Information Processing and Management 56 (2019) 102097

Contents lists available at ScienceDirect

Information Processing and Management

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

An image-text consistency driven multimodal sentiment analysis approach for social media

Ziyuan Zhao[a], Huiying Zhu[a], Zehao Xue[a], Zhao Liu[a], Jing Tian[a], Matthew Chin Heng Chua[a], Maofu Liu[*,b]

[a] Institute of Systems Science, National University of Singapore, Singapore 119615, Singapore
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430081, China

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Textual sentiment
Visual sentiment
Social media

ABSTRACT

Social media users are increasingly using both images and text to express their opinions and share their experiences, instead of only using text in the conventional social media. Consequently, the conventional text-based sentiment analysis has evolved into more complicated studies of multimodal sentiment analysis. To tackle the challenge of how to effectively exploit the information from both visual content and textual content from image-text posts, this paper proposes a new image-text consistency driven multimodal sentiment analysis approach. The proposed approach explores the correlation between the image and the text, followed by a multimodal adaptive sentiment analysis method. To be more specific, the mid-level visual features extracted by the conventional SentiBank approach are used to represent visual concepts, with the integration of other features, including textual, visual and social features, to develop a machine learning sentiment analysis approach. Extensive experiments are conducted to demonstrate the superior performance of the proposed approach.

1. Introduction

With the rapid growth of the social media, users tend to share their opinions in social media platforms such as Twitter, Facebook and Sina Weibo. These user-generated content is moving toward a diversification of content and formats, where people tend to post text embedded images, namely *image-text posts* (Soleymani et al., 2017; Yu, Qiu, Wen, Lin, & Liu, 2016). The posts are more informative since they contain visual content in addition to texts, unlike the conventional text-only posts. Sentiment analysis aims to automatically uncover the underlying attitude of the posts. Due to the rich sentiment cues that can be found in images, sentiment analysis of visual content can contribute more towards extracting user sentiments and understand user behavior, stock market forecasting and voting for politicians (Jiang et al., 2017; Nie, Peng, Wang, Zhao, & Su, 2017; Peng, Shen, & Fan, 2013). Taking the examples of some popular posters, as illustrated in Fig. 1, it can be seen that some posters record their time and express their expectations for the next period. Fig. 1(b) shows a dandelion with the words, 'Goodbye November', and show a beautiful tree with Chinese lanterns hanging from it. These kinds of posters can conjure up a positive sentence of confidence about the future. Fig. 1(c) - a scene from New York, posters can help them record valuable travelling experience in certain cities, like this photo in New York topic.

The major challenge of sentiment analysis for social media lies in effective feature extraction and representation for both text

* Corresponding author.
E-mail addresses: zhaoziyuan@u.nus.edu (Z. Zhao), zhuhuiying@u.nus.edu (H. Zhu), e0267594@u.nus.edu (Z. Xue), liuzhao@u.nus.edu (Z. Liu), tianjing@nus.edu.sg (J. Tian), mattchua@nus.edu.sg (M.C.H. Chua), liumaofu@wust.edu.cn (M. Liu).

https://doi.org/10.1016/j.ipm.2019.102097
Received 8 May 2019; Received in revised form 9 July 2019; Accepted 4 August 2019
Available online 12 August 2019

Fig. 1. Examples of image-text posts: (a) Go New York Giants!#newyorkgiants #football; (b) Goodbye November, Hello December; (c) Love this city #New York.

content and visual content. This challenge has drawn attention in the field of computer vision and especially retrieval and emotional semantic image retrieval, which applies computer vision technology to eliminate the affective gap between low-level features and the emotional content of an image (Machajdik & Hanbury, 2010). In these conventional approaches, low-level visual features, such as color histogram, are directly used into sentiment analysis with textual features. This has caused a great loss of emotional information from image, and consequently, there still exists a great semantic gap between low-level features and emotional content in the images. In view of this challenge, Borth, Ji, Chen, Breuel, and Chang (2013) proposed a more scientific *SentiBank* approach which models mid-level representations based on visual concepts, called *Adjective Noun Pairs* (ANPs), such as "cute cat" and "happy girl", where both the sentimental strength of adjectives and detectability of nouns are considered. This approach has been proven to be useful in detecting emotions depicted in the images.

To tackle the challenge of analyzing both text content and visual content in image-text posts, a text-image consistency driven multimodal sentiment analysis approach is proposed in this paper. The proposed approach is motivated by these two observations: firstly, low-level visual features like color-based features have proved to be simple yet effective for image emotions (Chen, Eldeen, He, Kan, & Lu, 2015). Different colors have different sentiment effects; for example, colors like red, orange and yellow are warm color and gives positive energy and feelings. In view of this, these low-level visual features should be considered in multimodal sentiment analysis. Secondly, the relationship between images and text is very important for multimodal sentiment analysis. In free social media platforms, people can post image-text posts freely without the limitation of image and text consistency, so there exist fake posts which can mislead the sentiment analysis, as seen in Fig. 2(a). Also, to depict certain moods or ideas, people may use satiric expression for their strong sentiment. For instance in Fig. 2(b), the man fared poorly in his exam and he says, 'what a nice day' and wears an unhappy expression which indicates his depressed mood. In Fig. 2(c), the poster says "I am a big fan of red apple", but in this context, the word 'apple' refers to a technology brand instead of a fruit. It is difficult to determine the true meaning just from this short context. In response to this problem, a new image-text correlation model is developed to examine the relationship between images and text. Furthermore, low-level visual features and different textual features are combined as enriched features to derive a multimodal sentiment analysis approach.

The contributions of this paper are two-fold.

- First, to effectively exploit the information from both visual content and textual content from image-text posts, the proposed

Fig. 2. Examples of image-text posts: (a) Bird in the sky; (b) What a nice day; (c) I am a big fan of red apple!.

approach explores the correlation between the image and the text, where the mid-level visual features extracted by the conventional SentiBank approach are used to represent visual concepts, with the integration of other features, including textual, visual and social features.

- Second, the proposed approach performs a multimodal adaptive sentiment analysis by incorporating the aforementioned image-text correlation model into the conventional SentiBank framework. First, for the related image-text data, four types of features (basic text feature, social feature, OCR feature from image and Adjective Noun Pairs (ANP) feature from image) are exploited, while for the unrelated image-text data, only the conventional ANPs features from image is used. By this way, the proposed approach is able to adaptively adjust features used for sentiment analysis.

The rest of this paper is organized as follows: First, a brief literature review is provided in Section 2. Then the proposed multimodal sentiment analysis approach is proposed in Section 3. This is further evaluated in extensive experimental results in Section 4. Finally, Section 5 concludes this paper.

2. Related work

Sentiment analysis, sometimes known as opinion mining, aims to judge emotional orientation (e.g., positive, negative or neural) based on user-generated content (Pang & Lee, 2008). Traditional sentiment analysis concentrates on textual sentiment analysis. However, research on visual sentiment analysis is relatively much less done. In recent years, much research has been done on visual sentiment analysis due to the exponential growth in Internet use. In this section, we will briefly discuss the related work in areas of textual sentiment analysis, visual sentiment analysis and multimodal sentiment analysis in social media.

2.1. Textual sentiment analysis

A brief review on existing textual sentiment analysis approaches is provided in this section. In the existing body of research, most of the sentiment information come from Web, blog and twitters, where the text posts are studied for sentiment analysis (Dave, Lawrence, & Pennock, 2003; Pang & Lee, 2004; Wilson et al., 2005; Yu & Hatzivassiloglou, 2003). In most of these works, textual features are directly extracted from original texts, and then used in sentiment analysis. To further reduce the influence of noise and improve the precision of classification, text preprocessing is needed in textual sentiment analysis (Dave et al., 2003). These special treatments of preprocessing decrease the accuracy of sentiment analysis; therefore, when identifying social data especially subjective sentences, most methods enter emotional words, assisting with various vocabulary and word frequency information into the machine learning classifiers (Dave et al., 2003; Pang & Lee, 2004; Wilson et al., 2005; Yu & Hatzivassiloglou, 2003). A few common methods of feature extraction and pattern mining approach are developed in Nasukawa and Yi (2003); Turney (2002) and Liu, Zhang, Liu, Hu, and Fang (2017) for word frequency and semantic features. In Nasukawa and Yi (2003) and Turney (2002), the sentiment words mining contributes to sentiment analysis. However, only relying on sentiment words may also cause a large deviation, especially for such comprehensive sentences as double negative sentences. For example, the “bad words“ of many negative emotions in horror movies unnecessarily denote the negative emotions of the reviewers. To extract deeper level semantic features, Liu et al. studied the characteristics of Weibo text features, including text length, noun density, verb density and named entity density four-dimensional text features, which contribute to association analysis and sentiment analysis (Liu et al., 2017).

2.2. Visual sentiment analysis

A brief review on existing visual sentiment analysis approaches is provided in this section. Initially,the study of visual sentiment analysis was based on image aesthetic quality assessment (Datta, Joshi, Li, & Wang, 2006; Ke, Tang, & Jing, 2006; Marchesotti, Perronnin, Larlus, & Csurka, 2011)and emotional semantic image retrieval (Colombo, Bimbo, & Pala, 1999; Wang & He, 2008; Zhao, Yao, Yang, & Zhang, 2014). Marchesotti et al. (2011) proposed to classify the aesthetic quality of images by using the support vector machine classifier with common image features including *Bag-Of-Visual Words* (BOVW) (Csurka, Dance, Fan, Willamowski, & Bray, 2004) and *Fisher Vector* (FV) (Perronnin & Dance, 2007). Zhao et al. (2014) proposed the feature representation approach by combining common low-level features, aesthetic features and mid-level features. Hayashi and Hagiwara (1998) proposed a high-accuracy rate based on the back-propagation neural network to construct the mapping relationship between visual features and impression keywords. These methods partly eliminate the gap between low-level features and the emotional content, but how to represent the features of an image is still a problem in visual sentiment analysis.

Apart from common low-level features such as color, texture and contour (Li, Feng, Xiong, & Hu, 2012; Siersdorfer, Minack, Deng, & Hare, 2010), mid-level feature representation, including facial emotion feature (Vonikakis & Winkler, 2012), Sentribute (Yuan, Mcdonough, You, & Luo, 2013) or *Adjective Noun Pairs* (ANPs) (Borth et al., 2013) are also widely used in visual sentiment analysis. More specifically, the ANPs mid-level feature representation was proposed by Borth et al., who also built a large-scale *Visual Sentiment Ontology* (VSO) database and ANP detectors respectively. They first collected labels containing emotional words from Flickr and YouTube, and then got the ANPs after processing. Finally a large-scale Visual Sentiment Ontology of more than 3000 ANPs was constructed. These ANPs can be upload into sentiment classifiers, and these mid-level features have higher accuracy than low-level features. Chen, Borth, Darrell, and Chang (2014) proposed using the pre-trained deep convolutional neural network model for visual concept classification. Jou et al. (2015) enriched the usage of SentiBank in different languages and proposed a large-scale multilingual sentiment concept ontology.

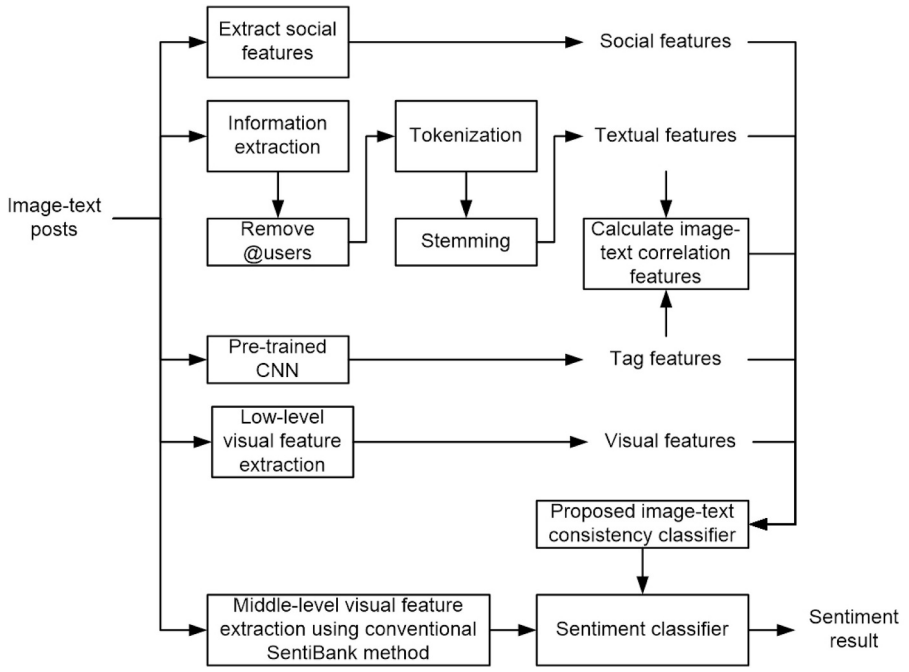

Fig. 3. An overview of the proposed multimodal sentiment analysis approach.

2.3. Multimodal sentiment analysis

Sentiment recognition expressed in social media from multimodal signals, including visual, audio and textual information has been studied and multimodal sentiment analysis is an emerging area (Soleymani et al., 2017; Wang, Qi, Gao, Zhao, & Wang, 2016; Zhao, Ding, Gao, & Han, 2017; Zhao, Gao, Ding, & Chua, 2018; Zhao, Yao, Gao, Ding, & Chua, 2016; Zhao, Yao, Gao, Ji, & Ding, 2017). Social media user generated text is always posted with an accompanying image or short video, and this adds one more channel of information in user sentiment expression. Cao, Ji, Lin, and Li (2016) extracted textual posts with related images extracted from Sina Weibo and conducted sentiment analysis by combing the prediction results of using n-gram textual features and mid-level features (Borth et al., 2013). Xu and Mao (2017) proposed extracting deep semantic features of images by identifying objects and scenes as salient, and then fusing features, which show that deep semantic features they extracted demonstrate high correlations with sentiments. Chen et al. (2015) explored the image-text correlation from multiple view. Liu et al. (2017) proposed to use different textual features (topic, social, etc.) extracted from social media and did feature mapping for image-text correlation.

3. Proposed image-text consistency driven multimodal sentiment analysis approach

In this section, the proposed image-text consistency driven multimodal sentiment analysis approach is presented. The proposed approach, as illustrated in Fig. 3, consists of four critical components, which are briefly described as follows.

- Preprocessing: In the preparation stage, some natural language processing methods e.g. stop-word removal, tokenization, stemming are used to process text data.
- Feature extraction: Three main types of features, i.e. textual feature, visual feature, and image-text similarity, are extracted from the preprocessed image-text post.
- Proposed image-text consistency classifier: A machine learning model is trained using three main types of features from image-text posts to make a binary decision on whether the image content and the text content are consistent with each other.
- Sentiment classifier: Performs sentiment classification on the input image-text posts.

3.1. Preprocessing

Data preprocessing is a critical step in the proposed approach, especially for user generated data from social media platforms, where the data is raw and unstructured. The proposed approach consists the following steps in preprocessing.

- Information extraction: Raw dataset is composed of different kinds of information, including time, source, image width, etc. Based on the system goal and common sense, messages titles and descriptions are selected as text data for further analysis.
- Special symbol removal: In social media platforms, users often post their messages and direct them to others or comments on

others' posts by using symbol @. The information after this symbol is relative to the privacy of users and useless in sentiment analysis, so the words after @ need to be deleted.
- Stop word removal: In natural language processing, some words, also known as "stop words" are filtered out. As such, common stop words were deleted.
- Tokenization: Tokenization is the process of breaking a text corpus most commonly into words, but also into phrases and meaningful elements, all of which are known as tokens. The tokens become the basic units for further text processing.
- Stemming: Stemming and lemmatization aim to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For example, the word *love* in the data set considering the other posts may appear similar *loves, loving*, and *lovable*, we can put them all into root love, to reduce the amount of data, and the only word, also won't lose a lot of information.

3.2. Feature extraction

3.2.1. Texture feature extraction

Text feature extraction plays a crucial role in text analysis, directly influencing the accuracy of sentiment classification. *Word2vec* has garnered a lot of interest in the text mining community (Lilleberg, Zhu, & Zhang, 2015; Mikolov, Chen, Corrado, & Dean, 2013). The model derives a supervised learning task from the corpus itself using either the continuous bag-of-words model or continuous skip-gram model. On the other hand, it is considered unsupervised in the sense that one can provide any large corpus of one's own choice.

In this paper, the public pre-trained words and phrase vectors *GoogleNews-vectors-negative300* is used in the proposed approach. It contains 300-dimensional vectors for 3 million words and phrases. It is sufficient enough to contain all our twitter data set corpus. The vector for each word is a semantic description of how that word is used in context, so two words that are used similarly in text will get similar vector representations. It has been shown that the word vectors capture many linguistic regularities, for example vector operations vector('Paris') - vector('France') + vector('Italy') result in a vector that is very close to vector('Rome'), and vector ('king') - vector('man') + vector('woman') is close to vector('queen'). Once mapping words into vector space, one can then use vector math to find words that have similar semantics. After changing each word into 300-dimensional vectors, Twitter text can be seen as a representation of a word serialization. To avoid dimensional disasters and the variant length of the input text, the proposed approach uses the word addition (Maas et al., 2011) method. In this way, we calculate the summation of each dimension to get 300-dimension vectors as a representation of a text, which can be expressed as

$$[v_1, v_2, v_3, \ldots, v_{300}] = \sum_{i=1}^{n} [v_{i1}, v_{i2}, \ldots, v_{i300}]/n. \tag{1}$$

3.2.2. Social feature extraction

The social features can reflect the social characteristics of image-text posts to some extent and generally can be found or derived from the text or its surrounding information. The number of likes, the number of comments, the number of forwards and topic could be considered as social features. Importantly, social media can benefit from knowing how influential a topic will be so that they can determine the amount of coverage they are willing to give to a specific news. A combination of these three measurements will help to gauge the value of a topic:

- Lifespan: to determine if the topic is time specific or long term and how long it actually lasted.
- Emotion transition: determine whether the emotions evolved over time.
- Reach: quantify how many different users got involved in the discussion.

As such, we consider topics only as social features. The proposed approach extracts topic (#) (see Fig. 4) from the image-text twitter, to represent the social information. The social feature extraction is represented as 300-dimensional vectors by word2vec too.

3.2.3. Visual feature extraction

For low-level image features, the proposed approach extracts the low-level image features and combine them. Together, they form the basic image features. In image processing and photography, a color histogram is a representation of the distribution of colors in an image. For digital images, a color histogram represents the number of pixels that have colors in each of a fixed list of color ranges, that span the image's color space, the set of all possible colors. The proposed approach first calculates generic visual features: a 3-by-256 dimension color histogram extracted from the RGB color channels, a 512-dimensional GIST descriptor that has been shown to be useful for detecting scenes like beautiful landscape, a 53-dimensional *Local Binary Pattern* (LBP) descriptor suitable for detecting textures and faces, and a Bag-of-Words quantized descriptor using a 1000 word dictionary with a 2-layer spatial pyramid and max pooling. Middle-level features, *Adjective Noun Pair*s (ANPs) are used as mid-level feature representation in sentiment analysis by SentiBank (Borth et al., 2013). Visual learning of adjectives is understandably difficult due to its abstract nature and high variability. Therefore, we use adjective nouns combinations to be the main semantic concept elements of the image. The advantage of using ANPs, as compared to nouns or adjectives only, is the feasibility of turning a neutral noun into a strong sentiment ANP. Such combined concepts also make the concepts more detectable, compared to adjectives only. The above described ANP structure shares certain similarity with the recent trend in computer vision and multimedia concept detection. To extract the high level semantic

beautiful #love #my #dog #white #like #followme

Fig. 4. The insightful information (after '#') provided in the image-text posts.

Rainbow over today # Sea #NYC sky offers glimpse of hope after Hurricane #Sandy

VGG16	VGG19	Resnet	Inception V3	Xception
seashore	seashore	water_bottle	clog	safety_pin
lakeside	sandbar	monitor	pitcher	mixing_bowl
breakwater	lakeside	jigsaw_puzzle	web_site	sea_urchin
promontory	breakwater	lakeside	saltshaker	toilet_tissue
water_bottle	promontory	seashore	flatworm	hog
sandbar	container_ship	spotlight	groom	sorrel
container_ship	fireboat	syringe	binoculars	guinea_pig
pier	liner	cup	sock	beaver
fireboat	speedboat	screen	leafhopper	marmot
liner	bathing_cap	web_site	lacewing	fox_squirrel

Fig. 5. A sample set of top 10 tags extracted by various visual feature models.

features from images, we use five pre-trained CNN-based models to extract the top 10 tags from images individually. These models are VGG16, VGG19 (Simonyan & Zisserman, 2014), Xception (Chollet, 2017), Inception V3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016) and Resnet (He, Zhang, Ren, & Sun, 2016), which are the state-of-the-art image classification models, and these models are pre-trained by image net which is a large visual database designed for use in visual object recognition research. Over 14 million images have been hand-annotated by ImageNet (Deng et al., 2009) (see Fig. 5), and to make sure that accuracy and diversity of tags, we combine the results from each model and also use the word2vec model into 300-dimensional vectors for further use.

3.2.4. Image-text similarity feature extraction

Similarity feature extraction is carried out to calculate similarities between text and image from the image-text posts. Image features are represented by image tags extracted from 5 models mentioned in the previous section, which are *VGG16, VGG19, Xception, Inception V3* and *ResNet*. The similarity between basic text feature and image tag feature is calculated using the cosine distance between the vector as follows

$$svw(\overrightarrow{term}_i, \overrightarrow{term}_j) = \frac{\overrightarrow{term}_i \bullet \overrightarrow{term}_j}{||\overrightarrow{term}_i|| * ||\overrightarrow{term}_j||}, \tag{2}$$

where $\overrightarrow{term}_i$ represents the vector of image tag feature, and $\overrightarrow{term}_j$ represents the vector of image tag feature. We implement the calculation by using Word2vec. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are

positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. Therefore, the higher results represent stronger relationship in semantic.

3.3. Proposed image-text consistency classifier

In the proposed image-text consistency classifier, the text features (basic text features and topic features) and image features (image basic features and image tag features), as well as and text-image similarity features are concatenated and incorporated into a support vector machine classifier to train a machine learning model (*Support Vector Machine* (SVM) is used in the proposed approach) to decide whether the image content and the text content is correlated to each other or not. The output is *yes* or *no*. If the output is *yes*, that means the text feature and image feature has correlation, then we put both of them into the sentiment classifier. By doing this, we can enhance sentiment prediction accuracy compared with only using text or image feature. If the output of SVM is *no*, this means that the correlation between text and image is not strong, and they may even have opposite sentiments.

3.4. Sentiment classifier

In the final sentiment prediction module of the proposed approach, two SVM-based models are trained for two different conditions: related image-text data and unrelated image-text data. First, for the related image-text data, we input four types of features (basic text feature, social feature, OCR feature from image and ANPs feature from image) from the training data into the SVM-based model and the four types of features from related testing data will be used to predict the sentiment in related image-text post. On the other hand, for the unrelated image-text data, we only put ANPs features from image of training data into another SVM-based model and ANPs feature from images of unrelated testing data will be used to predict the sentiment in unrelated posts.

4. Experimental results

4.1. Dataset description

The dataset used in this is paper is the benchmark data of Visual Sentiment Ontology (Borth et al., 2013). It contains 603 images in total, covering a diverse set of over 21 topics, and there is a corresponding emotional value ground truth. For a further scoop of correlation relationship, it is necessary to make multiple labels for the datasets. Based on the datasets characteristics, six categories of labels are created, including (i) *Noun* means both text and image shares the same object like *newphew*, (ii) *Name* means both text and image appeared the same name like *Obama* in words, (iii) *Word* means both text and image contains the same word, phrase or sentence, like *I am falling in love with you*, (iv) *Verb* means there is a verb in text and image shows a corresponding activity, (v) *Scene* means the text contains a scene and in the image portraying a similar message as that of the text, and (vi) *YorN* means there is some relationship for the image with text or not. Multiple labeling can also occur in the dataset at the same time.

4.2. Performance metric

The following performance metric are used in experiments in this paper, including *precision, recall, f1-score, accuracy*. These four metric are defined as

$$Precision = \frac{TP}{TP + FP}, \tag{3}$$

$$Recall = \frac{TP}{TP + FN}, \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \tag{6}$$

where TP is True Positive and FP is False Positive.

4.3. Implementation

The detailed implementation of the proposed approach is presented in this section. First, after checking each post in the original dataset of 603 posts, we find that 36 posts did not have text. So we only used the remaining 567 posts for the study. We randomly choose 400 samples as training data and the left 157 samples as test data in later model building and performance evaluation.

Next, from these 567 text data, we extracted 567 textual features that are descriptions written by users and 397 social features that are represented topics in the post, since some posts don't have topics. After tokenization, removing stopwords and stemming, each textual feature and social feature was represented by several words respectively. As for these 567 image data, we extracted 4 low-level image features, including color histogram, GIST, LBP, BoW, and combined them as our image basic features, which are

represented by 2000-dimensional vectors. Moreover, we used 5 pre-trained CNN-based models to extract the top 10 tags from images individually, such that each image has 50 tags. We also extracted middle-level visual features by conventional SentiBank method (Borth et al., 2013), which used 1200 ANPs. We got similarity coefficient by using 50 image tags and textual data.

Then, we built an image and text consistency classifier using the SVM approach, based on social features, textual features, similarity coefficients, tag features and visual features. Finally, two SVM sentiment classifiers using RBF kernel were built. The first model is used for related text-image posts based on social features, textual features and ANPs features to the model, while the second model is used for unrelated text-image posts based on ANPs features only.

4.4. Experimental results

The proposed approach is evaluated with conventional approaches, as described in Table 1, the performance evaluation is presented in Table 2. A more detailed performance comparison between the proposed approach with the state-of-the-art approach (Borth et al., 2013) is presented in Table 3. As seen from these tables, the proposed approach achieves overall better performance than conventional approaches, since it is able to adaptively determine the correlation between visual information and the textual information of the image-text posts, and provide a better semantic feature for more accurate sentiment classification.

Table 1
A short description of various sentiment analysis approaches used in the experimental evaluation.

Approach	Description
Visual model	Logistic regression on deep visual features from pre-trained CaffeNet model.
Textual model	Logistic regression on paragraph feature vectors of text.
CCR (You, Luo, Jin, & Yang, 2016)	Cross-modality consistent regression (CCR).
SentiBank (Borth et al., 2013)	Using Sentibank model with ANP analysis.
T-LSTM (You, Cao, Jin, & Luo, 2016)	Using T-LSTM model on the text.

Table 2
The performance evaluation of various sentiment analysis approaches.

Approach	Precision	Recall	F-score	Accuracy
Visual model	0.76	0.72	0.74	0.68
Textual model	0.83	0.64	0.72	0.69
CCR (You, Luo et al., 2016)	0.85	0.76	0.80	0.80
SentiBank (Borth et al., 2013)	0.87	0.84	0.83	0.84
T-LSTM (You, Cao et al., 2016)	1.00	0.81	0.89	0.88
Proposed approach	0.88	0.88	0.88	0.87

Table 3
The performance comparison between the proposed approach and the state-of-the-art approach (Borth et al., 2013).

Category	Method	Precision	Recall	F-score
Positive Sentiment	SentiBank (Borth et al., 2013)	**0.98**	0.67	0.80
	Proposed approach	0.91	**0.83**	**0.87**
Negative Sentiment	SentiBank (Borth et al., 2013)	0.76	**0.99**	0.86
	Proposed approach	**0.86**	0.92	**0.89**

5. Conclusions

An image-text consistency driven multimodal sentiment analysis approach has been proposed in this paper for social media. The proposed approach exploits a image-text consistency approach to decide whether the image content and the text content are consistent with each other, and then adaptively further merge the textual features and the visual features used in the conventional SentiBank to provide more accurate sentiment analysis for image-text posts.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2019.102097.

References

Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S. F. (2013). *Large-scale visual sentiment ontology and detectors using adjective noun pairs. ACM int. conf. on multimedia, Oct*223–232.

Cao, D., Ji, R., Lin, D., & Li, S. (2016). Visual sentiment topic model based microblog image sentiment analysis. *Multimedia Tools and Applications, 75*(15), 8955–8968.
Chen, T., Borth, D., Darrell, T., & Chang, S. F. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. arXiv: 1410.8586.
Chen, T., Eldeen, H. M. S., He, X., Kan, M.-Y., & Lu, D. (2015). *Velda: Relating an image tweet's text and images.* 30–36.
Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. arXiv: 1610-02 357.
Colombo, C., Bimbo, A. D., & Pala, P. (1999). Semantics in visual information retrieval. *IEEE MultiMedia, 6*(3), 38–53.
Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). *Visual categorization with bags of keypoints. European conf. on computer vision, no. 1-22, Sep*1–2.
Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). *Studying aesthetics in photographic images using a computational approach. European conference on computer vision, may*288–301.
Dave, K., Lawrence, S., & Pennock, D. M. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Int. conf. on world wide web, Mar*519–528.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei Fei, L. (2009). *Imagenet: A large-scale hierarchical image database. IEEE int. conf. on computer vision and pattern recognition, Jun*248–255.
Hayashi, T., & Hagiwara, M. (1998). Image query by impression words-the IQI system. *IEEE Transactions on Consumer Electronics, 44*(2), 347–352.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition. IEEE int. conf. on computer vision and pattern recognition, Jun*770–778.
Jiang, B., Yang, J., Lv, Z., Tian, K., Meng, Q., & Yan, Y. (2017). Internet cross-media retrieval based on deep learning. *Journal of Visual Communication and Image Representation, 48*, 356–366.
Jou, B., Chen, T., Pappas, N., Redi, M., Topkara, M., & Chang, S. F. (2015). *Visual affect around the world: A large-scale multilingual visual sentiment ontology. ACM int. conf. on multimedia, Oct*159–168.
Ke, Y., Tang, X., & Jing, F. (2006). *The design of high-level features for photo quality assessment. IEEE int. conf. on computer vision and pattern recognition, Jul*419–426.
Li, B., Feng, S., Xiong, W., & Hu, W. (2012). *Scaring or pleasing: Exploit emotional impact of an image. ACM int. conf. on multimedia, Nov*1365–1366.
Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). *Support vector machines and Word2vec for text classification with semantic features. IEEE int. conf. on cognitive informatics & cognitive computing, Jul*136–140.
Liu, M., Zhang, L., Liu, Y., Hu, H., & Fang, W. (2017). Recognizing semantic correlation in image-text weibo via feature space mapping. *Computer Vision and Image Understanding, 163*, 58–66.
Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Learning word vectors for sentiment analysis. Annual meeting of association for computational linguistics*142–150.
Machajdik, J., & Hanbury, A. (2010). *Affective image classification using features inspired by psychology and art theory. ACM int. conf. on multimedia, Oct*83–92.
Marchesotti, L., Perronnin, F., Larlus, D., & Csurka, G. (2011). *Assessing the aesthetic quality of photographs using generic image descriptors. IEEE Int. Conf. on Computer Vision, Nov*1784–1791.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv: 1301.3781.
Nasukawa, T., & Yi, J. (2003). *Sentiment analysis: Capturing favorability using natural language processing. Int. conf. on knowledge capture, Oct*70–77.
Nie, W.-Z., Peng, W.-J., Wang, X.-Y., Zhao, Y.-L., & Su, Y. T. (2017). Multimedia venue semantic modeling based on multimodal data. *Journal of Visual Communication and Image Representation, 48*, 375–385.
Pang, B., & Lee, L. (2004). *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Annual meeting association for computational linguistics, Jul*271.
Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135.
Peng, J., Shen, Y., & Fan, J. (2013). Cross-modal social image clustering and tag cleansing. *Journal of Visual Communication and Image Representation, 24*(7), 895–910.
Perronnin, F., & Dance, C. (2007). *Fisher kernels on visual vocabularies for image categorization. IEEE int. conf. on computer vision and pattern recognition, Jun*1–8.
Siersdorfer, S., Minack, E., Deng, F., & Hare, J. (2010). *Analyzing and predicting sentiment of images on the social web. ACM int. conf. on multimedia, Oct*715–718.
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556.
Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing, 65*, 3–14.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the inception architecture for computer vision. IEEE int. conf. on computer vision and pattern recognition, Jun*2818–2826.
Turney, P. D. (2002). *Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. Annual meeting on association for computational linguistics, Jul*417–424.
Vonikakis, V., & Winkler, S. (2012). *Emotion-based sequence of family photos. ACM Int. Conf. on Multimedia, Nov*1371–1372.
Wang, F., Qi, S., Gao, G., Zhao, S., & Wang, X. (2016). Logo information recognition in large-scale social media data. *Multimedia Systems, 22*(1), 63–73.
Wang, W., & He, Q. (2008). *A survey on emotional semantic image retrieval. IEEE Int. Conf. on Image Processing, Dec*117–120.
Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., et al. (2005). *Opinionfinder: A system for subjectivity analysis. HLT/EMNLP on interactive demonstrations*34–35.
Xu, N., & Mao, W. (2017). *Multisentinet: A deep semantic network for multimodal sentiment analysis. ACM int. conf. on information and knowledge management*2399–2402.
You, Q., Cao, L., Jin, H., & Luo, J. (2016). *Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. ACM int. conf. on multimedia conference*1008–1017.
You, Q., Luo, J., Jin, H., & Yang, J. (2016). *Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. ACM int. conf. on web search and data mining*13–22.
Yu, H., & Hatzivassiloglou, V. (2003). *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. Int. conf. on empirical methods in natural language processing*129–136.
Yu, R., Qiu, H., Wen, Z., Lin, C. Y., & Liu, Y. (2016). A survey on social media anomaly detection. *ACM SIGKDD Explorations Newsletter, 18*(1), 1–14.
Yuan, J., Mcdonough, S., You, Q., & Luo, J. (2013). *Sentribute: Image sentiment analysis from a mid-level perspective. Int. workshop on issues of sentiment discovery and opinion mining, Aug*10:1–10:8.
Zhao, S., Ding, G., Gao, Y., & Han, J. (2017). *Approximating discrete probability distribution of image emotions by multi-modal features fusion. International joint conference on artificial intelligence*4669–4675.
Zhao, S., Gao, Y., Ding, G., & Chua, T. (2018). Real-time multimedia social event detection in Microblog. *IEEE Transactions on Cybernetics, 48*, 3218–3231.
Zhao, S., Yao, H., Gao, Y., Ding, G., & Chua, T. S. (2016). Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing, 9*(4), 526–540.
Zhao, S., Yao, H., Gao, Y., Ji, R., & Ding, G. (2017). Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Transactions on Multimedia, 19*(3), 632–645.
Zhao, S., Yao, H., Yang, Y., & Zhang, Y. (2014). *Affective image retrieval via multi-graph learning. ACM int. conf. on multimedia, Nov*1025–1028.