Journal of Optimization of Soft Computing (JOSC)

Vol. 2, Issue 3, pp: (56-68), Autumn-2024
Journal homepage: https://sanad.iau.ir/journal/josc

Paper Type (Research paper)

Transformer-based Meme-sensitive Cross-modal Sentiment Analysis Using Visual-Textual Data in Social Media

Zahra Pakdaman [1], Abbas Koochari[1*] and Arash Sharifi[1]

1. Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.

Article Info

Article History:
Received: 2024/10/31
Revised: 2024/12/01
Accepted: 2024/12/08

DOI:

Keywords:
Visual Sentiment Analysis, Textual Sentiment Analysis, Vision Transformer, LDA, SBERT Bi-encoder

**Corresponding Author's Email Address: koochari@srbiau.ac.ir*

Abstract

Analyzing the sentiment of the social media data plays a crucial role in understanding users' intentions, opinions, and behaviors. Given the extensive diversity of published content (i.e., image, text, audio, and video) leveraging this variety can significantly enhance the accuracy of sentiment analysis models. This study introduces a novel meme-sensitive cross-modal architecture designed to analyze users' emotions by integrating visual and textual data. The proposed approach distinguishes itself by its capability to identify memes within image datasets, an essential step in recognizing context-rich and sentiment-driven visual content. The research methodology involves detecting memes and separating them from standard images. Form memes, embedded text is extracted and combined with user-generated captions, forming a unified textual input. Advanced feature extraction techniques are then applied: Vision Transformer (ViT) is employed for extracting visual features, while SBERT Bi-encoder is utilized to obtain meaningful textual embeddings. To address the challenges posed by high-dimensional data, Linear Discriminant Analysis (LDA) is used to reduce feature dimensionality while preserving critical classification information. A carefully designed neural network, consisting of two fully connected layers, processes the fused feature vector to predict sentiment classes. Experimental evaluation demonstrates the efficiency of the proposed method, achieving up to 90% accuracy on the MVSA-Single dataset and 80% accuracy on the MVSA-Multiple dataset. These results underscore the model's ability to outperform existing state-of-the-art approaches in cross-modal sentiment analysis. This study highlights the importance of integrating meme recognition and multi-modal feature extraction for improving sentiment analysis, paving the way for feature research in this domain.

1. Introduction

Sentiment analysis is a method to extract the user's real feelings from the data (text, image, video, and audio) that he/she publishes on the web platform. Considering that until the last few years, textual content was the most popular and most frequently published content on websites and social media, textual data attracted the attention of researchers. As a result, several methods were presented in text sentiment analysis, and the methods proposed in this field have become very rich and fruitful.

In recent years, when image-oriented platforms such as Facebook, Instagram, etc., came to light, a large amount of image data was published by users. In this respect, the importance of analyzing the sentiment of this type of data was

also clarified for researchers. Therefore, with the appearance of convolutional neural networks (CNN), it becomes possible to provide more accurate methods of visual sentiment analysis. Yet, visual sentiment analysis faces certain difficulties. For instance, individuals often employ memes on platforms like Twitter to convey concepts. The utilization of this content has experienced a notable rise in recent years [2]. Many memes possess an ironic quality that cannot be accurately and fully captured by computer vision methods (Samples of the ironic meme are shown in Figure 1.). To address this issue, the article employs meme detection. Consequently, both the image and the embedded text undergo analysis, along with the user's caption. Moreover, sometimes the published text is not complete, expressive, and clear (i.e., the text only includes hashtags) to be able to understand the user's real sentiment. Therefore, to provide a more accurate analysis, cross-modal approaches – methods that use both text and image- are used to analyze the sentiment of social media content.

Social media images may include a wide range of topics (more than just a picture of a face or nature) which is a substantial challenge in visual sentiment analysis. As an efficient solution, the analysis of the published text along with the image will help to provide a more accurate analysis of the user's sentiment. This article presents a cross-modal approach that analyzes social media data with a wide range of topics. Also, as stated above, most of the bi-modal or cross-modal schemes -that use image and text for sentiment analysis- analyze different content in parallel; thus, the complexity of the architecture increases. However, the proposed method analyzes the content in an integrated way (combining the features from the beginning and obtaining the result of the sentiment analysis of those features in an integrated manner). Therefore, it will significantly reduce the dimension of the feature vector and the complexity of the architecture while accuracy will impressively increase.

In this article, to assess the sentiment of the image with the assistance of auxiliary data, the initial step involves determining whether the input image is a regular image or a meme. In memes, the embedded text is extracted and concatenated to the published text. Subsequently, for all input data, both image and text features are extracted. then, the dimension reduction method and the resultant vectors are concatenated. Ultimately, after passing the vector through the fully connected layers, the final sentiment label is predicted. The main contributions of the paper are highlighted as follows:

For more precise analysis, a meme detector is employed to differentiate memes from other content. Hence, the embedded text in this content, which plays a crucial role in comprehending the genuine semantic and sentiment, will be extracted.

In this article, a transformer-based method is proposed. More accurate sentiment analysis is achieved thanks to the multi-head attention architecture used in the structure of transformers. This architecture makes it possible to capture both low-level features and high-level semantic information. Therefore, using transformers in both the image and text sides will help to solve the sentence ambiguity (like texts that only contain hashtags) and grasp the sense of irony or sarcasm of the image or text to predict the final sentiment.

Given to integrated architecture of the proposed method, the visual and textual features are combined from scratch, followed by a unified sentiment analysis operation on the concatenated feature vector.

In order to reduce the dimensionality of the extracted features, a dimension reducer (LDA) has been used. It improves the performance by better separation of sentiment categories.

In section 0 some sentiment analysis algorithms are reviewed. The proposed method is explained in section **Error! Reference source not found.**. In section 0, the results of the proposed method are reported compared to state-of-the-art methods in this field. Finally, the conclusion of this paper will be presented in section 0.

2. Related works

Please be sure your sentences are complete and that there is continuity within your paragraphs. Check the numbering of your graphics and make sure that all proper references are included.

Various categories are presented for different types of content sentiment analysis schemes. One of the introduced taxonomies is based on the number of contents used in the analysis process. Based on this classification criteria, existing methods are divided into unimodal, bi-modal, and multi-modal. In unimodal methods, only one content is analyzed for sentiment analysis. In the bi-modal (cross-modal), two types of content are used, and in multi-modal ones, more than two contents are used to estimate the final sentiment. Since a cross-modal approach is presented in this article, cross and multi-modal methods will be reviewed in this section.

Figure 1 Examples of ironic memes. (a) and (b) embedded text changes the sentiment of images. (c) the image has both positive and negative content.

2.1. Cross-modal approaches

The published data in social media includes a wide range of topics, including person's faces, landscapes, memes, abstract images, etc., which turns "analyzing their sentiments" into a challenge. Moreover, the presence of sarcasm and irony in viral images and memes on social media can create another challenge in analyzing accurate sentiment. One of the ways to improve the results of visual sentiment analysis is to use the auxiliary data (often text) published along with images in social media. The use of auxiliary data can provide a more comprehensive result regarding the genuine sentiment of the published data. These methods are called bi-modal (cross-modal).

Zhang et al. [2] presented a multimodal method in which CNN and GloVe are used to extract image and text features, respectively. The extracted features are given to IDLSTM as an image-text pair to finally obtain the result of sentiment analysis. Lu et al. [4] used BERT to process text and image in parallel. Since BERT was proposed for text analysis, in their article, by applying changes to this network, it is used for text and image analysis simultaneously. They used two BERTs in parallel. BERTs process regions of image and pieces of text. The difference between the BERT structure in this paper and the original BERT is that in their paper, a co-attention transformer layer is used to cross-feed image regions and text pieces to the Visual and linguistic transformer. Xu et al. [5] believe that the relationship between the contents of image and text is a bidirectional one. For this reason, they used text and image together. In this way, they can report more accurate results for sentiment analysis. The authors proposed a model called Bi-directional Multi-level Attention. In the architecture of this method, two attention networks have been used in parallel form. The complex relationship between the image and its corresponding text – which includes a description of that image- will be analyzed. At last, the final label will be predicted based on the obtained information from the image and text. In the proposed method by Huang et al. [6] sentiment analysis will be done through three parallel axes. In their approach, the sentiment labels of the text, image, and text-image are predicted separately. Eventually, the final label will be obtained in a late fusion way. Li et al. [7] extracted features of the image and translated the image into text using LSTM. Then, using the SentiBank dataset, a number of Adjective-Noun Pairs (ANP) will be generated for each image and the generated caption will be combined with ANPs. Finally, the sentiment of the obtained text will be predicted by a bi-directional LSTM. The proposed method by Serra et al. [8] used a teacher-student architecture which will bring about the textual data sentiment using a pre-trained BERT model. In the student section, ViT is used to obtain the sentiment label of the image data. Eventually, by setting a threshold for the loss function, the final sentiment label will be obtained. Zhu et al. [9] have focused on extracting relevant regions of text and image. First, a pre-trained Faster R-CNN is used to recognize image regions and their corresponding representation. In order to embed each word in a vector and summarize the text information in the sentence, a pre-trained BERT and a bi-directional GRU are used, respectively. Also, to extract information from the image and relevant features in both text and image, pre-trained ResNet18 and cross-modal attention are employed, respectively. Finally, the final sentiment is predicted by concatenating the features and passing through Softmax. Yadav et al. [10] extract image features by two types of attentions (channel and spatial

attentions) and text features by LSTM and GRU. Then, the final label is estimated using the attention mechanism. In [11], the visual and textual features are extracted by pre-trained BERT. Then, the sentiment of each data is obtained by two contrastive learning structures (label and data-based contrastive learning). The proposed scheme by Peng et al. [12] consists of three modules. The first module is the feature extraction module, which extracts text and image features. The second is the feature attention module, which is responsible for learning text and image attention features. The third is the cross-modal hierarchical fusion module, which combines the features of within and between modals. In [13], Xu and Mao used two pre-trained Object-VGG and Scene-VGG networks to extract object and scene features of the image, respectively. To extract text features, they used LSTM. At last, the sentiment label is predicted by passing the extracted features through fusion and Softmax layers.

2.2. Multi-modal approaches

Most multimodal sentiment analysis methods deal with datasets of speech videos of different people. Therefore, unlike the online published content, which covers a wide range of topics, they have a limited scope. In these methods, the person's speech text, audio, and video have been used to provide the final sentiment label. In such datasets, real emotion can be easily obtained by analyzing the tone of voice, laughter, crying, facial expression, etc.

Wen et al. [14] presented a method in which text, audio, and video features are first extracted by GloVe, COVAREP, and facet respectively. In order to modify the features, attention structure is used. After merging and modifying the extracted features two by two, the final label will be predicted using an attention transformer and a fully connected layer. This scheme has only been tested to analyze the sentiment of people's speech. Yang et al. [15] proposed a method that tried to analyze the sentiment of the text more precisely by using audio and video. In their method, using GRU, audio and video are translated into text and analyzed together with the existing text. In fact, like using Seq2Seq methods – which consist of an encoder and a decoder part- one language can be translated to another; these methods can be used to translate audio or video into text. This approach is also limited to the sentiment analysis of people's speech. The authors used an encoder-decoder architecture to design their proposed model. The proposed method by Rahman et al. [16] is a BERT-based [17] method, which processes text, audio, and video altogether through a mechanism called Multimodal Adaptation Gate (MAG) among BERT layers. This approach used two BERT architectures in a row to analyze people's feelings during the speech. Hazarika et al.' method [18] used text and audio as auxiliary content for more detailed speech video analysis. To increase the accuracy of the analysis, two categories of features are extracted; common and specific ones. By defining and extracting these two categories of features, the model will be more comprehensive because both specific and common features of different contents are considered. They used a pre-trained BERT model to extract the text feature and an sLSTM to extract the audio and video features. Mittal et al. [19] presented another multimodal method for speech sentiment analysis, the highlight of which is reducing the impact of noisy data and being able to make decision in the presence of noise. In this method, audio, video, and text features are extracted. If there is noise in the extracted features, the features are changed to noise-free data. Also, effective features are separated from ineffective ones so that effective ones can be used. The audio, video, and text features are individually fed into two-layer networks. Then the output of these networks is given to an attention layer. Finally, after several fully connected layers, the result of sentiment analysis will be predicted.

3. Proposed method

As Figure 2 shows, to enhance the accuracy of predicting the sentiment class of the image, a cross-modal approach is introduced, incorporating the published text alongside the image. initially, a binary classifier is applied to identify memes. Memes need to be preprocessed before entering the proposed architecture. Then, visual and textual features are extracted using the transformers. To reduce dimensionality and enhance classification, the dimensions of the features are reduced by Linear Discriminant Analysis (LDA). Ultimately, the reduced feature vectors are concatenated, and by passing through two fully connected layers, their sentiment label will be estimated. In the following, the steps of the proposed method are explained in detail.

3.1. Meme detection

Since both the image and embedded text play crucial roles in the meme's sentiment label prediction, and a substantial portion of published images on social media constitute memes, this

article includes a preprocessed section dedicated to meme detection and extracting embedded text.

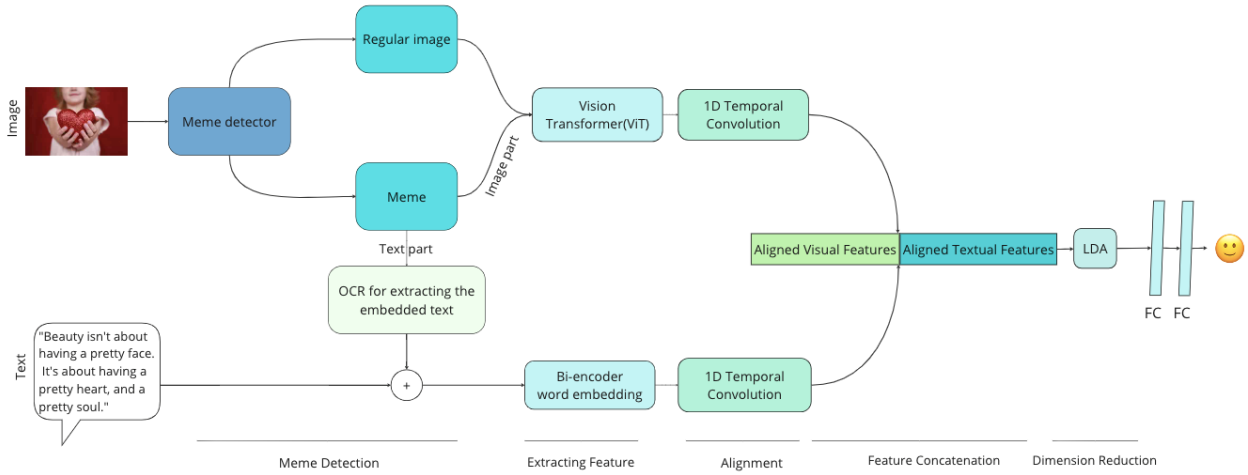


Figure 2 The proposed method flowchart, the final label is predicted after extracting features, reducing dimensionality, and passing through two fully connected layers.

To differentiate between a meme and a regular image, a binary classifier is employed. This classifier utilizes a pre-trained Xception as a base model, followed by three fully connected layers with 128, 64, and 32 hidden units to classify an input image into either a meme or regular class. The network is trained using a meme dataset introduced in section 0.

As depicted in Figure 2, if an image is identified as a meme, the embedded text is retrieved using Optical Character Recognition (OCR). The retrieved text part is then concatenated to the user's caption, while the image part is utilized to extract features. The recognition and extraction of embedded text in a meme are performed using EasyOCR module, an open-source tool known for its speed and efficiency in word recognition within images, along with its adaptability in handling various typeface and text layouts.

3.2. Feature extraction and alignment

In the second step, the text and image features are calculated using the transformer. Visual Transformer (ViT) and SBERT Bi-encoder are used respectively to extract image and text features. Then, in order to align these features, 1D temporal convolution will be applied to the text and image features separately. How to extract features using ViT and Bi-encoder are explained below.

3.2.1 Image feature extraction by ViT

The main idea of the transformer is to use the attention architecture to process the data. The use of transformers has found a special place in natural language processing problems, and they are used in almost all proposed methods. But Dosovitsky et al. [20] showed that this level of dependence on CNNs is not necessary in image processing applications and decided to use transformer in image processing areas.

The vision transformer learns a hierarchical representation of the image by stacking many layers of patch embeddings, multi-head self-attention, and feedforward networks. This capability makes it possible to capture both low-level features and high-level semantic information [21]. Using this method, tasks such as image classification can be performed well with only one transformer applied directly to a sequence of image patches. As Figure 3 shows, the overall architecture of this method is summarized in the following steps:

Step 1: The input image (x with resolution (H, W)) is divided into non-overlapping patches (x_p with resolution (P, P)). The number of patches is $N = HW/P^2$.

Step 2: The patches are flattened.

Step 3: From flattened patches, linear projection with lower dimensions is produced.

Step 4: Position embedding is added to each projected patch. Like *[class]* token of BERT, a learnable embedding is prepended to the sequence of patches ($z_0^0 = x_{class}$).

Step 5: The generated sequence is given as input to a Transformer encoder.

In the encoder, after performing normalization (Layer Norm (LN)) on the embedded patches (flattened patches with position embedding), the output is given to the multi-head self-attention unit (MSA). In this unit, each patch produces three vectors. The first vector is Query, which is based on the patch's meaning. Each patch creates this vector to announce its requirements to other

patches or to the model. The next vector is Key. Through this vector, each patch tells other patches what information it brings. The Value vector is the headline of a document, and based on this headline, it is decided how much information can be obtained from a document. After obtaining the Key, Value, and Query vectors, embedded patches are added to the result of MSA. This operation will be done in order to prevent the creation gradient vanishing problem and re-adding position embedding information after applying attention. Then, normalization will be done again, and the result will be passed through a Multi-Layer Perceptron (MLP). Finally, the MLP result will be added to the previous addition result. Passing through the Transformer encoder layer will be done L times. The operation of the Transformer encoder will be formulated as follows:

$$z_0 = [x_{class};\ x_p^1 E;\ x_p^2 E;\ \dots;\ x_p^N E] + E_{pos} \quad (1)$$

$$z_l' = MSA(LN(z_{l-1})) + z_{l-1}, \quad l = 1, \dots, L \quad (2)$$

$$z_l = MLP(LN(z_l')) + z_l', \quad l = 1, \dots, L \quad (3)$$

$$y = LN(z_L^0) \quad (4)$$

Where z_L^0 is the output of the Transformer encoder and *y* is the image representation.

Step 6: In this step, the features can be extracted or the output class can be estimated through MLP.

3.2.2. Sentence embedding by Bi-encoder

For generating sentence embedding of the published text with the image, a Sentence BERT/RoBERTa (SBERT) architecture, which is called Bi-encoder, is used. This sentence embedding method, as a transformer-based method, is much more efficient and faster than other similar ones [22]. Sentence embedding has several advantages over other common NLP tasks [23]:

The firstly mentioned (sentence embedding) is more comprehensive compared to other common NLP tasks, thanks to examining the whole sentences rather than word level feature extraction (common in other NLP approaches).

Since sentence embedding captures the meaning and context of a sentence in a numerical form, it can easily be used as an input to models and helps to understand the content of the text. Thus, performance will be improved.

Sentence embedding represents a sentence or a phrase as a single vector with a fixed length. So, the dimensionality of data will be reduced and it will be easier to work with.

The sentence embedding is resistant to the presence of noise and variability (such as spelling errors and variation in language usage). This property makes it suitable for some tasks like test classification and sentiment analysis, where the input sentences may not always be clear and uniform.

As Figure 4 shows, first, the input sentence (s_i) is tokenized: $T = [T_1, \dots, T_n]$. Tokens are fed to a BERT language model to obtain fixed-size word embeddings. As stated in the previous section, the attention mechanism has three qualities: Value (*V*), Key (*K*), and Query (*Q*). In SBERT architecture, the values of *V*, *K*, and *Q* for each token are calculated as follows:

$$v = W_v^T T_i \quad v \in \mathbb{R}^p \quad V = [v, \dots, v_n]_{p \times n} \quad (5)$$

$$k = W_k^T T_i \quad k \in \mathbb{R}^p \quad K = [k_1, \dots, k_n]_{p \times n} \quad (6)$$

$$q = W_q^T T_i \quad q \in \mathbb{R}^p \quad Q = [q_1, \dots, q_n]_{p \times n} \quad (7)$$

Where, W_v^T, W_k^T, and W_q^T are transformation matrices.

After calculating the values of *V*, *K*, and *Q*, MSA is applied to them. Passing through the normalization and feedforward layers, the transformer result will be generated. Finally, in order to reduce the dimension, it will be passed through a pooling layer.

$$Multihead(Q, K, V) = W^0 concat(head_1, \dots, head_h) \quad (8)$$

Where, W^o is a linear transformation. $head_i$s are the output of different heads. *H* is the number of repeating blocks.

3.2.3. Feature alignment

After extracting the image and text features, to ensure that each element of the extracted feature has sufficient awareness of its neighboring elements, they are passed through a 1D temporal convolution layer, as follows [24]:

$$\hat{X}_{\{Im,TxT\}} = Conv1D(X_{\{Im,TxT\}}) \quad (9)$$

Where, $\{Im, TxT\}$ are visual and textual content, respectively. $X_{Im}$is the extracted feature of the image. X_{TxT} is the extracted feature of the text.

Therefore, it is expected that the convolved result contains the local information of the feature. In addition, the extracted features that have different dimensions will be projected to the same dimension by 1D temporal convolution.

3.3. Dimension reduction by LDA

In this stage of the proposed method, after extracting the features and performing the alignment, the resulting vectors are concatenated.

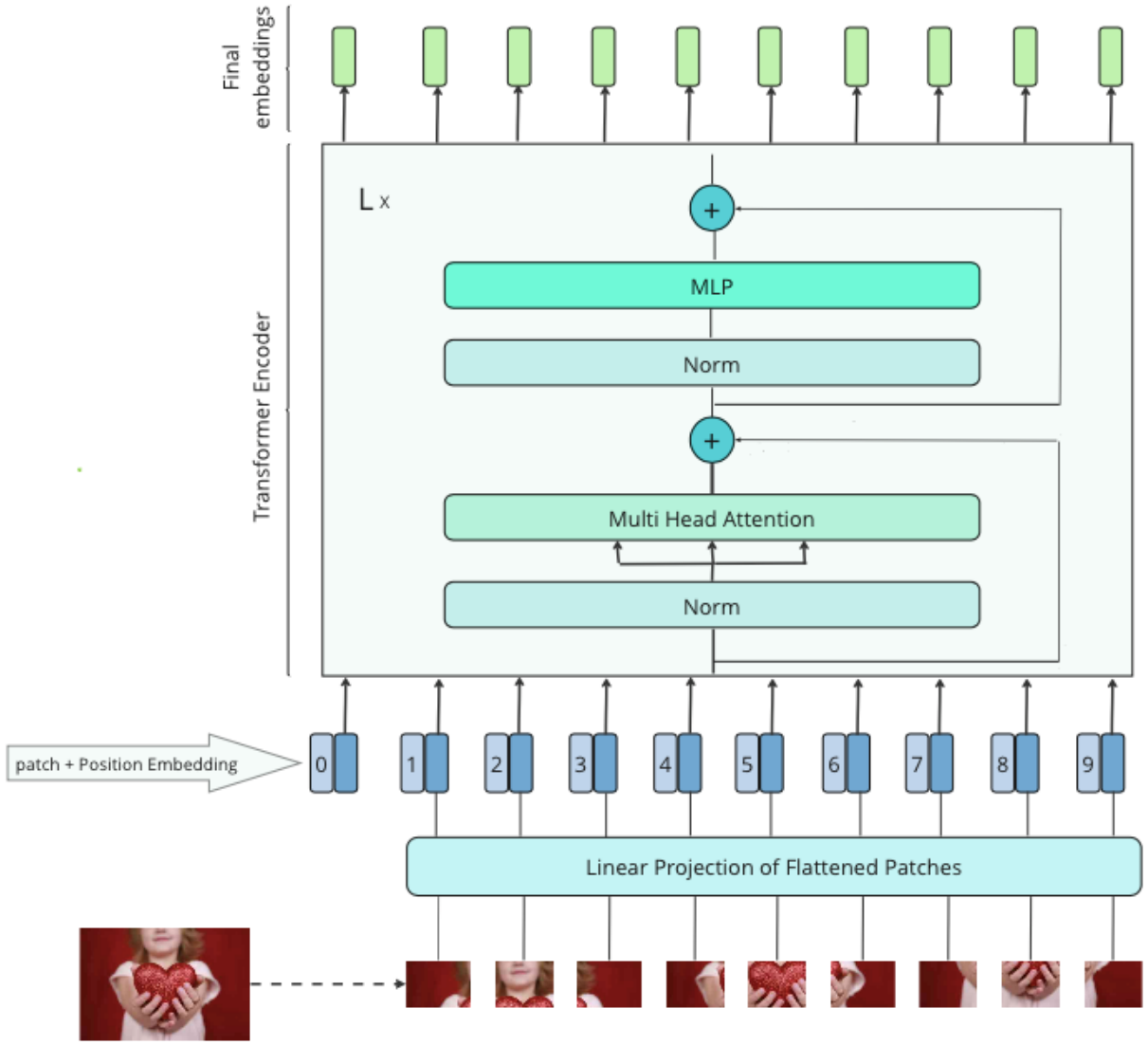


Figure 3 Visual feature extraction; Features of the input image are extracted by ViT. The input image is divided into non-overlapping patches. The patches are flattened. From flattened patches, linear projection with lower dimensions is produced. Position embedd

In order to improve the performance by better separation of sentiment categories, LDA is used to reduce the dimensionality of the concatenated vector.

The presence of high dimensions in the feature space can mean that the points in the space often represent very small and non-representative samples. This problem, which is known as the curse of dimensionality, can greatly affect the performance of the proposed method. This problem can often be solved by reducing the dimensions of the input features.

LDA is a predictive modeling algorithm used for multi-class classification. This algorithm can be used as a dimension reduction approach. This transformation projects the features from the higher dimensional space to the lower one while maintaining the most important structure or relations between the observed variables.

LDA first calculates the inter-class variance to reduce the dimensions of the features. For this purpose, it calculates the distance between the averages of different classes in order to obtain the amount of separation between classes.

$$S_b = \sum_{i=1}^{C} N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \quad (10)$$

Where c is the number of classes, N_i is the number of samples of class i, $\bar{x}_i$ is the average of i-th class samples and $\bar{x}$ is the average of all data points.

Then the intra-class variance is calculated. To do this, the distance between the average and sample of each class will be computed.

$$S_w = \sum_{k=1}^{C} S_k \,, S_k = \sum_{x(n)\in C_i} (x(n) - \bar{x}_i)\,(x(n) - \bar{x}_i)^T \tag{11}$$

Finally, LDA creates a space with lower dimensions that maximize the inter-class variance and minimize the intra-class variance. In this way, P is considered as the lower dimensional space.

$$P_{lda} = \underset{P}{\arg max} \frac{|P^T S_b P|}{|P^T S_w P|} \tag{12}$$

According to what was discussed about the method and benefits of dimension reduction by LDA, in the proposed method, this approach is used to reduce the feature vector resulting the concatenated of image and text features.

3.4. Label estimation

After extracting the visual and textual features, aligning and concatenating the features, and reducing the dimensions of the result vector, the final emotion label should be estimated at this stage. Therefore, two fully connected layers have been used. The reduced vector is fed into these two layers, and the final label of the sentiment class is estimated.

4. Experimental results

In order to estimate the sentiment of social media content, a meme-sensitive cross-modal method is proposed in this article. According to this approach, visual and textual content are used to determine the final sentiment class label.

To implement this method, first, image and text features are extracted using the ViT model and SBERT Bi-encoder method, respectively. Then, in order to align the extracted feature vectors, 1D temporal Convolution was used. Aligned vectors are concatenated, and the resulting dimensionality is reduced by LDA transformation. Then, they have been passed through two fully connected layers with ReLU activation function, which have 64 and 32 hidden units, respectively. Eventually, in the last layer, the sentiment label is estimated by the Softmax activation. The implementation is run on an NVIDIA GeForce RTX 2060. Also, the proposed method will be evaluated through Precision, Recall, F1, and Accuracy measures.

4.1. Datasets

In examining the results of the proposed method and comparing it with the performance of other state-of-the-arts, the following datasets have been used:

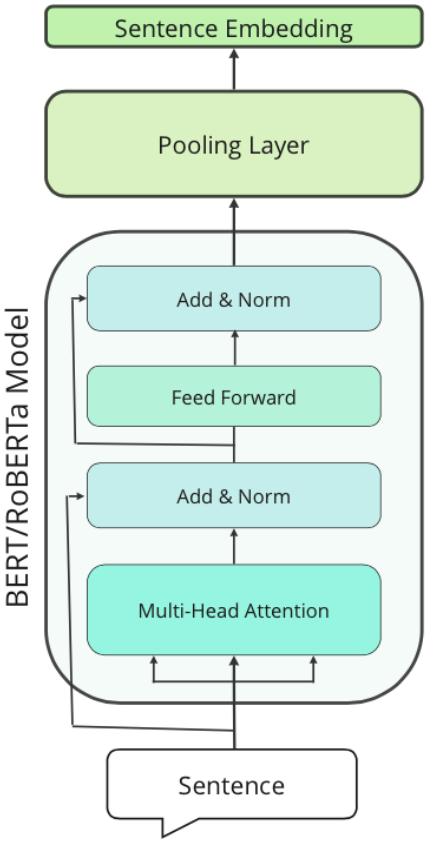


Figure 4 Sentence embedding

MVSA-Single and **MVSA-Multiple** [25]: In these datasets, text and image content of Twitter have been collected. The data are annotated into three categories: positive, negative, and neutral sentiment. The MVSA-Single containing 5,129 text-image pairs, is labeled by one annotator while the MVSA-Multiple with 19,600 text-image pairs is labeled by three annotators. In order to make the comparison fair, the content of both datasets was processed through the method proposed in [13]. According to this method, pairs that have different sentiment labels for text and image will be removed. Also, if one has a positive (or negative) label and the other has a neutral label, the positive (or negative) sentiment will be considered as the final label. Therefore, the number of pairs will be reduced to 4,511 in MVSA-Single and 17,024 in MVSA-Multiple.

Meme dataset: To collect a dataset for meme detection, the Hateful Memes dataset [26], consisting of 10,000 image memes, is utilized. This multimodal dataset is designed for detecting hateful memes, but the study does not consider the nature of these memes (whether they are hateful or not). All images from this dataset are treated as memes. Additionally, 10,000 random images from the COCO dataset are collected for regular image samples.

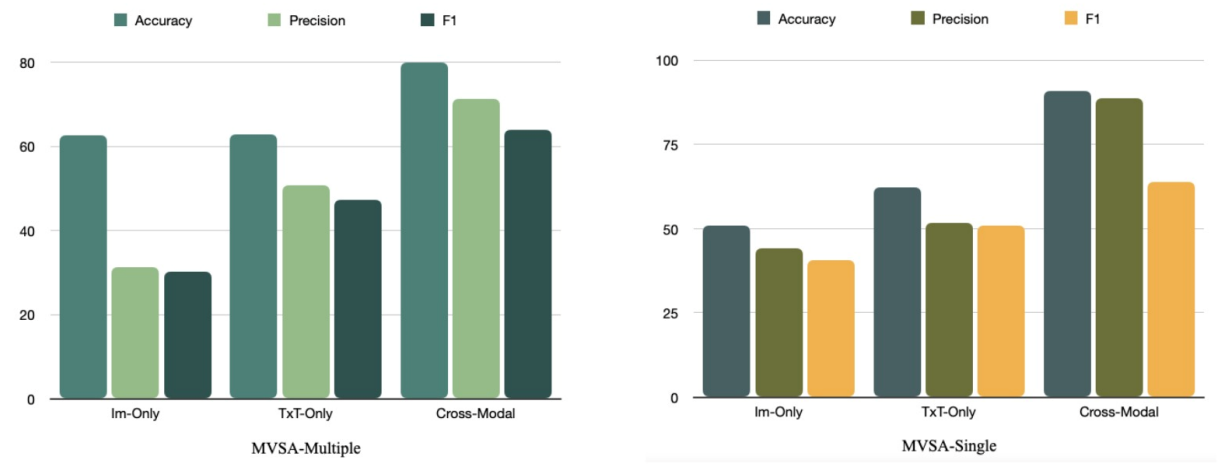


Figure 5 Results of comparison Accuracy between the Visual and Textual sentiment Analysis and the proposed method on MVSA datasets

4.2. The effect of using both image and text (cross-modal)

In this article, a cross-modal approach is proposed that uses visual and textual data in order to achieve a more accurate sentiment label. In this section, an experiment has been designed to prove that the use of text or image data alone has a significant effect in estimating the final result more accurately.

Im-Only: In this method, image features are first extracted using ViT. The dimensions of the features are reduced by LDA transformation and the final label is predicted after passing through two fully connected layers.

TxT-Only: First, sentence embeddings of input texts are calculated by SBERT RoBERTa model. After performing dimension reduction, the final text class is estimated using two fully connected layers.

As Figure 5 shows, on MVSA datasets, cross-modal has helped to more accurately estimate the sentiment of content in social media. This approach has a better performance than when only one type of content (text or image) is used.

4.3. The effect of using different feature extractors

This section will examine the effect of different feature extractors on textual and visual content.

4.3.1. Performance of different Visual feature extraction approaches

In the proposed method, ViT has been used to extract image features. Due to the fact that this approach is based on the transformer and its essence of attention architecture, it will extract better features and ultimately lead to a more accurate sentiment label prediction.

In this section, in order to compare the performance of different image feature extraction schemes, a number of well-known feature extractors of mid and high-level features, will be examined. In order to extract mid-level features, SIFT, ORB, and LBP, and to extract high-level features, pre-trained VGG16 and ResNet50 have been used. As mentioned in literatures [27], features extracted from deep networks are more efficient than low and mid-level features. In Table 1 the comparison results of different feature extractors on MVSA datasets have been compared. As expected, the results of ViT feature extraction are much better than other methods, even the feature extraction by pre-trained CNNs. Since in this method, the attention mechanism is used, the extracted features have higher efficiency. Therefore, they will have a significant impact on the final estimation.

4.3.2. Performance of different textual feature extraction approaches

As stated in 0, the SBERT Bi-encoder method has been used to obtain sentence embedding. This feature extraction method based on Sentence BERT/RoBERTa has a much better performance compared to traditional word embedding methods. As mentioned in [22], BERT or RoBERTa model can be used in this approach. According to the results [22], they do not have significant differences. However, in this section, in order to prove the better performance of these two models (Sentence BERT/RoBERTa) than the other pre-trained models, the applying results are compared with other embedding models;

MiniLM model [28]: It is a sentence transformer model that maps sentences or paragraphs to the vector space. This model, which is pre-trained on more than one billion training data, provides decent quality while being fast.

Universal Sentence Encoder (USE) [29]: This model is also a transformer-based model that

Table 1 Results of the different visual feature extractors on MVSA datasets (In all cases, SBERT-RoBERTa method is used for textual feature extraction and LDA is used for dimensionality reduction.)

Methods	MVSA-Multiple				MVSA-Single			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
SIFT	0.6386	0.5098	0.4787	0.4707	0.7284	0.7045	0.6789	0.6731
LBP	0.6351	0.5139	0.4815	0.4724	0.7284	0.6561	0.6108	0.6112
ORB	0.6433	0.4558	0.4201	0.3960	0.6220	0.6257	0.6151	0.5757
CNN-VGG16	0.5628	0.4357	0.4404	0.4132	0.6685	0.6908	0.6565	0.6504
CNN-ResNet50	0.5622	0.4295	0.4314	0.4061	0.6763	0.7216	0.6744	0.6755
ViT	0.8043	0.7171	0.6484	0.8060	0.9089	0.8860	0.8527	0.9078

encodes sentences into the vector space using a transformer encoder and deep averaging network.

Doc2Vec [30]: this method is an unsupervised one for generating a fixed-length feature vector for a piece of text (like a sentence, paragraph, or document). This method is introduced as an extension of Word2Vec. In Doc2Vec, a three-layer deep neural network is used. This scheme is not a unified method like Bag Of Word, but has two different variations of Skip-Gram and CBOW.

The results of applying different pre-trained text feature extraction approaches on datasets are shown in Table 2. As can be seen, the transformer-based methods provide better results than the Doc2Vec. Among the transformer-based schemes, the SBERT feature extractor and the RoBERTa model gives the best performance. The RoBERTa is superior than BERT model due to being trained in larger datasets, using a more effective training procedure, and using a dynamic masking pattern instead of the static one used in BERT.

4.4. The effect of using different dimension reduction methods

In the proposed method, LDA transformation is used in order to reduce the dimension of the vector resulting from concatenation of image and text features. In this section, the efficiency of using or not using the dimensionality reduction method (LDA) is presented. As can be seen in Table 3, when LDA is used to reduce the dimensionality of feature vectors, the values of measures on MVSA datasets have grown significantly compared to not using dimension reduction.

PCA Transformation: It is an unsupervised linear transformation technique that identifies patterns in data based on correlations between features. In fact, PCA seeks to find the directions of maximum variance in high-dimensional data and projects it into a new subspace that has dimensions less than or equal to the original space.

SVD Transformation: This transformation is a popular method for dimensionality reduction. This method is a data-driven one because patterns are extracted from data without any expert knowledge or intuition. Computing SVD consists of finding the eigenvalues and eigenvectors. The SVD is a steady numerical transformation. It creates a representation of data based on dominant correlation within data.

In Figure 6, the ROC plots of different dimensionality reduction approaches on MVSA-Single are illustrated. As can be seen, the results of LDA transformation have been much better compared to the other two transformations.

4.5. Performance comparison

In this section, the proposed method has been compared with various schemes that have been proposed in the field of sentiment analysis in social media.

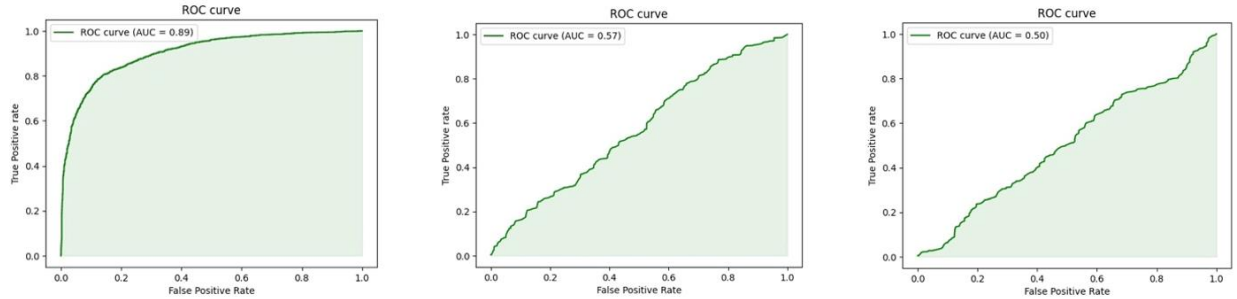
The comparison of the results of different methods on MVSA (Single and Multiple) is shown in Table 4. As it is clear, the proposed method has achieved significant results compared to other methods. In Xu's Method [13], text and image features are extracted by LSTM and pre-trained CNNs, respectively. Eventually, the sentiment label will be obtained through Softmax. Although Zhu et al. [9] also, have used CNN and GRU to extract features, they used a Faster R-CNN network to extract the corresponding features in both text and image. In the proposed method by Yadav et al. [10], Similar to previous methods, CNN and RNN were used for feature extraction. However, it used several attention structures to estimate the sentiment class with higher accuracy. The difference between Li's method [11] and the other aforementioned methods is the use of two contrastive learning structures to estimate the sentiment class. The three-module method of Peng et al. [12], in addition to extracting features by usual methods, uses attention structure and cross-modal hierarchical fusion to extract corresponding and specific features.

Table 2 Results of the different textual feature extractors on MVSA datasets (In all cases, ViT approach is used for visual feature extraction and LDA is used for dimensionality reduction.)

Methods	MVSA-Multiple				MVSA-Single			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Doc2Vec	0.7462	0.5130	0.4398	0.6958	0.7537	0.7053	0.6832	0.7552
USE	0.7743	0.7289	0.7691	0.7294	0.8571	0.8515	0.8487	0.8562
MiniLM	0.7704	0.6408	0.5667	0.7283	0.8350	0.7848	0.7427	0.8357
SBERT-BERT	0.7926	0.6851	0.5990	0.7470	0.8892	0.8758	0.8538	0.8885
SBERT-RoBERTa	0.8043	0.7171	0.6484	0.8060	0.9089	0.8860	0.8527	0.9078

Table 3 Results of using and not using LDA on MVSA datasets (In all cases, ViT and SBERT-RoBERTa are used for visual and textual feature extraction, respectively.)

Datasets	Without LDA				With LDA			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
MVSA-Multiple	0.6204	0.4625	0.4378	0.6186	0.8043	0.7171	0.6484	0.8060
MVSA-Single	0.6053	0.5657	0.5174	0.6090	0.9089	0.8860	0.8527	0.9078



MVSA-Single-LDA MVSA-Single-SVD MVSA-Single-PCA

Figure 6 ROC plot - AUC of the different dimension reduction methods on MVSA-Single dataset

Table 4 Comparison between the results of the proposed method and other state-of-the-arts MVSA (Single-Multiple) datasets

Method	MVSA-Single		MVSA-Multiple	
	Accuracy	F1	Accuracy	F1
Xu's Method [13]	0.6984	0.6963	0.6886	0.6811
Zhu's Method [9]	0.7519	0.7497	0.7352	0.7349
Yadav's Method [10]	0.7959	0.7947	0.7526	0.7789
Li's Method [11]	0.7533	0.7346	0.7200	0.6983
Peng's Method [12]	0.7361	0.7503	0.7045	0.7477
Proposed Method	**0.9089**	**0.9078**	**0.8043**	**0.8060**

5. Conclusion and future works

Since the sentiment analysis of published content in social media is an abstractive issue and its content cover a wide range of topics (full of ironic, metaphorical, and abstractive concepts), using a single content may not lead us to the exact sentiment of the content. In this respect, the use of auxiliary data helps to predict a more accurate result. Therefore, in this article, a cross-modal approach is introduced to predict the sentiment class of published content. Rather than relying on a single content type, both the image and the text published by the user are employed to determine the final label. Additionally, recognizing the growing prevalence of memes on social media, this method includes meme detection and preprocessing to ensure a more accurate analysis of the final sentiment. Since transformers have made significant progress in both visual and textual data analysis in recent years, this scheme has been used to extract features from pre-trained models based on transformers. Also, in the proposed method, LDA transformation has been

used to reduce the dimensionality of the extracted features and enhance classification. Finally, as reported in the experimental section, the performance of this method is much better compared to other state-of-the-arts. To improve the proposed method in the future, the corresponding relationships between visual and textual content can be used. In this case, it is possible to extract features that can express intra and extra-content relations in the best way. Eventually, the performance of the method can be increased, especially, for the analysis of content with ironic and metaphorical concepts.

References

[1] Jain R, Rai RS, Jain S, Ahluwalia R, Gupta J (2023) Real time sentiment analysis of natural language using multimedia input. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-15213-3

[2] Ramamoorthy S, Gunti N, Mishra S, Suryavardan S, Reganti A, Patwa P, Das A, Chakraborty T, Sheth A, Ekbal A, Ahuja C (2022) Memotion2: Datset on sentiment and emotion analysis Memes. In Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR.

[3] Zhang K, Zhu Y, Zhang W, Zhu Y (2021) Cross-modal image sentiment analysis via deep correlation of textual semantic. Knowledge-Based Syst 216:. https://doi.org/10.1016/j.knosys.2021.106803

[4] Lu J, Batra D, Parikh D, Lee S (2019) ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv Neural Inf Process Syst 32:1–11

[5] Xu J, Huang F, Zhang X, Wang S, Li C, Li Z, He Y (2019) Visual-textual sentiment classification with bi-directional multi-level attention networks. Knowledge-Based Syst 178:61–73. https://doi.org/10.1016/j.knosys.2019.04.018

[6] Huang F, Zhang X, Zhao Z, Xu J, Li Z (2019) Image–text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Syst 167:26–37. https://doi.org/10.1016/j.knosys.2019.01.019) (0123456789,-().vol

[7] Li Z, Sun Q, Guo Q, Wu H, Deng L, Zhang Q, Zhang J, Zhang H, Chen Y (2021) Visual sentiment analysis based on image caption and adjective–noun–pair description. Soft Computing. https://doi.org/10.1007/s00500-021-06530-6(0123456789().,-volV

[8] Serra A, Carrara F, Tesconi M, Falchi F (2023) The Emotions of the crowd: Learning image sentiment from Tweets via cross-modal distillation. https://doi.org/10.48550/arXiv.2304.14942

[9] Zhu T, Li L, Yang J, Zhao S, Liu H, Qian J (2022) Multimodal sentiment analysis with image-text interaction network. IEEE Transactions on Multimedia 25: 3375 – 3385

[10] Yadav A, Vishwakarma D (2020) A deep multi-level attentive network for multimodal sentiment analysis. https://doi.org/10.48550/arXiv.2012.08256

[11] Li Z, Xu B, Zhu C, Zhao T (2022) CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. https://doi.org/10.48550/arXiv.2204.05515

[12] Peng C, Zhang C, Xue X, Gao J, Liang H, Niu Z (2022) Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification. TSINGHUA SCIENCE AND TECHNOLOGY 27: 664:679

[13] Xu N, Mao W (2017) MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis. In Proceeding of the 2017 ACM on Conference on Information and Knowledge Management (CIKM'17) 2399-2402

[14] Wen H, You S, Fu Y (2021) Cross-modal context-gated convolution for multi-modal sentiment analysis. Pattern Recognit Lett 146:252–259. https://doi.org/10.1016/j.patrec.2021.03.025

[15] Yang B, Shao B, Wu L, Lin X (2022) Multimodal Sentiment Analysis with Unidirectional Modality Translation. Neurocomputing. https://doi.org/10.1016/j.neucom.2021.09.041

[16] Rahman W, Hasan MK, Lee S, Zadeh A, Mao C, Morency L, Hoque E (2020) Integrating multimodal information in large pretrained transformers. Proc Annu Meet Assoc Comput Linguist 2359–2369. https://doi.org/10.18653/v1/2020.acl-main.214

[17] Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf 1:4171–4186

[18] Hazarika D, Zimmermann R, Poria S (2020) MISA: Modality-Invariant and - Specific Representations for Multimodal Sentiment Analysis. MM 2020 - Proc 28th ACM Int Conf Multimed 1122–1131. https://doi.org/10.1145/3394171.3413678

[19] Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. AAAI 2020 - 34th AAAI Conf Artif Intell 1359–1367. https://doi.org/10.1609/aaai.v34i02.5492

[20] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929

[21] Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A (2023) A survey on visual transformer. https://doi.org/10.1109/TPAMI.2022.3152247

[22] Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using siamese BERT-networks. EMNLP-IJCNLP 2019 - 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process Proc Conf 3982–3992. https://doi.org/10.18653/v1/d19-1410

[23] Wang B, Kuo C (2020) SBERT-WK: A Sentence embedding method by dissecting BERT-based word models. https://doi.org/10.48550/arXiv.2002.06652

[24] Tsai, Y, Bai S, Liang P, Kolter J, Morency L, Salakhutdinov R (2019) Multi-modal transformer for unaligned multimodal language sequences. Proceedings of the Annual Meeting of the Association for Computational Linguistics 6558–6569

[25] Niu T, Zhu S, Pang L, Saddik A (2016) Sentiment analysis on multi-view social data. In International Conference on Multimedia Modeling 15–27

[26] Keila D, Firooz H, Mohan A, Goswami V, Singh A, Ringshia P, Testuggine D (2020) The hateful memes challenge: Detecting hate speech in multimodal memes. Annual Conference on Neural Information Processing Systems 2611-2624

[27] Onita D, Dinu L, Adriana B (2019) From image to text in sentiment analysis via regression and deep learning. Proceedings of Recent Advances in Natural Language Processing 862–868

[28] Wang W, Wei F, Dong L, Bao H, Yang N, Zhau M (2020) MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers

[29] Cer D, Yang Y, Kong S, Hua N, Limtiaco N, John R, Constant N, Guajardo-Ce´spedes M, Yuan S, Tar C, Sung Y, Strope B, Kurzweil R (2018) Universal sentence encoder. https://doi.org/10.48550/arXiv.1803.11175, 2018

[30] Le Q, Mikolov T (2014) Distributed Representations of Sentences and Documents. https://doi.org/10.48550/arXiv.1405.4053