IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 July 2023, accepted 6 August 2023, date of publication 23 August 2023, date of current version 30 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3307716

RESEARCH ARTICLE

Interpretable Multimodal Sentiment Classification Using Deep Multi-View Attentive Network of Image and Text Data

ISRAA KHALAF SALMAN AL-TAMEEMI[1], MOHAMMAD-REZA FEIZI-DERAKHSHI[1], SAEID PASHAZADEH[2], (Member, IEEE), AND MOHAMMAD ASADPOUR[2]

[1]Computerized Intelligence Systems Laboratory, Department of Computer Engineering, University of Tabriz, Tabriz 5166616471, Iran
[2]Department of Computer Engineering, University of Tabriz, Tabriz 5166616471, Iran

Corresponding author: Mohammad-Reza Feizi-Derakhshi (mfeizi@tabrizu.ac.ir)

ABSTRACT Multimodal data can convey user emotions and feelings more effectively and interactively than unimodal content. Thus, multimodal sentiment analysis (MSA) research has recently acquired great significance as a field of study. However, most current approaches either acquire sentimental features independently for each modality or simply combine multiple modal features. Thus, semantic details pertinent to sentiment analysis and the relationship between visual and textual content are neglected. Furthermore, most available multimodal datasets are sentiment-annotated, although user emotions are usually rich and unlimited. Motivated by these observations, this paper proposes a novel deep multi-view attentive network (DMVAN) for robust multimodal sentiment and emotion classification. The DMVAN model has three phases: feature learning, attentive interaction learning, and cross-modal fusion learning. During the feature learning phase, visual features from a multi-view perspective (region and scene) and textual features from various levels of analysis (word, sentence, and document) are extracted to capture information effectively for accurate classification. In the attentive interaction learning phase, the image-text interaction learning mechanism is employed to enhance visual and textual information interaction by extracting sentimental and discriminative visual features and utilizing the textual information to guide the learning process of image features. Moreover, a cross-modal fusion learning module is developed to incorporate different features into a comprehensive framework that takes advantage of the complementary aspects of multiple modalities. Then, a multi-head attention mechanism is employed to extract and merge sufficient data from the intermediate features, thereby aiding in developing a robust joint representation. Finally, a multi-layer perceptron with multiple stacking-fully connected layers is used to deeply fuse the modal features, thereby enhancing sentiment classification performance. An interpretable multimodal sentiment classification model is further developed utilizing the local interpretable model-agnostic explanation model (LIME) to ensure the model's explainability and strength. To perform a multimodal emotion classification, an image-text emotion dataset named Emotion-Getty (EMO-G) is constructed from Getty Images and labeled by distinct emotions. The proposed model is tested on three real-world datasets, attaining 99.801% accuracy on Binary_Getty (BG), 96.867% on Twitter, and 96.174% on the EMO-G dataset. These results show that the suggested model outperforms single-model techniques and current state-of-the-art methodologies based on model evaluation criteria.

INDEX TERMS Multimodal sentiment analysis, attention mechanism, deep learning, deep multi-view attentive network, interpretability.

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman.

I. INTRODUCTION

The widespread adoption of mobile Internet and smartphones results in the accumulation of vast collections of

user-generated multimodal content, including text, images, and videos, covering a variety of subjects and entities, thereby providing researchers with a valuable resource. The extraction and analysis of sentiments embedded in this data garner significant interest from academic communities [1], [2], and commercial sectors due to their potential for broad application. Despite the promising results of previous research, the current literature predominantly concentrates on unimodal data tasks, such as text sentiment classification [3], [4], [5] and image sentiment recognition [6], [7], [8]. However, multimodal data's valuable and complementary sentiment information is often neglected. The multimedia mining community is very interested in investigating the potential of multimodal sentiment tasks, which simultaneously incorporate emotional features from multiple modalities [9], [10]. However, MSA presents a formidable challenge as it involves the analysis of diverse modal data that may contain specific emotional information. Despite the existence of excellent deep learning (DL) models for MSA, most current approaches learn the representations of each modality independently before combining the acquired multimodal characteristics at an elevated level of the neural network. In addition, little research has been conducted on cross-modal interactions involving various modalities like text and images.

The present study concentrates on multimodal sentiment classification based on image-text pairs from social media postings. Figure 1 depicts several examples that serve as sources of motivation. Image A conveys a negative sentiment of anger through the male's facial expressions, the presence of flames, and the textual reference to ''angry.'' The image and accompanying text complement each other in representing this emotional state. Likewise, in image B, the smiling girl and the sunlight convey a sense of happiness, while the term ''happy'' reinforces the concept of positive sentiment. Regarding image C, the crying eyes, specific facial expressions, and a rainy background collectively convey a sense of sadness and negative feelings. In the last example, the woman's disgusted expression, the smoke, and the explicit use of the word ''disgusted'' all convey dislike and a negative feeling.

Generally, the interaction of various visual cues guided by certain textual words or phrases typically affects users' sentiments. The correlation between human emotions and visual information is significant, making it a valuable tool for comprehending user sentiment in multimodal data. Despite the significant advancements demonstrated by the current literature on MSA, this task still presents challenges for the following reasons:

- The conveyed emotions by the texts and images are not restricted to a single data modality. Instead, they are interrelated and serve to convey the users' sentiments and emotions in a complementary manner. Therefore, to develop a robust SA approach that can effectively bridge the gap across various modalities, extracting deep and discrete information from each modality is essential to the sentiment classification task.
- People tend to concentrate on specific regions of an image that capture their interest rather than allocating equal attention to the whole image's content. Similar to the emotive words in the text, specific regions and scenes in the pictures—like the smiling girl and the sunlight in Figure 1-B—provide substantial evidence of the emotion required for this task. Despite the successful utilization of various visual features in some SA techniques [11], [12], the integration of multi-view features within a unified framework has not been considered yet.
- The relationships between visual representations and written language are complex and multi-level. The visual elements of an image can be associated with a single term, a phrase, or the entire document's textual content. As shown in Figure 2, the image of a crying bride is related to the term ''bride'' and the sentence ''emotional tears.'' Similarly, the region of gorgeous flowers is associated with the term ''flowers''. Thus, the entire sentence must be considered to have a complete understanding of the image's sentiment.
- Previous research prioritized the attention mechanism to identify significant image regions and emotional language to produce effective multimodal features. Nevertheless, these techniques concentrate on attention based on regions and do not exploit channel information to construct visual characteristics, which is significant for identifying critical patterns within a given image. Moreover, the effective visual components can be evaluated for their visual content and the accompanying textual information. As a result, it is critical to the efficacy of SA that the model prioritizes emotional regions and scenes while simultaneously considering both visual content and textual information.
- DL models are ''black boxes'' that are ambiguous, with complex hidden layers. Their logic, dynamics, and decision-making are poorly explained. The prevalence of this issue is higher in multimodal systems because of the complex interrelationships among different input streams. This results in significant challenges related to interpretability and explainability.
- The task of multimodal emotion analysis is more challenging than multimodal sentiment polarity analysis. The reason for this is the rise in the number of emotion categories and the constraints of current models. Also, the insufficiency of extensive training datasets for multimodal DL models increases the task's difficulty.

To address these challenges, a dataset of images and textual descriptions is constructed and annotated with emotional labels named Emotion-Getty (EMO-G) from the Getty Images website. Furthermore, a novel deep multi-view attentive network (DMVAN) is proposed for robust multimodal sentiment and emotion classification. The DMVAN model has three phases: feature learning, attentive interaction learning, and cross-modal fusion learning. In the feature learning phase, the informative visual and textual features are extracted for accurate classification. The image-text

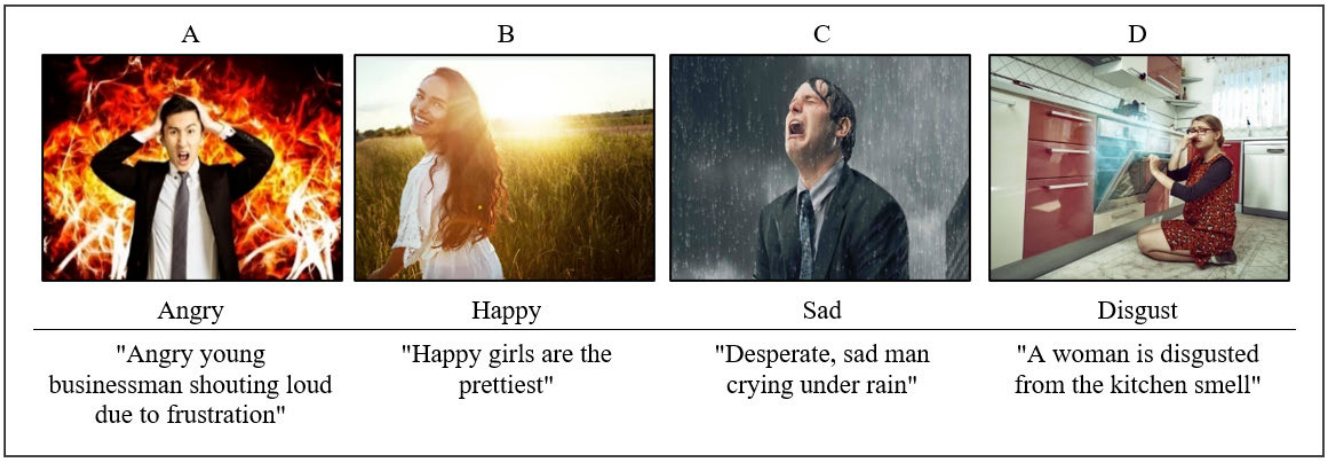


FIGURE 1. Examples of emotionally-labeled image-text multimodal data.

FIGURE 2. An example of image and text correlation. The image's emotional areas strongly match the text's emotional terms. Investigating the connection between these modalities helps to understand the image's sentiment.

interaction learning approach improves visual and textual data interactivity during attentive interaction learning. Cross-modal fusion learning integrates features into a single framework to capture distinct modalities' complementing attributes. Next, a multi-head attention mechanism extracts and integrates adequate information from intermediate features. Finally, the characteristics are extensively fused using a multi-layer perceptron (MLP) to conduct sentiment and emotion classification. In addition, an interpretable multimodal sentiment classification model is developed to explain further the underlying model process, which leverages local interpretable model-agnostic explanations (LIME) to ensure model trust and resilience.

The present study's main contributions are:

- An image-text dataset called Emotion-Getty (EMO-G) is developed and labeled based on emotions to conduct multimodal emotion analysis.
- A novel deep multi-view attentive network (DMVAN) is proposed for multimodal sentiment and emotion classification. Our model initially extracts visual features from a multi-view perspective (i.e., region and scene) and textual features from multiple levels of analysis (i.e., word level, sentence level, and document level), which aims to leverage the associations between the visual perspectives and the semantic levels of the textual representations through a unified scheme.
- An attentive interaction learning approach is proposed to obtain discriminative and emotional visual features with the guidance of textual data. The study incorporates two distinct visual attention branches to investigate visual characteristics at the region and scene levels. This approach aims to enhance the interaction between the two modalities. In particular, it enables textual information to guide the learning process for visual features and vice versa.
- Cross-modal fusion learning is designed to integrate the various features within a comprehensive framework to identify the complementary aspects across multiple modalities, followed by multi-head attention, which facilitates the extraction and fusion of adequate information in the intermediate features. Finally, an MLP with stacking-fully connected layers is used to classify sentiment by deeply fusing the modal features.
- An interpretable multimodal sentiment classification model based on LIME is developed to expose the internal model dynamics and visualize the association between the instance's characteristics and the model's prediction.

Extensive experiments on real-world sentiment and emotion datasets are carried out to demonstrate the proposed model's effectiveness through comparisons with previous models.

The paper's remaining sections are structured as follows: Section II presents the related literature. Section III describes

the proposed model in depth. Section IV provides the results of the experiments. Section V concludes the study and presents potential areas for future work.

II. LITERATURE REVIEW

A. MULTIMODAL SENTIMENT ANALYSIS

SA attempts to determine how people feel about social posts by combining textual and visual information that may help understand how users feel and act. In [13], an exhaustive overview of multimodal SA was presented, which investigated visual and linguistic information shared on social media websites and covered the common techniques for data fusion, the challenges involved, and the applications of sentiment. Several approaches for multimodal sentiment categorization were proposed that aim to include various modalities. Jindal and Aron [14] suggested a new VIsual-TExtual SA (VITESA) for polarity classification. A brownian novement-based meerkat clan algorithm centered densenet (BMMCA-DenseNet) was presented to merge written and visual data, enabling powerful SA using VITESA. The visual and textual characteristics were identified using an improved coyote optimization algorithm (ICOA) and an adaptable embedding for language models (Elmo). The suggested classifier categorized the data as positive or negative by assigning SentiWordNet polarity and extracting emoticon and non-emoticon features. You et al. [15] presented a cross-modality consistent regression (CCR) model to extract textual and visual sentimental information using a paragraph vector model and a convolutional neural network (CNN). A multi-modality regression model was then applied on top of them, aiming to achieve consensus between the sentiment labels forecasted by the text and image characteristics. Huang et al. [16] developed attention-based modality-gated networks (AMGN) for exploiting textual and visual content interactions. In particular, they proposed a modality-gated long short-term memory (LSTM) for defining multimodal properties by adjusting to the modality that provided the most reliable expression of emotion. This was followed by a semantic self-attention model, which focused on differentiating features for sentiment classification. The primary drawback of this study was that it assumed a highly detailed correlation existed between the visual and textual components because of the visual-semantic attention model. However, specific pairs might not possess a robust cross-modal correlation. Zhou et al. [17] proposed a hierarchical cross-modality interaction model (HCIM) that emphasized consistency and interdependence among modalities. This model employed a hierarchical approach to extract sentiment and semantic relationships between an image and text while also tackling the challenges posed by noise and joint understanding. Yadav and Vishwakarma [18] proposed a deep multi-level attentive network (DMLANet) to enhance multimodal learning. The study used semantic attention to simulate the connection between word meanings and visual regions, which was accomplished by identifying textual aspects associated with bi-attentive visual traits. Then, a self-attention method was used to acquire multimodal, sentimentally rich data for effective sentiment categorization. Xu et al. [11] proposed a novel approach called the bi-directional multi-level attention (BDMLA) model to conduct a collaborative sentiment classification of both visual and textual elements. This approach was designed to leverage both complementary and comprehensive information. The interaction between regional characteristics of an image and distinct conceptual levels of text was determined by the visual attention network, which ultimately determined the observed visual characteristics. The semantic attention network engaged with the semantic attributes of the textual content alongside diverse visual levels of the image in order to extract the attended semantic facets. The attributes of the two attention networks were subsequently incorporated within a comprehensive framework designed to classify sentiment in visual and textual data.

Yang et al. [19] created a model to identify sentiment in text and images using multi-channel graph neural networks. To capture hidden representations, a variety of modalities were encoded. Then, a graph neural network with multiple channels was developed to acquire knowledge of multimodal representations. The sentiment of image-text pairs is finally predicted using a multimodal fusion with a multi-head attention method. Yu and Jiang [20] presented a target-oriented multimodal BERT model (TomBERT), where the target-sensitive textual representations were initially obtained using BERT. Then, a target attention mechanism was designed to generate target-sensitive visual representations. Although a series of self-attention layers were built on top to record the multimodal interactions, they neglected textual information's impact on the picture. Zhang et al. [21] developed a hybrid fusion network (HFN) to obtain intra- and intermodal attributes. The visual characteristics were used to derive emotional data from written content via multi-head visual attention. Several base classifiers were then taught to acquire discriminative data from various modal representations. The main drawback of this approach was that choice diversity and classification accuracy clash as the model approached convergence. Meanwhile, Khan and Fu [22] developed a two-stream model named EF-CapTrBERT-DE, which used an object-aware transformer to translate images and non-auto-regressive text synthesis. An auxiliary sentence for a language model was then made using the translation. However, the significant variance in the utility of the visual modality and the complexity of the scene were significant limitations that restrict the efficacy of this approach.

Zhang et al. [23] introduced a novel approach named cross-modal semantic content correlation (SCC) to establish the connection between captions and images. A mixed attention network was devised to establish the semantic correlation between an image and its associated written representation. In order to obtain additional cross-modal nonlinear connections for sentiment prediction, a class-aware distributed feature vector was sent to an inner-class dependency long short-term memory (IDLSTM) network utilizing the image and text data as a query. However, this model suffered from

excessive memory overhead due to its lengthy execution time. Cao et al. [24] proposed various syncretic co-attention networks (VSCN) to investigate multi-level matching correlations across multimodal information and consider each modality's specific characteristics for integrated sentiment classification. However, the emotion polarity is frequently unclear because visual components convey more information than text, causing the model to generate incorrect predictions occasionally. Hu and Yamamura [25] proposed a neural network that assessed local and global fusion features for predicting user feelings. The approach first generated global modality-based fusion characteristics from attention modules and established local fusion features via coarse-to-fine fusion learning. Finally, these features are integrated to generate more precise forecasts.

Al-Tameemi et al. [26] proposed a new multi-model fusion (MMF) model for SA that optimally used a hybrid fusion technique to capture vital data and the natural interaction between visual and textual components. Hu and Yamamura [27] proposed a two-phase attention-based fusion neural network to classify sentiment based on textual and visual data. Yang et al. [28] introduced a multi-view attention network-based model for the SA, combining scene-text and object-text fusion. Li et al. [29] proposed a contrastive learning and multi-layer fusion network for detecting sentiment. To improve the correlation between images and text, scene and object extraction techniques were employed to extract more image details. The MultiSentiNet model was proposed by Xu and Mao [30], which involved the extraction of significant semantic characteristics from an image. These features were then utilized to facilitate the acquisition of text features. Huang et al. [31] created a deep multimodal attention fusion model (DMAF) for image-text SA by combining several attention methods and fusion techniques. Xing et al. [32] introduced a new and efficient approach for enhancing unpaired low-light images (LLIE). This method, called CLEGAN, used a single deep generative adversarial network (GAN) framework and employed self-similarity contrastive learning (SSCL) to maximize the mutual information between low-light and reformed images. An et al. [33] suggested a complete approach for improving targeted multimodal sentiment categorization (ITMSC) using semantic image descriptions. The model automatically used semantic explanations of images and text similarity relations to change the significance of images in the fusion representation. However, specific image descriptions might not accurately match the visual content, influencing the results of semantic similarity computations and reducing the model's accuracy. Kiaei et al. [34] developed an emotion analytic system to extract and visualize emotions. The data consisted of Persian comments from Instagram that were acquired using a custom-built web crawler. The research findings and lexicon-based analysis of "Rouhani" indicated a significant presence of trust, rage, and disgust. Kumar et al. [35] created a new interpretability method that utilized the divide-and-conquer strategy to compute shapely values that represent the importance of each speech and image component. Similarly, [36] introduced a new method for achieving interpretability called k-average additive explanation (KAAP) to pinpoint the crucial verbal, written, and visual cues for predicting a specific emotion category. Lyu et al. [37] created an original explanation by decomposing the model through unimodal contributions (UC) as well as multimodal interactions (MI). The disentangled multimodal explanations (DIME) approach could maintain generalizability across diverse modalities while promoting the precise and comprehensive examination of multimodal models.

Nevertheless, the studies that currently exist possess certain limitations. First, it is necessary for most current methods to adequately consider the deep semantic characteristics of images, which may serve as useful cues of emotions from various perspectives. Second, due to images' abstract and subjective nature compared to text, most research emphasizes text while disregarding the correlation between text and image. Third, most MSA models based on DL function as black boxes, making it challenging to comprehend their internal workings. Fourth, DL models heavily rely on large-scale training data. However, the majority of the current utilized datasets for MSA are labeled solely with positive, negative, and neutral labels [12], [30], [38]. Several limited datasets containing emotion labels are currently accessible [39], [40], [41].

Motivated by these observations, in this paper, a novel model is proposed for interpretable multimodal sentiment and emotion classification based on a deep multi-view attentive network (DMVAN). In this model, the visual and textual features are exploited from a multi-view perspective and at multiple levels of analysis, aiming to capture more efficient features that accurately reflect the sentiment of the image-text information within a unified scheme. Moreover, incorporating attentive interaction learning and cross-modal fusion learning modules enhances the interaction between two modalities, thereby facilitating the acquisition of complementary information from both modalities. This ultimately leads to enhanced results in the context of MSA while also explaining how the various modalities contribute and interact. An image-text multimodal emotion dataset is further constructed to facilitate multimodal emotion analysis performance and overcome the emotion dataset scarcity problem.

III. PROPOSED MODEL

A. OVERVIEW

Textual and visual information frequently coexist on social media. However, a single text may occasionally refer to numerous photos, complicating matters. The present study focuses on social data in which an image corresponds to a text. The problem of classifying sentiment in image-text multimodal data is described as follows: Given image-text pairs $P = \{(V^1, T^1), \ldots (V^i, T^i), \ldots (V^n, T^n)\}$ and the associated label set $L = \{L^1, \ldots L^i, \ldots L^n\}$, where V^i indicates a single image, T^i represents the related text, L^i represents the sentiment or emotion label, and n denotes the entire

number of pairs in the given set. The objective of multimodal sentiment prediction is to discover the mapping function $f : (V, T) \rightarrow L$ using the multimodal training set $\{(V^i, T^i, L^i)|0 \leqslant i \leqslant n-1\}$.

Therefore, a new deep multi-view attentive network (DMVAN)[1] is proposed for accurate multimodal sentiment and emotion classification. The framework of the DMVAN model is displayed in Figure 3, which includes three phases: feature learning, attentive interaction learning, and cross-modal fusion learning. During the feature learning phase, visual features from a multi-view perspective (region and scene) and textual features from multiple levels of analysis (word, sentence, and document) are retrieved to effectively represent multimodal social data. These in-depth visual semantic features are considered supplemental data when performing MSA. In the attentive interaction learning phase, the image-text interaction learning mechanism is employed to improve the interplay between the visual and textual data, considering the text as the main modality to guide learning the attention networks for the region and scene visual features. Specifically, to leverage region- and scene-level image features with the guidance of textual information, a regional attention network and a scene attention network are designed into two branches to extract emotion-related visual features that are more discriminative. A convolutional block attention module (CBAM) and a textual-guided attention module are developed in each branch. The primary objective of CBAM is to acquire informative characteristics by integrating cross-channel and spatial information. This approach effectively produces resilient region and scene visual features by acquiring knowledge on which information to emphasize or reduce. Simultaneously, the textual-guided attention module facilitates the acquisition of emotional visual features closely associated with textual information. In cross-modal fusion learning, the visual features from attended regions and scenes, as well as the attended textual features, are integrated within a comprehensive framework. This approach aims to capture the complementary features that exist between multiple modalities. Then, a multi-head attention mechanism is introduced to facilitate the fusion and refinement of intermediate feature information. Finally, the features are extensively fused to classify the sentiment using an MLP that incorporates stacking-fully connected layers to achieve an improved F1 score and accuracy for multimodal classification.

B. DEEP MULTI-VIEW ATTENTIVE NETWORK

This study presents a multimodal sentiment and emotion classification model based on a deep multi-view attentive network (DMVAN) to capture the complementarity between textual and visual information. The model comprises three phases: feature learning, attentive interaction learning, and cross-modal fusion learning. The proposed model is explained comprehensively in the subsequent sections.

[1]Implementation code can found at: https://github.com/cominsys/DMVAN

1) FEATURE LEARNING

The consistency level between textual and visual contents exhibits significant variation. For example, the salient areas of an image may correspond to various hierarchical levels of textual information, ranging from individual words to complete sentences or even larger scopes. Likewise, sentimental discourse may encompass multiple viewpoints of a visual representation, wherein specific images' regions and scenes serve as effective cues of emotion that are important to the task. However, the majority of earlier methods merely considered single-level characteristics. Thus, the first stage of our proposed approach is to extract multi-level textual and multi-view visual features from each modality to identify more diverse correspondences between the two modalities and convey their properties from several perspectives.

Step1: Multi-level Textual Feature Extraction:

The stratified textual representations are obtained by extracting textual features at the word, sentence, and document levels, as shown in Figure 4. Hence, the bidirectional encoder representation from transformers (BERT) [42] is utilized in our study for creating the word embedding. Currently, BERT is considered the most effective vectorization model for extracting semantic, contextual, locational, and grammatical features from texts. The goal is to pre-train deep bidirectional text representations on huge amounts of unannotated text while simultaneously considering left and right contexts. BERT is a transformer-based encoder with a multi-layer bidirectional structure [43]. It incorporates multi-head attention, which separates the model into several heads and creates various subspaces. As a result, the model can concentrate on different information aspects and fully integrate the sentence's contextual knowledge, while parallel processing is also possible. In BERT design, the raw text word is embedded by adding the token, position, and segment embeddings. The model is pre-trained with a massive unlabeled text corpus, such as Wikipedia or the Book Corpus; as a result, it can acquire a deeper and more intimate understanding of how language functions. After going through the necessary text preprocessing steps for the multi-layer transformer encoder, the pre-trained BERT model first maps the input text words to a vector representation with 768-dimensional word embeddings.

$$E = TW^e, \quad E \in R^{l_T \times n} \tag{1}$$

where T represents the text content, W^e denotes the weighted matrix, E refers to the word embedding, l_T indicates the length of every text string, The symbol R denotes the set of real numbers, and in this context, $R^{l_T \times n}$ represents the set of text strings containing n-word vectors' dimensions. The pre-trained word embedding E is then transformed into d-dimensional space using a dense linear layer of 256 neurons to get word-level embedding, as described below.

$$T^w = Relu(WE + b) \tag{2}$$

where T^w represents the textual features at the word-level, while W and b are trainable parameters. In contrast to a single

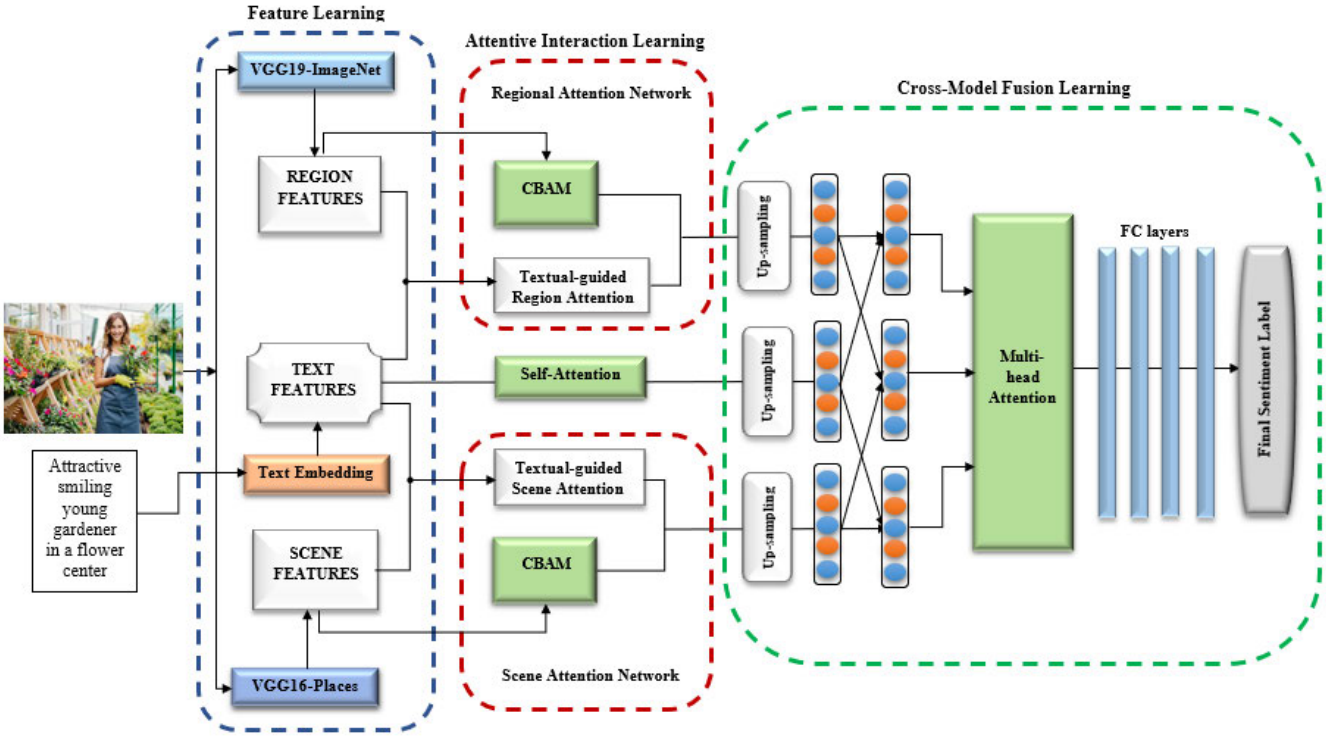


FIGURE 3. Deep multi-view attentive network (DMVAN) model.

word, which conveys the semantic attributes of distinct words in a particular sentence and can be employed to assess the correlation between word and image, a phrase represents the context-specific information conveyed by every phrase in the written material, which reflects the association between the sentence and the image. The word embedding is subjected to two 1-D convolution layers with window sizes of 3 and 4, each comprising 256 filters. This results in a representation that produces local features corresponding to 3-gram and 4-gram text strings. In order to enhance the learning speed of the network and provide some regularization, a batch normalization layer is incorporated after the convolution layer. Then, the max-pooling layer is utilized to capture crucial information by retaining the maximum value as the ultimate feature acquired by the filter. Finally, these features are concatenated to form a fixed-dimensional feature vector. The convolution operation yields local features that eliminate redundant terms, and the essential features of the sentences are retained to acquire phrase-level embedding in the following manner:

$$F_{CNN} = f_{CNN}(E; \theta_t^c) \tag{3}$$

where, f_{CNN} indicates the CNN operation, which consists of the convolution and maximum pooling operations and θ_t^c is a CNN parameter. To provide document-level embedding, the LSTM network is utilized, a subset of the recurrent neural network family [44]. LSTM features specialized ''memory cells'' that can maintain information for extended periods. Three gates—input, forget, and output—help to control the memory cells. These gates are responsible for managing the inflow and outflow of information. It is highly effective in modeling complex sequential data and can generate high-level representations accurately reflecting the data's structure. Two LSTM layers, each comprising 256 neurons, are employed to encode the entire sequence at the document level. This facilitates obtaining long-term text information that aids in comprehending the text description in-depth. The document-level embedding is identified as the LSTM hidden vector.

$$H = f_{LSTM}(E; \theta_t^{LS}), H \in R^{l_T \times d} \tag{4}$$

where $H = h_0, h_1, \ldots, h_i, \ldots, h_{d-1}$ represents the output of the LSTM, θ_t^{LS} retains the LSTM parameters, and d represents the number of hidden units within the LSTM. A unified textual feature that covers all three levels must be constructed to examine the relationship between the various semantic levels in a written description and the image regions. This is accomplished by concatenating the textual features from the various levels, as demonstrated:

$$F^t = f_{concat}(T^w, F_{CNN}, H), F^t \in R^{l_T \times D_T} \tag{5}$$

where F^t indicates the joint textual feature and D_T represents the dimension of the connected textual characteristics.

Our approach relies on the idea that certain vital emotional words in the input sequence are essential in determining the sentiment. Therefore, to define the contribution of each element in the fused textual representation, a self-attention mechanism is used to highlight the sentimental elements of the textual representation for accurate SA. In particular,

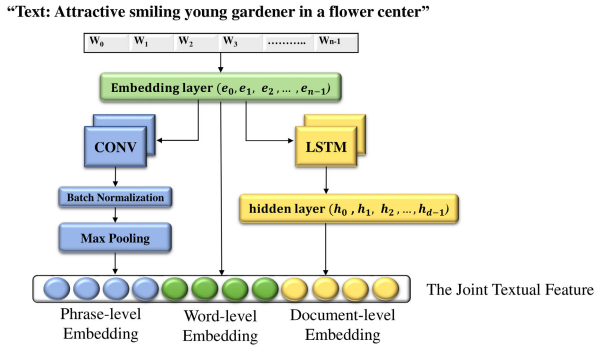


FIGURE 4. The process of extracting textual features.

the network automatically calculates the critical weights for each sequence based on the fused feature vector through a nonlinear process, resulting in the acquisition of a normalized attention score, as shown below:

$$e_f = \varphi(W * F^t + b) \tag{6}$$

$$\alpha^f = \frac{exp(e_f)}{\sum_f (e_f)} \tag{7}$$

where e_f is the un-normalized attention score indicating how accurately the vector F^t reflects the sentiment, α^f is used to normalize the attention throughout the sequence by utilizing the SoftMax function; W and b represent the learnable parameters, whereas φ denotes the nonlinear activation function (e.g., tanh). As a result, distinct input features in a self-attention network can interact with one another (''self'') to determine which input receives greater attention. The attention scores are then utilized to modify the attention intensity across the various textual representations. Finally, the attended textual feature is obtained by computing the weighted average throughout the text sequence, as described below:

$$F^{at} = \sum_f \alpha^f * F^t \tag{8}$$

The attended textual feature F^{at} is more effective in representing significant features compared to the original joint textual feature F^t. This improvement in representation helps to enhance sentiment prediction.

Step2: Multi-view Visual Feature Extraction:

The process of acquiring visual attributes from multiple viewpoints is illustrated in Figure 5, which can be classified into two distinct categories: region-level and scene-level. The scene feature can express a broad spectrum of emotions, unlike the region feature, which only identifies the semantic information associated with every image region at a higher level. For instance, in image C of Figure 1, the region of the crying eyes and the rainy background collectively create a sense of sadness and negative feelings, serving as excellent indications for the image's overall attitude, which is critical to the SA task. As a result, the region and scene features of the target image are analyzed to provide a comprehensive representation of the image's content.

Initially, the images are resized to attain the dimensions of 224×224 pixels. Then, the VGG19 network [45], which is pre-trained on ImageNet [46], is employed to derive the region features. The structure of this network contains five convolutional blocks and three fully connected layers, which exhibit superior performance in image classification tasks. Like [47], the region features are derived from the ''conv5-4'' layer of VGG-19 networks that can be represented as $V^r \in R^{512\times14\times14}$, indicating that there are 14×14 regions, each with 512 visual feature channels.

Similarly, scene features are extracted using the state-of-the-art Scene-VGG16 network [48]. The model is pre-trained using the massively popular Place365 dataset, which includes many images for classifying 365 distinct scene categories. The scene features are represented as $V^s \in R^{512\times14\times14}$, extracted from the ''conv5-3'' layer of the VGG-16 network. It is observable that each image comprises 196 distinct regions with a feature dimension of 512. A dense layer of d neurons (256) transmits the visual semantics and scene features to a higher-level space.

$$F^r = ReLu(W^r V^r + b^r) \in R^{d\times14\times14} \tag{9}$$

$$F^s = ReLu(W^s V^s + b^s) \in R^{d\times14\times14} \tag{10}$$

where (F^r, F^s) represents the extracted region and scene visual features, which have the same dimension, (W^r, W^s) represents the weights, and (b^r, b^s) is the bias, which is trainable parameters. $R^{d\times14\times14}$ represents the set of image regions, each with a feature dimension of d.

2) ATTENTIVE INTERACTION LEARNING

The attentive interaction learning module focuses on the auxiliary information that exists between the textual and visual elements, aiming to enhance the overall quality of MSA. The present module explores the correlation between text and image by iteratively querying their respective visual and textual features. Specifically, to leverage the region- and scene-level visual features with the guidance of textual information, a regional attention network and a scene attention network are designed into two branches to derive additional emotional-related distinguishing visual cues. These features can aid in comprehending the internal associations within the visual content. This, in turn, enhances the association between visual and textual data, leading to the accurate classification of sentiments. The details of these networks are discussed in the following section.

Step 1: Regional Attention Network:

The multi-level embedding approach outlined above generates textual features with a wealth of structural information. Given the dual impact of the visual and textual contents, the next challenge is to obtain additional emotional and discriminating visual characteristics for SA. Motivated by the attention mechanism's notable achievements in various vision-based tasks [49], [50], our goal is to highlight the significant emotional segments of an image using the perspective of regions and scenes, respectively.

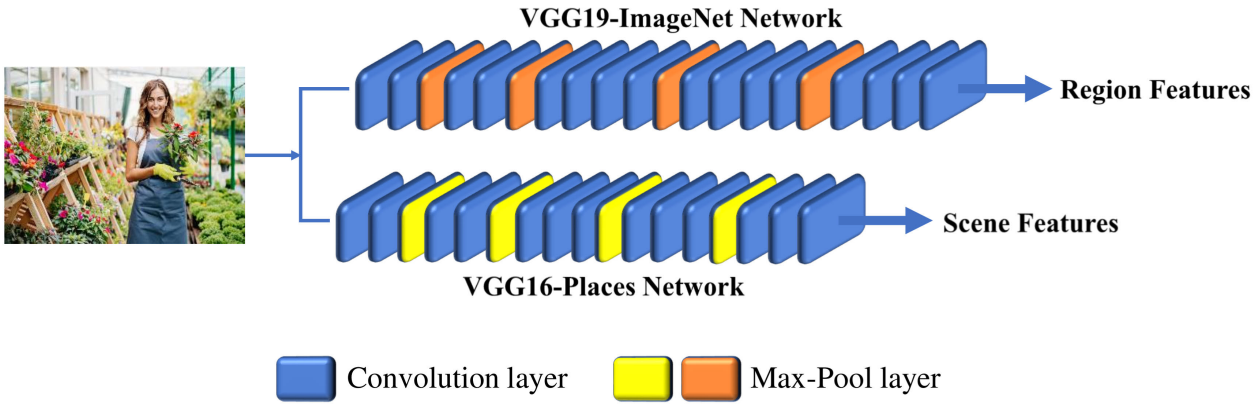


FIGURE 5. The process of extracting visual features.

First, the image is analyzed from a regional perspective, wherein a textual-guided region attention module is proposed to facilitate the collaboration of textual and visual elements. This module enables the identification of emotional regions with the guidance of textual information. To match the spatial dimension of the region feature $F^r \in R^{256\times14\times14}$, the joint textual feature F^t is first passed to the GlobalAveragePooling1D layer to get $\overrightarrow{F^t} \in R^d$ and then spatially replicated 12×12 times to form $\widehat{F^t} \in R^{d\times14\times14}$. As a result, the region's visual feature F^r and the resulting textual feature $\widehat{F^t}$ have the same dimension, allowing them to be combined to produce the joint region-textual feature m^f as follows:

$$m^f = (F^r \odot \widehat{F^t}) \tag{11}$$

where $\odot$ represents the element-wise multiplication of two vectors, then the fused vector m^f is subjected to a nonlinear operation that involves a SoftMax function, resulting in the acquisition of the normalized attention score, which has the potential to modulate the level of attentional intensity across distinct visual areas:

$$\alpha^f = \frac{exp(\varphi(W * m^f + b))}{\sum_f(\varphi(W * m^f + b))} \tag{12}$$

where W and b are the learnable parameters, φ is the nonlinear activation function (e.g., tanh). The visual characteristics that were observed are computed by taking the weighted average of the entire region's features based on the following formula:

$$\widehat{F^{er}} = \sum_f \alpha^f * F^r \tag{13}$$

In contrast to the original shared visual features F^r, the emotionally attended region-textual features $\widehat{F^{er}}$ are more effective in capturing the emotional image regions that are relevant to the textual feature $\widehat{F^t}$.

Second, the convolutional block attention module is employed to improve the model's representation power by selectively focusing on important region characteristics and discarding unimportant ones. This module aims to highlight the significant region features along the channel and spatial axes. In order to accomplish this task, a sequential application of channel and spatial attention is employed, utilizing the region feature map $V^r \in [H * W * C]$, which was extracted using the VGG19 network from each image, where H, W, and C represent the height, width, and number of channels for the region feature map. This module facilitates the acquisition of information by each branch regarding "what" and "where" to concentrate on the channel and spatial dimensions. The following section describes each attention module in detail.

A. Channel attention module The generation of the channel attention map is accomplished by leveraging the inter-channel correlation inherent in the features. The idea behind channel attention involves treating each channel within a feature map as a distinct feature detector [51]. This aids in identifying "what" is the essential and meaningful region with an input image. To achieve efficient computation of channel attention, the first step entails gathering spatial data from the feature map using the average and max pooling functions. Two distinct spatial context descriptors are generated: F^c_{avg} and F^c_{max}, representing average-pooled and max-pooled characteristics. These are then transmitted to a common network, creating a channel attention map denoted as $M_c = (1 * 1 * 256)$. Once the shared network is used for each descriptor, which includes an MLP with a single hidden layer, the resulting feature vectors are combined by performing element-wise summation. In summary, the process for computing channel attention can be outlined as follows:

$$\begin{aligned} M_c &= \sigma(MLP(AvgPool(V^r)) + MLP(MaxPool(V^r))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \end{aligned} \tag{14}$$

where σ represents the sigmoid function while V^r defines the region feature map. Notably, W_0 and W_1 of the MLP weights are shared between the two inputs.

B. Spatial attention module The generation of the spatial attention map is accomplished by utilizing the inter-spatial correlation of features. Spatial attention differs from channel attention as it determines "where" the informative region within an image. This is achieved by identifying the relevant

areas of the image based on the attended channel-based features. The generation of a refined feature map F involves the element-wise multiplication of the input feature map V^r and the channel attention map M_c. The channel refined feature map is then aggregated utilizing two pooling operations, resulting in two maps: average pooled F^s_{avg} and max pooled F^s_{max}, representing average-pooled and maximum-pooled channel features. These features are combined and convolved through a convolutional layer employing a 7×7 kernel size. The resulting output represents the spatial attended features, denoted as M_s, which encode the areas that should be given more or less attention. The transformation of the channel refined feature map F to spatial attended features M_s can be formulated as follows:

$$\begin{aligned} M_s &= \sigma(F^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(F^{7\times7}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{15}$$

where σ represents the sigmoid function, while $f^{7\times7}$ denotes a convolution operation that employs a filter size of 7×7. On this basis, the attended regional features F^{ar} are derived by performing an element-wise multiplication of $\overline{\overline{M_s}}$ and $\overline{\overline{F^r}}$, as follows:

$$\overline{\overline{F^r}} = ReLu(W^r \widehat{F^{er}} + b^r) \in R^d \tag{16}$$

$$\overline{\overline{M_s}} = ReLu(W^s M_s + b^s) \in R^d \tag{17}$$

$$F^{ar} = (\overline{\overline{M_s}} \odot \overline{\overline{F^r}}) \tag{18}$$

where (W^r, W^s) represents the weights and (b^r, b^s) is the bias, which represent the trainable parameters. The symbol $\odot$ denotes the element-wise multiplication between two vectors, and R^d denotes the set of features, each with a dimension of d that represents the number of units of the dense layer, which equals 256. The implementation of the regional attention network allows for identifying and retrieving significant emotional and discriminative characteristics essential for understanding the emotional context of the entire image region. Figure 6 illustrates the details of this network.

Step2: Scene Attention Network:

Although region features can emphasize the interconnected nature of regions and uncover implicit emotional information, the image scene also provides useful cues that can assist in comprehending the user's sentiment. Just like the regional attention network, the significant emotional aspects of an image are emphasized from a scene's perspective. The image is first analyzed regarding its scenes, wherein a textual-guided scene attention module is developed to encourage the association of textual and visual elements. This is achieved by focusing on the significant scenes, guided by textual information. To match the spatial dimension of the scene feature $F^s \in R^{256\times14\times14}$, the joint textual feature F^t is first passed to the GlobalAveragePooling1D layer to get $\overrightarrow{F^t} \in R^d$ and then spatially replicated 12×12 times to form $\widehat{F^t} \in R^{d\times14\times14}$. As a result, the scene's visual feature F^s and the textual feature $\widehat{F^t}$ have the same dimension, allowing them to be combined to produce the joint scene-textual feature Z^f as follows:

$$Z^f = (F^s \odot \widehat{F^t}) \tag{19}$$

here, $\odot$ represents the element-wise multiplication of two vectors. The normalized attention score is then obtained by feeding the fused feature vector Z^f through a nonlinear process with a SoftMax operation:

$$\alpha^f = \frac{exp(\varphi(W * Z^f + b))}{\sum_f (\varphi(W * Z^f + b))} \tag{20}$$

where W and b are the learnable parameters, φ is the nonlinear activation function (e.g., tanh). The attentive strength over various visual scenes can therefore be regulated using the attention scores. The weighted average of the scene's overall features is used to determine the attended visual features as follows:

$$\widehat{F^{es}} = \sum_f \alpha^f * F^s \tag{21}$$

In contrast to the original shared visual features F^s, the emotionally attended scene-textual features $\widehat{F^{es}}$ are more representative in capturing the emotional scene regions associated with the textual feature $\widehat{F^t}$.

Second, the convolutional block attention module is also applied to highlight the significant scene characteristics across the channel and spatial dimensions. In order to perform this task, a series of channel and spatial attention modules are utilized sequentially. These modules operate on the scene feature map $V^s \in [H * W * C]$, extracted using the Scene-VGG16 network from each image, where H, W, and C represent the height, width, and number of channels for the scene feature map. This enables each branch to acquire knowledge on both "what" and "where" to concentrate on the channel and spatial dimensions with respect to the scene features. The following section describes each attention module in detail.

A. Channel attention module. To attain effective computation of channel attention, the first step involves gathering spatial data from the feature map by utilizing average and max-pooling operations. The outcome of this process entails the production of two different spatial context descriptors, namely F^c_{avg} and F^c_{max}. These descriptors are subsequently transmitted to a shared network, which generates a channel attention map denoted as Z_c, with dimensions of $(1 * 1 * 256)$. After implementing the shared network, the resultant feature vectors are combined through element-wise summation. In summary, the computation of channel attention can be expressed as follows:

$$\begin{aligned} Z_c &= \sigma(MLP(AvgPool(V^s)) + MLP(MaxPool(V^s))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \end{aligned} \tag{22}$$

where σ denotes the sigmoid function and V^s specifies the scene feature map, whereas W_0 and W_1 of the MLP weights are shared between the two inputs.

B. Spatial attention module. In contrast to channel attention, spatial attention is concerned with identifying the specific location of informative features within an image. The

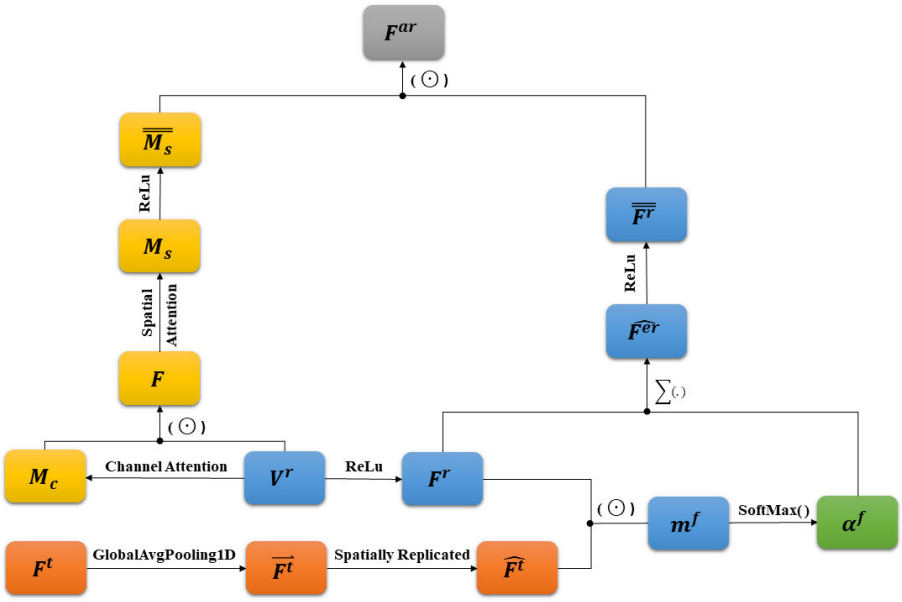


FIGURE 6. The regional attention network.

feature map V^s is subjected to an element-wise multiplication with the channel attention map Z_c, generating the channel-refined feature map K, which is then aggregated utilizing two pooling operations. The output of this process results in generating two maps: average pooled F^s_{avg} and max pooled F^s_{max}, representing average-pooled and maximum-pooled channel features. These features are combined and undergo a convolutional layer with a kernel size of 7×7, producing a spatial attention map denoted as Z_s, which encode the regions that require further or less attention. The transformation of the channel-refined feature map K into the spatial attended features Z_s can be formulated as follows:

$$\begin{aligned} Z_s &= \sigma(F^{7\times7}([AvgPool(K); MaxPool(K)])) \\ &= \sigma(F^{7\times7}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{23}$$

where σ indicates the sigmoid function, while $f^{7\times7}$ implies a convolution operation utilizing a filter size of 7×7. As a result, the attended scene features can be derived by element-wise multiplication of $\overline{\overline{Z_s}}$ and $\overline{\overline{F^s}}$, as shown below.

$$\overline{\overline{F^s}} = ReLu(W^s\widehat{F^{es}} + b^s) \in R^d \tag{24}$$

$$\overline{\overline{Z_s}} = ReLu(W^z Z_s + b^z) \in R^d \tag{25}$$

$$F^{as} = (\overline{\overline{Z_s}} \odot \overline{\overline{F^s}}) \tag{26}$$

where (W^s, W^z) represents the weights and (b^s, b^z) is the bias, which represent the trainable parameters, and d represents the number of units for the dense layers, equal to 256. The successful implementation of the scene attention network facilitates capturing the scene's significant emotional and discriminative features. These features are essential for understanding the interaction of various regions within a scene in conjunction with textual words or sentences. Indeed, visual content incorporates valuable semantic information, including regions and scenes. Moreover, human sentiments correlate highly with this visual information, so they help understand users' sentiments in multimodal information. The details of this network are shown in Figure 7.

3) CROSS-MODAL FUSION LEARNING

The cross-modal fusion learning module aims to establish a comprehensive framework that integrates three distinct characteristics: attended textual features, attended region visual features, and attended scene visual features, to efficiently capture the complementary features shared across multiple modalities. This module comprises three distinct layers: a feature up-sampling layer, a cross-modal fusion layer, and a classifier layer.

The feature up-sampling layer is intended to uniformly up-sample three feature vectors: $F^{at}; F^{ar}; F^{as}$, which is accomplished by adding a dense layer with 256 neurons and a ReLu activation function. This process facilitates the accurate integration of these feature vectors. The formulas are expressed as follows:

$$\widetilde{F^{at}} = ReLu(W^t F^{at} + b^t) \in R^d \tag{27}$$

$$\widetilde{F^{ar}} = ReLu(W^r F^{ar} + b^r) \in R^d \tag{28}$$

$$\widetilde{F^{as}} = ReLu(W^s F^{as} + b^s) \in R^d \tag{29}$$

The cross-modal fusion layer conducts three fusion operations, namely, $(\widetilde{F^{at}}, \widetilde{F^{ar}})$ between the attended textual features and the attended region visual features, $(\widetilde{F^{at}}, \widetilde{F^{as}})$ between the attended textual features and the attended scene visual features, and $\widetilde{F^{ar}}, \widetilde{F^{as}})$ between the attended region visual features and the attended scene visual features. The fusion operation implemented through element-wise

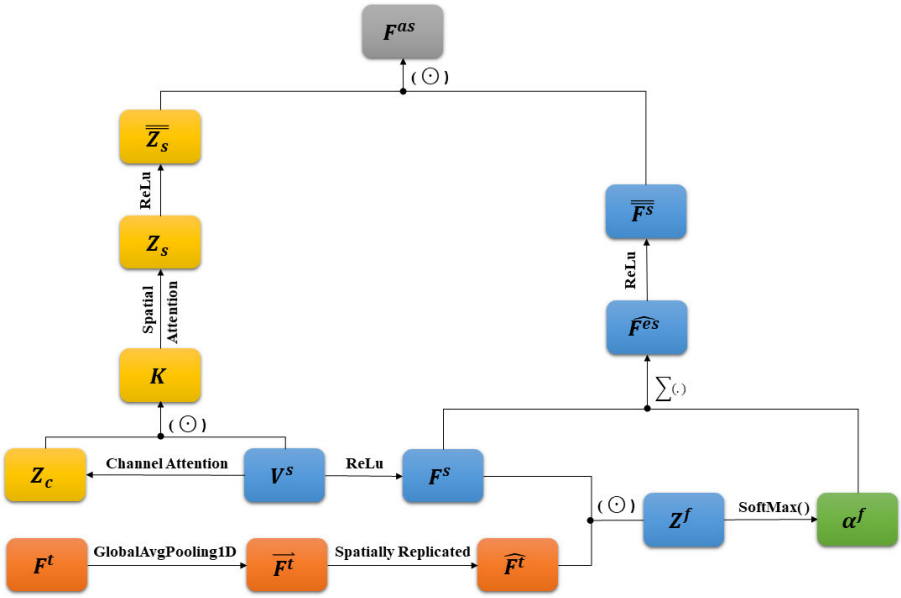


FIGURE 7. The scene attention network.

multiplication and L2 normalization is used to limit the magnitude of the representation as expressed in Eqs. 30,31,32, respectively.

$$F^{tr} = Norm_2(\widetilde{F^{at}} \odot \widetilde{F^{ar}}) \tag{30}$$

$$F^{ts} = Norm_2(\widetilde{F^{at}} \odot \widetilde{F^{as}}) \tag{31}$$

$$F^{rs} = Norm_2(\widetilde{F^{ar}} \odot \widetilde{F^{as}}) \tag{32}$$

Following the implementation of the cross-model fusion layer, which attempts to capture the complementary nature among the various multimodal features, the attention mechanism is used to obtain the necessary data from the shallow fusion characteristics and achieve improved multimodal feature fusion. The attention operation involves an input comprising queries Q and keys K of dimension d_k, and values V of dimension d_v. In our study, the obtained features from the joint representations: F^{rs}, F^{tr}, F^{ts}, are exploited as the queries, keys, and values, respectively. Initially, the dot products are calculated between the query and all keys, followed by the division of each product by $\sqrt{d_k}$. Then, a SoftMax function is applied to derive the weights assigned to the values.

$$Attention(Q, K, V) = Softmax(QK^T/\sqrt{d_k})V \tag{33}$$

Rather than running a single attention function with keys, values, and queries of d_{model}-dimensions, it is preferable to run linear projections of the queries, keys, and values h times. Thus, multi-head attention is used in the present study to perform the attention function in parallel on every projected variant of queries, keys, and values, resulting in output values of d_v- dimensions. Finally, the values are concatenated and projected again, leading to the final values, as shown below.

$$O = Concat(O_1, O_2, \ldots, O_h)W^O \tag{34}$$

For each head i, the output

$$O_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{35}$$

here $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_v}$, and $W^O \in R^{hd_v \times d_{model}}$ are parameter matrices that need to be learned during training. In this study, we employ the multi-head attention class, which is imported from the keras_multi_head package with $h = 4$ parallel attention layers or heads. For each of these, $d_k = d_v = d_{model}/h$. The parameter W^O and the multi-head output feature O maintain the same dimensions as the input features.

Finally, the obtained output feature from the multi-head attention is flattened and sent to the classifier layer to perform the sentiment classification. Specifically, an MLP consisting of four stacked fully connected layers, each with 256, 256, 256, and 128 units, and ReLu as an activation function are employed to enable the acquisition of deeply fused features, with the network weights being shared across the four stacked layers. Followed by a dropout layer with a 50% probability to prevent overfitting. The final representation F^{se} is fed into the SoftMax classifier to predict the final sentiment as follows:

$$p(s) = softmax(W^s, F^{se}) \tag{36}$$

$$L_{CE} = -\sum log(p(s), y) \tag{37}$$

where W^s denotes the SoftMax layer parameters, $p(s)$ is the sentiment prediction probability distribution, and y is the actual sentiment label of the training data. The comprehensive network is trained to minimize cross-entropy (CE) loss to attain maximum efficiency.

IV. EXPERIMENTAL SETUP

A. DATASETS

The proposed DMVAN model is evaluated on three social media datasets to determine its efficacy. The datasets are divided into training, validation, and testing sets, with the proportions being 60:20:20. The complete statistical information for the datasets is displayed in Tables 1 and 2, and the datasets are also discussed in detail as follows:

1) **Getty Images:** Getty Images [52] provides creative photographs, videos, and audio to businesses and consumers, with over 477 million resources in its collection. The main advantages of Getty Images are its user-friendly, efficient query-based search engine and its formal yet descriptive image descriptions. In particular, 3244 adjective-noun pairs (ANPs) from the visual sentiment ontology [53] are used as keywords to collect two types of datasets. The first dataset comprises a total of 20,127 image-text samples that are related to sentiment classification. The dataset contains two classes, namely Positive and Negative. The dataset, which comprises images, relevant textual explanations, and labels, is named ''Binary-Getty'' (BG).
   The second dataset pertains to emotion classification and comprises 19,732 image-text samples, divided into four classes: Angry, Disgust, Happy, and Sad. The dataset, comprising images, textual explanations, and labels, is named ''Emotion-Getty'' (EMO-G).
   Initial labeling for the sentiment dataset is done using sentiment scores associated with ANP keywords. To achieve strong labeling, we used a valence-aware dictionary and sentiment reasoner (VADER) [54], a lexicon, and a rule-based SA tool [55] to label the preprocessed textual description. Then, only the text samples with identical ANP and VADER sentiment scores are chosen. Due to the close relationship between Getty Images' text and image content, the image samples are classified based on the accompanying textual labeling. Finally, three volunteers are chosen to assess our data sets' quality. Each image-text sample is graded 1 (suitable) or 0 (unsuitable). The results show that 95% of the samples are suitable and 5% are unsuitable; we only considered the samples with grade 1 (suitable) and ignored the others.
   For the emotion dataset, the initial labeling is accomplished using the national research council of canada (NRC) [56], a lexicon and rule-based emotion analysis tool [57], to label the preprocessed textual description. Due to the close relationship between the text and image content of Getty Images, we classified the image samples based on the accompanying textual labeling. Finally, three volunteers are chosen to evaluate our data sets' quality. Each image-text sample is graded 1 (suitable) or 0 (unsuitable). The results show that 93% of the samples are suitable and 7% are unsuitable; we only considered the samples with grade 1 (suitable) and ignored the others.
2) **Twitter Dataset:** Additionally, we gathered a new dataset from Twitter. English tweets with text and photos are specifically gathered using the Twitter streaming application programming interface (API) [58], with user-generated hashtags as keywords. We carefully filtered out duplicated, low-quality, pornographic photos and all text that was too short (less than five words) or too long (more than 100 words). To speed up the labeling process, VADER is used to predict text sentiment polarity. Then, a visual SA model [59] is employed, utilizing the T4SA [60] dataset to forecast the polarity of the visual sentiment. Based on the projected sentiment polarity and visual-textual content, the tweets are manually categorized as having Positive, Negative, and Neutral sentiments. Finally, high-quality tweets containing 17,073 image-text pairs are obtained.
   As text data typically contains numerous irrelevant characters for SA, the three data sets' text data is preprocessed in the following manner: (1) Lowercase involves changing all text to lowercase. (2) Removing irrelevant information, including punctuation, special characters (e.g., $, &, and %), hashtags, additional spaces, URL references, @username, stop words, and numbers. (3) Emoticon translation involves translating all emoticons into their respective terms. (4) Spelling correction involves correcting the spelling of words to reflect their intended meanings accurately. (5) Language translation involves converting each text to English using Google Translate [61].

B. IMPLEMENTATION DETAILS

The proposed model is developed in Python 3.7.13, utilizing the Keras library in the Google Colaboratory environment. It is trained with a learning rate of 0.0001 and a default batch size of 32 using the stochastic gradient descent (SGD) optimizer. The advantage of using SGD is that it generalizes better, resulting in greater overall performance. It is computationally efficient and scalable to massive datasets. It updates the model parameters in small batches, making it suited for large-scale training. SGD uses substantially less memory than batch gradient descent since it processes data points in small batches. This is especially crucial when dealing with memory-intensive models and massive datasets. SGD's stochastic character allows it to converge faster, especially when dealing with noisy or high-dimensional data. To ensure a safe upper bound, the proposed model is trained for 50 epochs with early stopping using a patience value of 4. The model is evaluated using accuracy metrics and a loss function based on cross-entropy. The research used an NVIDIA A100 graphics processing unit (GPU) and 25 GB of random-access memory (RAM).

C. RESULTS AND ANALYSIS

The following evaluation metrics are utilized to assess the effectiveness of the proposed model and compare it to prior

TABLE 1. The complete statistics for the sentiment datasets.

Dataset	Positive	Negative	Neutral	Total	Train	Valid	Test	Max. #Words	Min. #Words	Avg. #Words
BG	10098	10029	–	20127	12880	3221	4026	283	1	13.76
Twitter	6075	5228	5770	17073	10926	2732	3415	117	1	8.21

TABLE 2. The complete statistics for the emotion dataset.

Dataset	Angry	Disgust	Happy	Sad	Total	Train	Valid	Test	Max. #Words	Min. #Words	Avg. #Words
EMO-G	5102	3989	5740	4901	19,732	12628	3157	3947	94	1	14.69

research: precision, recall, F1-score, and accuracy. These measurements are explained and computed as follows:

Accuracy is the proportion of accurate predictions to the total number of examined instances, which indicates the model's overall performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{38}$$

Precision is the proportion of accurate positive predictions to all positive predictions generated by the classifier. It evaluates the model's ability to identify only the relevant instances accurately.

$$Precision = \frac{TP}{TP + FP} \tag{39}$$

Recall, also called sensitivity, is the proportion of accurate positive outcomes to the total number of actual positive outcomes (the sum of true positives and false negatives). It evaluates the model's capacity for identifying every relevant instance.

$$Recall = \frac{TP}{TP + FN} \tag{40}$$

The F1-score represents the harmonic mean of accuracy and recall. It seeks to strike a balance between these two metrics and provides a single score that reflects the model's overall performance. This measurement ranges from 0 to 1. The classifier returns a value of 1 when all samples are correctly classified, indicating a high level of classification success.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{41}$$

where TP = true positive, TN = true negative, FP = false positive, and FN = false negative. All evaluation measures range from 0 to 100%, with higher values indicating great model performance. The metrics mentioned earlier provide a thorough comprehension of the model's performance. Figure 8 reports the overall outcomes the proposed DMVAN model accomplished on the sentiment and emotion datasets. The proposed DMVAN model on the BG dataset achieves an average accuracy of 99.801%, precision of 99.802%, recall of 99.801%, and F1-score of 99.801%. Also, with the Twitter dataset, the proposed DMVAN method attains an average accuracy of 96.867%, precision of 96.875%, recall of 96.959%, and F1-score of 96.913%. In addition, the presented DMVAN approach achieves an average accuracy of 96.174%,

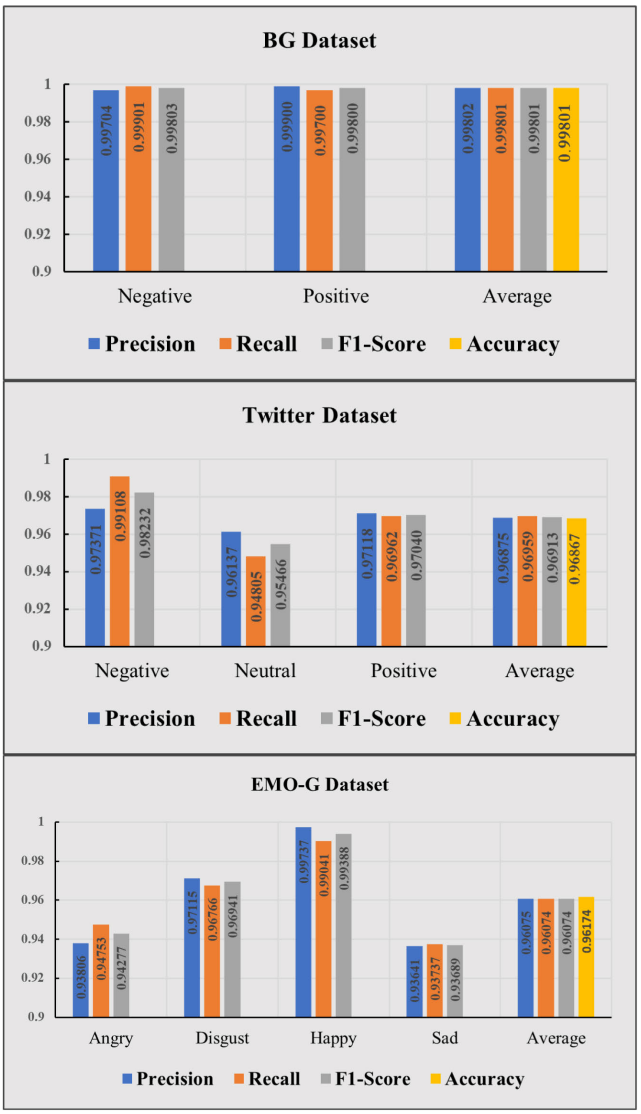


FIGURE 8. Experimental results on the datasets.

precision of 96.075%, recall of 96.074%, and F1-score of 96.074% using the EMO-G dataset.

The proposed model's performance on the sentiment datasets is illustrated in Figures 9 and 10, which display the values of training accuracy (TA), validation accuracy (VA), training loss (TL), and validation loss (VL). The experimental results in Figure 9 demonstrate that the DMVAN approach, as applied to the BG dataset, produces the maximum TA and

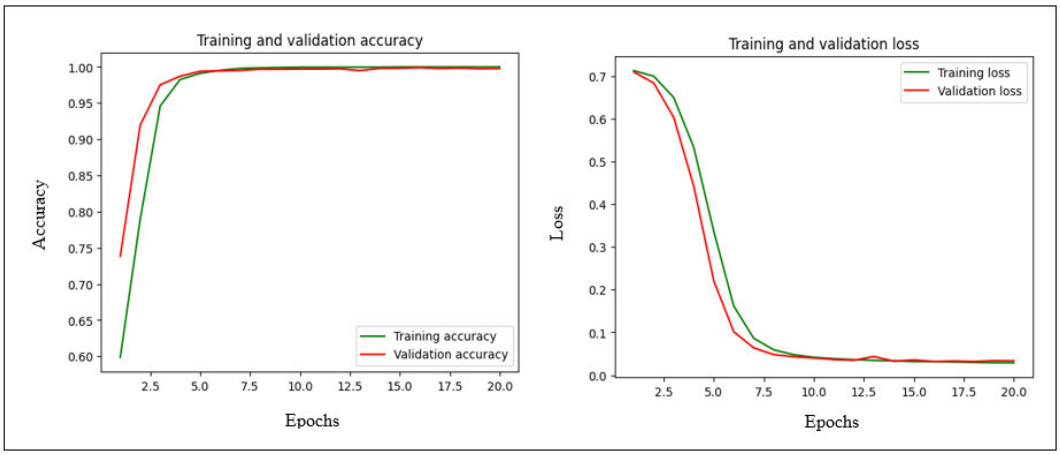


FIGURE 9. The accuracy and loss curves for the BG dataset during training and validation.

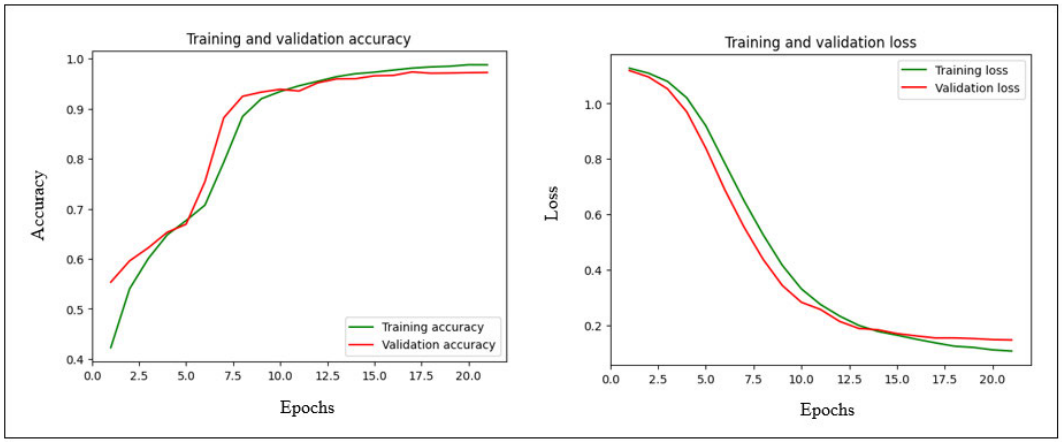


FIGURE 10. The accuracy and loss curves for the Twitter dataset during training and validation.

VA values. Notably, the VA values exceed the TA values until epoch 5, when both values converge. In addition, the approach yields the minimum TL and VL values, with the VL values remaining lower compared to the TL values until epoch 10, where both values converge. It is important to mention that the proposed model's overall training time is 1 hour, 19 minutes, and 25.99 seconds.

The experimental results in Figure 10 indicate that the proposed DMVAN approach on the Twitter dataset achieves the maximum TA and VA values. In contrast, the VA values exceed the TA values until epoch 10, when both values converge. At the same time, it achieves the minimum TL and VL values, wherein the VL values are lower compared to the TL values until reaching near epoch 12.5, where both values converge. The overall training time for the proposed model is 1.0 hours, 42.0 minutes, and 46.83 seconds.

On the other hand, the performance of the proposed method on the emotion dataset is shown in Figure 11. The findings demonstrate that the model reaches its maximum TA and VA values, with VA values initially exceeding TA values until epoch 10, at which point both values converge. The model also produces the lowest TL and VL values, with VL values continuing to be lower than TL values until epoch 10, when both values converge. The model's training time is 1.0 hours, 16.0 minutes, and 33.87 seconds.

Meanwhile, a comparison between the confusion matrix of the DMVAN model on all the datasets is shown in Figure 12. It can be noticed that the proposed DMVAN model on the BG dataset performs better in accurately classifying 99.901% of the negative polarity while achieving 99.7% in detecting the positive polarity. It correctly identifies the 2023 samples as negative and the 1995 samples as positive.

Similarly, the proposed model on the Twitter dataset fares better in correctly classifying 99.108% and 96.962% of the negative and positive polarities while achieving 94.805% in detecting the neutral polarity. Specifically, it correctly classifies 1000 samples as the negative class, 1213 as the positive class, and 1095 as the neutral class.

On the other hand, it is demonstrated that the proposed DMVAN using the EMO-G dataset performes better in correctly classifying 99.041%, 96.766%, and 94.753% of the happy, disgusting, and angry emotions while achieving

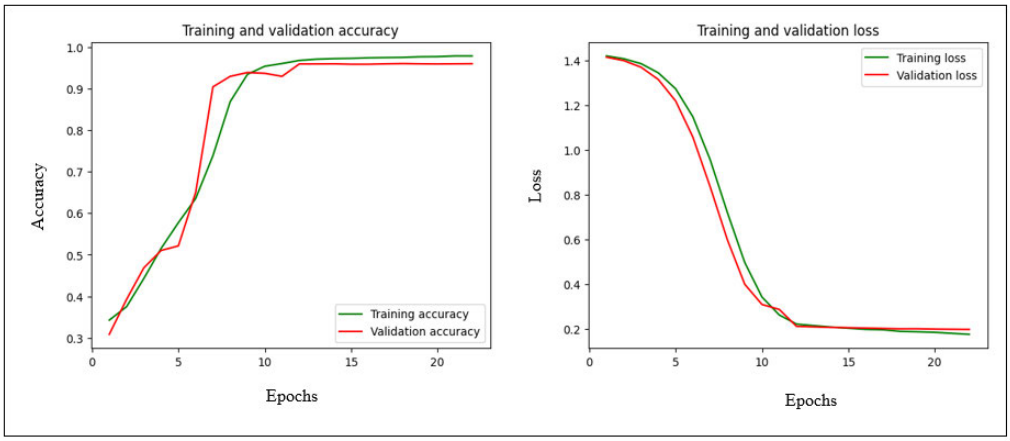


FIGURE 11. The accuracy and loss curves for the emotion dataset during training and validation.

93.737% in detecting the sad emotion. It correctly classifies 1136 samples under the happy class, 808 under the disgusting class, and 939 under the angry class, while correctly classifying 913 samples from the sad class.

D. COMPARED METHODS AND BASELINES

The proposed model is evaluated in comparison to unimodal and multimodal baselines and other recent literature on visual-textual SA.

1) UNIMODAL SENTIMENT BASELINES

Our study assesses the effectiveness of unimodal sentiment analysis techniques to emphasize the benefits of incorporating multimodal feature fusion. For textual modality, Single Textual Model [11]. LR-BERT and SVM-BERT models utilize logistic regression (LR) and support vector machine (SVM) to classify the sentiment based on textual features retrieved using BERT and CNN, LSTM [44], and CNN [62]. Hybrid-ACL predicts sentiment by combining the CNN and LSTM models with an attention mechanism. For visual modality, Single Visual Model [11], Inception-V3 [63], ResNet50 [64], VGG19 [45], and SC-IMG predict the sentiment by combining the scene and region visual features.

2) MULTIMODAL SENTIMENT BASELINES

Different approaches have been proposed; **Early Fusion-1** [11], **Early Fusion-2**: an LR classifier predicts sentiment by combining visual and textual features extracted using VGG19 and BERT with CNN. **Late Fusion-1** [11], **Late Fusion-2**: a classification-based approach using SVM on a single visual and textual model, with both models classified using LR, where the visual and textual features are extracted using VGG19 and BERT with CNN.

3) COMPARISON TO CURRENT VISUAL-TEXTUAL SA RESEARCH

Our results were compared to several robust baseline methods reported in the literature on visual-textual SA. Although other articles have used different datasets, particularly the BG and Twitter datasets, it is still essential to shed light on the factors they considered and the categorization strategy they employed with their results. The comparative findings are shown in Tables 3 and 4, which are discussed in Section IV-E. To the best of our knowledge, no model has been published that uses emotion datasets obtained from the Getty Images website. As a result, the proposed model was only compared to the previously described baseline models.

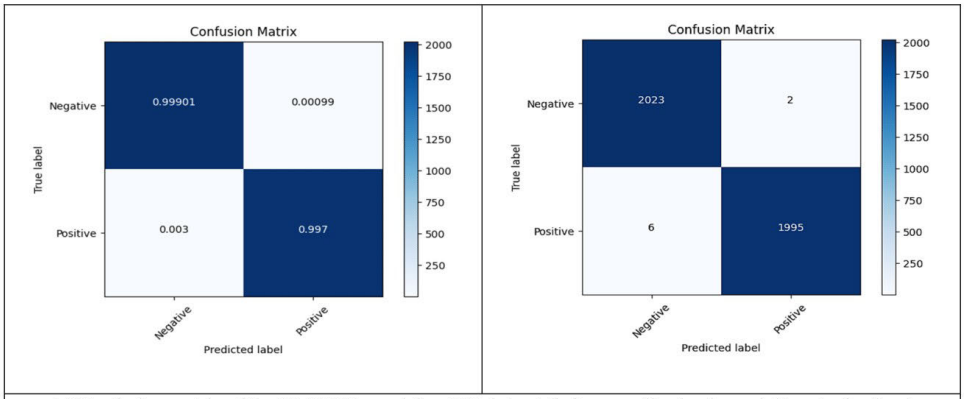
E. COMPARATIVE RESULTS AND DISCUSSION

Tables 3 and 4 demonstrate the results of the proposed model utilizing the three datasets in comparison to the unimodal and multimodal sentiment baseline models and other recent methods reported in the literature, which indicate the following observations:

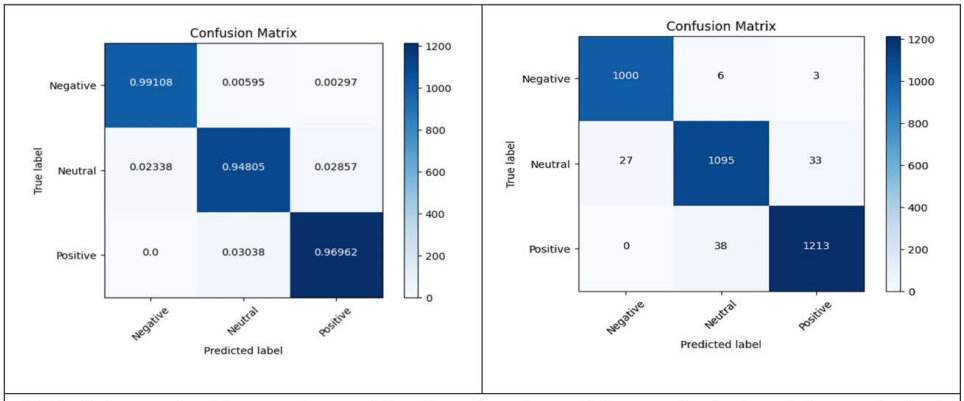
Firstly, the unimodal baselines based on image data demonstrate the poorest performance among all the datasets. The primary reason is that images lack the contextual information required for a more accurate interpretation. Unlike words, visuals cannot directly describe emotions. Thus, adding additional information, such as visual cues based on scene information and textual data, improves the efficacy of SA.

Secondly, the unimodal baselines based on text demonstrate better performance than the image-based analysis models. This can be attributed to the superior efficacy and informative nature of emotional cues in textual content compared to visual information and the success of BERT models in acquiring knowledge from extensive datasets, which enhances their effectiveness in extracting task-relevant features.

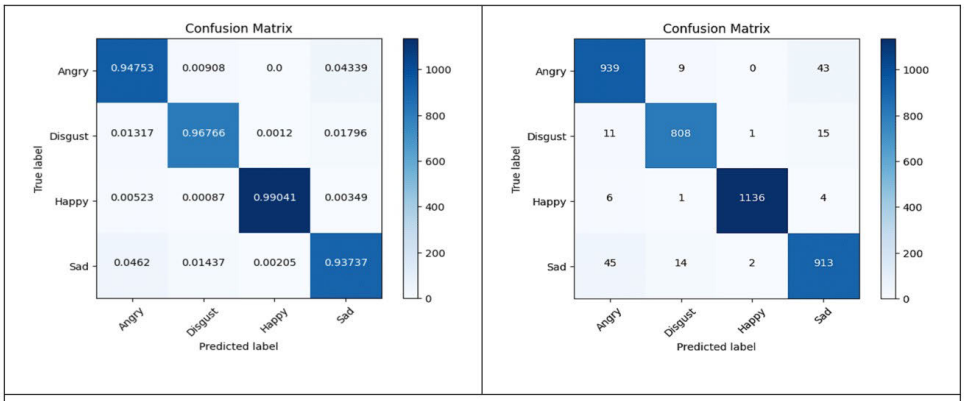
Thirdly, it is observed that MSA models outperform most single-modal SA models by a significant margin on the three datasets. This demonstrates that relying only on textual or visual elements is usually inadequate for SA. In contrast, combining several modalities can aid in capturing semantic



(a) Confusion matrix of the DMVANs model on BG dataset (left: normalized values, right: actual values)



(b) Confusion matrix of the DMVANs model on the Twitter dataset (left: normalized values, right: actual values)



(c) Confusion matrix of the DMVANs model on the EMO-G dataset (left: normalized values, right: actual values)

FIGURE 12. **A comparison between the confusion matrix of the DMVAN model.**

characteristics and natural relationships through data integration, thereby acquiring more information.

In order to assess the effectiveness of our proposed approach, a comparative analysis between our investigation was undertaken utilizing the BG dataset and the current literature: [11], [16], [17], [18], [23], [24]. The comparison results in Group 1 of Table 3 demonstrate that, regarding the F1-score (92.60%) and accuracy (92.65%), the DMLANet model outperformed AMGN. In contrast, AMGN showed better performance compared to BDMLA and VSCN. It had an F1-score of 88.7% and 88.% accuracy. On the other hand, the latter achieved accuracies of 86.5% and 85.6%, respectively. The SCC model exhibited the least optimal performance compared to the other models. This was demonstrated by its F1-score of 81.0% and accuracy rate of 80.6%. Although HCIM showed remarkable performance compared to other techniques, with an F1-score of 93.2% and an accuracy rate of 93.6%, our model outperforms it. According to the findings, our model demonstrates superior performance compared to the current leading approach by a significant

TABLE 3. Comparing the outcomes of several techniques on the sentiment datasets, PR is the precision, RE denotes recall, F1 refers to the F1-score, and ACC is the accuracy.

Modality	Models	BG Dataset				Twitter Dataset			
		PR	RE	F1	ACC	PR	RE	F1	ACC
Unimodal Text Baselines	Single Textual Model	87.56	87.51	87.50	87.51	75.48	75.46	75.47	75.55
	LR-BERT	94.93	94.94	94.93	94.93	84.69	84.49	84.58	84.39
	SVM-BERT	95.20	95.21	95.21	95.21	84.62	84.45	84.53	84.36
	GRU	97.52	97.51	97.52	97.52	88.71	88.71	88.71	88.58
	CNN	97.00	97.00	96.99	96.99	87.53	87.46	87.49	87.32
	Hybrid-ACL	98.86	98.85	98.86	98.86	91.24	91.27	91.25	91.16
Unimodal Image Baselines	Single Visual Model	72.68	72.65	72.64	72.65	71.30	71.50	71.39	71.22
	InceptionV3	76.31	76.31	76.30	76.30	70.68	70.35	70.47	70.40
	ResNet50	69.06	68.91	68.82	68.88	71.28	71.29	71.06	71.01
	VGG19	77.17	77.17	77.17	77.17	72.71	72.74	72.72	72.62
	SC-IMG	78.88	78.67	78.60	78.64	73.82	73.90	73.85	73.76
Multimodal Baselines	Early Fusion-1	89.87	89.88	89.87	89.87	81.50	81.48	81.49	81.49
	Early Fusion-2	97.24	97.25	97.24	97.24	86.87	86.81	86.84	86.73
	Late Fusion-1	89.28	89.28	89.27	89.27	81.78	81.80	81.79	81.82
	Late Fusion-2	96.94	96.93	96.94	96.94	86.44	86.35	86.38	86.24
Current Literature	SCC [23]	83.2	79.1	81.0	80.6	-	-	-	-
	VSCN [24]	85.9	84.7	85.3	85.6	-	-	-	-
	BDMLA [11]	87.1	85.4	86.2	86.5	-	-	-	-
	AMGN [16]	89.8	87.6	88.7	88.2	-	-	-	-
Group 1	DMLANet [18]	-	-	92.60	92.65	-	-	-	-
	HCIM [17]	92.8	93.6	93.2	93.6	-	-	-	-
Group 2	ITMSC [33]	-	-	-	-	-	-	68.40	70.28
	TomBERT-17 [20]	-	-	-	-	-	-	68.04	70.50
	HFN-17 [21]	-	-	-	-	-	-	68.52	71.35
	EF-CapTrBERT-DE-17 [22]	-	-	-	-	-	-	70.2	72.3
	DMVAN (Ours)	**99.802**	**99.801**	**99.801**	**99.801**	**96.875**	**96.959**	**96.913**	**96.867**

margin. It achieves the maximum accuracy and an F1-score of 99.801%.

A comparison of our study utilizing the Twitter dataset was further carried out with the following current literature: [20], [21], [22], [33]. Based on the comparative outcomes presented in Group 2 of Table 3, one can infer that EF-CapTrBERT-DE performed better than HFN. This was supported by its F1-score of 70.2% and accuracy rate of 72.3%. The TomBERT architecture performed inferiorly, achieving 68.04% for the F1-score and a 70.50% accuracy rate. In contrast, the HFN model displayed a notable enhancement over the TomBERT, achieving 68.52% and 71.35% in F1-score and accuracy, respectively. The model known as ITMSC exhibited the least optimal performance compared to the other models. This was demonstrated by its F1-score of 68.40% and accuracy rate of 70.28%. In terms of F1-score (96.913%) and accuracy (96.867%), our proposed model surpasses the state-of-the-art by a significant margin.

To further illustrate the advantages of our model, we present a comparative analysis of its outcomes on the EMO-G dataset, as displayed in Table 4. The extensive and varied nature of the EMO-G dataset ensures that the models remain relatively unaffected by various factors, yielding robust outcomes. The results indicate that the DMVAN model exhibits a competitive level of performance, achieving an F1-score of 96.074% and an accuracy of 96.174% compared to the baseline methodologies. The model is believed to show a superior ability to detect emotions such as happiness, disgust, and anger, potentially due to users' explicit expression of these emotions through text and images. The models exhibit inferior performance in detecting sadness compared to their performance in detecting other emotions, which may be attributed to the implicit nature of users' expressions of sadness.

The results outlined above illustrate the proposed model's superiority. It integrates deep semantic visual and textual features from various perspectives and levels, enabling it to extract more efficient features that accurately reflect the sentiment of the image-text information. Furthermore, attentive interaction learning enhances the interaction between two modalities, facilitating the acquisition of discriminative and emotional visual features by utilizing text information. Moreover, incorporating the cross-modal fusion learning module helps to capture the complementary nature of multiple modalities, followed by utilizing multi-head attention to gather sufficient information while facilitating the development of an effective joint representation of intermediate features. Finally, using the stacking-fully connected layers within the MLP enables the features to be deeply fused, ultimately leading to improved outcomes in the context of MSA.

F. INTERPRETABLE MULTIMODAL SENTIMENT CLASSIFICATION MODEL

Most related literature focused on new designs to improve this task, with few attempts to explain these models' decisions. This study presents an interpretable multimodal sentiment classification model using LIME to define an explainable model over an interpretable illustration that is locally accurate for any classifier's predictions. After dividing the input into features, it randomly perturbs each feature S times and

TABLE 4. Comparing the outcomes of several techniques on the emotion dataset.

Modality	Models	EMO-G Dataset			
		Precision	Recall	F1-score	Accuracy
Unimodal Text Baselines	Single Textual Model	74.02	73.95	73.97	74.92
	LR-BERT	86.57	86.31	86.41	86.88
	SVM-BERT	86.81	86.59	86.69	87.13
	LSTM	92.91	92.93	92.90	93.06
	CNN	90.73	90.66	90.69	90.93
	Hybrid-ACL	93.67	93.60	93.62	93.77
Unimodal Image Baselines	Single Visual Model	60.67	60.65	60.57	60.83
	InceptionV3	64.13	63.77	63.91	63.95
	ResNet50	62.94	62.67	62.37	63.24
	VGG19	67.29	67.15	67.19	67.32
	SC-IMG	69.21	69.02	68.97	69.22
Multimodal Baselines	Early Fusion-1	81.65	81.66	81.64	81.99
	Early Fusion-2	89.66	89.66	89.65	89.94
	Late Fusion-1	81.46	81.02	81.20	81.66
	Late Fusion-2	88.74	88.72	88.73	89.00
	DMVAN (Ours)	**96.075**	**96.074**	**96.074**	**96.174**

FIGURE 13. Interpretable multimodal sentiment classification model.

analyzes the model's output logits for class c. LIME then produces a linear model that maps each feature's perturbations to their logits of c. The linear model's weights explain each feature: a positive weight supports class c, while a negative weight opposes it. Furthermore, the higher the absolute value of the weight, the more significant its contribution.

To explain the model results, the fundamental LIME method is enhanced to make it applicable to our proposed DMVAN model using LIME Explainer for textual and visual content. In the case of image data, the explanations are generated by creating a new dataset of perturbations surrounding the instance that needs to be explained. For this purpose, a simple linear iterative clustering (SLIC) segmentation algorithm [65] is employed that effectively groups pixels in the unified 5-dimensional color and picture plane space to construct condensed, relatively uniform superpixels. The generated model is then used to forecast the class of the recently created images. The importance (weight) of each perturbation in predicting the related class is calculated using cosine similarity and weighted linear regression. Finally, LIME explains the image regions (superpixels) and the most important words that considerably influence the image–text instance's assignment to a specific class.

Figure 13 displays some of the explanations provided by LIME using the Twitter dataset; as can be seen, the explanation model has effectively highlighted the most critical terms in the text section and the essential image regions (pros in green, cons in red), which contribute more to the final correct prediction and have greater weight. Where (P1, P2), (NU1, NU2), and (NG1, NG2) represent the original and interpreted images for the Positive, Neutral, and Negative classes, respectively.

V. CONCLUSION

In this study a novel deep multi-view attentive network (DMVAN) was proposed for multimodal sentiment and emotion classification. Our model could extract visual features from multiple viewpoints, including region and scene, as well as textual features from various levels of analysis, such as word, sentence, and document levels, which aimed to leverage the associations between the visual perspectives and the semantic aspects of the text description in a unified framework. An attentive interaction learning module was proposed to improve the interaction between the visual and textual characteristics; this module aimed to capture the discriminative and emotional visual features by utilizing text information to guide the learning process for image features and vice versa. Moreover, a cross-modal fusion learning module was created to incorporate various features into a comprehensive framework that acknowledged the complementary nature of multiple modalities—followed by multi-head attention—constructed to gather sufficient information from the fusion of shallow features while facilitating the development of an effective joint representation of intermediate features. Finally, an MLP that incorporates stacking-fully connected layers was utilized to deeply fuse the modal features, thereby enhancing the efficiency of sentiment classification. To facilitate the implementation of multimodal emotion analysis, an image-text dataset (Emotion-Getty) was further developed and annotated with emotional categories.

The experimental results from the analysis of three real-world datasets indicated that multimodal approaches produced significantly better results in terms of model evaluation criteria than their corresponding unimodal baseline and current literature techniques and achieved the highest accuracy using the BG dataset with 99.801%. Thus, it could be concluded that relying solely on textual or visual cues for sentiment classification is usually insufficient and that incorporating diverse modalities might provide more comprehensive information; this validated our strategy for improving decision-making and results.

For future work on interactive learning, we intend to consider the object features in addition to the scene features, which focus on a specific object in the image, by developing an algorithm that can consistently and accurately describe the content of an image. This could be useful in multimodal emotion analysis because the objects' significance changes depending on the scenes in which they are discovered. In addition, we intend to build a model that utilizes the benefits of pre-trained vision-language models to analyze sentiment precisely and deliver more precise outcomes than the current models. One notable aspect of our future work will be the evaluation of the scalability of the proposed model in handling large datasets. The evaluation will be crucial for understanding the model's capacity to maintain high performance, even when dealing with extensive data. Furthermore, our objective is to modify our model to incorporate other types of multimodal data, including audio and video, thus expanding its usefulness to various fields.

REFERENCES

[1] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 8, no. 4, pp. 1–25, Jul. 2018.

[2] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A survey of sentiment analysis from social media data," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 2, pp. 450–464, Apr. 2020.

[3] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Intl. Conf. Mach. Learn. (ICML)*, vol. 4, May 2014, pp. 2931–2939.

[4] R. Obiedat, R. Qaddoura, A. M. Al-Zoubi, L. Al-Qaisi, O. Harfoushi, M. Alrefai, and H. Faris, "Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022.

[5] S. T. Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, Jul. 2022, Art. no. 100157.

[6] H. Ou, C. Qing, X. Xu, and J. Jin, "Multi-level context pyramid network for visual sentiment analysis," *Sensors*, vol. 21, pp. 1–20, Mar. 2021.

[7] A. Yadav and D. K. Vishwakarma, "A deep learning architecture of RA-DLNet for visual sentiment analysis," *Multimedia Syst.*, vol. 26, no. 4, pp. 431–451, Aug. 2020.

[8] H. Xiong, Q. Liu, S. Song, and Y. Cai, "Region-based convolutional neural network using group sparse regularization for image sentiment classification," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, p. 30, Dec. 2019.

[9] G. Chandrasekaran, T. N. Nguyen, and J. Hemanth D., "Multimodal sentimental analysis for social media applications: A comprehensive review," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 5, Sep. 2021, Art. no. e1415.

[10] R. Kaur and S. Kautish, "Multimodal sentiment analysis: A survey and comparison," *Int. J. Service Sci., Manage., Eng., Technol.*, vol. 10, no. 2, pp. 38–58, 2019.

[11] J. Xu, F. Huang, X. Zhang, S. Wang, C. Li, Z. Li, and Y. He, "Visual-textual sentiment classification with bi-directional multi-level attention networks," *Knowl.-Based Syst.*, vol. 178, pp. 61–73, Aug. 2019.

[12] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 929–932.
[13] I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "A comprehensive review of visual-textual sentiment analysis from social media networks," 2022, *arXiv:2207.02160*.
[14] K. Jindal and R. Aron, "A novel visual-textual sentiment analysis framework for social media data," *Cognit. Comput.*, vol. 13, no. 6, pp. 1433–1450, Nov. 2021.
[15] Q. You, J. Luo, H. Jin, and J. Yang, "Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia," in *Proc. 9th ACM Int. Conf. Web Search Data Mining*, Feb. 2016, pp. 13–22.
[16] F. Huang, K. Wei, J. Weng, and Z. Li, "Attention-based modality-gated networks for image-text sentiment analysis," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 3, pp. 1–19, Jul. 2020.
[17] T. Zhou, J. Cao, X. Zhu, B. Liu, and S. Li, "Visual-textual sentiment analysis enhanced by hierarchical cross-modality interaction," *IEEE Syst. J.*, vol. 15, no. 3, pp. 4303–4314, Sep. 2021.
[18] A. Yadav and D. K. Vishwakarma, "A deep multi-level attentive network for multimodal sentiment analysis," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–19, 2022.
[19] X. Yang, S. Feng, Y. Zhang, and D. Wang, "Multimodal sentiment detection based on multi-channel graph neural networks," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2021, pp. 328–339.
[20] J. Yu and J. Jiang, "Adapting BERT for target-oriented multimodal sentiment classification," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5408–5414.
[21] S. Zhang, B. Li, and C. Yin, "Cross-modal sentiment sensing with visual-augmented representation and diverse decision fusion," *Sensors*, vol. 22, no. 1, p. 74, Dec. 2021.
[22] Z. Khan and Y. Fu, "Exploiting BERT for multimodal target sentiment classification through input space translation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3034–3042.
[23] K. Zhang, Y. Zhu, W. Zhang, and Y. Zhu, "Cross-modal image sentiment analysis via deep correlation of textual semantic," *Knowl.-Based Syst.*, vol. 216, Mar. 2021, Art. no. 106803.
[24] M. Cao, Y. Zhu, W. Gao, M. Li, and S. Wang, "Various syncretic co-attention network for multimodal sentiment analysis," *Concurrency Comput., Pract. Exper.*, vol. 32, no. 24, pp. 1–17, Dec. 2020.
[25] X. Hu and M. Yamamura, "Global local fusion neural network for multimodal sentiment analysis," *Appl. Sci.*, vol. 12, no. 17, p. 8453, Aug. 2022.
[26] I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "Multi-model fusion framework using deep learning for visual-textual sentiment classification," *Comput. Mater. Continua*, vol. 76, no. 2, pp. 1–32, 2023.
[27] X. Hu and M. Yamamura, "Two-stage attention-based fusion neural network for image-text sentiment classification," in *Proc. 4th Int. Conf. Image, Video Signal Process.*, Mar. 2022, pp. 1–7.
[28] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Trans. Multimedia*, vol. 23, pp. 4014–4026, 2021.
[29] Z. Li, B. Xu, C. Zhu, and T. Zhao, "CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection," in *Proc. Findings Assoc. Comput. Linguistics (NAACL)*, 2022, pp. 2282–2294.
[30] N. Xu and W. Mao, "MultiSentiNet: A deep semantic network for multimodal sentiment analysis," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 2399–2402.
[31] 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, Mar. 2019.
[32] L. Xing, H. Qu, S. Xu, and Y. Tian, "CLEGAN: Toward low-light image enhancement for UAVs via self-similarity exploitation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610714.
[33] J. An, W. M. N. W. Zainon, and Z. Hao, "Improving targeted multimodal sentiment classification with semantic description of images," *Comput., Mater. Continua*, vol. 75, no. 3, pp. 5801–5815, 2023.
[34] S. F. Kiaei, M. D. Rouzi, and S. Farzi, "Designing and implementing an emotion analytic system (EAS) on Instagram social network data," *Int. J. Web Res.*, vol. 2, no. 2, pp. 9–14, 2019.
[35] P. Kumar, S. Malik, and B. Raman, "Interpretable multimodal emotion recognition using hybrid fusion of speech and image data," 2022, *arXiv:2208.11868*.
[36] P. Kumar, S. Malik, and B. Raman, "Hybrid fusion based interpretable multimodal emotion recognition with insufficient labelled data," 2022, *arXiv:2208.11450*.
[37] Y. Lyu, P. P. Liang, Z. Deng, R. Salakhutdinov, and L.-P. Morency, "DIME: Fine-grained interpretations of multimodal models via disentangled local explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2022, pp. 455–467.
[38] N. Xu, "Analyzing multimodal public sentiment based on hierarchical semantic attentional network," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2017, pp. 152–154.
[39] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018.
[40] 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.
[41] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1033–1038.
[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.
[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2017, pp. 5999–6009.
[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR) Conf. Track*, Sep. 2014, pp. 1–14.
[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[47] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015, pp. 2048–2057. [Online]. Available: https://dl.acm.org/doi/proceedings/10.5555/3045118
[48] 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. 2018.
[49] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 457–468.
[50] D. B. D. Parikh, J. Lu, and J. Yang, "Hierarchical question-image co-attention for visual question answering," *Proc. 30th Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 90, 2016, pp. 343–348.
[51] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8689, Nov. 2013, pp. 818–833.
[52] *Royalty Free Stock Photos, Illustrations, Vector Art, and Video Clips—Getty Images*. [Online]. Available: https://www.gettyimages.com/
[53] 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. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 223–232.
[54] C. E. Hutto and Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014, p. 18.
[55] *GitHub—cjhutto/vaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a Lexicon and Rule-Based Sentiment Analysis Tool That is Specifically Attuned to Sentiments Expressed in Social Media, and Works Well on Texts From Other Domains*. [Online]. Available: https://github.com/cjhutto/vaderSentiment
[56] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013, doi: 10.1111/j.1467-8640.2012.00460.x.
[57] *GitHub—metalcorebear/NRCLex: An Affect Generator Based on TextBlob and the NRC Affect Lexicon*. [Online]. Available: https://github.com/metalcorebear/NRCLex
[58] *Use Cases, Tutorials, & Documentation | Twitter Developer Platform*. [Online]. Available: https://developer.twitter.com/en
[59] *GitHub—Fabiocarrara/Visual-Sentiment-Analysis: For Visual Sentiment Analysis Pre-Trained on the T4SA Dataset*. [Online]. Available: https://github.com/fabiocarrara/visual-sentiment-analysis

[60] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 308–317.
[61] S. Han. *Googletrans PyPI*. [Online]. Available: https://pypi.org/project/googletrans/
[62] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.
[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[65] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

ISRAA KHALAF SALMAN AL-TAMEEMI received the B.Sc. degree in computer and software engineering from Mustansiriya University, Baghdad, Iraq, and the M.Tech. degree in computer and software engineering from Osmania University, Hyderabad, India. She is currently pursuing the Ph.D. degree in computer engineering, the major of artificial intelligence, and the field of natural language processing with the University of Tabriz, Tabriz, Iran. Her research interests include natural language processing, deep learning, and computer vision.

MOHAMMAD-REZA FEIZI-DERAKHSHI received the B.S. degree in software engineering from the University of Isfahan, Iran, and the M.Sc. and Ph.D. degrees in artificial intelligence from the Iran University of Science and Technology, Tehran, Iran. He is currently a Professor with the Faculty of Computer Engineering, University of Tabriz, Iran. His research interests include natural language processing, optimization algorithms, deep learning, social network analysis, and intelligent databases.

SAEID PASHAZADEH (Member, IEEE) received the B.Sc. degree in computer engineering from the Sharif University of Technology, Tehran, Iran, in 1995, and the M.Sc. and Ph.D. degrees in computer engineering from the Iran University of Science and Technology, Tehran, in 1998 and 2010, respectively. He is currently an Associate Professor with the Department of Information Technology, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran. His research interests include formal verification, software engineering, modeling and verification, performance evaluation, and distributed systems.

MOHAMMAD ASADPOUR received the B.S. degree in electrical engineering from the University of Tabriz, Iran, in 1993, the M.S. degree in telecommunication engineering from the Ferdowsi University of Mashhad, Mashhad, Iran, in 1996, and the Ph.D. degree in telecommunication engineering from the University of Tabriz, in February 2015. His research interests include wireless communication, signal processing, channel estimation, power line communication (PLC) channels, and OFDM systems.