Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.039763
Article

Leveraging Vision-Language Pre-Trained Model and Contrastive Learning for Enhanced Multimodal Sentiment Analysis

Jieyu An[1,*], Wan Mohd Nazmee Wan Zainon[1] and Binfen Ding[2]

[1]School of Computer Sciences, Universiti Sains Malaysia, Penang, 11800, Malaysia
[2]Jiangxi University of Applied Science, Nanchang, 330000, Jiangxi, China
*Corresponding Author: Jieyu An. Email: anjieyu@student.usm.my
Received: 15 February 2023; Accepted: 13 April 2023; Published: 23 June 2023

Abstract: Multimodal sentiment analysis is an essential area of research in artificial intelligence that combines multiple modes, such as text and image, to accurately assess sentiment. However, conventional approaches that rely on unimodal pre-trained models for feature extraction from each modality often overlook the intrinsic connections of semantic information between modalities. This limitation is attributed to their training on unimodal data, and necessitates the use of complex fusion mechanisms for sentiment analysis. In this study, we present a novel approach that combines a vision-language pre-trained model with a proposed multimodal contrastive learning method. Our approach harnesses the power of transfer learning by utilizing a vision-language pre-trained model to extract both visual and textual representations in a unified framework. We employ a Transformer architecture to integrate these representations, thereby enabling the capture of rich semantic information in image-text pairs. To further enhance the representation learning of these pairs, we introduce our proposed multimodal contrastive learning method, which leads to improved performance in sentiment analysis tasks. Our approach is evaluated through extensive experiments on two publicly accessible datasets, where we demonstrate its effectiveness. We achieve a significant improvement in sentiment analysis accuracy, indicating the superiority of our approach over existing techniques. These results highlight the potential of multimodal sentiment analysis and underscore the importance of considering the intrinsic semantic connections between modalities for accurate sentiment assessment.

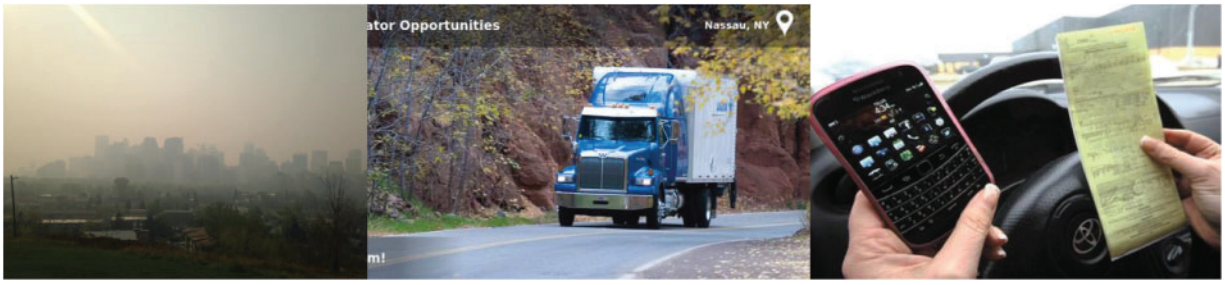
Keywords: Multimodal sentiment analysis; vision–language pre-trained model; contrastive learning; sentiment classification

1 Introduction

Sentiment analysis, also known as opinion mining, involves the computational examination of individuals' perceptions, assessments, evaluations, attitudes, and emotions in relation to products,

services, and events [1]. In recent years, as social media and user-generated content have proliferated, sentiment analysis has become an important area of research within the fields of natural language processing and machine learning [2–7]. Traditional sentiment analysis models, however, rely primarily on text data, ignoring the vast amount of information that can be extracted from other modalities such as images, audio, and video. To address this limitation, researchers have become increasingly interested in developing multimodal sentiment analysis approaches to improve performance [8].

The field of multimodal sentiment analysis has seen a significant increase in research efforts in recent years, intending to develop models that can effectively leverage multimodal data for sentiment analysis. As illustrated in Fig. 1, it has been demonstrated that relying solely on either text or image modalities for sentiment detection may not consistently yield an accurate inference of the sentiment polarity of a tweet. In light of this, multimodal sentiment analysis has been proposed as a solution. This approach incorporates information from multiple modalities, thereby improving the performance of sentiment analysis. Consistently, empirical evidence has demonstrated that multimodal models are more robust and accurate than single-modal models for sentiment analysis [9–12].



(a)That's smoke, not fog. (neutral)

(b) Just got around to scanning some of the summer's great catches for the new book. (positive)

(c) Hefty traffic fines kick in Tuesday. (negtive)

Figure 1: Several image-text pairs extracted from twitter along with their respective sentiment polarities, as indicated in parentheses

Multimodal sentiment analysis is challenged by the need to capture information from multiple modalities. The prevalent methods for effectively incorporating information from multiple modalities in multimodal sentiment analysis involve the utilization of unimodal pre-trained models specifically designed for feature extraction. Examples of these models include the visual geometry group network (VGG) [13] and Residual Networks (ResNet) [14] models for images, as well as the Bidirectional Encoder Representation from Transformers (BERT) [15], the Generative Pre-trained Transformer (GPT) [16] models for text. These models have been pre-trained on large datasets and can effectively capture high-level features from both images and text. Despite their effectiveness in single modality data, pre-trained models may not be able to adequately handle the complexities of multimodal data or capture the interactions between different modalities. In light of this, the development of vision-language pre-trained models is a major advance in the field of artificial intelligence. These models can be trained on multimodal data to learn the representation of visual and textual information. Examples of such models include the Contrastive Language-Image Pre-training (CLIP) model [17] and the Visual-Linguistic BERT (VL-BERT) model [18]. Although vision-language pre-trained models have shown promise in capturing interactions between different modalities, their potential in multimodal sentiment analysis remains largely unexplored. This underscores the necessity for further investigation into the effectiveness of these models and their ability to enhance the accuracy of sentiment analysis.

In addition to vision-language models, contrastive learning is another promising approach for representation learning that has gained popularity in the fields of computer vision and natural language processing [19–23]. Contrastive learning is notable for its focus on learning representations by comparing and contrasting different examples. This learning process necessitates the optimization of feature embeddings within the feature space, where similar examples are drawn towards each other, while dissimilar examples are pushed away. This strategy enables the model to effectively distinguish between diverse examples and acquire meaningful representations. Despite the demonstrated effectiveness of contrastive learning, its application to multimodal sentiment analysis has received relatively little attention in the research community. Leveraging the abundance of information available across multiple modalities presents a promising opportunity to improve the performance of sentiment analysis.

Motivated by the above observations, our work integrates the vision-language pre-trained model with our proposed multimodal contrastive learning approach for multimodal sentiment analysis. Specifically, we utilize the CLIP model, a state-of-the-art vision-language pre-trained model, to extract representations from image-text pairs. To capture the rich semantic information contained within these pairs, we introduce the Transformer architecture. We then apply our proposed multimodal contrastive learning method to enhance these representations, which allows us to gain a more holistic and resilient understanding of information from different modalities. To the best of our knowledge, our work represents the first effort to integrate these approaches for multimodal sentiment analysis, and we anticipate that our findings will stimulate further research in this direction. Towards achieving this goal, we make the following contributions in this paper:

- We propose a novel approach to multimodal sentiment classification, the VLMSA architecture. The VLMSA architecture capitalizes on the advantages of a vision-language pre-trained model and the utilization of contrastive learning techniques, ultimately resulting in improved accuracy in sentiment classification tasks.
- To further improve the performance of the VLMSA, we propose a multimodal contrastive learning approach. The method involves separate data augmentation for different modalities in an image-text pair, forming two new sample pairs for contrastive learning. This approach has the potential to enhance representation learning in situations with limited labeled data.
- In order to validate the efficacy of our proposed method, we conducted extensive experiments and comparisons on publicly available datasets. We demonstrate that a substantial improvement can be achieved for multimodal sentiment classification without the need for additional input features, such as object bounding boxes or face detection, even with limited training epochs.

The structure of this article is as follows: Section 2 provides an overview of the literature on unimodal and multimodal sentiment analysis. In Section 3, we present our VLMSA model in detail. Section 4 describes our experimental results on two publicly available datasets. Lastly, Section 5 concludes our study and outlines potential areas for future research.

2 Related Work

2.1 Unimodal Sentiment Analysis

Textual sentiment analysis, which is sometimes known as opinion mining, is a highly dynamic field of study that has garnered significant interest in the natural language processing community [24–26]. Its primary objective is to identify and extract subjective information from a given text, with the ultimate goal of determining the overall sentiment or emotion that is being conveyed.

Traditionally, sentiment analysis methods have relied on lexicon-based approaches, which associate words or phrases in a text with pre-assigned sentiment scores [27,28]. These methods are contingent upon the existence of lexicons or dictionaries that have been manually annotated with sentiment information. However, the effectiveness of this method is constrained by its reliance predefined sentiment lexicons that may not be comprehensive or applicable to the specific domain of the text being analyzed. An alternative approach is the utilization of machine learning algorithms in sentiment analysis. Several machine learning techniques, including support vector machines (SVMs) [29] and naive Bayes classifiers (NBs) [30], have been explored in the literature. Despite advancements in these methods, their performance is still heavily dependent on the amount and quality of the data used for training. In recent times, with the emergence of deep learning techniques, researchers have proposed the use of neural network-based models for sentiment analysis [31–33]. The most recent trend in the field involves the use of pre-trained language models for sentiment analysis. These models, trained on extensive corpora, have been shown to effectively learn generic language representations, thereby eliminating the need for training from scratch when tackling downstream NLP tasks. Hoang et al. [34] investigated the efficacy of using the BERT model, along with a fine-tuning approach that incorporates supplementary generated text, to mitigate the challenge of out-of-domain aspect-based sentiment analysis. A study by Singh et al. [35] used the BERT model on tweets to study the effects of coronavirus on social life.

Visual sentiment analysis is a rapidly growing field of research that aims to understand the emotions and attitudes conveyed by images. With the increasing popularity of social media platforms and the proliferation of visual content, there has been a growing need to develop techniques for automatically analyzing the sentiment conveyed by images.

The investigation of image sentiment analysis has undergone numerous developments over the years. One of the earliest methods of image sentiment analysis was rooted in the utilization of low- or mid-level image features to make inferences regarding the sentiment conveyed by the image. For example, Machajdik et al. [36] aimed to classify images by considering the psychological and aesthetic aspects of images. The study proposed a method for extracting low-level features such as composition, texture, and color from images and using them to predict the emotions conveyed by the image. Borth et al. [37] introduced SentiBank, a method that recognizes adjective-noun pairs (ANPs) from visual content, which can be considered as mid-level visual features that reflect semantic concepts. While these earliest approaches to image sentiment analysis have relied on manual feature extraction and rule-based algorithms, deep learning techniques have emerged as a promising alternative. For instance, You et al. [38] proposed a new method for visual sentiment analysis that leverages a progressive fine-tuning procedure in the context of transfer learning. Song et al. [39] presented a novel approach to visual sentiment analysis that involves the incorporation of an attention mechanism within a convolutional neural network (CNN) framework.

2.2 Multimodal Sentiment Analysis

The domain of multimodal sentiment analysis, with a particular emphasis on visual-textual sentiment analysis, has seen a significant increase in scholarly attention in recent years. This can be attributed to the widespread use of social media platforms that enable individuals to post multimedia content along with the written text. The visual and textual components often complement each other and supply additional contextual information, which is particularly significant in the realm of social media posts. Integrating visual and textual information in such contexts provides a more comprehensive understanding of the sentiment expressed in the media.

Early works in this domain adopted feature-based methods to integrate information from different modalities. For instance, Cao et al. [40] proposed a cross-media sentiment analysis approach that leveraged both textual and visual features. To extract textual features, they employed a sentiment dictionary, while for visual sentiment ontology from images, they relied on adjective-noun pairs (ANP). Despite their effectiveness, these methods require significant effort and time for feature engineering. With the advent of deep learning, neural network-based models have emerged, achieving notable progress in this field. Yu et al. [41] pre-trained convolutional neural networks (CNNs) on text and image data to extract feature representations. These multimodal features are then fused and used to train a logistic regression model for sentiment classification. Xu et al. [42] proposed an approach for sentiment analysis that incorporates a co-memory attentional mechanism. This mechanism enables the interactive modeling of the relationship between text and image data, thus enhancing the accuracy of the sentiment analysis results. Li et al. [43] proposed a label-based and data-based contrastive learning approach to capture the complementary information between modalities and a multi-layer fusion mechanism to integrate multimodal features effectively. Recent advancements in vision-language models have led to remarkable progress in multimodal tasks, such as Cheema et al. [44] applied CLIP to multimodal sentiment analysis, demonstrating its potential as a powerful baseline for sentiment prediction tasks in tweets. However, the utilization of vision-language pre-trained models in multimodal sentiment analysis remains under-explored in the literature.

Compared to prior unimodal or multimodal sentiment analysis methods, our approach has two distinctive features. Firstly, instead of relying on unimodal pre-trained models with limited sensitivity for sentiment detection, we adopt transfer learning by utilizing a vision-language pre-trained model to extract visual and textual representations in a unified framework. These representations are then integrated using a Transformer architecture to capture deep semantic information in image-text pairs. Secondly, our proposed multimodal contrastive learning method can expand the original sample size to obtain richer representation information. This approach allows us to effectively leverage multimodal data and extract more comprehensive features for sentiment analysis.

3 Methodology

In this section, we provide a brief overview of multimodal sentiment analysis and introduce the proposed VLMSA architecture, as shown in Fig. 2. Subsequently, we elaborate on the methodology we have introduced for multimodal sentiment classification. Our approach deviates from conventional methods by incorporating a vision-language pre-trained model and our proposed multimodal contrastive learning approach, aimed at improving the performance of sentiment analysis.

Task Formulation: Our goal in this research is to determine the sentiment polarity expressed in a multimodal post that consists of both an image and text. A multimodal sentiment sample is represented as $M = (I, T)$, where I denotes the visual modality, and T represents the textual modality. Our objective is to develop a mapping function that classifies M into one of the predefined categories y, which includes the sentiments of neutral, negative, or positive.

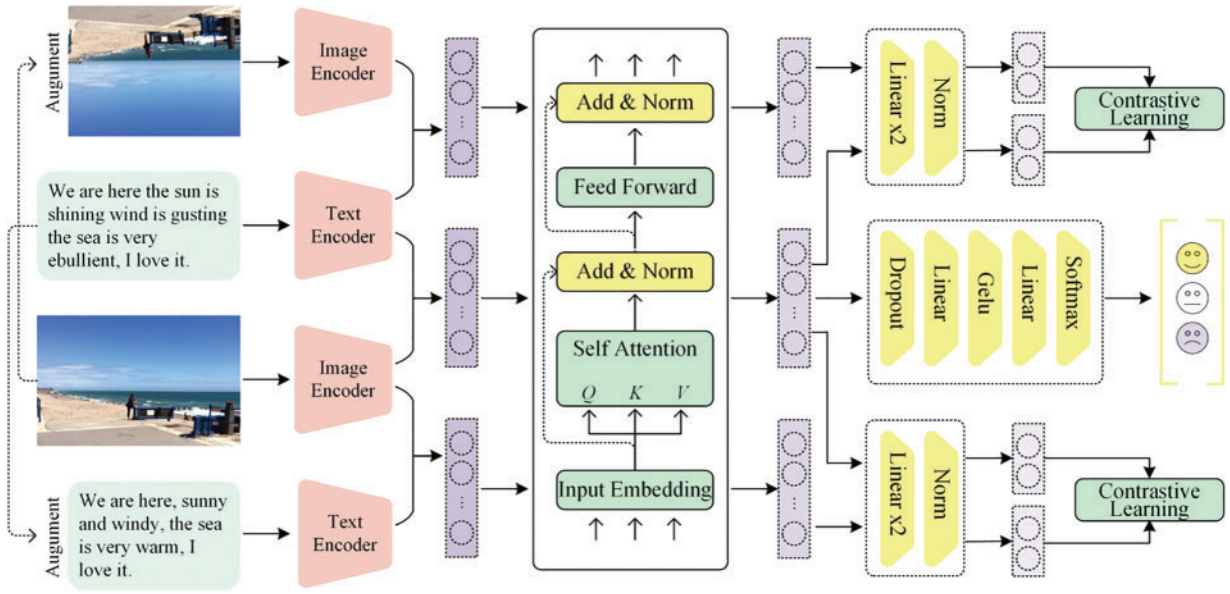


Figure 2: The overview of our proposed VLMSA architecture

3.1 Representations Extraction with the Vision-Language Pre-Trained Model

In this study, we adopted the CLIP architecture as a feature extractor for sentiment classification. CLIP encompasses a visual encoder for images and a textual encoder for text, allowing for effective extraction of meaningful representations from both modalities. We used the pre-trained weights from the CLIP model, keeping them frozen, to generate image and text representations as inputs for our sentiment classification model. This approach addresses limitations of traditional unimodal pre-trained models in capturing complex relationships between different modalities and improving feature extraction performance. The process can be succinctly expressed as follows:

$$I_O = CLIP_{img}\left(I_{original}\right), \tag{1}$$

$$T_O = CLIP_{txt}\left(T_{original}\right), \tag{2}$$

$$I_A = CLIP_{img}\left(I_{augument}\right), \tag{3}$$

$$T_A = CLIP_{txt}\left(T_{augument}\right), \tag{4}$$

where both $I_{original}$ and $T_{original}$ are original visual modality and textual modality, respectively; both $I_{augument}$ and $T_{augument}$ are generated through data augmentation techniques. The data augmentation methods employed include the back-translation strategy, as adopted in previous works [43,45] for text, and RandAugmentation [46] for images; $CLIP_{img}$ and $CLIP_{txt}$ refer to the visual and textual encoders of the CLIP model, respectively.

3.2 Multimodal Representations Learning

While it is possible to utilize extracted visual and text representations from CLIP for sentiment analysis by concatenating image and text features (as in [44]), we argue that direct fusion may not adequately capture modalities' relationship, potentially decreasing accuracy and reliability. Therefore, we propose a method that incorporates self-attention to model interdependence between image and text. Self-attention is a mechanism that facilitates the calculation of a weighted sum of values, where the weights are determined by mapping a query and its corresponding key to a set of key-value pairs. This mechanism enables us to effectively capture the relevance between the image and text data, thereby enhancing the performance of multimodal sentiment analysis.

$$Att\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{5}$$

where the query (Q), key (K), and value (V) are utilized in computing the attention scores, while the factor $\sqrt{d_k}$ is incorporated to alleviate the vanishing gradient issue that arises in the softmax function when the inner product becomes excessively large. Our methodology utilized a single-layer Transformer Encoder [47] to perform the self-attention mechanism. This selection was motivated by its ability to capture the interactions between different input sequences, making it suitable for our task of multimodal sentiment analysis.

$$F_u = Att\left(Concat\left(I_A, T_O\right)\right), \tag{6}$$

$$F_o = Att\left(Concat\left(I_O, T_O\right)\right), \tag{7}$$

$$F_v = Att\left(Concat\left(I_O, T_A\right)\right) \tag{8}$$

where F_u, F_o and F_v are the output from Transformer Encoder Layer computed through the application of a feed-forward layer to the $Att\ (Q,K,V)$ vectors. The Transformer Encoder layer enhances the capability of the model to selectively attend to salient information in both visual and textual modalities, thereby improving the quality of the representations generated.

3.3 Multimodal Contrastive Learning

Despite its effectiveness in training models with limited labeled data, the potential of contrastive learning as an approach in multimodal sentiment analysis has not been fully explored in previous research. In scenarios where there is a scarcity of labeled data, leveraging contrastive learning techniques holds promise in significantly improving the efficacy of sentiment analysis models. As such, exploring the implications of contrastive learning in multimodal sentiment analysis presents a promising opportunity for advancing the field and unlocking new possibilities.

In this research, we propose an innovative approach that augments original image-text pairs for each modality separately, generating two new input pairs. These pairs are then learned in contrast to the original image-text pair, with the two augmented pairs serving as positive examples and all other pairs in the batch as negative examples. Our approach aims to maximize the similarity between the original pair and its augmented pairs, while minimizing the similarity between the augmented pairs and all other pairs in the batch, as depicted in Fig. 3. This novel approach differs significantly from previous works in the field, such as [43], and holds the potential to advance multimodal contrastive learning. Specifically, our method eliminates the need for labeled data and expands the sample number based on the original image-text pair, making it a more efficient and scalable solution for multimodal sentiment analysis.

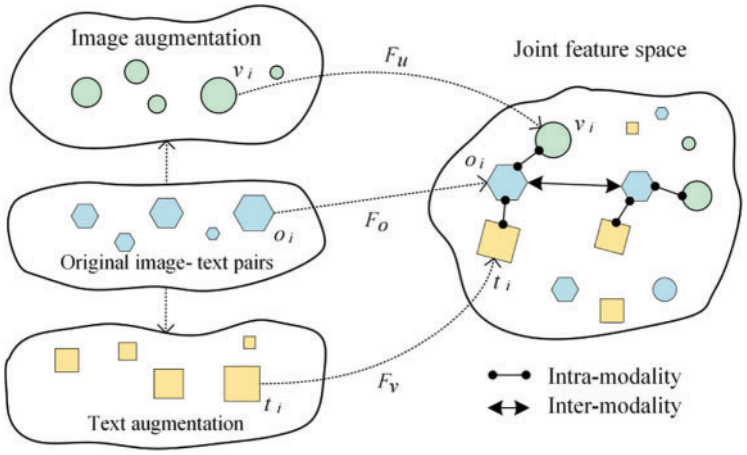


Figure 3: Multimodal contrastive learning aiming to maximize the similarity between the original pair and its augmented pairs, while minimizing the similarity between the augmented pairs and all other pairs

In order to effectively compare and learn from the different modalities, we utilize a two-layer perceptron (MLP) with ReLU activation to transform the modalities F_u, F_o and F_v into a 512-dimensional vector space, followed by the *L2*-Normalization, as illustrated in Eqs. (9)–(11). By implementing these preprocessing steps, we aim to produce robust and meaningful representations of the input data, which are less susceptible to variations.

$$u = Norm\left(MLP\left(F_u\right)\right), \tag{9}$$

$$o = Norm\left(MLP\left(F_o\right)\right), \tag{10}$$

$$v = Norm\left(MLP\left(F_v\right)\right), \tag{11}$$

In our proposed multimodal contrastive learning approach, we present a novel approach that leverages the power of contrastive learning. Specifically, as illustrated in Fig. 2, our approach comprises two distinct contrastive learning sections located in the upper and lower right corners, respectively. The aim of our training objective is to optimize two primary loss functions, namely the $u_i \rightarrow o_i$ contrastive loss and $v_i \rightarrow o_i$ contrastive loss, for each i-th pair within a minibatch of N input pairs from the training data:

$$\mathcal{L}_c = -\frac{1}{2}\log\frac{\exp\left(\langle u_i, o_i\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle u_i, o_j\rangle/\tau\right)} - \frac{1}{2}\log\frac{\exp\left(\langle v_i, o_i\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle v_i, o_j\rangle/\tau\right)}, \tag{12}$$

where $\langle u_i, o_i\rangle$ and $\langle v_i, o_i\rangle$ represent the cosine similarity, i.e., $\langle u, o\rangle = u^T o/||u||\,||o||$; and τ represents a temperature parameter. The $\mathcal{L}_c$ loss function is designed to optimize the dissimilarity between the representations of two new pairs and their corresponding original pairs within the same class, while concurrently maximizing the divergence between two new pairs and the original pairs from distinct classes. The implementation of this loss function enabled the model to acquire semantically significant representations, which facilitated a more precise prediction of sentiment.

3.4 Classification and Objective Function

The aim of multimodal sentiment analysis is to accurately classify the sentiment expressed in a given image-text pair, represented as *(I, T)*. To achieve this, we employ another MLP that comprises two fully connected layers designed to process the input representations F_o generated from the feature extraction steps outlined in Section 3.2. The first fully connected layer performs a linear transformation of the input representations, followed by the application of a non-linear activation function such as GeLu. After the initial fully connected layer, the output is then fed into another fully connected layer. Subsequently, the output of these layers is passed through a Softmax layer, which transforms the output of the previous layers into a probability distribution over the possible sentiment labels. The Softmax layer is defined by the equation:

$$y_{pre} = Softmax\left(MLP\left(F_o\right)\right). \tag{13}$$

Furthermore, the optimization of our model is achieved through the utilization of the Adam optimization algorithm, in combination with the minimization of cross-entropy loss:

$$\mathcal{L}_s = CrossEntropyLoss\left(y_{pre}, y\right), \tag{14}$$

where y denotes the ground truth sentiment label.

Finally, in order to effectively predict sentiment, we employ a joint optimization method that simultaneously optimizes the losses of the classification loss $\mathcal{L}_s$ and multimodal contrastive learning $\mathcal{L}_c$ to obtain correctly predicted sentiment, thereby ensuring that the model can acquire a diverse set of features and avert overfitting to a specific corpus of inputs. To train the model for multimodal sentiment analysis, we formulate the total loss as follows:

$$\mathcal{L}_{total} = \mathcal{L}_s + \mathcal{L}_c. \tag{15}$$

4 Experiment

This section presents the experimental evaluation of the proposed VLMSA model for multimodal sentiment analysis. Initially, we outline the intricate details of the data preparation, model implementation, and comparative methods employed in the experiment. Subsequently, we verify the efficacy of our approach by conducting evaluations on two publicly available datasets, accompanied by a comprehensive analysis of the results. Finally, we investigate supplementary assessments, such as attention visualization and ablation study, to gain additional insights into the model's workings.

4.1 Datasets

In this study, we have utilized two publicly available datasets, MVSA-Single and MVSA-Multi, provided by the Sentiment Analysis on Multi-view Social Data (MVSA) project, which was established by Niu et al. [48]. These datasets have been collected from the popular social media platform Twitter and include a total of 5129 and 196,000 image-text pairs, respectively. The MVSA project is a significant development in the field of multimodal sentiment analysis, as it provides a standardized benchmark that allows researchers to compare the performance of their models against that of other models, ensuring a fair and reliable evaluation. The datasets are labeled with sentiment polarity forms, including *positive*, *neutral*, and *negative*.

In order to ensure a fair comparison, we have adopted a consistent preprocessing strategy for the training, validation, and test sets. Furthermore, we have divided the datasets into three segments, following the methodology employed by prior studies [43], with a ratio of 8:1:1. These key

characteristics of the datasets have been summarized in Table 1, which is provided for reference in our experimental section.

Table 1: The statistics of the databases for MVSA-Single and MVSA-Multiple

Lable	MVSA-Single			MVSA-Multiple		
	Train	Validation	Test	Train	Validation	Test
Positive	2147	268	268	9056	1131	1131
Neutral	376	47	47	3528	440	440
Negative	1088	135	135	1040	129	129
Total	3611	450	450	13624	1700	1700

4.2 Experimental Settings

The experiments for the VLMSA model were conducted using a Tesla V100 Graphics Processing Unit (GPU), which provided high-performance computing power for training and testing the model. The substantial 32 GB of Random Access Memory (RAM) on this GPU allowed us to process large amounts of data efficiently and effectively. In our research on multimodal sentiment analysis, we leveraged the capabilities of Google Colaboratory, a cloud-based platform offering free access to powerful computing resources. With its support for the PyTorch framework, we could effortlessly implement our proposed model and take advantage of the platform's intuitive interface for handling data and code management.

In order to enhance the performance of the VLMSA model, we took a deliberate approach in selecting the hyper-parameters, considering the balance between computational efficiency and model performance. The resulting optimal values have been consolidated and presented in Table 2. Through the optimization of hyper-parameters, our VLMSA model was able to attain outstanding performance on the task of multimodal sentiment analysis.

Table 2: Settings of important parameters

Hyperparameters	MVSA-Single	MVSA-Multiple
Image size	224	224
Text length	77	77
Batch size	32	64
Learning rate	1e-4	5e-4
Maximum epochs	6	6
Pretrained CLIP model	ViT-B/32	ViT-B/32

4.3 Compared Methods

To further validate the efficacy of the proposed VLMSA model for multimodal sentiment analysis, we compare our approach to a variety of existing competitive methods.

SentiBank & SentiStrength [37] employs a mid-level representation of an image based on the extraction of 1200 adjective-noun pairs using SentiBank. Subsequently, the sentimentality of the resulting text is quantified utilizing the SentiStrength algorithm.

MultiSentiNet [49] identifies objects and scenes as semantic features of images. These semantic features are utilized to guide the learning of textual representations through attention mechanisms and are subsequently aggregated with textual features for sentiment analysis.

Co-Memory [42] takes into account the reciprocal relationship between visual content and textual words and models the dynamic interactions between these two modalities for multimodal sentiment analysis.

CLMLF [43] introduces a transformer-based fusion approach that employs contrastive learning. Two distinct forms of contrastive learning, label-based and data-based, are proposed as training strategies to facilitate the model's acquisition of salient features relevant to sentiment analysis.

Se-MLNN [44] explores the utilization of various sophisticated visual features in combination with a textual model for the purpose of conducting comprehensive multimodal sentiment analysis.

4.4 Results and Analysis

In our research, we compared the accuracy and weighted-F1 score of the proposed VLMSA approach against other compared methods in Table 3. Through our experimentation, we have drawn several significant conclusions: (1) the SentiBank & SentiStrength model demonstrated the lowest level of accuracy in determining the sentiment polarity of tweets. This model, which relied solely on traditional statistical features, was unable to effectively extract the crucial internal features of text and images; (2) the MultiSentiNet model was suboptimal compared to other models. Even though it considered the influence of visual data on text, it didn't fully consider the impact of textual data on images, resulting in a superficial connection between the two modalities; (3) the Co-Memory model surpasses MultiSentiNet in multimodal sentiment analysis by incorporating a fusion module that facilitates mutual influence between modalities; (4) the CLMLF model not only addresses the relationship between images and text but also leverages contrast learning to enhance the representational learning capability, thereby improving the accuracy of sentiment analysis. These findings suggest the potential of contrastive learning in this domain; (5) the Se-MLNN, which incorporates CLIP for feature extraction and fusion, is a straightforward and efficient methodology for sentiment analysis. Nevertheless, the direct feature fusion approach was inadequate in the MVSA-Multiple dataset, emphasizing the need for a deeper understanding of the complex relationships between images and text; and (6) the VLMSA model leverages vision-language pre-trained models with contrastive learning techniques to enhance multimodal sentiment analysis accuracy. Our findings demonstrate that this approach yields a significant improvement in the accuracy of sentiment analysis tasks. These results are promising and represent an important step forward in this field of research.

Table 3: Performance comparison of MVSA-Single and MVSA-Multiple. Results denoted with † were obtained from Cheema et al. [44], while the remaining results were obtained from their respective sources

Methods	MVSA-Single		MVSA-Multiple	
	Accuracy	F1	Accuracy	F1
SentiBank & SentiStrength†	52.05	50.08	65.62	55.36
MultiSentiNet	63.27	59.12	63.08	59.12
Co-Memory	70.51	70.01	69.92	69.83
CLMLF	75.33	73.46	72.00	**69.83**
Se-MLNN†	75.33	73.76	66.35	61.89
VLMSA (Ours)	**76.44**	**75.27**	**72.29**	67.02

4.5 In-Depth Analysis

Attention Visualization. Our research paper delves into the examination of the impact of incorporating a vision-language pre-trained model in a multimodal sentiment analysis task. Our inquiry employs the use of an attention mechanism to highlight the CLIP model's capacity to extract semantically relevant data from image-text pairs. The results, as depicted in Fig. 4, reveal that for a given textual input, the pre-trained model can identify the relevant regions in the image and assign higher attention weights to them, resulting in improved text-image feature fusion. As an illustration, Fig. 4b exemplifies the successful alignment of the term 'amazing' in the text and the 'girl's face' in the image, demonstrating the vision-language pre-trained model's capacity to perform semantic alignment with great accuracy. Given the robust cross-modal alignment capabilities of the vision-language pre-trained model, we have extended its sentiment analysis capabilities by integrating a Transformer Encoder Layer. This integration enables the investigation of the interdependencies between text and images, as discussed in Section 3.2.

Ablation Study. We conducted ablation experiments to assess the impact of the CLIP model, the Transformer Encoder Layer, and Multimodal Contrastive Learning on the performance of sentiment analysis. Table 4 presents the findings of our ablation study, which aimed to enhance our comprehension of the individual impact of each component on the overall effectiveness of our multimodal sentiment analysis system. This information was essential in guiding the development of the model and informing future research directions.

Our analysis revealed several key insights: (1) the evaluation of the VLMSA model across two datasets showed that the optimal performance was achieved with all modules intact. The removal of any single component resulted in a noticeable decline in accuracy and F1 score, thus confirming the interdependence of the modules and their crucial role in driving optimal performance. The loss of a critical component from the system, due to the removal of a single module, would inevitably lead to a degradation in the overall performance of the model; (2) the incorporation of the Transformer Encoder Layer in multimodal data fusion was found to enhance performance significantly. The results demonstrate that the Transformer Encoder Layer has the ability to effectively capture complex dependencies between different modalities, thus yielding improved performance compared to traditional methods; (3) the integration of multimodal contrastive learning has demonstrated its

potential in further improving the performance of sentiment analysis models. This approach enables the model to learn the common features indicative of sentiment and differentiate between instances of differing sentiment through the application of contrastive learning. As a result, the model possesses a strengthened understanding of the relationships between visual and textual features, leading to improved accuracy in the sentiment classification of multimodal data.

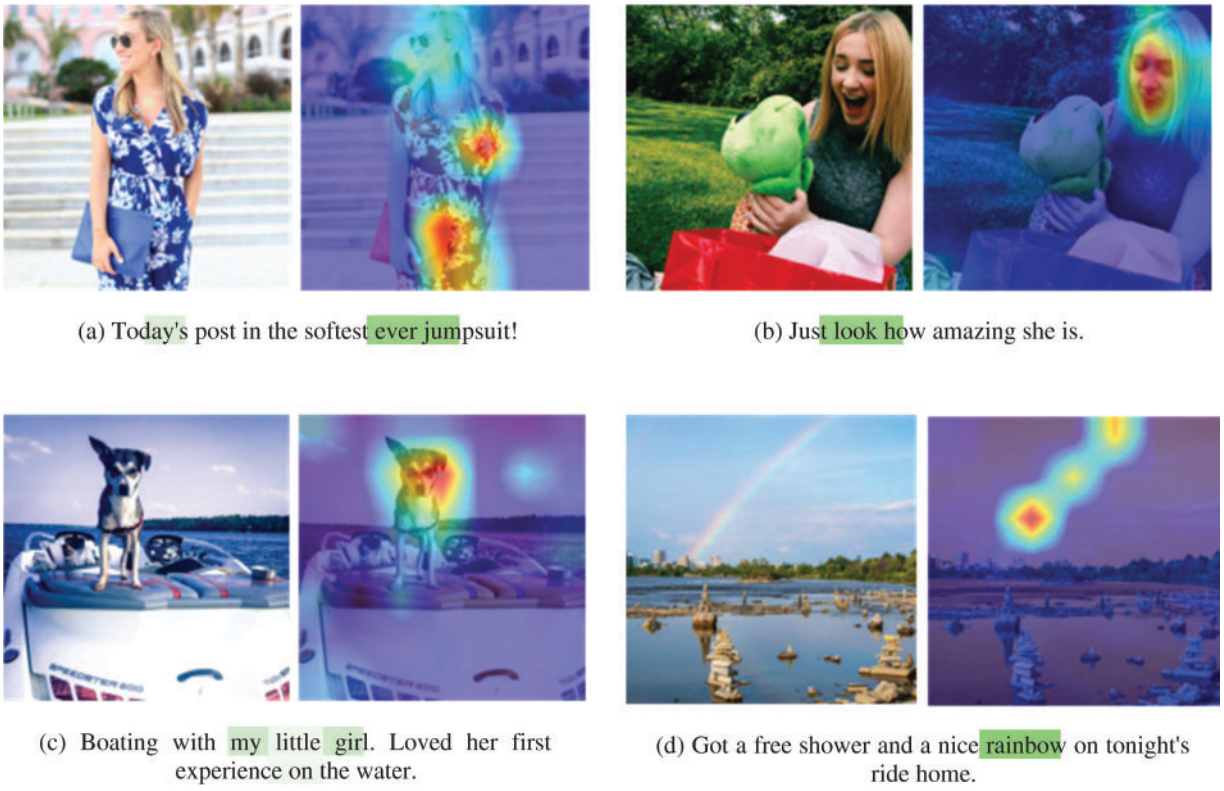


Figure 4: Attention visualization of some image-text pairs. Given a textual input, the model can identify the relevant regions in the image and assign higher attention weights to them

Table 4: Ablation study results on two MVSA datasets

Methods	MVSA-Single		MVSA-Multiple	
	Accuracy	F1	Accuracy	F1
Pre-trained CLIP model	72.67	70.73	70.94	65.51
+ Transformer Encoder Layer	75.33	73.92	71.12	63.64
+ Multimodal contrastive learning	76.44	75.27	72.29	67.02

5 Conclusion

Performing multimodal sentiment analysis on social media is a challenging task due to the varied and complex nature of user-generated content. While previous research in this field has mainly focused

on using unimodal pre-trained models to extract features from either visual or textual modalities, our approach differs in its utilization of vision-language pre-trained models and our proposed multimodal contrastive learning approach to effectively integrate information gained from different modalities. The efficacy of the proposed methodology has been assessed on two publicly available datasets. The experimental outcomes manifest its potency and supremacy in comparison to other contemporary techniques. To the best of our knowledge, this is the first study to integrate vision-language pre-trained models with contrastive learning for multimodal sentiment analysis, and we expect that our findings will inspire further research in this field.

Despite the promising performance, our proposed approach still has several limitations. One limitation is that it relies on the quality of vision-language pre-trained models. Another limitation of our proposed approach is that it may not perform well on tasks involving fine-grained sentiment analysis, such as determining the sentiment of a specific phrase or aspect within a text. As part of our ongoing research efforts, we plan to undertake a deeper analysis of the interplay between image and text modalities in sentiment analysis. Additionally, we aim to devise a robust and advanced fusion technique to augment our current methodology.

Funding Statement: This research project was supported by Science and Technology Research Project of Jiangxi Education Department. Project Grant No. GJJ2203306.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] D. S. Khafaga, M. Auvdaiappan, K. Deepa, M. Abouhawwash and F. K. Karim, "Deep learning for depression detection using Twitter data," *Intelligent Automation & Soft Computing*, vol. 36, no. 2, pp. 1301–1313, 2023.

[3] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," *Journal of Information and Telecommunication*, vol. 5, no. 1, pp. 1–15, 2021.

[4] S. Saranya and G. Usha, "A machine learning-based technique with intelligent wordnet lemmatize for Twitter sentiment analysis," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 339–352, 2023.

[5] S. Paliwal, A. Kumar Mishra, R. Krishn Mishra, N. Nawaz and M. Senthilkumar, "Xgbrs framework integrated with word2vec sentiment analysis for augmented drug recommendation," *Computers, Materials & Continua*, vol. 72, no. 3, pp. 5345–5362, 2022.

[6] V. Arya, A. K. M. Mishra and A. González-Briones, "Analysis of sentiments on the onset of Covid-19 using machine learning techniques," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 11, no. 1, pp. 45–63, 2022.

[7] H. Khan, A. Srivastav and A. Kumar Mishra, "Multiclass intent analysis: Beyond the conventional polarities," *ECS Transactions*, vol. 107, no. 1, pp. 7119, 2022.

[8] R. Kaur and S. Kautish, "Multimodal sentiment analysis: A survey and comparison," *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, vol. 10, no. 2, pp. 38–58, 2019.

[9] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh *et al.,* "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *2017 IEEE Int. Conf. on Data Mining (ICDM)*, New Orleans, LA, USA, pp. 1033–1038, 2017.

[10] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-Based Systems*, vol. 161, pp. 124–133, 2018.

[11] X. C. Ju, D. Zhang, R. Xiao, J. H. Li, S. S. Li *et al.,* "Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection," in *Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, pp. 4395–4405, 2021.
[12] X. Yan, H. Xue, S. Jiang and Z. Liu, "Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 2000688, 2022.
[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int. Conf. on Learning Representations, ICLR 2015*, San Diego, CA, USA, Conference Track Proceedings, 2015.
[14] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, pp. 770–778, 2016.
[15] J. Devlin, M. -W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, (Long and Short Papers), vol. 1, pp. 4171–4186, 2019.
[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei *et al.,* "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, pp. 9, 2019.
[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh *et al.,* "Learning transferable visual models from natural language supervision," in *Proc. of the 38th Int. Conf. on Machine Learning, ICML 2021*, Virtual Event, vol. 139, pp. 8748–8763, 2021.
[18] W. J. Su, X. Z. Zhu, Y. Cao, B. Li, L. Lu *et al.,* "VL-BERT: Pre-training of generic visual-linguistic representations," in *8th Int. Conf. on Learning Representations*, Addis Ababa, Ethiopia, 2020.
[19] T. Chen, S. Kornblith, M. Norouzi and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of the 37th Int. Conf. on Machine Learning, ICML 2020*, Virtual Event, vol. 119, pp. 1597–1607, 2020.
[20] K. He, H. Fan, Y. Wu, S. Xie and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR 2020*, Seattle, WA, USA, pp. 9726–9735, 2020.
[21] T. Gao, X. Yao and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing*, Virtual Event/Punta Cana, Dominican Republic, pp. 6894–6910, 2021.
[22] H. Zeng and X. Cui, "SimCLRT: A simple framework for contrastive learning of rumor tracking," *Engineering Applications of Artificial Intelligence*, vol. 110, pp. 104757, 2022.
[23] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang *et al.,* "Multimodal contrastive training for visual representation learning," in *IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2021*, virtual, pp. 6995–7004, 2021.
[24] R. Jain, A. Kumar, A. Nayyar, K. Dewan, R. Garg *et al.,* "Explaining sentiment analysis results on social media texts through visualization," *Multimedia Tools and Applications*, vol. 82, pp. 1–17, 2023.
[25] J. Y. -L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong and W. K. Cheng, "State of the art: A review of sentiment analysis based on sequential transfer learning," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 749–780, 2023.
[26] J. Mutinda, W. Mwangi and G. Okeyo, "Sentiment analysis of text reviews using lexicon-enhanced bert embedding (LeBERT) model with convolutional neural network," *Applied Sciences*, vol. 13, no. 3, pp. 1445, 2023.
[27] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. -F. Chang *et al.,* "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.
[28] M. Taboada, J. Brooke, M. Tofiloski, K. D. Voll and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.
[29] S. Naz, A. Sharan and N. Malik, "Sentiment classification on twitter data using support vector machine," in *2018 IEEE/WIC/ACM Int. Conf. on Web Intelligence*, Santiago, Chile, pp. 676–679, 2018.

[30] J. Song, K. T. Kim, B. J. Lee, S. -Y. Kim and H. Y. Youn, "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis," *KSII Transactions on Internet and Information Systems*, vol. 11, no. 6, pp. 2996–3011, 2017.
[31] J. Zhao, X. Gui and X. Zhang, "Deep convolution neural networks for Twitter sentiment analysis," *IEEE Access*, vol. 6, pp. 23253–23260, 2018.
[32] M. Umer, I. Ashraf, A. Mehmood, S. Kumari, S. Ullah *et al.,* "Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model," *Computer Intelligence*, vol. 37, no. 1, pp. 409–434, 2021.
[33] R. Ni, X. Liu, Y. Chen, X. Zhou, H. Cai *et al.,* "Negative emotions sensitive humanoid robot with attention-enhanced facial expression recognition network," *Intelligent Automation & Soft Computing*, vol. 34, no. 1, pp. 149–164, 2022.
[34] M. Hoang, O. A. Bihorac and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proc. of the 22nd Nordic Conf. on Computational Linguistics*, Turku, Finland, pp. 187–196, 2019.
[35] M. Singh, A. K. Jakhar and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 33, 2021.
[36] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. of the 18th Int. Conf. on Multimedia*, Firenze, Italy, pp. 83–92, 2010.
[37] D. Borth, R. Ji, T. Chen, T. M. Breuel and S. F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM Multimedia Conf., MM '13*, Barcelona, Spain, pp. 223–232, 2013.
[38] Q. You, J. Luo, H. Jin and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. of the Twenty-Ninth AAAI Conf. on Artificial Intelligence*, Austin, Texas, USA, pp. 381–388, 2015.
[39] K. Song, T. Yao, Q. Ling and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, 2018.
[40] D. Cao, R. Ji, D. Lin and S. Li, "A cross-media public sentiment analysis system for microblog," *Multimedia Systems*, vol. 22, pp. 479–486, 2016.
[41] Y. Yu, H. Lin, J. Meng and Z. Zhao, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks," *Algorithms*, vol. 9, no. 2, pp. 41, 2016.
[42] N. Xu, W. Mao and G. Chen, "A co-memory network for multimodal sentiment analysis," in *The 41st Int. ACM SIGIR Conf. on Research & Development in Information Retrieval, SIGIR 2018*, Ann Arbor, MI, USA, pp. 929–932, 2018.
[43] Z. Li, B. Xu, C. Zhu and T. Zhao, "CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States, pp. 2282–2294, 2022.
[44] G. S. Cheema, S. Hakimov, E. Müller-Budack and R. Ewerth, "A fair and comprehensive comparison of multimodal tweet sentiment analysis methods," in *Proc. of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding*, Taipei, Taiwan, pp. 37–45, 2021.
[45] Q. Xie, Z. Dai, E. H. Hovy, T. Luong and Q. Le, "Unsupervised data augmentation for consistency training," in *Advances in Neural Information Processing Systems 33: Annual Conf. on Neural Information Processing Systems*, Virtual Event, 2020.
[46] E. D. Cubuk, B. Zoph, J. Shlens and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR Workshops 2020*, Seattle, WA, USA, pp. 3008–3017, 2020.
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conf. on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 5998–6008, 2017.

[48] T. Niu, S. Zhu, L. Pang and A. El-Saddik, "Sentiment analysis on multi-view social data," in *MultiMedia Modeling–22nd Int. Conf.*, Miami, FL, USA, Proceedings, Part II, vol. 9517, pp. 15–27, 2016.
[49] N. Xu and W. Mao, "Multisentinet: A deep semantic network for multimodal sentiment analysis," in *Proc. of the 2017 ACM on Conf. on Information and Knowledge Management, CIKM 2017*, Singapore, pp. 2399–2402, 2017.