example, from the image of "Neg_3", a withered mushroom is focused on. It looks like the negative class. However, the background colors of words like "very", "haggard", "frail", and "broken" in the text are very deep, which means strong emotion of negative. With the text attention, the one certain thing is that "Neg_3" is negative. From the two visualization examples, we can further verify that the visual content and textual content have complementary information, which can help make reasonable sentiment prediction.

5. Conclusion and future work

In this paper, we propose a Bi-Directional Multi-Level Attention (BDMLA) model to exploit the bi-direction attentions and multi-level correlations between the visual and textual content for sentiment classification. To focus on emotional image regions related to the corresponding text description, a visual attention network is proposed first. Here region features interact with multiple semantic levels of text (word, phrase, and sentence level) to excavate the multi-level correlations. Then, a semantic attention network is also put forward to highlight the emotional words related to the corresponding image. Here semantic features interact with multiple visual levels of image (global and local level) to excavate the multi-level correlations. Experimental results show that our approach outperforms existing state-of-the-art models on three real-world datasets consistently. In future work, we plan to explore the effect of social links among social images on sentiment analysis of social images.

Acknowledgments

This work was supported in part by Beijing Natural Science Foundation of China (No. 4182037), National Natural Science Foundation of China (No. U1636210, No. U1636211, No. 61602237), Natural Science Foundation of Jiangsu Province of China (No. BK20171420), Guangdong Provincial Special Funds for Applied Technology Research and Development and Transformation of Important Scientific and Technological Achieve, China (No. 2017B010124002), Natural Science Foundation of Guangdong Province, China (No. 2017A030313334).

References

[1] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8, http://dx.doi.org/10.1016/j.jocs.2010.12.007.

[2] X. Li, H. Xie, L. Chen, J. Wang, X. Deng, News impact on stock price return via sentiment analysis, Knowl.-Based Syst. 69 (2014) 14–23, http://dx.doi.org/10.1016/j.knosys.2014.04.022.

[3] V. Kagan, A. Stevens, V.S. Subrahmanian, Using twitter sentiment to forecast the 2013 Pakistani election and the 2014 Indian election, IEEE Intell. Syst. 30 (1) (2015) 2–5, http://dx.doi.org/10.1109/MIS.2015.16.

[4] M. Ibrahim, O. Abdillah, A.F. Wicaksono, M. Adriani, Buzzer detection and sentiment analysis for predicting presidential election results in a twitter nation, in: IEEE International Conference on Data Mining Workshop, ICDMW 2015, Atlantic City, NJ, USA, November 14–17, 2015, IEEE Computer Society, 2015, pp. 1348–1353, http://dx.doi.org/10.1109/ICDMW.2015.113.

[5] C. Caragea, A.C. Squicciarini, S. Stehle, K. Neppalli, A.H. Tapia, Mapping moods: Geo-mapped sentiment analysis during hurricane sandy, in: S.R. Hiltz, L. Plotnick, M. Pfaf, P.C. Shih (Eds.), 11th Proceedings of the International Conference on Information Systems for Crisis Response and Management, University Park, Pennsylvania, USA, May 18–21, 2014, ISCRAM Association, 2014.

[6] S. Yadav, A. Ekbal, S. Saha, P. Bhattacharyya, Medical sentiment analysis using social media: Towards building a patient assisted system, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018, European Language Resources Association (ELRA), 2018.

[7] C.N. dos Santos, M. Gatti, Deep convolutional neural networks for sentiment analysis of short texts, in: J. Hajic, J. Tsujii (Eds.), 25th International Conference on Computational Linguistics, COLING 2014, August 23–29, 2014, Dublin, Ireland, in: Proceedings of the Conference: Technical Papers, ACL, 2014, pp. 69–78.

[8] H. Saif, Y. He, M. Fernández, H. Alani, Contextual semantics for sentiment analysis of Twitter, Inf. Process. Manag. 52 (1) (2016) 5–19, http://dx.doi.org/10.1016/j.ipm.2015.01.005.

[9] Q. You, J. Luo, H. Jin, J. Yang, Robust image sentiment analysis using progressively trained and domain transferred deep networks, in: B. Bonet, S. Koenig (Eds.), Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25–30, 2015, Austin, Texas, USA, AAAI Press, 2015, pp. 381–388.

[10] M. Sun, J. Yang, K. Wang, H. Shen, Discovering affective regions in deep convolutional neural networks for visual sentiment prediction, in: IEEE International Conference on Multimedia and Expo, ICME 2016, Seattle, WA, USA, July 11–15, 2016, IEEE Computer Society, 2016, pp. 1–6, http://dx.doi.org/10.1109/ICME.2016.7552961.

[11] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: F. Bonchi, J. Domingo-Ferrer, R.A. Baeza-Yates, Z. Zhou, X. Wu (Eds.), IEEE 16th International Conference on Data Mining, ICDM 2016, December 12–15, 2016, Barcelona, Spain, IEEE, 2016, pp. 439–448, http://dx.doi.org/10.1109/ICDM.2016.0055.

[12] L. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: harvesting opinions from the web, in: H. Bourlard, T.S. Huang, E. Vidal, D. Gatica-Perez, L. Morency, N. Sebe (Eds.), Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14–18, 2011, ACM, 2011, pp. 169–176, http://dx.doi.org/10.1145/2070481.2070509.

[13] V. Pérez-Rosas, R. Mihalcea, L. Morency, Utterance-level multimodal sentiment analysis, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long Papers, ACL 2013, 4–9 August 2013, Sofia, Bulgaria, The Association for Computer Linguistics, 2013, pp. 973–982.

[14] D. Cao, R. Ji, D. Lin, S. Li, A cross-media public sentiment analysis system for microblog, Multimedia Syst. 22 (4) (2016) 479–486, http://dx.doi.org/10.1007/s00530-014-0407-8.

[15] M. Wöllmer, F. Weninger, T. Knaup, B.W. Schuller, C. Sun, K. Sagae, L. Morency, Youtube movie reviews: Sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (3) (2013) 46–53, http://dx.doi.org/10.1109/MIS.2013.34.

[16] Q. You, J. Luo, H. Jin, J. Yang, Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia, in: P.N. Bennett, V. Josifovski, J. Neville, F. Radlinski (Eds.), Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22–25, 2016, ACM, 2016, pp. 13–22, http://dx.doi.org/10.1145/2835776.2835779.

[17] A. Zadeh, M. Chen, S. Poria, E. Cambria, L. Morency, Tensor fusion network for multimodal sentiment analysis, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, Association for Computational Linguistics, 2017, pp. 1103–1114.

[18] L. Pang, C.-W. Ngo, Mutlimodal learning with deep Boltzmann machine for emotion prediction in user generated videos, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 619–622.

[19] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17–23 May 2010, Valletta, Malta, European Language Resources Association, 2010.

[20] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 19-24 June, 2011, Portland, Oregon, USA, in: Proceedings of the Conference, The Association for Computer Linguistics, 2011, pp. 142–150.

[21] N. Howard, E. Cambria, Intention awareness: improving upon situation awareness in human-centric environments, Hum.Cent. Comput. Inf. Sci. 3 (2013) 9, http://dx.doi.org/10.1186/2192-1962-3-9.

[22] S. Kiritchenko, X. Zhu, S.M. Mohammad, Sentiment analysis of short informal texts, J. Artificial Intelligence Res. 50 (2014) 723–762, http://dx.doi.org/10.1613/jair.4272.

[23] S.W.K. Chan, M.W.C. Chong, Sentiment analysis in financial texts, Decis. Support Syst. 94 (2017) 53–64, http://dx.doi.org/10.1016/j.dss.2016.10.006.

[24] H. Kanayama, T. Nasukawa, Fully automatic lexicon expansion for domain-oriented sentiment analysis, in: D. Jurafsky, É. Gaussier (Eds.), Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP 2007, 22–23 July 2006, Sydney, Australia, ACL, 2006, pp. 355–363.

[25] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Predicting elections with twitter: What 140 characters reveal about political sentiment, in: W.W. Cohen, S. Gosling (Eds.), Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23–26, 2010, The AAAI Press, 2010.

[26] M. Taboada, J. Brooke, M. Tofiloski, K.D. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2) (2011) 267–307, http://dx.doi.org/10.1162/COLI_a_00049.

[27] X. Wang, F. Wei, X. Liu, M. Zhou, M. Zhang, Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach, in: C. Macdonald, I. Ounis, I. Ruthven (Eds.), Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24–28, 2011, ACM, 2011, pp. 1031–1040, http://dx.doi.org/10.1145/2063576.2063726.

[28] R. Remus, Asvuniofleipzig: Sentiment analysis in twitter using data-driven machine learning techniques, in: M.T. Diab, T. Baldwin, M. Baroni (Eds.), Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14–15, 2013, The Association for Computer Linguistics, 2013, pp. 450–454.

[29] E. Cambria, T. Mazzocco, A. Hussain, C. Eckl, Sentic medoids: Organizing affective common sense knowledge in a multi-dimensional vector space, in: D. Liu, H. Zhang, M.M. Polycarpou, C. Alippi, H. He (Eds.), Advances in Neural Networks - ISNN 2011 - 8th International Symposium

on Neural Networks, ISNN 2011, Guilin, China, May 29-June 1, 2011, in: Proceedings, Part III Lecture Notes in Computer Science, vol. 6677, Springer, 2011, pp. 601–610, http://dx.doi.org/10.1007/978-3-642-21111-9_68, https://doi.org/10.1007/978-3-642-21111-9_68.

[30] D. Tang, F. Wei, B. Qin, T. Liu, M. Zhou, Coooolll: A deep learning system for twitter sentiment classification, in: P. Nakov, T. Zesch (Eds.), Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23–24, 2014, The Association for Computer Linguistics, 2014, pp. 208–212.

[31] A. Severyn, A. Moschitti, Twitter sentiment analysis with deep convolutional neural networks, in: R.A. Baeza-Yates, M. Lalmas, A. Moffat, B.A. Ribeiro-Neto (Eds.), Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9–13, 2015, ACM, 2015, pp. 959–962, http://dx.doi.org/10.1145/2766462.2767830.

[32] M. Dragoni, G. Petrucci, A neural word embeddings approach for multi-domain sentiment analysis, IEEE Trans. Affect. Comput. 8 (4) (2017) 457–470.

[33] S. Siersdorfer, E. Minack, F. Deng, J.S. Hare, Analyzing and predicting sentiment of images on the social web, in: A.D. Bimbo, S. Chang, A.W.M. Smeulders (Eds.), Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25–29, 2010, ACM, 2010, pp. 715–718, http://dx.doi.org/10.1145/1873951.1874060.

[34] B. Li, S. Feng, W. Xiong, W. Hu, Scaring or pleasing: exploit emotional impact of an image, in: N. Babaguchi, K. Aizawa, J.R. Smith, S. Satoh, T. Plagemann, X. Hua, R. Yan (Eds.), Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29–November 02, 2012, ACM, 2012, pp. 1365–1366, http://dx.doi.org/10.1145/2393347.2396487.

[35] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, J. Tang, How do your friends on social media disclose your emotions? in: C.E. Brodley, P. Stone (Eds.), Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, 2014, QuéBec City, QuéBec, Canada, AAAI Press, 2014, pp. 306–312.

[36] J. Yuan, S. Mcdonough, Q. You, J. Luo, Sentribute: image sentiment analysis from a mid-level perspective, in: E. Cambria, B. Liu, Y. Zhang, Y. Xia (Eds.), Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013, Chicago, IL, USA, August 11, 2013, ACM, 2013, pp. 10:1–10:8, http://dx.doi.org/10.1145/2502069.2502079.

[37] D. Borth, R. Ji, T. Chen, T.M. Breuel, S. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D.A. Shamma, M. Worring, R. Zimmermann (Eds.), ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21–25, 2013, ACM, 2013, pp. 223–232, http://dx.doi.org/10.1145/2502081.2502282.

[38] C. Xu, S. Cetintas, K. Lee, L. Li, Visual sentiment prediction with deep convolutional neural networks, CoRR abs/1411.5731 (2014) arXiv:1411.5731.

[39] Q. You, H. Jin, J. Luo, Visual sentiment analysis by attending on local image regions, in: S.P. Singh, S. Markovitch (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA, AAAI Press, 2017, pp. 231–237.

[40] M. Wang, D. Cao, L. Li, S. Li, R. Ji, Microblog sentiment analysis based on cross-media bag-of-words model, in: H. Wang, L. Davis, W. Zhu, S. Kopf, Y. Qu, J. Yu, J. Sang, T. Mei (Eds.), International Conference on Internet Multimedia Computing and Service, ICIMCS '14, Xiamen, China, July 10–12, 2014, ACM, 2014, p. 76, http://dx.doi.org/10.1145/2632856.2632912.

[41] L. Li, D. Cao, S. Li, R. Ji, Sentiment analysis of chinese micro-blog based on multi-modal correlation model, in: 2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27–30, 2015, IEEE, 2015, pp. 4798–4802, http://dx.doi.org/10.1109/ICIP.2015.7351718.

[42] S. Poria, E. Cambria, N. Howard, G. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59, http://dx.doi.org/10.1016/j.neucom.2015.01.095.

[43] S. Poria, H. Peng, A. Hussain, N. Howard, E. Cambria, Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis, Neurocomputing 261 (2017) 217–230, http://dx.doi.org/10.1016/j.neucom.2016.09.117.

[44] Q. You, L. Cao, H. Jin, J. Luo, Robust visual-textual sentiment analysis: when attention meets tree-structured recursive neural networks, in: A. Hanjalic, C. Snoek, M. Worring, D.C.A. Bulterman, B. Huet, A. Kelliher, Y. Kompatsiaris, J. Li (Eds.), Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15–19, 2016, ACM, 2016, pp. 1008–1017, http://dx.doi.org/10.1145/2964284.2964288.

[45] N. Xu, W. Mao, Multisentinet: A deep semantic network for multimodal sentiment analysis, in: E. Lim, M. Winslett, M. Sanderson, A.W. Fu, J. Sun, J.S. Culpepper, E. Lo, J.C. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, V.S. Tseng, C. Li (Eds.), Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06–10, 2017, ACM, 2017, pp. 2399–2402, http://dx.doi.org/10.1145/3132847.3133142.

[46] N. Xu, W. Mao, A residual merged neutral network for multimodal sentiment analysis, in: Big Data Analysis (ICBDA), 2017 IEEE 2nd International Conference on, IEEE, 2017, pp. 6–10.

[47] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, 2016, pp. 289–297.

[48] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016, The Association for Computational Linguistics, 2016, pp. 457–468.

[49] P. Lu, H. Li, W. Zhang, J. Wang, X. Wang, Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, 2018.

[50] A. Karpathy, F. Li, Deep visual-semantic alignments for generating image descriptions, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June, 7–12, 2015, IEEE Computer Society, 2015, pp. 3128–3137, http://dx.doi.org/10.1109/CVPR.2015.7298932.

[51] K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio, Show, attend and tell: Neural image Caption generation with visual attention, in: F.R. Bach, D.M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015, in: JMLR Workshop and Conference Proceedings, vol. 37, JMLR.org, 2015, pp. 2048–2057.

[52] F. Huang, X. Zhang, Z. Li, Learning joint multimodal representation with adversarial attention networks, in: S. Boll, K.M. Lee, J. Luo, W. Zhu, H. Byun, C.W. Chen, R. Lienhart, T. Mei (Eds.), 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22–26, 2018,, ACM, 2018, pp. 1874–1882, http://dx.doi.org/10.1145/3240508.3240614.

[53] F. Huang, X. Zhang, Z. Zhao, Z. Li, Bi-directional spatial-semantic attention networks for image-text matching, IEEE Trans. Image Process. 28 (4) (2019) 2008–2020, http://dx.doi.org/10.1109/TIP.2018.2882225.

[54] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR (2014) arXiv:1409.1556.

[55] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada, 2015, pp. 91–99.

[56] C. Zhang, C. Liu, X. Zhang, G. Almpanidis, An up-to-date comparison of state-of-the-art classification algorithms, Expert Syst. Appl. 82 (2017) 128–150.

[57] C. Zhang, J. Bi, S. Xu, E. Ramentol, G. Fan, B. Qiao, H. Fujita, Multi-imbalance: An open-source software for multi-class imbalance learning, Knowl.-Based Syst. (2019).

[58] M. Katsurai, S. Satoh, Image sentiment analysis using latent correlations among visual, textual, and sentiment views, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2016, pp. 2837–2841.

[59] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46.

[60] J.J. Randolph, Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa, Online Submiss. (2005).

[61] R.M. Borromeo, M. Toyama, Automatic vs. crowdsourced sentiment analysis, in: B.C. Desai, M. Toyama (Eds.), Proceedings of the 19th International Database Engineering & Applications Symposium, Yokohama, Japan, July 13–15, 2015, ACM, 2015, pp. 90–95, http://dx.doi.org/10.1145/2790755.2790761.

[62] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, a meeting of SIGDAT, a Special Interest Group of the ACL, ACL, Doha, Qatar, 2014, pp. 1532–1543.

[63] Q.V. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21–26 June 2014, in: JMLR Workshop and Conference Proceedings, vol. 32, JMLR.org, 2014, pp. 1188–1196.