LD-MAN: Layout-Driven Multimodal Attention Network for Online News Sentiment Recognition

Wenya Guo, Ying Zhang, Xiangrui Cai, Lei Meng, Jufeng Yang, and Xiaojie Yuan

***Abstract*—The prevailing use of both images and text to express opinions on the web leads to the need for multimodal sentiment recognition. Some commonly used social media data containing short text and few images, such as tweets and product reviews, have been well studied. However, it is still challenging to predict the readers' sentiment after reading online news articles, since news articles often have more complicated structures, e.g., longer text and more images. To address this problem, we propose a layout-driven multimodal attention network (LD-MAN) to recognize news sentiment in an end-to-end manner. Rather than modeling text and images individually, LD-MAN uses the layout of online news to align images with the corresponding text. Specifically, it exploits a set of distance-based coefficients to model the image locations and measure the contextual relationship between images and text. LD-MAN then learns the affective representations of the articles from the aligned text and images using a multimodal attention mechanism. Considering the lack of relevant datasets in this field, we collect two multimodal online news datasets, containing a total of 14,566 articles with 56,260 images and 251,202 words. Experimental results demonstrate that the proposed method performs favorably compared with state-of-the-art approaches. We will release all the codes, models and datasets to the community.**

***Index Terms*—Multimodal sentiment recognition, online news, attention mechanism, article layout.**

I. Introduction

SENTIMENT recognition is one of the fundamental tasks in artificial intelligence (AI). It is usually formulated as a classification problem to label the sentiment categories for the given content. Automatic assessment of sentiment benefits various AI applications, such as decision-making support [1], and personalized advertising [2]. In particular, online news, one of the mainstream social media items, plays a significant role in our daily lives. In addition to understanding the preferences of news readers, predicting readers' sentiment after they read news is of great value for governments and other organizations in their policy-making process.

Fig. 1. A screenshot of online news entitled "Deadly fire brings review" from China Daily. The news story consists of two images corresponding to different parts of the article. For example, the first image (in the orange box) indicates a hospital ward, which is closely relevant to the 2nd, 4th, and 5th sentences, while the second image (in the green box) relates to the 7th and the penultimate sentences.

Notably, online news usually includes unpaired text and images. As shown in Fig. 1, images in the online news may be mentioned in several paragraphs in different locations. This makes existing methods for multimodal sentiment analysis [3], [4] achieve a degraded performance, since they are designed to handle well-aligned short text and images. As such, new algorithms are required to align images to the informative text to better predict the readers sentiment. This problem has not been investigated in the literature.

To accurately predict readers' sentiment according to multimodal online news, we propose a layout-driven multimodal attention network (LD-MAN) to learn the affective representations of news articles. Rather than directly fusing the sentences and image features, LD-MAN attends to the informative contents from the dozens of sentences corresponding to the images through a multimodal attention module. The attention mechanism is used to learn the importance of the sentences from not only the relevance between text and images but also the news layouts that are associated with image locations. Specifically, for the attention weights learned with news layouts, we introduce

Manuscript received February 1, 2020; revised May 17, 2020; accepted June 15, 2020. Date of publication June 19, 2020; date of current version May 26, 2021. This work was supported in part by the Major Project for New Generation of AI Grant under Grant 2018AAA0100403, in part by NSFC under Grants 61876094, U1933114, U1836109, U1903128, and U1936206, in part by Natural Science Foundation of Tianjin, China under Grants 18JCYBJC15400 and 18ZXZNGX00110. The associate editor coordinating the review of this manuscript and approving it for publication was Raouf Hamzaoui. *(Corresponding author: Ying Zhang.)*

Wenya Guo, Ying Zhang, Jufeng Yang, and Xiaojie Yuan are with the Tianjin Key Laboratory of Network and Data Security Technology, College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: guowenya@dbis.nankai.edu.cn; yingzhang@nankai.edu.cn; yangjufeng@nankai.edu.cn; yuanxj@nankai.edu.cn).

Xiangrui Cai is with the College of Cyber Science, Nankai University, Tianjin 300350, China (e-mail: caixr@nankai.edu.cn).

Lei Meng is with the Senior Research Fellow with the NUS-Tsinghua-Southampton Center for Extreme Search (NExT++) School of Computing, National University of Singapore, Singapore 117417, Singapore (e-mail: lmeng@nus.edu.sg).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.3003648

the distance-based coefficients to measure the distance between sentences and images, which indicates how the news is organized. LD-MAN infers readers' sentiment from extracted news representations containing the contents of text and images along with the layout of the whole news article. The consideration of news layout comes from a common hypothesis of online news, *i.e.*, images are always described by the sentences around them. The contents and locations of images are helpful in important sentence learning.

To the best of our knowledge, there is no publicly available multimodal online news dataset. Therefore, we collect two online news datasets, namely, the Rappler Online News (RON) dataset and the Daily Mail Online News (DMON) dataset, which are detailed in Section III. Notably, each news article in the datasets contains one or more images.

Our contributions are summarized as follows:

- We propose a layout-driven multimodal attention network that extracts news features from both the multimodal contents of text and images and the layout of news articles.
- To better evaluate the proposed method, we created two datasets named RON and DMON. RON contains 10,081 news items with 25,693 images, while DMON contains 4,485 news items with 30,567 images.
- Experimental results demonstrate that the proposed method outperforms the state-of-the-art methods. We will release all the datasets, models and codes used in this paper upon acceptance.

The rest of this paper is organized as follows. In Section II, we review related work. In Section III, we explain how we collected and labeled RON and DMON and provide statistics of the two datasets. In Section IV, we describe the proposed multimodal sentiment analysis model for online news. Finally, we conduct extensive experiments and summarize the entire paper in Section V and Section VI, respectively.

II. Related Work

Sentiment analysis has been widely studied in AI. Users can express their opinions on various social media platforms. This results in plentiful resources for sentiment analysis and promotes the development of automatic sentiment analysis [5], [6]. Many methods have been applied to product comments [2], tweets [7] and movie reviews [8]. We focus on sentiment analysis for online news in this paper.

A. Single-Modality Sentiment Analysis

Because online news contains images and text, we review the sentiment analysis methods with the inputs of images and text, which are closely relevant to this paper.

1) Image-based Methods: Numerous methods have been developed based on images to detect the sentiment evoked by observers [9], [10]. Benefiting the successful application of CNNs in various visual fields, many works fine-tune the state-of-the-art CNNs pretrained on a large-scale general dataset for visual sentiment analysis [11], [12]. Campos *et al.* [12] found that deep CNNs can learn useful features for visual sentiment and even outperform state-of-the-art methods on some datasets. In [13], the emotions are predicted by integrating the features and dependencies among different CNN layers.

Recently, in addition to utilizing the overall features extracted from the deep CNNs, the relationship between some specific visual contents and human emotion has been further considered [14]–[16]. Lee *et al.* [17] used the visual scene to encode the context that comprehensively represents the emotional responses. Considering the significant influence of some local objects on people's emotions, some works have explored methods to recognize human emotions from these objects [18]. In [18], Yang *et al.* proposed automatically discovering affective regions by computing both the objectness score and the sentiment score, and then the selected regions were aggregated to produce the final prediction for the whole image. [19] proposed utilizing attention bias on emotion-evoking objects to help human attention investigation. Panda *et al.* [20] used a weakly supervised method to obtain discriminative emotion features and improve the generalization ability. [21] utilized different levels of visual features and proposed a multilevel region-based framework to discover the sentiment of local regions.

2) Text-based Methods: A number of methods have been proposed to recognize sentiment from textual contents [22], [23]. Early studies usually represent a document as a bag of words and classify the text into sentiment categories by nonneural methods such as SVM [24]. Considering that the sentiment of a document has close ties to the context words, Zhang *et al.* [25] introduced a conditional random fields-based model that considered the context to encode the reviews and mine the sentiment polarity to the products.

Benefiting from the rapid development of deep learning, there exist many massive simple but effective text classification methods that can be used to predict sentiment for textual contents [26], [27]. Iyyer *et al.* [26] presented a simple method in which feedforward layers take the average of the embeddings associated with an input sequence of tokens as input and then perform classification on the final layer's representation. Joulin *et al.* [27] constructed a magnitude fast method for training and evaluation. In addition, many studies have considered the characteristics of sentiments and design corresponding neural networks. From words to segments, sentiment can be inferred from various granularities. Dragoni *et al.* [28] defined a common-sense ontology based on SenticNet to precisely associate words with different sentiment values. Beyond this word-sentiment relationship, the topic-sentiment correlation has been considered by a segment-level model [29]. Moreover, association with other tasks can also be used to assist sentiment analysis. For example, Majumder *et al.* [30] found that sentiment is always expressed through touches of sarcasm, and they constructed a multitask framework to utilize the correlation between sentiment analysis and sarcasm detection. Recently, in contrast to analyzing the sentiment of a sentence or a document, revealing sentiment for multiple aspects has gradually attracted attention [31]. Tang *et al.* [32], [33] constructed a neural network that models user-comment and product-comment consistencies and rates numeric scores to products accordingly.

While neural networks are regarded as black boxes, debuggability and interpretability have been emphasized in recent years.

The attention mechanism has been widely explored to assist and explain how networks model words and sentences. [34] is one of the typical works modeling documents with hierarchical attention. Letarte *et al.* [35] proposed a flexible and interpretable architecture for text classification and verified the importance of attention in modeling insightful relations between words and enhancements to predictions.

B. Multimodal Sentiment Analysis

In contrast to sentiment analysis based on a single modality, multimodal sentiment analysis has gained increasing attention in recent years. There exist various multimodal datasets with sentiment labels. Most of the existing methods are designed for videos [36]–[38], tweets [39], microblogs [40], [41], or biological signal data [42]. Additionally, inspired by text-only aspect-level sentiment analysis, Xu *et al.* [43] proposed a new task, aspect-based multimodal sentiment analysis. For the large quantity and fast-changing online content, the methods developed for conventional datasets may not be applied to scenarios involving emerging events. Some works focus on real-time sentiment analysis. To deal with a large number of online videos, Tran *et al.* [44] proposed a real-time multimodal sentiment model in which an extreme learning machine is leveraged to improve the processing speed. In addition, Dou *et al.* [45] proposed a multimodal emotion recognition system to meet the real-time requirements of elderly communication robots. To the best of our knowledge, this paper is the first to propose multimodal online news datasets with sentiment labels. Compared with other multimodal datasets, *e.g.*, tweets and videos, news articles consist of much longer text and more images. Additionally, news articles are organized with certain layouts, where images are inserted in the text and closely relevant to their contexts.

The availability of multimodal data makes it possible to understand the sentiment of documents from multiple views, such as visual, textual and audio cues [46]. Numerous works have been designed to model and fuse unimodal representations into a unified multimodal representation [4], [47]. In [48], Cambria *et al.* proposed a sentic blending method that can continuously interpret semantics and sentics on the basis of the integration of affective knowledge and common sense within multiple modalities. Zadeh *et al.* [47] used three unimodal neural networks for three modalities (language, acoustic and visual) and combined the extracted features through tensor fusion. Arevalo *et al.* [49] proposed a gated multimodal feature fusion method that learns to decide how modalities influence the activation of the unit using multiplicative gates. Because images in social data usually have interconnections with each other, Xu *et al.* [50] used a hierarchical deep fusion model to model multimodal interactions at different levels while considering the linkages among social images. In contrast to simply combining different data modalities, Huang *et al.* [51] proposed a mixed method to fuse the discriminative features and the internal correlation between modalities. Analogously, Sunny *et al.* [52] utilized a unified network to extract both the common intramodal information and modality-specific information. Chaturvedi *et al.* [53] extended the affective common-sense vector space from the one for English to that of multiple modalities by projecting multimodal features into a common space. They also used a fuzzy logic classifier to address the partial or mixed sentiments expressed by humans.

In addition to directly fusing multiple unimodal features, many methods have been proposed to explore the interaction between multiple modalities and achieve multimodal representations with more adequate information [40], [54]. These methods are often designed considering the characteristics of the corresponding target datasets. You *et al.* [55] treated images and text in a structural fashion, and aligned words and regions by a semantic tree for robust sentiment analysis. Wang *et al.* [54] proposed a select-additive learning CNN to learn generalizable features across speakers. [40] identified the independence of sentiment in different modalities and introduced a robust sentiment prediction method. [56] explored the correlation between text and image and proposed an adaptive sentiment analysis approach. Du *et al.* [42] constructed a method to model the multimodal relationship with a shared latent space. Additionally, the interaction among different modalities is modeled via the attention mechanism [57], [58]. Zhu *et al.* [57] utilized cross-modality attention and bidirectional RNN to learn robust joint representation. [58] relied on visual information to identify important sentences in documents by an attention mechanism. Xu *et al.* [59] exploited the complementary and comprehensive information between text and images to extract emotional visual and textual features. Additionally, in [60], Poria *et al.* illustrated different factors that should be considered in multimodal sentiment analysis, such as the importance of different modalities and the generalizability of the algorithm.

C. Sentiment Analysis for Online News

Sentiment analysis for online news from the reader's perspective was investigated decades ago [61]. Effective reorganization of readers sentiment can help online news providers attract more attention and more clicks [62]. In the beginning, only the text was considered to infer readers emotions. Rao *et al.* [63] associated words with emotions by building an emotional dictionary. Li *et al.* [64] proposed some filtering schemes to decrease the original dataset according to the frequent terms and contextual polarity of a news article. In [65], readers evoked emotions are associated with latent topics in news articles. In addition to analyzing topics of news articles, Liu *et al.* [66] propose a probabilistic generative model to analyze the evolution of topics and sentiments over time. Because of the evolution of online news content, studies are focusing on news videos [67] and some key factors in news videos, such as important entities, news time, and places [68]. However, as stated in Section I, sentiments evoked by news images are closely related to both the text and images. Our work is the first to explore the images and the news layout on readers' moods.

In this paper, we focus on text and images to address the long text and the unique structure in online news, and we propose an attention-based multimodal sentiment analysis method. Compared with the existing attention-based methods, our method

considers more interaction between text and images. Specifically, in addition to the basic semantic correlation, our method models the layout of online news by adding the spatial relationship between sentences and images into the procedure of a multimodal attention mechanism. The readers' sentiment is recognized from a unified representation extracted by fusing informative textual and visual features.

III. Datasets

With the advent of social media, we have observed an increase in the use of images in online news. However, there is currently no online news dataset with images available for sentiment analysis. To better analyze the impact of these images on readers' moods, we collect two English online news datasets, named the Rappler Online News (RON) dataset and the Daily Mail Online News (DMON) dataset. The datasets can be used in many meaningful practical applications.

A. Online News Acquisition

Every news article in both datasets contains one or more images. To retain the locations of images, we mark their unique identifiers in the text. We reviewed the layout of news on different browsers on different devices. The retained locations of images are consistent with the layouts shown in the web pages. That is, we preserve the layout presented to readers in our datasets. The form of crawled news is the same as the sample in Fig. 1.

The RON dataset was crawled from a Philippine news website called Rappler (*www.rappler.com*). The Rappler website provides a labeling function of 8 emotions for each article and enables the readers to select one of the 8 emotions after reading the article. Therefore, the data and their corresponding labels are all available on the Rappler website. The 8 emotions in Rappler were used in our RON dataset, including happy, sad, annoyed, do not care, inspired, afraid, amused, and annoyed. For every news article, we selected the most voted emotional category as the label. We did not distinguish emotion tagging from sentiment analysis in the remainder of the paper for brevity.

The DMON dataset was collected from the Daily Mail website (*www.dailymail.com*). We invited 5 annotators (3 females and 2 males) to label the online news as positive, negative or neutral. To control the quality of crowdsourced annotations, we design a qualification test for the annotators. This test contains the standard Emotional Quotient test [69] and annotation for the *Yelp* dataset [58] which is a multimodal online review dataset (please refer to Section V-H1 for details about this dataset). One hundred reviews were randomly selected from *Yelp*. Some annotators were asked to label the sentiment for these samples. The annotation performance was evaluated by computing accuracy with the sentiments in the Yelp dataset. Only the annotators achieving an accuracy of over 90% were allowed to participate in our task. Finally, the 5 annotators participated in the dataset annotation. Each news article was assigned to 3 persons. We labeled news articles with the most selected sentiment out of the 3 annotations. A piece of online news was considered valid only when at least 2 annotators agreed on the exact sentiment. In addition, we also filtered out news relevant to policies for the sake of objectivity. Finally, 4,485 online news articles were retained.

To measure the annotation consistency between different annotators, we compute Fleiss' Kappa for the labels from 3 annotators in the DMON dataset. We obtained $\kappa > 0.65$, which shows a substantial degree of agreement in the annotation. Moreover, to verify the quality of the annotation, we randomly selected 100 news articles from the dataset. Five new annotators were employed to annotate the 100 news items again. We used the majority voting method to determine the newly annotated sentiment labels. 94% of the labels were consistent with the initial labels. We also computed Cohen's Kappa to measure the agreement with the original annotations, obtaining $\kappa > 0.80$, which indicates very good agreement. This proves that the DMON dataset is practicable.

TABLE I
Statistics of the Newly Collected RON and DMON Datasets. News in the DMON Dataset Contains Longer Text and More Images Than the Those in RON Dataset

Dataset	RON	DMON
#news	10,081	4,485
Average #images/news	2.55	6.82
Average #sentences/news	33	35
Average #words/sentence	20	22
#unique words	165,099	86,103

TABLE II
The Distribution of Sentiment of the RON and the DMON Datasets

RON				DMON	
Emotion	#news	Emotion	#news	Emotion	#news
happy	4621	inspired	1488	negative	3163
sad	1138	afraid	604	positive	861
angry	1894	amused	586	neutral	461
do not care	243	annoyed	307		

B. Statistics and Analysis

News articles in the RON dataset were labeled as eight emotional tags, and news articles in the DMON dataset were tagged as one of three sentiment polarities. Table II shows the distribution of sentiment of the RON and DMON datasets. As shown in Table I, the RON dataset contains 10,081 news articles, while there are 4,485 news articles in the DMON dataset due to substantial efforts. The average number of sentences and images in the DMON dataset (35, 6.82) is higher than that in the RON dataset (33, 2.55), which means that the news structure in the DMON dataset is more complex. Because there are more news articles in the RON dataset, the vocabulary size is larger compared to the DMON dataset.

To illustrate the complex structure of online news, we report the statistics in Fig. 2. Fig. 2(a) shows the distribution of image numbers in the RON and DMON datasets. The minimum image number is 1, and there are 40 images in several news articles of the DMON dataset. The number of sentences per news

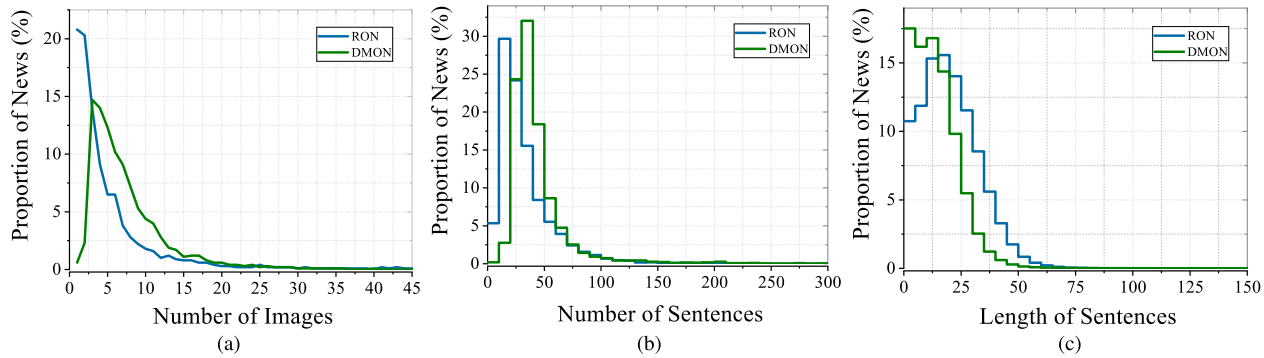

Fig. 2. Statistics of the RON and DMON datasets. Figure (a) shows the numbers of images per news article in the RON and DMON datasets. Figure (b) and (c) show the numbers and lengths of sentences in the RON and DMON datasets.

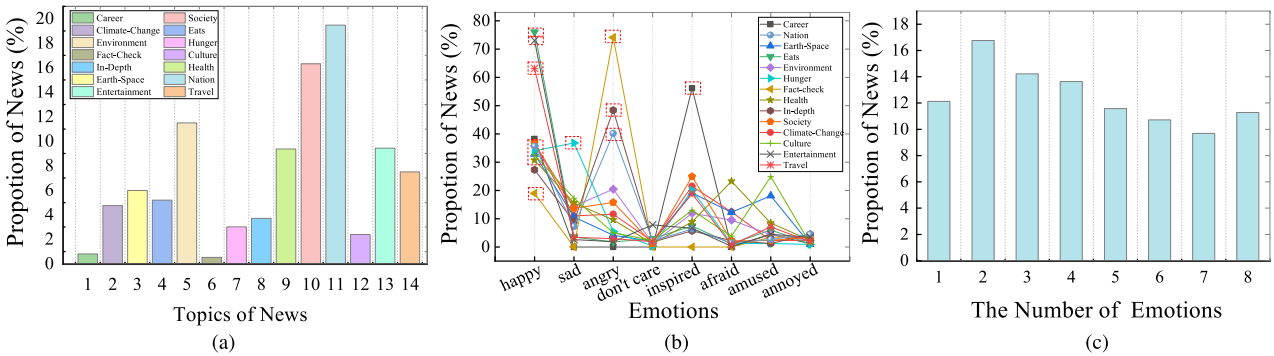

Fig. 3. Statistics of the RON dataset. Figure (a) shows the topic distribution in the RON dataset. Figure (b) shows the distribution of emotions for different news topics, and the red dotted boxes indicate the emotion with the highest proportion in every topic. Figure (c) demonstrates the number of emotion categories corresponding to a piece of news.

article and the length of sentences are shown in Fig. 2(b) and Fig. 2(c), respectively. We can see that over 50% of news articles in the RON dataset contain 10–30 sentences, while the number of sentences in approximately 55% of news of DMON datasets is between 20 and 40. Additionally, to explore whether there are undesirable biases, we show distributions of text length and the image numbers for different sentiments in Fig. 4 and Fig. 5, respectively. It can be observed that for news with different sentiments in the same dataset, the distribution of image numbers and text lengths are approximately the same. In addition, we calculated the Spearman correlation coefficients between the sentiment category and the text length (and the image number) for the RON and DMON datasets, obtaining $|\rho| < 0.1$, which means there is a very weak correlation between text length and sentiment category in the two datasets. Therefore, there is a trivial bias for the text length and the image number.

In addition, we provide more characteristics about the RON datasets in Fig. 3: 1) The news articles were collected from 14 different topics. The topic distribution of news in the RON dataset is shown in Fig. 3(b). These multiple topics diversify the news contents in the RON datasets. Fig. 3(b) shows the distributions of readers' emotions evoked by news articles with different topics. It can be seen that news with different topics causes different emotional distributions of readers, which means that readers' emotions are related to the topic of the news. 2) Because of the

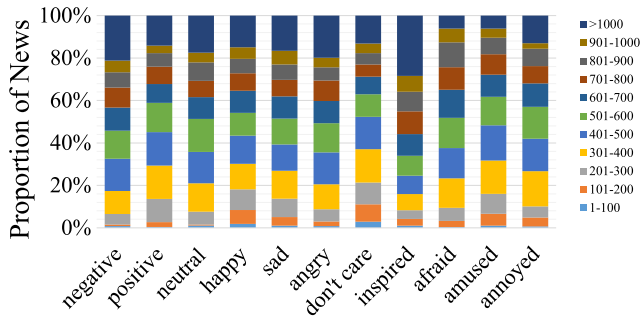

Fig. 4. The distribution of text length for different sentiments. The "negative", "positive" and "neutral" are for the DMON dataset, and the others are the emotions in the RON dataset.

subjectivity of humans' emotions, the same news can arouse different emotions of different readers. Fig. 3(c) shows the statistics that a piece of news contains different emotion labels, in which the horizontal axis represents the number of emotion categories in a news item, and the vertical axis represents the proportion of news items in the RON dataset. We can see that over 80% of the news articles in the RON dataset can evoke more than one emotion on readers. These two characteristics of the RON dataset can help in research on topic-aware sentiment analysis and emotional distribution learning for online news.

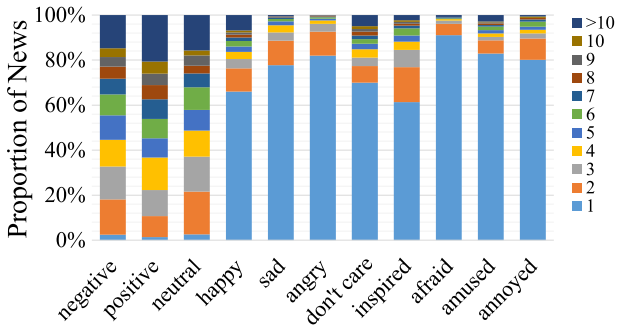

Fig. 5. Illustration of the distribution of image numbers for different sentiments. The "negative", "positive" and "neutral" are for the DMON dataset, and the others are the emotions in the RON dataset.

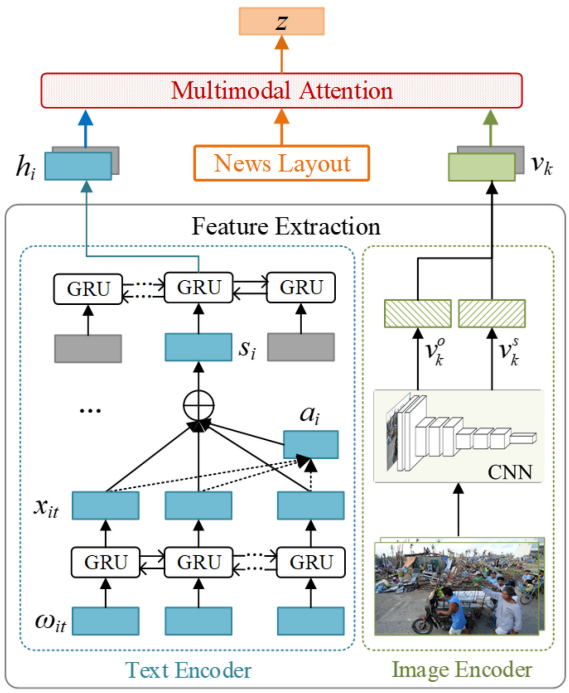

Fig. 6. Illustration of the proposed LD-MAN for online news sentiment analysis. The text and images in the input news were first fed into the feature extraction module. The text encoder and the image encoder were utilized to extract textual features for the sentences and visual features for the images. Then, the obtained features, h_i and v_k, and news layouts were fused into a unified news representation in the multimodal attention module. Finally, readers' sentiment was inferred from the obtained news representation.

IV. Methodology

We propose a layout-driven multimodal attention network (LD-MAN) to recognize readers' sentiment from text and images in news articles. As shown in Fig. 6, LD-MAN consists of two modules: a feature extraction module (including the text encoder and the image encoder), and the multimodal attention (MA) module. Sentence and image features are obtained from the text encoder and the image encoder, respectively. The obtained features are fused into the unified news representations in the MA module. Our method can predict readers' sentiment according to the representation obtained in the MA module.

A. Feature Extraction

As news articles contain both text and images, readers' sentiment can be evoked by either of them. Before analyzing the role of text and images in readers' sentiment, we should represent them properly. This module is used to extract representations for the text and images.

1) Text Encoder: As mentioned above, the text in the news is much longer than in videos or tweets. The long text is organized as a hierarchical structure, *i.e.*, words form sentences, sentences form documents. The hierarchical structure of the text is conducive to better presenting the semantics of a single word and its neighbors, as is the case for the sentences. We use the hierarchical attention network (HAN) proposed in [34] to extract sentence representations in this module.

At the word-to-sentence level, we first represent the words in the sentence with pretrained word vectors. The j-th word in the i-th sentence is denoted as $w_{ij} \in \mathbb{R}^{300}, j \in [0, n_i]$, where n_i is the number of words in the i-th sentence. Beyond the exact semantics of a single word, to represent the content of the whole sentence, we need to consider the relationship between the word and its context word. w_{ij} is fed into a bidirectional GRU [70] model. We concat the forward and backward hidden states as the further annotation for the j-th word in the i-th sentence, and the result is denoted as $x_{ij} \in \mathbb{R}^{200}, x_{ij} = [\overrightarrow{x}_{ij}, \overleftarrow{x}_{ij}]$. x_{ij} summarizes the information of the words centered around w_{ij}. Because not all words contribute equally to the sentence, the attention mechanism can help extract and aggregate informative words to form a sentence vector. We obtain the representation $s_i \in \mathbb{R}^{200}$ for the i-th sentence as

$$s_i = \sum_j a_{ij} x_{ij} \tag{1}$$

where a_{ij} is the attention weight of x_{ij} through a softmax function of the word hidden representation e_{ij}, *i.e.*, $e_{ij} = W_w(\tanh(W_x x_{ij} + b_x))$ and $a_{ij} = \mathrm{softmax}(e_{ij})$.

Analogously, for the sentence-to-document level, we first obtain the representation for the i-th sentence, h_i, using a bidirectional GRU with the input of s_i, *i.e.*, $h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i]$. h_i summarizes the neighbor sentences around the i-th sentence. Then, h_i is used to learn a unified news representation with consideration of the images and news layouts in the multimodal attention module.

2) Image Encoder: It has been proven that the objects and scenes in the image are highly related to readers' moods [71]. For example, a cute dog or a wedding chapel in an image evokes readers positive sentiment, while a dilapidated house or a messy dump evokes negative sentiment. Therefore, we extract both the object and scene features from the images.

Specifically, we use ResNet-152 [72] pretrained on the large-scale Places365 dataset [73] as the scene feature extractor. The 2048 dimensional scene representation is extracted from the "avgpool" layer, denoted as $v_k^s \in \mathbb{R}^{2048}$. Analogously, another ResNet-152 pretrained on ImageNet is used as the object feature extractor. The 2048 dimensional vector extracted from the "avgpool" layer acts as the object representation of the image,

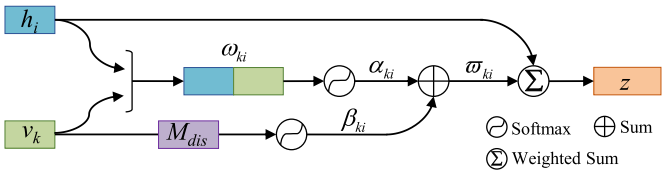

Fig. 7. Illustration of the multimodal attention module. The representation for the whole news, z, is the weighted sum of important sentences learned through a multimodal attention mechanism. The importance for each sentence ϖ_{ki} is the tradeoff of the multimodal semantic attention weight α_{ki} and the distance-based weight β_{ki} derived from the distance matrix M_{dis} between sentences and the image.

denoted as $v_k^g \in \mathbb{R}^{2048}$. We concatenate the object and scene features into a unified representation for the k-th image, denoted as v_k.

B. Multimodal Attention Module

In a news article, dozens of sentences and multiple images are used to describe an event. It is very significant to filter out informative content from the complex article. In this module, LD-MAN achieves multimodal news representations based on the sentence and image features. As an auxiliary of news content, images are more intuitive and can attract readers' attention than text. When representing news content, we utilize them to locate important sentences that can resonate with readers. The multimodal attention procedure is illustrated in Fig. 7.

Specifically, news representation z can be obtained from the previously described sentence and image representations, h_i and v_k, by performing attention from the joint multimodal representation over sentences as follows.

$$z = f_K(\gamma \times \sum_i \varpi_{ki} h_i + (1-\gamma) \times W_v v_k) \tag{2}$$

where $f_K(\cdot)$ is a max-pooling function over the K news annotations corresponding to K images in the news article and The hyperparameter γ controls the tradeoff between the textual and visual features, where γ and (1-γ) indicate how much the two kinds of features contribute to the news representation. As a result, the obtained news representation contains abundant semantic information in text and images. ϖ_{ki} denotes the generated attention weight for the i-th sentence corresponding to the k-th image. Except for the attention weight learned by the sentences denoted as τ_i, ϖ_{ki} consists of two other parts: One part is derived from the semantic relevance of the images and the text, denoted as α_{ki}; The other part is from the layout of the news, *i.e.*, the contextual correlation between images and the sentences which is modeled the distance between the k-th image and different sentences, denoted as β_{ki}.

$$\varpi_{ki} = \tau_i + \lambda_m \alpha_{ki} + \lambda_p \beta_{ki} \tag{3}$$

in which λ_m and λ_p are the weights for α_{ki} and β_{ki}.

1) Multimodal Relevance Attention: The semantic relevance attention weights α_{ki} can be calculated by a softmax function of the multimodal fusion of sentences and the k-th image as $\alpha_{ki} = \text{softmax}(W_m \omega_{ki})$, where ω_{ki} is calculated with the gated multimodal unit [49]:

$$\omega_{ki} = c \odot l_{v_k} + (1-c) \odot l_{t_i} \tag{4}$$

where $\odot$ refers to the elementwise product. c is the multimodal joint embedding, which can be calculated as follows.

$$c = \sigma(W_c[l_{v_k}, l_{t_i}]) \tag{5}$$

in which l_{v_k} and l_{t_i} are features transformed into the same space for the k-th image and features for the i-th sentence, $l_{v_k} = \tanh(W_{lv} v_k)$ and $l_{t_i} = \tanh(W_{lt} h_i)$.

2) Layout Based Attention: As mentioned above, online news contains a certain layout. The content of an image usually expatiates in its context, and images have different interactions with different sentences. In other words, the location reflects how the image interacts with different sentences. Therefore, we compute another distance-based attention weight β to indicate the contextual interaction between text and images. β_{ki} is the distance-based attention weight of the i-th sentence corresponding to the k-th image. The "distance" is the distance between the k-th image and every sentence. Before computing the distance, we number sentences and images based on their order in news and call the result position indexes. p_{v_k} and p_{s_t} are used to indicate the position indexes of the k-th image and the i-th sentence, respectively. Then, we introduce $M_{dis_k} \in \mathbb{R}^L$ as the distance matrix that represents the distance between the k-th image and different sentences, where L is the number of sentences in the given news article. M_{dis_k} is defined as follows.

$$M_{dis_k} = |p_{v_k} - p_{s_t}|,\ k = 1, 2, .., K,\ t = 1, 2, ..., L \tag{6}$$

To utilize this distance information in sentence importance learning, the distance matrix is embedded as a distance-based coefficient μ_k by:

$$\mu_k = e^{-\max(0, W_d M_{dis_k}^{\mathrm{T}} + b_d)} \tag{7}$$

β_k is the normalized result of μ_k. β_k can be used to model the layout in news when LD-MAN learns important sentences.

C. Model Training

Building upon the above modules, the proposed multimodal sentiment analysis for a given online news item can be summarized as follows. Except for the textual features in the images and visual features in text, the layout of the news is also considered. The news is represented as the embedding of the images and important sentences that are learned through the visual contents and the contextual interaction among sentences and images. Based on the multimodal news representation z, we use a fully connected neural network to predict the sentiment label. The likelihood function in this module is defined as follows.

$$\arg\max_{\phi}\ p(\phi|z; W_s) = \arg\max_{\phi}\ f(z; W_s) \tag{8}$$

where ϕ is the predicted sentiment label and $f(\cdot)$ is a softmax probability function making use of cross-entropy loss. W_s conditioned on z is the weight. In the experiments, we set the dimension of W_s as a different value to adapt to the different numbers of label types in different datasets. Because all of the parameters

TABLE III
PERFORMANCE OF THE RECENTLY PROPOSED VISUAL SENTIMENT METHODS FOR THE NEWS IMAGES. THE ABBREVIATIONS REPRESENT DIFFERENT TYPE OF ACCURACIES: "AVG."= THE AVERAGE ACCURACY, "HAP."=HAPPY, "ANG."=ANGRY, "DC."=DON'T CARE, "INS."=INSPIRED, "AFR."=AFRAID, "AMU."=AMUSED, "ANN."=ANNOYED, "NEG."=NEGATIVE, "POS."=POSITIVE, AND "NEU."=NEUTRAL

Methods	RON									DMON			
	Hap.	Sad	Ang.	DC.	Ins.	Afr.	Amu.	Ann.	Avg.	Neg.	Pos.	Neu.	Avg.
SVM	47.71	21.43	25.11	0.0	16.44	23.08	12.96	0.0	31.04	62.50	24.68	14.63	51.57
RF	55.01	20.78	26.12	5.88	10.38	21.52	10.81	6.06	33.40	62.50	36.36	27.59	54.67
VGG [74]	83.93	6.48	6.76	3.60	14.45	10.17	6.12	2.33	52.51	93.18	17.20	5.76	69.26
PDANet [75]	78.42	18.89	28.98	8.47	17.57	21.59	2.33	4.17	54.14	90.75	19.92	9.20	69.82
WSCNet [76]	78.82	14.44	30.25	8.47	15.97	11.36	3.49	4.17	53.72	94.63	14.46	5.60	71.22

in our model can be derived, we use Adam to minimize the loss function.

V. EXPERIMENTS

In this section, we conduct a series of experiments to evaluate the proposed LD-MAN along multiple dimensions. LD-MAN achieves promising performance for multimodal online news sentiment analysis compared to the state-of-the-art methods. We also demonstrate the ability of LD-MAN in selecting informative words and sentences.

A. Implementation Details

Given a news article, we gathered all sentences and mark locations of the images. For the text, we first converted all words to lower cases and split the text into words by the Natural Language Toolkit (NLTK). Each word was transformed into word embeddings (dim=300) by GloVe pretrained on the Wikipedia corpus. The number of hidden units in Bi-GRU was set t0 100. We employed ResNet-152 pretrained on Place365 and ImageNet as the scene and object feature extractor, respectively. LD-MAN was optimized by Adam. The learning rate was initialized as 0.001 and decreased by a factor of 10 every 10 epochs. Both the RON and DMON datasets were split randomly into training, validation and testing sets by 80% : 5% : 15%. In addition, to alleviate the influence of the unbalanced labels on model training, we used the strategy of random oversampling [77] to oversample the minority classes. Note that all the baselines were trained on the oversampled training set. Our framework was implemented by PyTorch [78]. All of our experiments were performed on an NVIDIA GTX 1080ti with 11 GB onboard memory. We used the accuracies of every sentiment and the average accuracy as the evaluation metrics.

B. Image Sentiment Verification

The experiment in this section was conducted to verify whether news images contain sentiment. The sentiment of the news article that contains the image was used as the ground-truth labels for the images. We evaluated some recently proposed visual sentiment approaches on news images. SVM and RF are two nonnumerical methods. PDANet [75] integrates the attention mechanism into a CNN with a constraint for the emotion polarity. Since PDANet was proposed for the task of emotion regression, we altered the last layer to fit the sentiment classification tasks in this paper. WSCNet [76] is our method that utilizes both holistic and localized information to extract robust representation. As reported in Table III, these methods achieved satisfactory results, which shows that the news images contain sentiment clues. Therefore, images can help analyze the sentiment of the whole news item.

C. Performance Comparison

In this section, we compare the performance of LD-MAN with existing sentiment analysis methods. We evaluate the performance of some nonneural methods, such as SVM and RF [38]. Other compared neural methods can be grouped into 2 classes: text-based methods and multimodal methods.

Text-based methods: We compared with the aforementioned HAN [34], which considers the hierarchical structure of long text. We also compared with some simple but efficient methods, such as DAN [26] and fastText [27], which are widely used in text classification. Other CNN-based methods, such as TextCNN [79] and DPCNN [80], were also used.

Multimodal methods: We report the performance of the effective multimodal fusion method GMU [49] and the typical sentiment analysis method for videos such as TFN [47] and LMF [81]. These methods fuse textual and visual features and feed them to a classification layer. The visual features are obtained by max-pooling or avg-pooling, and the corresponding methods have suffixes of "-m" and "-a", respectively. We also compared with VistaNet [58], which is the state-of-the-art method for data including text and images.

We also report human-level performance. Following [82], we invited two persons to report their sentiments for the online news in the test split. The results of humans were computed by averaging the accuracies achieved by two people.

1) Results: The performance of LD-MAN and previously mentioned neural and nonneural methods are reported in Table IV. Because neural models can automatically learn more informative features, the neural methods outperformed all the nonneural models. LD-MAN achieved the best performance in both datasets. Random indicates a sentiment was selected randomly. As expected, it performed the worst compared with the other methods. We can draw observations that text-based methods perform better than image-based methods, which illustrates the major role of the text and the subsidiary role of images in readers' sentiment. Methods with multiple modalities yielded

TABLE IV
PERFORMANCE OF LD-MAN AND THE COMPARED METHODS FOR SENTIMENT ANALYSIS ON RON AND DMON DATASETS. THE COMPARED METHODS CAN BE DIVIDED INTO 2 CLASSES: TEXT-BASED METHODS ("T"), AND MULTIMODAL METHODS FOR SENTIMENT ANALYSIS ("I+T"), AMONG WHICH LD-MAN ACHIEVED THE BEST PERFORMANCE IN BOTH OF ACCURACY (%). AND THE ABBREVIATIONS HAVE THE SAME MEANING AS TABLE III

	Methods	RON									DMON			
		Hap.	Sad	Ang.	DC.	Ins.	Afr.	Amu.	Ann.	Avg.	Neg.	Pos.	Neu.	Avg.
	Human	94.78	95.49	97.15	87.79	89.61	93.82	89.87	93.38	94.05	97.45	89.98	91.70	95.43
	Random [38]	17.85	19.47	10.96	10.71	12.33	12.31	11.11	22.22	15.15	30.79	27.27	26.83	29.82
	SVM-T [38]	35.47	10.62	12.33	50.00	2.05	27.69	42.59	3.70	23.23	41.77	76.62	53.66	48.88
	RF-T [38]	81.46	9.73	27.85	0.0	16.44	13.85	3.70	0.0	42.52	96.95	12.99	4.88	73.99
	HAN [34]	60.95	34.42	59.79	2.94	35.52	34.18	16.22	3.03	48.16	89.81	56.57	24.14	77.75
T	DAN [26]	76.91	15.58	45.70	5.88	9.84	36.71	5.41	9.09	46.35	85.65	51.51	22.41	73.68
	FastText [27]	79.12	33.12	26.11	2.94	13.66	40.51	5.41	3.03	45.65	94.21	49.49	15.52	78.95
	TextCNN [79]	57.38	27.27	42.61	5.88	42.08	39.24	18.92	9.09	43.91	90.74	33.33	20.69	74.19
	DPCNN [80]	67.74	27.27	25.43	2.94	25.14	13.91	4.05	3.03	40.15	93.28	15.15	29.31	73.85
	SVM-MD-m [38]	51.26	22.12	34.25	0.0	22.60	23.08	11.11	0.0	34.71	66.16	27.27	19.51	55.16
	RF-MD-m [38]	40.41	33.12	49.14	2.94	59.56	41.77	9.46	18.18	40.92	96.95	9.09	2.44	73.09
	SVM-MD-a [38]	56.03	31.17	48.80	5.88	33.33	39.24	14.86	3.03	43.56	74.54	61.62	34.48	68.42
	RF-MD-a [38]	47.19	42.86	56.70	2.94	37.16	36.71	10.81	6.06	42.94	98.40	9.09	1.72	75.04
	TFN-m [47]	74.53	11.04	68.73	2.94	21.86	48.10	4.05	3.03	51.43	89.12	57.58	18.97	76.91
I	LMF-m [81]	76.57	26.61	54.98	5.88	16.39	37.97	6.76	3.03	50.10	90.51	59.60	22.41	78.61
+	GMU-m [49]	62.99	31.17	46.05	5.88	42.62	53.16	17.57	3.03	47.94	96.99	29.29	6.90	76.74
T	TFN-a [47]	69.44	26.62	58.76	1.24	32.24	21.52	12.16	9.09	49.34	90.51	47.47	22.41	76.57
	LMF-a [81]	55.01	35.06	48.45	2.94	46.99	34.18	12.16	3.03	44.75	93.97	41.41	20.69	77.92
	GMU-a [49]	68.25	34.42	49.83	18.03	46.84	51.90	9.46	3.03	47.32	91.67	37.37	8.62	74.36
	VistaNet [58]	71.99	21.43	57.04	2.94	12.57	68.35	1.35	9.09	49.06	91.72	58.81	14.89	77.35
	LD-MAN (Ours)	67.01	26.62	71.48	4.04	38.80	59.49	5.41	3.03	**53.51**	92.82	69.69	10.34	**80.81**

better performance than methods with a single modality. Therefore, it is important to consider the images in online news articles.

For single-modality methods, HAN outperformed other text-based methods because the hierarchical structure in HAN captured more document-level characteristics than other methods. For multimodal methods, because the text in news is much longer than that in videos, ignoring the hierarchical structure of the text limits the performance of TFN for online news sentiment analysis. The performance of VistaNet is limited due to ignoring the layout of news.

2) Further Discussion: As reported in Table IV, on the RON dataset, almost all of the methods obtain poor performance for the type of "DC.", "Amu." and "Ann.". A similar situation occurred in the DMON dataset for news with a type of "Neu.". There could be 2 reasons: First, as reported in Table II, the number of some categories (*e.g.*, "neutral") was smaller than other categories in the same dataset, which hindered the models learning the characteristics of news articles with these sentiment categories. Second, the features for some emotions, such as "amused" and "happy", were so similar that it was difficult to distinguish for the models. The human-level performance was better than other machine-based methods (including LD-MAN). Extracting effective sentimental clues from such a complex structure of online news is one of the major research directions to be explored.

D. Ablation Study

The experiments conducted in this section evaluated the influence of the visual contents and image positions on the sentiment analysis for online news. The results are shown in Table V. Our baseline was HAN in which only the text and its hierarchical structure were considered (*i.e.*, the weighted sum of important sentence representations were used to predict readers' sentiment). From Table V, we can draw the following conclusions. Using image features improves the accuracy while combining image locations further improves the performance, which indicates the necessity of considering images and the news layout. In addition, from row-3 and row-4, we can see that the contents and locations of images had a similar effect on the performance, and image locations were slightly more effective than image contents. Our proposed LD-MAN achieved the best performance by considering both the contents and locations of images, which indicates that both the textual and visual contents and the news layout are important for predicting readers' sentiment.

TABLE V
ABLATION STUDY ON THE RON AND DMON DATASET. THE BASELINE IS HAN WHICH ONLY CONSIDERS THE HIERARCHICAL TEXT CONTENTS. THE "VISUAL CONTENT" DENOTES A METHOD THAT USES IMAGE CONTENT TO HELP LEARN IMPORTANT SENTENCES. AND THE "NEWS LAYOUT" DENOTES THE METHOD THAT USES THE LAYOUT OF NEWS TO HELP LEARN IMPORTANT SENTENCES

HAN	Visual Content	News Layout	RON	DMON
✓			48.16	77.75
	✓		40.99	73.34
✓	✓		49.62	78.95
✓		✓	49.97	79.29
✓	✓	✓	**53.51**	**80.81**

E. Hyperparameter Analysis

In this section, we analyze the performance of the proposed method when the image attention weights, *i.e.*, λ_m and λ_p in Equation (3), were set to various values. We report the performance of sentiment classification on the RON dataset in Fig. 8. As shown in the figure, setting $\lambda_m = 0.8$ and $\lambda_p = 0.7$ achieved

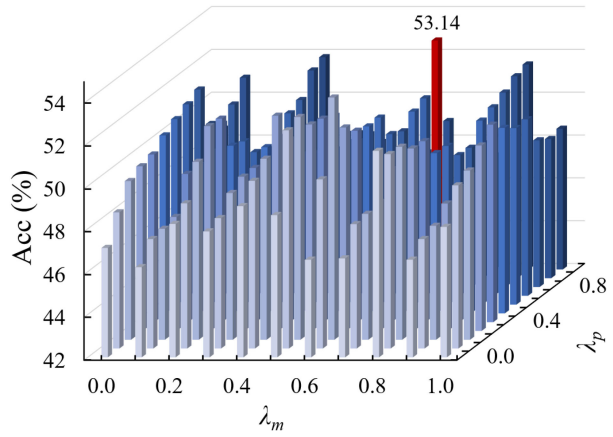

Fig. 8. Impact of different λ_m and λ_p in Equation (3) on the performance of LD-MAN. The results were evaluated on the validation set of the RON dataset.

the best average accuracy on the validation dataset. Only using image content attention (*i.e.*, $\lambda_m = 1.0\,\lambda_p = 0$), or only using image location attention (*i.e.*, $\lambda_m = 0\,\lambda_p = 1.0$) limits the performance. Therefore, it is necessary to consider both the image content and locations in the multimodal attention module.

We also report the performance when γ in Equation (2) is set with different values. In addition, to explore whether the images are relevant to the text above them, below them or both, we show the results when images act on different sentences. We adjusted the distance matrix in Equation (6). By default, the distance matrix in Equation (6) takes the distance between sentences above and below the images into account. For the situation where the distances between the sentences above the image were considered, we set M_{diskt} as 0 when $p_{v_t} > p_{s_t}$; otherwise, we set M_{diskt} to 0 when $p_{v_t} < p_{s_t}$. As shown in the figure, LD-MAN achieved better results when considering the contextual interaction with the sentences above and below images and LD-MAN. Our method achieved the best performance when $\gamma = 0.1$. Therefore, we set $\lambda_m = 0.8\,\lambda_p = 0.7\,\gamma = 0.1$ for the RON dataset. Analogously, we set $\lambda_m = 0.8\,\lambda_p = 1.0\,\gamma = 0.8$ for the DMON dataset.

F. Effect of Online News Layouts

We conducted an experiment to investigate the influence of the layout on the performance of LD-MAN. We identified different positions of image I_k by setting p_{v_k} in Equation (6) to different values. We set $p_{v_k} = 1$ to indicate that image k was at the beginning of the news, whereas $p_{v_k} = L$ indicates that image k was at the tail of the news. The results when images are set in the head, tail and 5 random positions of the news articles are reported in Table VI. It can be observed that the layout has certain practical significance for LD-MAN. Keeping the layout the same as original news articles led to the best accuracy. Different random image locations yielded different performances. Setting images in the head and tail led to better performance than other random positions because sentences at the beginning and end of news articles are usually more comprehensive. Analogously, sentences at the end of news articles usually summarize the whole news, which contains abundant information, so putting images at the end of news articles has less impact on the effect. The effect of the layout on the sentiment is interesting, and we will study this issue in future work.

TABLE VI
INFLUENCE OF CHANGING THE LAYOUT ON RON AND DMON DATASETS. "#CHANGE" DENOTES THE NUMBER OF SENTIMENT FLIPS ACROSS DIFFERENT SENTIMENT CATEGORIES. THE TEST SETS CONTAIN 1,108 AND 448 NEWS ARTICLES IN THE RON AND DMON DATASET, RESPECTIVELY

Layout	RON		DMON	
	#change	Acc (%)	#change	Acc (%)
Original	-	**53.51**	-	**80.81**
Head	385	52.66	118	80.40
Tail	399	52.21	118	80.11
Random	402	52.02	120	79.93
	407	51.81	121	79.62
	410	51.65	125	79.34
	425	51.47	129	79.00
	431	50.05	132	78.63

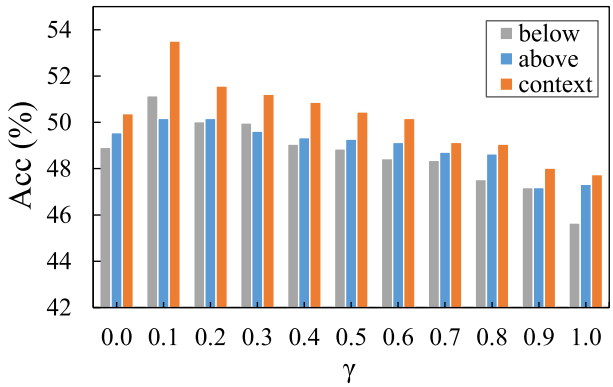

Fig. 9. Impact of different γ in Equation (2) on the performance of LD-MAN. The results were evaluated on the validation set of the RON dataset.

G. Case Study

To validate the ability of our model to select informative sentences and words in news articles, we visualize the importance of sentences and informative words from different methods for RON and DMON datasets in Fig. 10. The sentence selection results are shown in Fig. 10(a) and 10(c), which display the attention weights of the sentences and the image locations. Figures 10(b) and 10(d) show the selected keywords in sentences with the top-5 attention weights.

For important sentence selection, as shown in Figures 10(a) and 10(c), when only the text was considered, the model only extracted clues from the complicated relationship among the sentences. The attention weights were learned from the sentence semantics. After the images and news layouts were considered, the model inferred the relevance between text and images and learned important sentences from these rich clues. The attended words and sentences were more accurate. It is more likely to obtain the correct prediction from this accurate information.

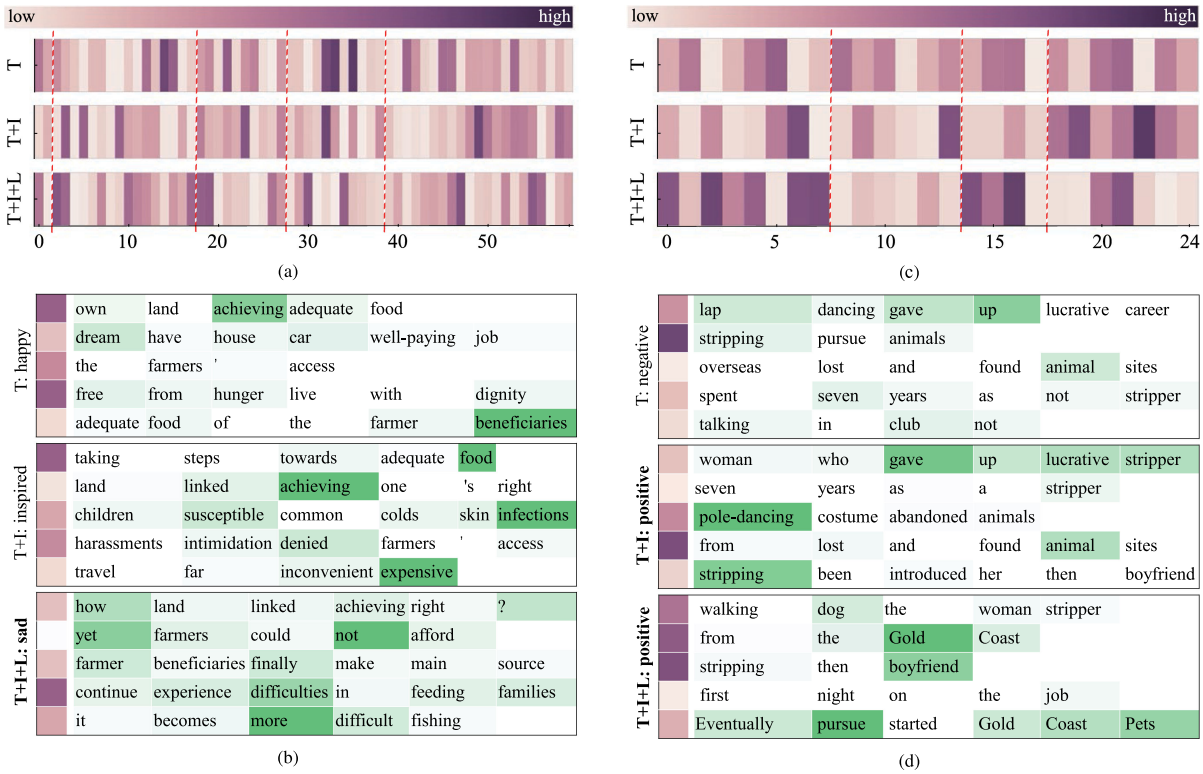

Fig. 10. Visualization of sentence weights and informative words. Results for RON and DMON datasets are given in (a, b) and (c, d), respectively, where 'T'=text-only methods, 'T+I'=methods with separate text and images and 'T+I+L'=LD-MAN. The weights of different sentences in the attention operation are shown in (a) and (c), in which red dotted lines indicate positions of images. Informative words and sentiments predicted by different methods are shown in (b) and (d), in which the predicted labels are listed. The text in bold denotes the methods reached correct sentiments.

For selected keywords, as shown in Fig. 10(b), for the news entitled "Hacienda Matias: Taking steps towards the right to adequate food" (*https://www.rappler.com/move-ph/issues/hunger/1001 57-hacienda-matias-right-food*), the text-only method selects words carrying positive sentiment such as "achieving" and "beneficiaries", which leads to the prediction of "happy". The method with text and image content selected more meaningful words such as "susceptible" and "infections" and obtained "inspired" with the pictures of some people. LD-MAN considered the contextual relationship of the image and different sentences and figures out some other sentences and words such as "not" and "experience difficulties" and recognizes the correct sentiment. In the sample news (*http://www.dailymail.co.uk/news/article-3170627/Vanessa-O-Brien-gave-career-stripper-follow-passion-animals-rescued-thousand-lost-dogs-pets.html*) from the DMON dataset, there are images of cute animals and strippers. Directly ignoring the images caused incorrect predictions for the text-only method. Moreover, the method with text and images considers the images and select words corresponding to images such as "stripper" and "animal", but fail to obtain the contextual relationship between the text and images, which leads to the incorrect prediction.

For the incorrectly classified sample in Fig. 11, the news with the title of "How one worker escaped the HTI fire in Cavite" contains two aspects, *i.e.*, the factory is on fire and a worker successfully escapes from the fire. Readers' emotion after they

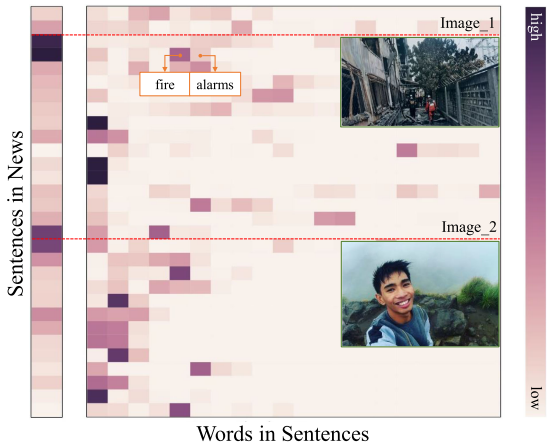

Fig. 11. Heatmap for a wrongly classified sample from the RON dataset. The left heatmap in the bar shows the attention weight of each sentence in the news, and the right heatmap in the rectangle shows the weight of each word in the sentences. The red dashed lines denote the two images.

read the news is "happy". As shown in Fig. 11, our model assigns high attention weights to the informative sentences around the images. The sentences used to describe the fire and how the worker escaped were selected. However, the description of the fire and the corresponding words (*e.g.*, "fire" and "alarms") were assigned to higher attention than that about the escape, which leads to the incorrect prediction, "sad". As stated in Section IV-B1, image content plays a significant role in multimodal

TABLE VII
PERFORMANCE COMPARISON TO EXISTING METHODS ON THE DATASET USED IN [58]. THE '-M' AND '-A' SUFFIXES OF ALGORITHMS INDICATE APPLYING MAX-POOLING OR AVERAGE-POOLING AS THE LAST LAYER OF VGGNET FOR VISUAL FEATURE EXTRACTION

Methods	BO	CH	LA	NY	SF	Avg.
TFN-a	46.35	43.69	43.91	43.79	42.81	43.89
TFN-m	48.25	47.08	46.70	46.71	47.54	46.87
BiGRU-a	51.23	51.33	48.99	49.55	48.60	49.32
BiGRU-m	53.92	53.51	52.09	52.14	51.36	52.20
HAN-a	55.18	54.88	53.11	52.96	51.98	53.16
HAN-m	56.77	57.02	55.06	54.66	53.69	55.01
VistaNet	**63.81**	**65.74**	**62.01**	61.08	**60.14**	**61.88**
LD-MAN	61.90	64.00	61.02	**61.57**	59.47	61.22

attention. The first image is the scene after the fire, which has a strong correlation with the description of the "fire", while the second image is a picture of the surviving worker. The stronger correlation is the reason for the higher attention weight for the description of "fire" than for "escape".

H. Performance on Other Datasets

1) Dataset: We evaluated our method on the dataset used in [58]. This dataset is a collection of online reviews for the food and restaurant categories of *Yelp.com*. The reviews covered 5 cities, including Boston (BO), Chicago (CH), Los Angeles (LA), New York (NY), and San Francisco (SF). The ratings of *Yelp* reviews on the scale 1 to 5 were treated as 5 sentiment levels. The training details of our method on this dataset were set the same as those in [58]. Another dataset was the Twitter dataset used in [83]. There were 24,635 English tweets, and every tweet contained an image. The task for this dataset was multimodal sarcasm detection. Tweets with some special hashtags (*e.g.*, *#sarcasm*, *etc.*) were treated as positive, and those without these tags were treated as negative. This dataset was divided into training, validation, and test sets by $8:1:1$ as in [83].

2) Performance: The accuracies for every city and the average accuracy are listed in Table VII. Our proposed LD-MAN achieved comparable performance compared with the state-of-the-art method. All of the compared methods were based on both textual and visual features. Because there is no layout information in *Yelp* reviews, LD-MAN was the method that used image and text features to perform sentiment analysis. The informative textual features were selected by the visual attention mechanism in our method and VistaNet, which improved the performance compared to other fusion-based methods. The main distinction between our method and VistaNet lies in that our method uses both the visual and textual features when inferring the sentiment rather than only using the image to attend important textual contents in VistaNet. As stated in [58], in the *Yelp* reviews, images play a supporting role in text. Therefore, considering visual features when inferring sentiment decreases the effect of our method.

The results on the Twitter dataset are shown in Table VIII. "Text" is one of the most popular methods for text classification, which uses CNN as the textual feature extractor. "Image" is a method that uses the features extracted from ResNet to predict the labels. "Attr" predicts tweets whether sarcastic or not through the image attributes. "Concat" concatenates features of text and image along with the image attributes. Because most of the tweets in this dataset only contain one sentence, we did not consider the hierarchical structure of text or add the image locations. Our method outperformed other single-modality methods (*i.e.*, "Text", "Image" and "Attr") because both visual and textual features were considered, which is effective for recognizing sentimental clues in this multimodal scenario. Since we did not use the attributes of the image as HFN and "Concat", the performance of LD-MAN was not as effective.

TABLE VIII
PERFORMANCE ON THE TWITTER DATASET USED IN [83]

Approach	F-score	Pre	Rec	Acc
Text	75.32	74.29	76.39	80.03
Image	61.53	54.41	70.80	64.76
Attr	63.34	56.06	72.78	66.46
Concat	78.74	73.36	**84.98**	81.74
HFM	**80.18**	76.57	84.15	**83.44**
LD-MAN (Ours)	75.31	**76.73**	73.93	80.70

The novelty of LD-MAN is the consideration of the layout of the online news, which is not contained in the *Yelp* and Twitter datasets. We modify LD-MAN according to the characteristics of the two datasets. The above experimental results show that the multimodal attention mechanism adopted in LD-MAN is effective for sentiment analysis on multimodal data, including text and image.

VI. CONCLUSION

In this paper, we presented a novel layout-driven multimodal attention network (LD-MAN) to recognize the sentiment of online news. It models the relevance between news content and reader sentiment. In addition to the text and images, the layout of articles was also considered in our method. Specifically, LD-MAN models the layout as the image locations to align images with the corresponding text, and it employs the distance-based coefficients to learn the unified news representation using a multimodal attention module. In addition, we collected two datasets due to the lack of multimodal online news datasets. Extensive evaluations were carried out in the newly crawled datasets. The superior performance of LD-MAN was observed over several state-of-the-art and alternative approaches. These results demonstrate the merits of the proposed scheme for multimodal sentiment prediction. Future work of this line of research will include a better multimodal representation scheme by extracting superior representation with feature selection techniques to avoid using images without any sentiment.

REFERENCES

[1] L. Luo, X. Ao, F. Pan, J. Wang, T. Zhao, N. Yu, and Q. He, "Beyond polarity: interpretable financial sentiment analysis with hierarchical query-driven attention," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 4244–4250.

[2] X. Li, L. Bing, P. Li, and W. Lam, "A unified model for opinion target extraction and target sentiment prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6714–6721.

[3] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *Proc. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2018, pp. 929–932.

[4] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[5] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Proc. Mining Text Data*, 2012, pp. 415–463.

[6] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, Mar./Apr. 2013.

[7] Y. Chen, J. Yuan, Q. You, and J. Luo, "Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 117–125.

[8] S. Angelidis and M. Lapata, "Multiple instance learning networks for fine-grained sentiment analysis," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 17–31, 2018.

[9] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 223–232.

[10] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7584–7592.

[11] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction," in *ACM Int. Conf. Multimedia*, 2015, pp. 57–62.

[12] V. Campos, B. Jou, and X. Giro-i Nieto, "From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction," *Image Vision Comput.*, vol. 65, pp. 15–22, 2017.

[13] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, "Dependency exploitation: a unified cnn-rnn approach for visual emotion recognition," in *Proc. the Int. Joint Conf. Artificial Intell.*, 2017, pp. 3595–3601.

[14] X. Jia, X. Zheng, W. Li, C. Zhang, and Z. Li, "Facial emotion distribution learning by exploiting low-rank label correlations locally," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 9841–9850.

[15] X. Yao, D. She, S. Zhao, J. Liang, Y.-K. Lai, and J. Yang, "Attention-aware polarity sensitive embedding for affective image retrieval," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 1140–1150.

[16] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 1151–1160.

[17] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 10 143–10 152.

[18] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2513–2525, Sep. 2018.

[19] M. O. C. II, S. Fan, Z. Shen, and M. S. Kankanhalli, "Emotion-aware human attention prediction," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1140–1150.

[20] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: Understanding and overcoming dataset bias," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 579–595.

[21] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429–439, 2019.

[22] S. Zhu, S. Li, and G. Zhou, "Adversarial attention modeling for multi-dimensional emotion regression," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 471–480.

[23] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," *Inf. Sci.*, vol. 471, pp. 216–232, 2019.

[24] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. Int. Conf. Empirical Methods Natural Lang. Process.*, 2002, pp. 79–86.

[25] K. Zhang, Y. Xie, Y. Yang, A. Sun, H. Liu, and A. Choudhary, "Incorporating conditional random fields and active learning to improve sentiment identification," *Neural Netw.*, vol. 58, pp. 60–67, 2014.

[26] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 1681–1691.

[27] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. Int. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 427–431.

[28] M. Dragoni, S. Poria, and E. Cambria, "Ontosenticnet: A commonsense ontology for sentiment analysis," *IEEE Intell. Syst.*, vol. 33, no. 3, pp. 77–85, May/Jun. 2018.

[29] Q. Yang, Y. Rao, H. Xie, J. Wang, F. L. Wang, and W. H. Chan, "Segment-level joint topic-sentiment model for online review analysis," *IEEE Intell. Syst.*, vol. 34, no. 1, pp. 43–50, Jan./Feb. 2019.

[30] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. F. Gelbukh, "Sentiment and sarcasm classification with multitask learning," *IEEE Intell. Syst.*, vol. 34, no. 3, pp. 38–43, May/Jun. 2019.

[31] A. Weichselbraun, S. Gindl, F. Fischer, S. Vakulenko, and A. Scharl, "Aspect-based extraction and analysis of affective knowledge from social media streams," *IEEE Intell. Syst.*, vol. 32, no. 3, pp. 80–88, May/Jun. 2017.

[32] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proc. Int. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1422–1432.

[33] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 1014–1023.

[34] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Int. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2016, pp. 1480–1489.

[35] G. Letarte, F. Paradis, P. Giguère, and F. Laviolette, "Importance of self-attention for sentiment analysis," in *Proc. EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. NLP*, 2018, pp. 267–275.

[36] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 973–982.

[37] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53, May/Jun. 2013.

[38] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.

[39] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," in *Proc. Int. Conf. Multimedia Model.*, 2016, pp. 15–27.

[40] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 997–1007, Apr. 2018.

[41] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1062–1075, Apr. 2018.

[42] C. Du *et al.*, "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 108–116.

[43] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 371–378.

[44] H. N. Tran and E. Cambria, "Ensemble application of ELM and GPU for real-time multimodal sentiment analysis," *Memetic Comput.*, vol. 10, no. 1, pp. 3–13, 2018.

[45] S. Dou, Z. Feng, X. Yang, and J. Tian, "Real-time multimodal emotion recognition system based on elderly accompanying robot," *J. Phys.: Conf. Ser.*, vol. 1453, 2020, Art. no. 012093.

[46] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 108–116.

[47] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[48] E. Cambria, N. Howard, J. Y. Hsu, and A. Hussain, "Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics," in *Proc. IEEE Symp. Comput. Intell. Human-like Intell.*, 2013, pp. 108–117.

[49] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–17.

[50] J. Xu *et al.*, "Sentiment analysis of social images via hierarchical deep fusion of content and links," *Appl. Soft Comput.*, vol. 80, pp. 387–399, 2019.

[51] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image-text sentiment analysis via deep multimodal attentive fusion," *Knowl. Based Syst.*, vol. 167, pp. 26–37, 2019.

[52] S. Verma, C. Wang, L. Zhu, and W. Liu, "Deepcu: integrating both common and unique latent information for multimodal sentiment analysis," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3627–3634.

[53] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognit. Lett.*, vol. 125, pp. 264–270, 2019.

[54] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 949–954.

[55] Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *Proc. ACM Conf. Multimedia*, 2016, pp. 1008–1017.

[56] Z. Zhao *et al.*, "An image-text consistency driven multimodal sentiment analysis approach for social media," *Inf. Process. Manage.*, vol. 65, no. 60, pp. 15–22, 2019.

[57] X. Zhu, B. Cao, S. Xu, B. Liu, and J. Cao, "Joint visual-textual sentiment analysis based on cross-modality attention mechanism," in *Proc. Int. Conf. Multimedia Model.*, 2019, pp. 264–276.

[58] Q.-T. Truong and H. W. Lauw, "Vistanet: Visual aspect attention network for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 305–312.

[59] J. Xu *et al.*, "Visual-textual sentiment classification with bi-directional multi-level attention networks," *Knowl. Based Syst.*, vol. 178, pp. 61–73, 2019.

[60] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. F. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov./Dec. 2018.

[61] K. H.-Y. Lin, C. Yang, and H.-H. Chen, "Emotion classification of online news articles from the reader's perspective," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2008, pp. 220–226.

[62] J. C. S. Dos Rieis, F. B. de Souza, P. O. S. V. de Melo, R. O. Prates, H. Kwak, and J. An, "Breaking the news: First impressions matter on online news," in *9th Int. AAAI Conf. Web Social Media*, 2015, pp. 357–366.

[63] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, "Building emotional dictionary for sentiment analysis of online news," *World Wide Web*, vol. 17, no. 4, pp. 723–742, 2014.

[64] J. Li, S. Fong, Z. Yan, and R. Khoury, "Hierarchical classification in text mining for sentiment analysis of online news," *Soft Comput.*, vol. 20, no. 9, pp. 3411–3420, 2016.

[65] Y. Rao, Q. Li, X. Mao, and L. Wenyin, "Sentiment topic models for social emotion mining," *Inf. Sci.*, vol. 266, pp. 90–100, 2014.

[66] P. Liu, J. A. Gulla, and L. Zhang, "Dynamic topic-based sentiment analysis of large-scale online news," in *Proc. Web Inf. Syst. Eng.*, 2016, pp. 3–18.

[67] J. G. Ellis, S. Karaman, H. Li, H. B. Shim, and S.-F. Chang, "Placing broadcast news videos in their social media context using hashtags," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 684–688.

[68] H. Li, B. Jou, J. G. Ellis, D. Morozoff, and S.-F. Chang, "News rover: Exploring topical structures and serendipity in heterogeneous multimedia news," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 449–450.

[69] Y. Groen, A. B. M. Fuermaier, A. E. Heijer, O. Tucha, and M. Althaus, "The empathy and systemizing quotient: The psychometric properties of the dutch version and a review of the cross-cultural stability," *J. Autism Developmental Disorders*, vol. 45, no. 9, pp. 2848–2864, 2015.

[70] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[71] N. Xu and W. Mao, "Multisentinet: A deep semantic network for multimodal sentiment analysis," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2017, pp. 2399–2402.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[73] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2017.

[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[75] S. Zhao, Z. Jia, H. Chen, L. Li, G. Ding, and K. Keutzer, "Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 192–201.

[76] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1358–1371, May 2020.

[77] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[78] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[79] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.

[80] R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 562–570.

[81] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[82] C. Wu, R. Socher, and C. Xiong, "Global-to-local memory pointer networks for task-oriented dialogue," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–19.

[83] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2506–2515.