



A new multimodal sentiment analysis for images containing textual information

Garvit Ahuja¹ · Alireza Alaei² · Umapada Pal³

Received: 8 April 2024 / Revised: 21 July 2024 / Accepted: 31 July 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Multimodal sentiment analysis on images with textual content is a research area aiming to understand the sentiment conveyed by visual and textual elements in the images. While multimodal sentiment analysis on images and text (reviews) has its own challenges, the combination of textual and visual content in the form of images presents new challenges as well as opportunities. In this research work, we proposed a multimodal sentiment analysis method that works on images incorporating textual elements. In the textual sentiment analysis model, we initially employed a recognition system to extract textual data from input images. Our proposed multimodal method is based on transfer learning, considering two pre-trained deep learning models, Xception, and RoBERTa, to extract features from both visual and textual content from multimedia images. We then implemented a fusion strategy to combine these two modalities (Visual Sentiment Analysis (VSA) and Textual Sentiment Analysis (TSA)) to enhance the accuracy of the proposed method and to provide a more comprehensive understanding of sentiment in multimedia content. In addition, we curated a custom dataset comprising images with associated text labels and sentiments. To ensure accurate labels, we conducted human evaluations involving thirty annotators. Our dataset includes images labeled with negative, neutral, and positive sentiments. Experimental results demonstrated the effectiveness of combining visual and textual features for sentiment analysis. The findings from this research hold promising implications for real-world applications, such as sentiment analysis in social media, product reviews, and marketing campaigns, where both images and text play a significant role in conveying emotional context.

Keywords Multimodal sentiment analysis · Image understanding · Deep learning · Transfer learning

1 Introduction

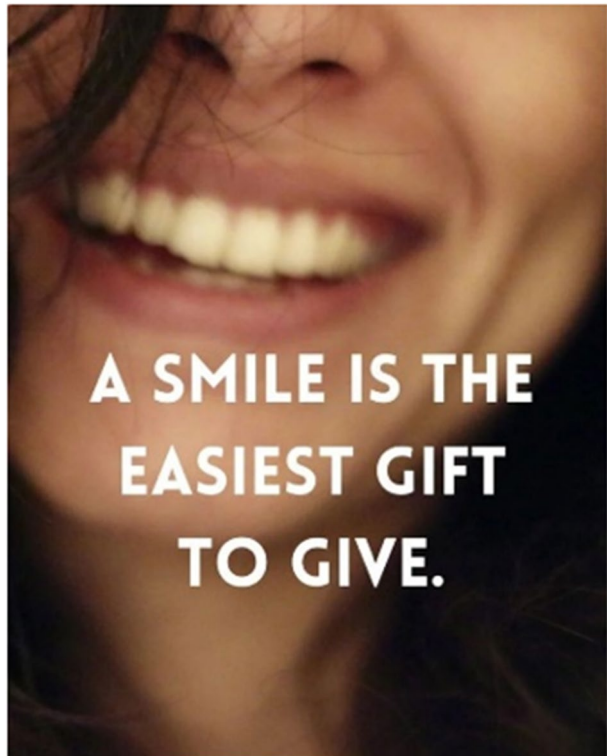
With the significant growth and use of social media platforms, people generate a vast amount of information and data daily. Most of the generated data was in text format at the earlier stage [40]. With the development of technology, images and videos have

Extended author information available on the last page of the article

also become a major part of the generated data. Images containing textual elements (multimedia images) as part of their content have become popular among people to convey their messages in a better/different way. An example of such images is shown in Fig. 1. Understanding the sentiment encapsulated in such images containing textual information is crucial in several applications, such as social media advertisement effect analysis, social media moderation, political campaign analysis, and brand monitoring [48–50]. For example, images containing textual elements have been used by people, brands, companies, and even politicians to gauge public sentiment on social media networks, such as Twitter, Instagram and Facebook [48–50]. The analysis of these multimedia data can provide deep insights into consumer sentiments and political opinions, offering valuable information for marketers, brand managers, and policymakers [48–50]. Sentiment extracted from these images can then be used to monitor and/or moderate a campaign by tailoring campaign materials and refining the messaging process [48]. However, the understanding and manual processing of people's opinions from a massive amount of data, including multimedia images, is not practical [40].

In the rapidly evolving landscape of artificial intelligence, sentiment analysis, or opinion mining, has become pivotal for automatically understanding people's opinions from data generated and stored in various formats. Sentiment analysis predominantly started with text based methods aiming to classify data (text) into three categories: positive, negative, or neutral. Several methods were developed, and thousands of articles about sentiment analysis were written in the literature [1–3, 40, 41]. As Yadollahi et al. [1] highlighted,

Fig. 1 A sample multimedia image containing a mix of visual and textual information in the image format



text-based sentiment analysis has seen considerable advancements in the literature. They underscored the extensive groundwork in the domain, categorizing sentiment analysis into opinion and emotion mining and emphasizing the need for clear definitions and logical frameworks [1]. Their review covers influential lexicons, datasets, and benchmarks, such as the Harvard General Inquirer and the Amazon dataset. Additionally, the review explores different models of basic emotions and automatic emotion classification techniques, offering a concise yet comprehensive snapshot of advancements in text sentiment analysis up to 2017. However, similar progress has not been observed for its visual sentiment analysis (VSA) counterpart, as indicated in [2]. This survey paper [2] details various problems associated with visual sentiment analysis, highlighting its infancy and the lack of significant advances in the literature. This situation is then intensified in multimodal sentiment analysis [3]. Gandhi et al. [3] stated that multimodal sentiment analysis, focusing on sentiment analysis using images, audio, and text, has gained popularity over the past few years. However, several challenges, such as the complexity of the data itself, are associated with these models [3].

From the sentiment analysis literature [1–3, 40, 41], we noted that the specific sentiment analysis problem involving images that incorporate textual data received limited attention, underscoring the importance and novelty of our research problem. Our research in this paper attempts to address this gap by proposing a new approach to visual sentiment analysis, leveraging the synergy between deep learning models designed for both image and text sentiment analysis. Our proposed method involves a multimodal approach, utilizing image and extracted text (from the input image) in two separate pipelines. The proposed text-based sentiment analysis model was designed using RoBERTa [4], along with a few additional layers. RoBERTa [4], as one of the best language models in the literature, is basically a reimplementation of the BERT [19] model with some modifications, such as removing the next-sentence pre-training objective and training with larger mini-batches and learning rates. We employed the Google Cloud Vision API (<https://cloud.google.com/vision/docs/ocr>), renowned for its substantial OCR capabilities, to extract textual information from images. This decision aligns with prior research findings [20], demonstrating the superior accuracy of the Google Vision OCR (<https://cloud.google.com/vision/docs/ocr>) compared to alternative tools. Moreover, implementing an automated text extraction framework in a Hadoop architecture leveraging Google Cloud Vision OCR (<https://cloud.google.com/vision/docs/ocr>) yielded notable efficiency gains, with the automatic extraction process being approximately two times faster than manual extraction. Additionally, the achieved text recognition accuracy approached nearly 100%, affirming the reliability and effectiveness of the selected OCR solution. The extracted text data underwent preprocessing to ensure uniformity and mitigate case sensitivity issues. Tokenization and sequence conversion were then performed to make the data suitable as the input of our RoBERTa-based text sentiment analysis model.

For visual sentiment analysis, a transfer deep learning model based on eXtreme Inception (Xception) [5] was proposed. The Xception deep learning architecture was initially designed for image classification tasks and gained popularity in computer vision applications [5]. Its capabilities to capture hierarchical features in complex visual scenes, high transfer learning ability, and novelty in terms of its application to visual sentiment analysis were the main reasons for choosing this model in our research work for VSA. As the fusion of different models commonly outperformed individual models, achieving the highest accuracy in the literature [45, 46], we introduced a weighted fusion technique, exploring various average weight combinations assigned to the image (visual) and text sentiment extracted from the proposed model to obtain the final sentiment for each input multimedia

image. The experimental results and comparative analysis further showcased the effectiveness of this fusion strategy in our proposed approach.

In addition, while studying this topic, we noticed the absence of a dedicated online publicly available dataset composed of multimedia images in the sentiment analysis literature. Therefore, through a well-defined data collection and annotation process, we created a reasonably large dataset of human-annotated images for sentiment analysis in this research.

In summary, our contributions in this research work are as follows:

- i) creating a new dataset for sentiment analysis composed of multimedia images and human annotated labels,
- ii) proposing a multimodal sentiment analysis framework suitable for multimedia images,
- iii) Unlike existing methods, our approach harnesses images as the sole source for extracting both visual and textual sentiment within our proposed multimodal analysis framework,
- iv) conducting several experiments considering several text and image-based sentiment analysis methods from the literature on our dataset and
- v) providing a comprehensive comparative analysis of the results obtained from our proposed model and state-of-the-art models, such as Visual Geometry Group19 (VGG19) [6], DenseNet201 [7], EfficientNetV2l [8], ResNet152V2 [9], InceptionV3 [10], InceptionResNetV2 [11], and MobileNetV2 [12] for VSA, and DistillBERT [13], ALBERT [14], and XLNet [15] for TSA on three different datasets.

The remainder of the paper is outlined as follows. Section 2 discusses the related work on image, text, and multimodal sentiment analysis models. Section 3 details a description of our data collection and annotation. Section 4 discusses our proposed multimodal framework. Section 5 provides experimental results and discussion around datasets, evaluation metrics, model training and fine-tuning, ablation study, and comparative analysis. Section 6 concludes the paper and provides future works.

2 Related work

2.1 Visual sentiment analysis

Visual information understanding has been a topic of research for years, and image and video recognition and captioning have facilitated visual understanding by transforming the information into natural language descriptions [51–54]. In recent advancements of dense captioning in visual scenes, various transformer-based approaches have been proposed to overcome limitations in handling sequential encoding and region prioritization [55], incorporating textual context and dynamic vocabulary diversification [56], and improving scale-invariant feature acquisition and eliminating redundancy in region interactions [57]. VSA, however, goes beyond traditional image analysis by incorporating expertise from image processing, machine learning, computer vision, data mining, and affective computing (sentiment/emotion analysis) domains to enable machines to comprehend and respond to the affective aspects of visual information in an image or video. VSA has emerged in recent years, focusing on the automatic understanding of sentiment/opinion conveyed through the visual content of images [2]. As a result, significant developments have been made in this field, driven by advancements in deep and transfer learning techniques [2, 17, 18, 29, 39, 42–44] over the past few years.

Borth et al. [29] introduced SentiBank, a concept detector library, and a Visual Sentiment Ontology (VSO) to predict sentiment in visual content [29]. The VSO provided a systematic, data-driven methodology to construct a large-scale sentiment ontology, facilitating a mid-level representation for bridging the affective gap in visual sentiment analysis [29]. It is worth noting that the method relies on the concept detectors in SentiBank that may be influenced by training data, the focus of the study, changes, and differences in emotion categories, affecting the generalisability of the method. You et al. [39] later proposed a Progressive CNN (PCNN) with a domain transfer learning method for VSA on Twitter images. The transfer learning from a small number of confidently labeled images in the target domain and progressive training strategy helped reduce the impact of noisy data on the training process, leading to improved performance and the generalisability of the model [39]. The method may, however, require a large amount of training data for optimal performance, which is challenging to obtain, especially in domains with limited labeled data. The reliance on domain transfer learning and the manually labeled data used for evaluation may also introduce biases from the source domain, impacting the applicability, effectiveness, and generalisability of the model in other domains [39]. While this study [39] demonstrated advancements through progressive training and transfer learning from labeled images, newer research suggests fine-tuning pre-trained models tailored for image classification tasks as a superior approach. Notably, Hassan et al. [18] adapted a pre-trained model using object and background details for multi-label sentiment analysis on disaster images shared on social media. Oversampling techniques, such as crowd-sourcing, were also used to address class imbalance in this method [18]. However, the complexity of disaster-related images with multiple objects and intricate backgrounds may pose challenges in accurately capturing and analyzing sentiments from such visual content. In addition, Jindal and Singh proposed a domain-specific fine-tuning approach using a deep convolutional neural network (CNN) for VSA from social media [42]. The approach effectively used transfer learning from large-scale image classification to sentiment prediction, improving the performance and generalisability of the proposed VSA. However, the network may fail to correctly distinguish between particular sentiment classes, such as very happy and happy or very sad and sad [42].

Later, Zhang et al. proposed a Visual Semantic Correlations Network (VSCNet) that combines deep learning with attention mechanisms and affective region discovery for VSA [43]. The model effectively captures visual semantics and emotional signs in images, improving sentiment detection accuracy. However, The proposed method may face challenges in accurately distinguishing between particular sentiments, especially negative ones. Moreover, different types of images (social networks dominated by objects and abstract paintings dominated mainly by color and texture) may affect the model to distinguish the sentiment at different levels [43]. A data-augmented transfer learning approach for VSA was also developed in the literature [44]. The method was a hybrid model based on VGG16 and SVM (Support Vector Machine), where the pre-trained VGG16 and SVM were used for feature extraction and sentiment classification, respectively [44]. VGG-19 and DenseNet121, as pre-trained CNN models, were also employed for extracting high-level features in VSA [17]. However, both methods [17, 44] primarily focus on positive and negative sentiments, overlooking neutral ones.

From the literature in VSA, it is noted that deep learning, domain-specific fine-tuning, attention mechanisms, and affective region discovery based models were designed to extract sentiment primarily based on visual emotions presented in images. Notably, none considered images with textual information that may be part of the image content for VSA. In addition, in most models, only two classes (positive and negative) were considered for

sentiment detection experiments. These issues led to justifying our proposed research in this paper and highlighted the importance of using the textual layer of information hidden in those images for sentiment analysis.

2.2 Textual sentiment analysis

Sentiment analysis from text has been an attractive research topic for over two decades, and several TSA methods were developed in the literature [1, 35–38, 40, 41]. The proposed methods can be grouped into lexical-based, machine learning, and hybrid categories. Lexical-based methods usually use dictionaries in which each word in a dictionary is associated with a predefined sentiment. For example, SentiStrength was a prominent dictionary-based model for sentiment analysis in social media [35]. Machine learning methods often rely on supervised machine learning approaches, which require labeled data to train them [40]. Hybrid models employ a combination of lexical and machine learning based models to extract sentiment from data [36]. For instance, Zhang et al. [36] proposed a hybrid approach combining lexicon and machine learning techniques for sentiment prediction from tweets. A detailed overview of text-based sentiment analysis methods can be found in [1, 38, 40, 41]. Though lexical-based, conventional machine learning and hybrid models have contributed significantly to text-based sentiment analysis, these approaches have several limitations [38, 40]. For instance, despite their simplicity, lexical-based methods often rely heavily on predefined dictionaries, and their effectiveness can be limited in different contexts/domains (context-dependent) and when faced with evolving language patterns and new words. Conventional machine learning models, on the other hand, require labeled datasets for training, making them less adaptable to diverse domains and potentially prone to overfitting. They also face challenges in handling a high-dimensional and sparse textual feature vector. While attempting to combine the strengths of lexical-based and machine learning approaches, hybrid models may also encounter difficulties in finding an optimal balance between the two components. As a result, the performance gains may not always be proportional to the added complexity. Additionally, the success of hybrid models depends on the availability of lexical resources, which may not always be the case for specific contexts or languages.

With significant advances in the field of deep and transfer learning, text-based sentiment analysis has experienced notable progress in recent years [1, 4, 13–15, 23, 28, 37]. Deep learning algorithms, particularly recurrent neural networks like LSTM models, have shown great success in capturing long-term dependencies and extracting meaningful features [23]. Attention-based RNNs also exhibited robust performance, underscoring the significance of preprocessing and feature extraction in sentiment and emotion analysis tasks. The transfer learning approach has also been effective in sentiment analysis, as it addresses the challenge of limited labeled data for specific domains [23, 28]. Considering recent advancements in deep neural networks, researchers like Rehman et al. [37] combined CNN and LSTM architectures to capture local features and long-term dependencies for sentiment detection from movie reviews. Researchers have further explored various pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers) [19], GPT (Generative Pre-trained Transformer), RoBERTa [4], DistillBERT [13], AIBERT [14], and XLNet [15], for sentiment analysis tasks in real-world applications. These models can capture contextual information and language patterns, improving sentiment analysis accuracies [1].

Despite significant progress in the TSA literature [1, 38], the limited consideration of visual information present in images is noticeable. Many traditional sentiment analysis methods focus exclusively on textual content, overlooking the valuable insights that can be derived from visual elements, especially in multimedia data. In contexts where text and images are jointly available, neglecting visual aspects can result in a poor understanding of sentiment. Therefore, as proposed in our framework, we considered textual data along with visual information for sentiment analysis.

2.3 Multimodal sentiment analysis

Multimodal sentiment analysis usually detects sentiment by combining sentiments expressed across different modalities, such as textual, image, video, and auditory information [3]. This approach enables a more comprehensive analysis of sentiments, expressions, and context, particularly in applications where sentiments are conveyed through diverse channels, such as images, videos, audio signals, and text [3, 27]. Recently, a few multimodal sentiment detection models were presented in the literature [27, 31–34, 45, 46]. A comprehensive literature review of the methods was also provided in [3]. Transfer learning was considered and explored in most of the recent multimodal sentiment analysis models in the literature [3].

Xu and Mao [31] presented a deep semantic network (MultiSentiNet) based multimodal sentiment analysis model, incorporating object and scene information from images. The model further considered a visual feature-guided attention mechanism to identify important words in tweets that contribute to understanding sentiment and aggregate these words with visual semantic features [31]. While MultiSentiNet captured detailed semantic information by correlating images and text information, its complexity and data dependency may pose challenges in specific contexts [31]. Xu [32] later proposed a hierarchical semantic attentional network for multimodal sentiment analysis using visual semantic and text features extracted from image captions and reviews, respectively [32]. Xu et al. [33] further introduced a co-memory network for multimodal sentiment analysis that considers the interrelation of visual and textual information by iteratively modeling the interactions between visual contents and textual words for better sentiment analysis [33]. However, the architecture of these models is complex, and they are sensitive to noise in image tagging. Obtaining high-quality captions is also difficult. Toledo and Marcacini [34] also proposed a fine-tuning strategy with an attention mechanism to select the most relevant features of each modality (text/image) for multimodal sentiment analysis. The model is a simpler and more cost-effective architecture as it uses simple concatenation strategies for multimodal learning [34]—however, early fusion at the feature level results in high dimensional features, resulting in complexity issues.

Recently, a Context-Sensitive Multimodal Dense Fusion (CS-MDF) framework based on visual and textual modalities was proposed to analyze sentiments at the utterance level by preserving sequential order and allowing contextual information [45]. The method involves extracting unimodal features independently at the first tier and fusing them at later stages to make predictions. The CS-MDF utilizes Bi-directional Gated Recurrent Units at the second tier to extract context-sensitive information from different modalities [45]. The framework highlighted the importance of each modality by incorporating a visual and semantic attention mechanism that helped align salient regions and words [45]. As this framework contains three tiered structures and multimodal fusion mechanisms, it can be computationally intensive, posing challenges for real-time applications. A graph convolution-based

heterogeneous fusion network was also proposed for multimodal sentiment analysis of data from diverse modalities, such as text, visual, and acoustic data [46]. This method utilized a graph structure, a convolutional aggregation module, attention mechanisms, and multi-task learning frameworks to overcome the noise problem caused by differences in modal information density [46]. The graph convolutional neural networks can capture inter-modal interactions and intra-modal feature dependencies, extracting complex interaction information between modalities [46]. Using graph convolution and dynamic routing may, however, introduce complexity to the model, potentially impacting computational efficiency. In a more recent paper, Bui et al. [27] introduced a holistic analysis of user sentiment using multimodal sentiment analysis, fusing image and text data. The model utilized the DenseNet201 model to extract visual features from images and the BERT transformer model, followed by an LSTM network to extract semantic and context-aware features from text data. These features were then combined using a hybrid fusion strategy to create the multimodal sentiment analysis model [27]. This model, however, achieved only slightly better results at a higher computational cost.

As indicated in the multimodal sentiment analysis literature, multimodal models have provided more accurate sentiment detection considering several modalities. Considering the scope of our problem in this research, limited research has been done to develop multimodal sentiment analysis based on only one modality (for example, images containing textual information). To address this gap in the literature, we proposed a new multimodal sentiment analysis in this research work to advance sentiment analysis in multimodal contexts. It is important to note that we fundamentally used only one modality (image) as the input to our sentiment analysis framework. However, we extracted several layers of information, including visual and text modalities, from the input image.

3 Data collection and annotation

Several datasets with annotated data have been created by experts or non-expert evaluators for sentiment analysis in the literature [1–3, 40]. Eighteen datasets containing mainly textual data were listed and explained in [40]. In addition, visual and multimodal datasets were discussed in [2, 3]. However, to our knowledge, no dataset in the literature combines textual and visual modalities exclusively in image format for sentiment analysis. Therefore, we created a custom dataset as a part of this research work. We initiated this process by manually sourcing 2,000 images from Google search, specifically targeting images containing visible text elements (multimedia images). This step was crucial to ensure the relevance of the dataset to our research objectives. The images with different sizes were all saved and stored in the JPG format. We named this dataset the Document Image Sentiment (DocImSent) dataset.

The 2,000 multimedia images were then manually categorized into three distinct classes, positive, negative, and neutral, by a knowledgeable individual. This process resulted in 670 positive, 670 negative, and 660 neutral images as the initial three subsets of our dataset. However, as this dataset was collected and annotated by only one person, there might be potential bias in the annotation process. To address this bias, we implemented a validation strategy to improve the credibility of our dataset. The validation strategy comprised 14 unique Google Forms, each containing 150 randomly selected images (50 positive, 50 negative, and 50 neutral images) from our initial dataset. Thirty undergraduate students were instructed to participate in annotating the images associated with these Google Forms. This

process was designed to ensure that each image was evaluated by at least five additional individuals, excluding the original annotator. To better understand the validation process, examples of the questionnaire and instruction forms given to students are shown in Figs. 2 and 3, respectively.

To proceed with the annotation process, two to three sets of Google Forms, along with instructions, were distributed to each individual within a group of five people, and their responses were subsequently recorded. After carefully assessing their responses, images with conflicting opinions were excluded (rejected), while those with an absolute majority sentiment were included (accepted) in the dataset. An example illustrating the rejection and acceptance of images based on annotator responses is presented in Fig. 4. As shown in Fig. 4, if all five annotators selected "positive," the image was accepted in our validated dataset. Conversely, as illustrated in the second pie chart, there was disagreement among annotators, with 3 indicating a neutral sentiment and 2 suggesting a positive sentiment; hence, images with such disagreements were excluded from our dataset.

Following this rigorous validation process, we successfully refined the initial dataset to a total of 1,717 images, comprising 627 positive images, 528 negative images, and 562 neutral images. This step played a crucial role in mitigating potential biases and enhancing the overall reliability and consistency of our dataset.

As we were also interested in textual data in each image, we utilized the Google Cloud Vision OCR (<https://cloud.google.com/vision/docs/ocr>) to accurately extract text content from each image. Subsequently, we manually verified the text extracted by the OCR to ensure the extracted text from each image was correct. The extracted text and the corresponding image name were stored in a Comma-Separated Values (CSV) file as part of our annotation process. The CSV includes three columns: one for the image name, another for the extracted text, and a third for the corresponding sentiment label assigned to each image based on our earlier annotation process. This CSV file thus encapsulated the textual

Question *



Positive Negative Neutral

Fig. 2 Questioner form designed to annotate images in the DocImSent dataset

Section 1 of 4

Select Sentiment for the image

Dear respondents,

Thank you for participating in my survey. I kindly request your valuable feedback on the following images.

Instructions:

1. Please carefully review each image and make your personal overall assessment.
2. Once you have formed an overall impression, select one sentiment (positive, negative, or neutral) that best represents your assessment for each image.
3. Your sentiment selection should be based on your personal evaluation of the image's content and context. Make the selection without being influenced by your emotions and try to look at the images as if you were in the third person.

I appreciate your thoughtful responses, as they will contribute significantly to my research. If you have any questions or need further clarification, please don't hesitate to contact me personally at [redacted].

Thank you for your participation!

Best regards,
Garvit Ahuja

Fig. 3 Instruction form designed to help annotate images in the DocImSent dataset

content alongside the key metadata, facilitating seamless integration of the text data into our dataset. The refined dataset associated with annotation for textual and visual information formed a strong foundation for the subsequent phases of our research.

4 Proposed model

The block diagram of our proposed framework is depicted in Fig. 5. As demonstrated in the figure, the proposed framework contains two pipelines for sentiment extraction from visual and textual data of the input image, followed by a fusion strategy. Each pipeline and its corresponding blocks are illustrated in the following subsections.

4.1 Visual sentiment analysis

The VSA pipeline comprises the image preprocessing step followed by the proposed visual sentiment analysis model discussed as follows.

4.1.1 Image preprocessing

To improve the quality of input images and simplify the subsequent feature extraction, we applied a series of preprocessing steps to raw images before feeding them into our proposed model. Initially, we resized all input images to a standard 256×256 pixels to maintain consistency across all the images in the dataset. This resizing process is crucial for uniformity

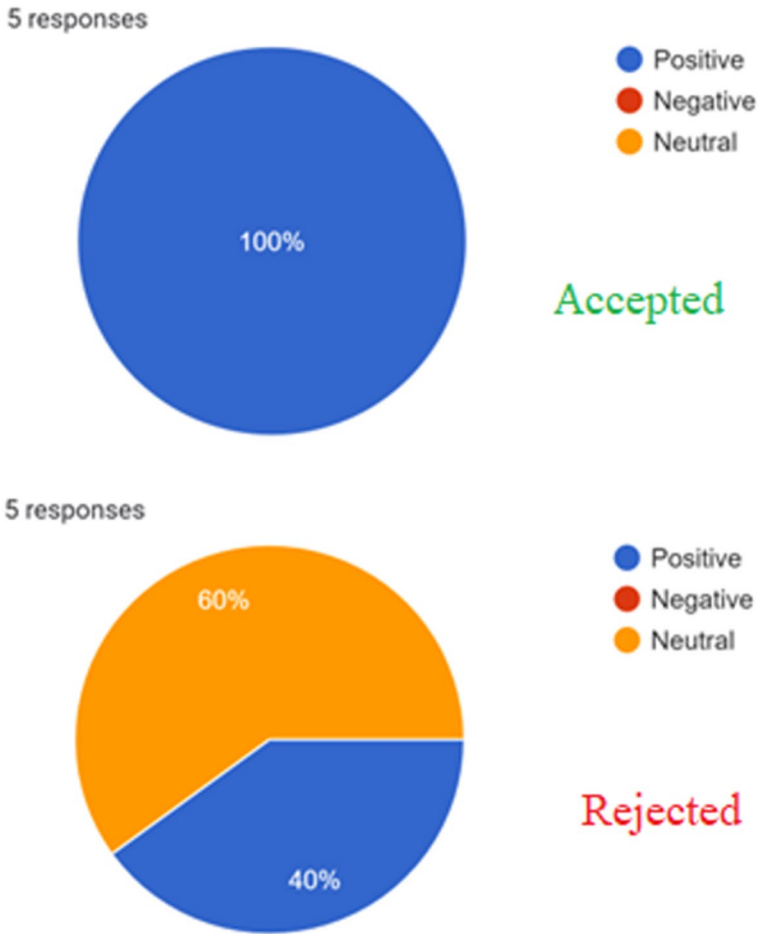


Fig. 4 An example of form responses received from annotators

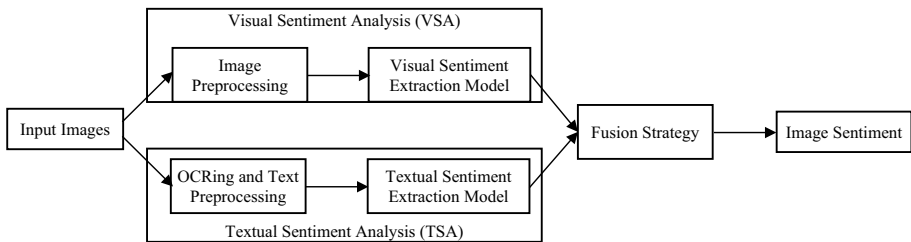


Fig. 5 The architecture of the proposed framework

and helps address potential issues arising from varying image sizes. We also scaled the pixel values of the images by dividing each value by 255. This normalization process brought the input features into the range [0, 1], enhancing the numerical stability and convergence of our model during the training. Together, these preprocessing techniques laid a

solid foundation for subsequent stages of our model, ensuring effective feature extraction and facilitating an efficient learning process.

4.1.2 Proposed transfer learning based visual sentiment extraction model

Several state-of-the-art image-based deep learning models, including Xception [5], VGG19 [6], DenseNet201 [7], EfficientNetV2L [8], ResNet152V2 [9], InceptionV3 [10], InceptionResNetV2 [11], and MobileNetV2 [12], were implemented and trained for common computer vision tasks, such as image classification, object detection and recognition in the literature (Table 1). Since constructing and training models from scratch for different tasks can be costly, capitalizing on the weights of pre-trained models, transfer learning has emerged as a practical solution to such an issue [5–12].

Therefore, we adopt transfer learning in our image sentiment analysis, utilizing pre-trained models that have gained complex features from extensive datasets, such as ImageNet. This strategy expedites the training of our sentiment analysis model by using previously acquired knowledge, substantially reducing the computational resources needed. In addition, the inherent ability of transfer learning to generalize allows our model to effectively handle new sentiment analysis tasks, even with limited task-specific data available. A study [16] highlighted the effectiveness of transfer learning in achieving impressive accuracy scores within limited computational resources. Despite computational power and time constraints, their validation of transfer learning's efficacy showed its broader application to sentiment analysis [2].

From Table 1, it is evident that there are several deep learning models in the literature with different parameters, depths, and sizes for image classification. To choose the most appropriate transfer learning for our proposed image-based sentiment analysis model, we considered the different factors outlined in Table 1. A careful consideration of all models and the factors presented in Table 1 revealed that the Xception model [5] is one of the best models with outstanding performance, showcasing a compelling combination of compact size and remarkable Accuracy. The Xception [5] surpasses many other models, with Top-1 and Top-5 accuracy rates of 79.0% and 94.5%, respectively. Furthermore, its modest size of 88 MB and a relatively low parameter count of 22.9 million make it a good choice, balancing between the model complexity and computational demands. It is also worth noting that the Xception [5] model is pre-trained using a diverse set of images from the ImageNet, consisting of millions of images. In addition, it draws inspiration from Google's Inception

Table 1 Performance of some of the state-of-the-art deep learning image classification models on the ImageNet (<https://keras.io/api/applications/>)

Model	Size (MB)	Parameters	Depth	Top-1 Accuracy	Top-5 Accuracy
Xception [5]	88	22.9 M	81	79.0%	94.5%
VGG19 [6]	549	143.7 M	19	71.3%	90.0%
DenseNet201 [7]	80	20.2 M	402	77.3%	93.6%
EfficientNetV2L [8]	479	119.0 M	-	85.7%	97.5%
ResNet152V2 [9]	232	60.4 M	307	78.0%	94.2%
InceptionV3 [10]	92	23.9 M	189	77.9%	93.7%
InceptionResNetV2 [11]	215	55.9 M	449	80.3%	95.3%
MobileNetV2 [12]	14	3.5 M	105	71.3%	90.1%

model and presents an extreme interpretation of Inception principles. Notably, to the best of our knowledge, this model has not been used for sentiment analysis in the literature.

Therefore, we chose the Xception [5] model as the basis of our image sentiment analysis task. This pre-trained model provided good results in classifying various object categories and demonstrates its adaptability in providing robust representations suitable for a wide range of images, including fruits [21]. The block diagram of the proposed Xception based architecture with additional layers is presented in Fig. 6. As demonstrated in Fig. 6, in our proposed Xception-based transfer learning visual sentiment analysis model, we initially initialized the Xception [5] model with pre-trained parameters and fed it with the pre-processed images as input. We further augmented the model architecture by adding three additional layers to facilitate fine-tuning and enhance its capabilities. Firstly, we flatten the multi-dimensional output obtained from the Xception [5] pre-trained model into a vector (one-dimensional array) to prepare the data for subsequent layers. Next, we introduced a dropout layer with a threshold of T1 as a regularization technique to mitigate overfitting and enhance the model’s ability to generalize to new and unseen data. Lastly, we appended a dense layer—a fully connected layer—with a softmax activation and an L2 kernel regularizer, which helps to prevent overfitting by discouraging large weight values, designed explicitly for our classification task with three classes, positive, negative, and neutral. This new architecture allowed us to explicitly tailor the Xception model [5] for our purposes and achieve optimal sentiment analysis based on image content. By incorporating these

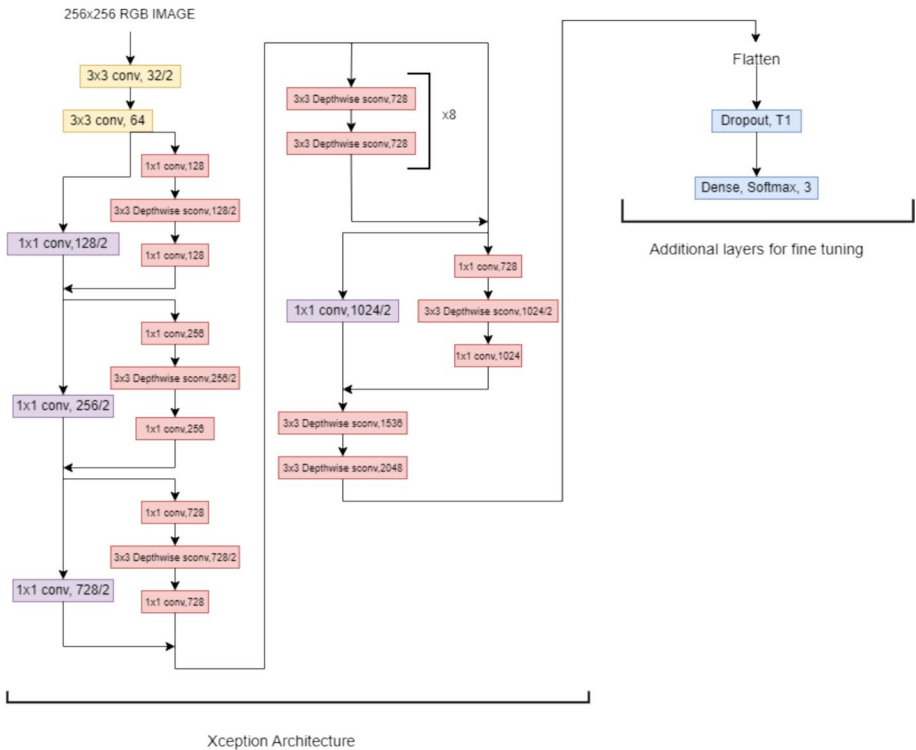


Fig. 6 Our proposed Xception-based architecture with additional layers for VSA

layers and techniques, our model can effectively generalize and provide accurate predictions across various sentiment categories. In Sect. 5.3 (Ablation Study), we also used other existing pre-trained models with our fine-tuning layers to compare results and select the one that is most suitable for our proposed framework.

4.2 Textual sentiment analysis

The TSA pipeline begins with applying an OCR on the input image to extract text, followed by the text preprocessing step, and finally, the proposed textual sentiment analysis model, as discussed below.

4.2.1 OCR and text preprocessing

Text extraction from images is a critical step in performing sentiment analysis in the TSA pipeline. OCR, a mature text recognition method, can usually fulfill this task proficiently. In our proposed framework, we chose Google Cloud Vision OCR due to its performance in accurately and efficiently extracting text from images (<https://cloud.google.com/vision/docs/ocr>). Google Cloud Vision OCR is known for appropriately handling text with different conditions, such as multiple scripts used within the text or distorted text. It can also be integrated with Google Cloud services, making it easy to scale up and use in our proposed solution (<https://cloud.google.com/vision/docs/ocr>).

The texts extracted from the Google Cloud Vision OCR were then subjected to several preprocessing methods to enhance the quality of the texts. For example, applying data transformation and filtering significantly improved classifications [22]. Therefore, we first converted all text to lowercase to ensure consistency and avoid issues arising from different capitalizations. Removing newline characters and breaks introduced during OCR further improved the readability of the text. Removing URLs and filtering out distracting web links were also essential to keep only meaningful text for further processes. English stopwords were then eliminated from the extracted text to refine it by excluding frequently used but less meaningful words. Removing non-alphanumeric characters and unnecessary symbols contributed to a better text representation. Lemmatization was further applied to reduce words to their base forms, ensuring consistency and capturing the basic meaning of words. Part-of-speech (POS) tagging was used to help with targeted lemmatization based on word categories, thereby enhancing precision. Finally, tokenization and rejoining of the text were applied to prepare the extracted text for further analysis. Each preprocessing method was chosen to address challenges, such as varied capitalization, noise, and less meaningful elements, introduced during OCR. Through these preprocessing methods, the extracted text was transformed into a clean, more accurate, and standard format suitable for subsequent analysis.

4.2.2 Proposed transfer learning based textual sentiment extraction model

In this research work, we explore the suitability of various deep transfer learning models, including RoBERTa [4], DistillBERT [13], ALBERT [14], and XLNet [15], each renowned for their power in understanding textual data. It is worth noting that a comparative analysis presented in the literature [23] evaluated text-based emotion recognition employing RoBERTa [4], DistillBERT [13], XLNet [15], and BERT [19]. Among these models, RoBERTa emerges as the top performer, achieving a high accuracy for seven emotion classifications

[23]. Similarly, RoBERTa consistently delivered superior results compared to various language models, such as DistilBERT and XLNet, on the GoEmotion dataset [24, 25], reaffirming its efficacy in emotion recognition tasks. As such, RoBERTa has also become a basis for developing successful NLP models and has gained popularity in research and industrial applications [4].

Therefore, in our proposed framework, we chose the Robustly optimized BERT approach (RoBERTa) [4] as the basis of our textual sentiment analysis. RoBERTa is a variant of the BERT [19] (Bidirectional Encoder Representations from Transformers) model developed by Facebook AI researchers [4]. While sharing some architectural similarities with BERT, RoBERTa distinguishes itself through its architecture and training procedures. One important modification in RoBERTa is the removal of the Next Sentence Prediction (NSP) objective, which involves training the model to predict whether observed document segments are from the same or distinct documents. Removing/adding NSP loss in different versions indicated that eliminating NSP either matches or slightly improves downstream task performance [4].

Furthermore, RoBERTa introduced training with larger batch sizes and longer sequences, involving 125 steps of 2 K sequences and 31 K steps with 8 K sequences of batch size, which offered several advantages, such as improved perplexity on the masked language modeling objective and enhanced end-task Accuracy. Larger batches also facilitated easier parallelization via distributed parallel training. In addition to these modifications, RoBERTa employed a dynamic masking technique during training, departing from BERT's static masking performed once during data preprocessing. Instead, training data was duplicated and masked ten times with different mask strategies over 40 epochs. This approach was compared with dynamic masking, where data generated by different masks every time was passed into the model. Despite these alterations, RoBERTa maintained its transformer-based language model functionality, utilizing self-attention mechanisms to process input sequences and generate contextualized representations of words within sentences. The model's training on a substantially larger dataset of 160 GB of text, along with its dynamic masking technique, contributes to its proficiency in natural language processing tasks, outperforming BERT and other models across various applications, such as sentiment analysis, language translation, text classification, question answering and text-based emotion classification [23, 24]. This compelling evidence underscores the suitability of RoBERTa as the choice of transformer model for our proposed text sentiment analysis framework.

The block diagram of the proposed RoBERTa-based architecture with additional layers is provided in Fig. 7. As demonstrated in Fig. 7, we began by implementing a 1D global average pooling to obtain a condensed output from the results of the pre-trained model. This technique involved averaging values along the temporal dimension, reducing spatial dimensions, and providing a concise representation of features. Following this step, we introduced a dense layer with 768 neurons and leveraged the Rectified Linear Unit (ReLU) activation function for feature transformation. The ReLU function introduced non-linearity, allowing the model to capture intricate patterns in the data. Subsequently, we incorporated another dense layer with 256 neurons and ReLU activation to further enhance the performance of our proposed model. In addition, we integrated a dropout layer with a threshold of T2 into the architecture to prevent overfitting and promote generalization. Dropout randomly deactivated a fraction of neurons during training, fostering robustness and preventing the model from overly relying on specific features. Increasing the architecture, we further introduced two additional dense layers. The first layer consists of 64 neurons with the ReLU activation function, aiding the model in learning hierarchical representations. The final layer comprises only three neurons

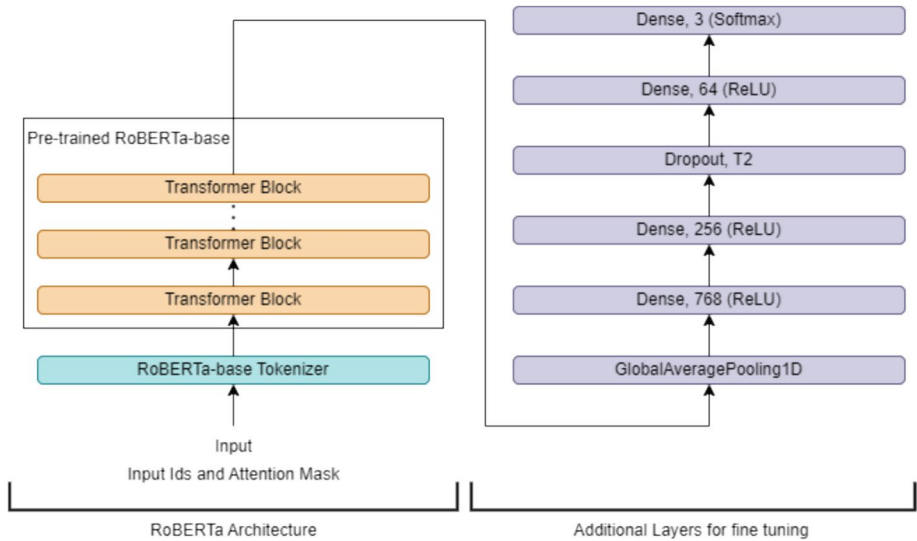


Fig. 7 Our proposed RoBERTa-based architecture with additional layers for TSA

with a softmax activation function, generating the probability distribution for three classes, positive, negative, and neutral, for textual sentiment analysis purposes. As with Xception, we used various existing transformers and added additional layers for experiments on our custom dataset fine-tuning in Sect. 5.3. We then combined the best-performing combinations of our text and image sentiment analysis models using our proposed fusion method to obtain the highest accuracy.

4.3 Weighted fusion strategy

There are several fusion strategies in the literature, and each fusion method performs differently depending on the applications [26]. Various fusion strategies for combining different classifiers were presented in [26], where the authors emphasized the critical role of selecting an appropriate fusion method to achieve optimal results from multiple classifiers [26]. As shown in Fig. 8, our proposed sentiment analysis method comprised two pipelines, one for image and the other for textual sentiment analysis, with a weighted fusion to help obtaining final sentiment results. We tried different weight combinations for image and text sentiment predictions to find the optimal mix per Eq. 1.

$$\text{combinePred} = [p1 * w1 + p2 * w2 | p1, p2 \in \text{zip}(\text{pred1}, \text{pred2})] \quad (1)$$

where $p1$ and $p2$ are the individual sentiment predictions from the image and text models, respectively. The image weight is represented by $w1$, and the weight for text is $w2$ (where $w1 + w2 = 1$). Equation (1) generates a list containing sublists, each comprising three probabilities: the first position representing the probability of a negative prediction, the second for neutral, and the third for positive.

$$\text{Prediction} = [\text{argmax}(\text{pred}) | \text{pred} \in \text{combinePred}] \quad (2)$$

We introduced Eq. (2) to transform this *combinePred* into a *Prediction* list suitable for evaluating the accuracy and other metrics. This Equation iterated through each sublist in the *combinePred* and replaced it with the index corresponding to the maximum value within that sublist. For instance, if a sublist contains values like [0.1, 0.1, 0.8], Eq. (2) would replace this sublist with the index 2 in the list. To assess how well this weighted fusion has performed, we compared our combined predictions to the actual labels (ytrue). We tested different weight combinations and found the best weights (w1 and w2) that provided the most accurate predictions when combining visual and textual information. This strategy helped the proposed model adapt to different situations, improving its ability to make more accurate sentiment predictions based on both visual and textual elements of images (Fig. 8)

5 Experimental results and discussion

5.1 Dataset and evaluation metrics

We used three different datasets to evaluate our proposed multimodal sentiment extraction framework. The first dataset was the DocImSent dataset developed in this research work. The DocImSent dataset is composed of 1,717 images, of which 627 were positive, 528 were negative, and 562 were neutral. A few examples of the images and their metadata from DocImSent are shown in Table 2.

The other two datasets utilized in our evaluation were MVSA datasets, namely MVSA-Single and MVSA-Multiple [28]. The MVSA datasets [28] consist of image-text pairs,

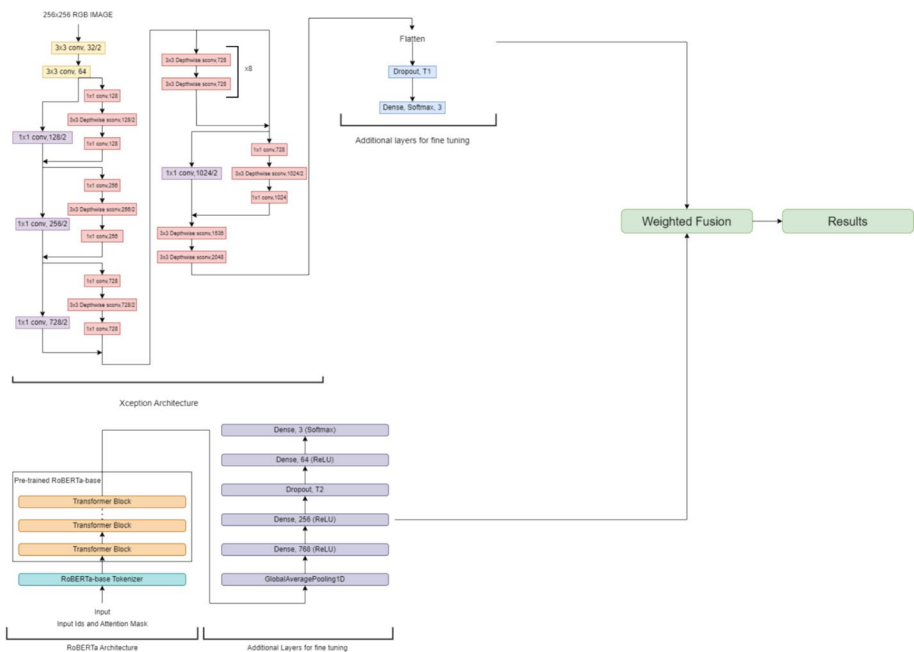











Fig. 8 Proposed fused model architecture for sentiment extraction from multimedia images




Table 2 Examples of image and textual data from the DocImSent dataset and respective prediction results obtained from our proposed framework

Sno.	Image	Text	True Label	Predicted Label
1		A smile is the easiest gift to give. routinely nomadic	Positive	Positive
2		Vibes don't lie.	Positive	Neutral
3		Happy people are always beautiful.	Positive	Positive
4		Meaningless	Neutral	Neutral
5		Avocado	Neutral	Neutral
6		Helicopter rope riddle	Neutral	Neutral
7		My silence is another word for my pain.	Negative	Negative
8		Journalist attacked assaulted r. news	Negative	Negative
9		Pray for Kerala	Negative	Positive

each comprising an image and a separate text. The MVSA-Single dataset comprises 4,869 pairs collected from Twitter, with respective annotations for both image and text by a single annotator. In contrast, the MVSA-Multiple dataset contains 19,600 pairs labeled by three annotators. The MVSA datasets contain images and the corresponding texts posted alongside them sourced from Twitter. While these datasets do not contain images with texts within themselves, they serve as valuable datasets to showcase the performance of our proposed framework across different settings. After removing invalid data based on the criteria outlined in [27] for the experimental study, we obtained 4,511 pairs (image, text) for MVSA-Single and 17,024 pairs for MVSA-Multiple, respectively. We partitioned the dataset according to the division specified in [27], allocating an 8:1:1 ratio for training, validation, and testing. We further refined this dataset based on the rules outlined in [27] to ensure alignment with the results presented in that paper for our comparative analysis. Table 3 shows examples of images and textual captions from the MVSA datasets.

In this research work, we considered four commonly used evaluation metrics, Accuracy, Precision, Recall, and F1-score [40], to assess the performance of our proposed model and

Table 3 Example of image and textual caption of MVSA datasets [28]

Sno.	Image	Text	Sentiment
1		The moment I find my favourite TV character	Positive
2		Depressed depression bully anxiety overdose addict drug pill cut cut	Negative
3		Author read to an enthusiastic audience	Neutral

compare its results to the state-of-the-art models. As we have three different classes (positive, negative, and neutral), Precision, Recall, and F1-score were calculated individually for each class. For instance, Precision for the positive class is the ratio of true positive predictions for the positive class to the total predicted positive instances, and similarly for Recall and F1-score [40]. To report an overall evaluation metric for the model, the average Precision, Recall, and F1-score were computed, and the results were displayed as a comprehensive performance measure for the entire model, considering its performance across all three classes. This ensured a balanced assessment considering the model's performance across different categories. We have further displayed the confusion matrix for each model to provide detailed results.

5.2 Model training and testing

As our main focus in this research was working on multimedia images, we first used our dataset for training, parameter setting (fine-tuning), and testing. We considered 1,237, 240, and 240 images for training, validation, and testing of our proposed method, respectively. The visual and textual sentiment extraction models underwent separate training procedures using images and their extracted text. Throughout the training phase, a validation procedure using the validation dataset was considered to assess the performance of models and prevent overfitting continually. The validation process further helped in terms of the generalization of the proposed models on unseen data. We chose sparse categorical cross-entropy for both image and text models as the loss function and utilized the Adam optimizer as the optimization function.

5.3 Ablation study

As several models in the literature could be used in our proposed pipelines for visual and textual sentiment analysis, we performed several experiments on our dataset. We further reported the results of only image (visual) and text (textual) models to compare how accurately both models performed individually in this research work. Tables 4 and 6 show sentiment results obtained from only visual sentiment and textual sentiment analysis models on our dataset. Tables 5 and 7 demonstrate the confusion matrix obtained from our experiments for the visual and textual sentiment models, respectively. When comparing visual sentiment analysis results presented

Table 4 Visual sentiment analysis results obtained from different VSA models designed based on various state-of-the-art deep learning models on our DocImSent dataset

Visual Sentiment Extraction Model	Accuracy (%)	Precision	Recall	F1-Score
VGG19 [6]	75.41	0.759	0.754	0.753
DenseNet201 [7]	77.08	0.771	0.771	0.771
EfficientNetV2l [8]	55.41	0.579	0.554	0.541
Xception [5]	78.30	0.782	0.783	0.783
ResNet152V2 [9]	73.75	0.737	0.737	0.737
InceptionV3 [10]	71.25	0.713	0.712	0.712
InceptionResNetV2 [11]	75.01	0.751	0.75	0.749
MobileNetV2 [12]	77.50	0.775	0.775	0.774

Table 5 The confusion matrix supporting the results presented in Table 4

Visual Sentiment Extraction Model	Classes	Negative	Neutral	Positive
VGG19 [6]	Negative	72.5%	11.25%	16.25%
	Neutral	12.5%	68.75%	18.75%
	Positive	3.75%	11.25%	85%
DenseNet201 [7]	Negative	77.5%	11.25%	11.25%
	Neutral	17.5%	73.75%	8.75%
	Positive	6.25%	13.75%	80%
EfficientNetV2l [8]	Negative	48.75%	21.25%	30%
	Neutral	13.75%	34.15%	51.25%
	Positive	3.75%	13.75%	82.5%
Xception [5]	Negative	81.25%	8.75%	10%
	Neutral	16.25%	73.75%	10%
	Positive	6.25%	13.75%	80%
ResNet152V2 [9]	Negative	76.25%	10%	13.75%
	Neutral	15%	73.75%	11.25%
	Positive	8.75%	20%	71.25%
InceptionV3 [10]	Negative	73.75%	16.25%	10%
	Neutral	21.25%	66.25%	12.5%
	Positive	7.5%	18.75%	73.75%
InceptionResNetV2 [11]	Negative	72.5%	16.25%	11.25%
	Neutral	15%	71.25%	13.75%
	Positive	5%	13.75%	81.25%
MobileNetV2 [12]	Negative	80%	11.25%	8.75%
	Neutral	15%	73.75%	11.25%
	Positive	6.25%	15%	78.75%

Table 6 Textual sentiment analysis results obtained from different TSA models designed based on various state-of-the-art NLP deep learning models on our DocImSent dataset

Text-based Sentiment Extraction Model	Accuracy (%)	Precision	Recall	F1-Score
RoBERTa [4]	88.75	0.897	0.887	0.887
DistilBERT [13]	85.83	0.863	0.858	0.857
AlBERT [14]	90.01	0.90	0.90	0.899
XLNet [15]	77.08	0.771	0.771	0.768

in Table 4, we noted that the highest accuracy for image sentiment classification of 78.3% was obtained from the Xception-based model. As demonstrated in Table 5, the Xception-based VSA model provided the best results in negative and neutral sentiment

Table 7 The confusion matrix supporting the results presented in Table 6

Text-based Sentiment Extraction Model	Classes	Negative	Neutral	Positive
RoBERTa [4]	Negative	90%	2.5%	7.5%
	Neutral	3.75%	80%	16.25%
	Positive	2.5%	1.25%	96.25%
DistilBERT [13]	Negative	90%	2.5%	7.5%
	Neutral	8.75%	77.5%	13.75%
	Positive	6.25%	3.75%	90%
AIBERT [14]	Negative	92.5%	2.5%	5%
	Neutral	7.5%	86.25%	6.25%
	Positive	2.5%	6.25%	91.25%
XLNet [15]	Negative	78.75%	8.75%	12.5%
	Neutral	22.5%	65%	12.5%
	Positive	2.5%	10%	87.5%

classification and a comparable result with other methods in positive sentiment analysis. These results were also in line with the results reported in the literature (<https://keras.io/api/applications/>).

Considering TSA results shown in Table 6, the highest results for only textual sentiment analysis were obtained from the AIBERT-based model. The RoBERTa-based model provided the second best performance. From Table 7, however, we noted that the RoBERTa demonstrated a better accuracy in positive sentiment classification and comparable results in negative and neutral sentiment classification with other methods. In addition, from Table 5, we noted that the Xception has performed better in negative and neutral sentiments but not in positive, and the VGG19 performed better in positive sentiments. Considering this information and as the RoBERTa complemented the Xception in terms of performance, we decided to combine/fuse the RoBERTa with Xception in our proposed framework. A careful observation of the results presented in Tables 4 and 6 further revealed that text elements of an image were better than visual elements in detecting the sentiment of the image more accurately. This may also be related to the fact that a massive amount of research has been done on text sentiment analysis and NLP. In contrast, VSA is lagging in terms of research compared to TSA.

5.4 Parameter settings and experimental results

As our proposed framework comprises a few parameters (T_1 , T_2 , learning rate, batch size, number of epochs, w_1 , and w_2), they were required to be fine-tuned during the training/validation process. After extensive experimentation with various values, ranging between 0 and 1 for parameters T_1 and T_2 , we determined that setting the dropout layer threshold to 0.4 yielded optimal results for both visual and textual deep learning models. As a result, T_1 and T_2 were fixed to 0.4. Additionally, we fine-tuned the Adam optimizer with a learning rate of 0.00001 for visual sentiment analysis models and 0.0001 for textual sentiment analysis models to minimize loss and enhance parameter refinement during the training

Table 8 Results obtained from the proposed framework when using different weights for the RoBERTa (for text) and Xception (for image) models on our DocImSent dataset

Weight 1 (Image)	Weight 2 (Text)	Accuracy (%)	Precision	Recall	F1-Score
0.5	0.5	92.08	0.925	0.920	0.920
0.02	0.98	89.16	0.9	0.891	0.891
0.04	0.96	89.58	0.903	0.895	0.895
0.54	0.459	91.25	0.915	0.913	0.913
0.68	0.319	85.41	0.854	0.854	0.854
0.98	0.02	78.75	0.787	0.788	0.787

Table 9 The accuracy confusion matrix obtained from the proposed framework using RoBERTa (for text)- and Xception (for image)- models on our DocImSent dataset

Classes	Negative	Neutral	Positive
Negative	93.75%	1.25%	5%
Neutral	3.75%	86.25%	10%
Positive	3.75%	0%	96.25%

process. All models underwent training for 50 epochs with a batch size of 32 to update model weights. The choice of the batch size of 32 aided in parallel processing during the training phase, which was crucial for managing memory constraints and speeding up convergence.

As mentioned, our proposed framework used weights (w_1 and w_2) to blend probabilities from distinct pipelines. To determine the most effective weight combinations for each classifier, we systematically explored all possibilities for w_1 and w_2 , ranging from 0 to 1 in increments of 0.01. Table 8 demonstrates the results achieved with various weight configurations from our proposed framework. Notably, the highest performance in terms of Accuracy, Precision, Recall, and F1-score was obtained when both RoBERTa- and Xception-based models were assigned equal weights of 0.5. Table 9 further displays the confusion matrix obtained from the fused model when w_1 and w_2 were set to 0.5, and the highest Accuracy of 92.08% was achieved.

A comparison of the results shown in Tables 4, 6 and 8 further revealed that the proposed fusion strategy combining the textual model with visual deep learning models has increased sentiment detection results, for example, improving the accuracy by more than 2%. It is also worth mentioning that textual pre-trained models used for experimentation were explicitly designed for sentiment classification. However, visual deep learning models were presumably designed for image classification and fine-tuned for sentiment classification in this research work. This indicates that if we can train a large model for image sentiment analysis with more annotated data, we can significantly increase the sentiment classification of images with text.

As demonstrated in our ablation study, we used eight visual-based deep learning models and four textual-based deep learning models, resulting in a combination of 32 models. To get an idea of their performance in our proposed framework, we performed an extensive set of experiments considering different weights on our dataset. Table 10 shows the best results obtained from each combination of models with associated weights. From Table 10, it is evident that the combination of RoBERTa and Xception provided the best results in the majority of the cases. In addition, we can observe that the second-best accuracy of

Table 10 The best results obtained from our proposed framework, considering different weights for various visual and textual sentiment detection models on our DocImSent dataset

Textual Method	Visual Method	Weight 1 (Visual)	Weight 2 (Textual)	Accuracy (%)	Precision	Recall	F1-Score
RoBERTa	VGG19	0.52	0.48	90.41	0.910	0.904	0.904
	DenseNet201	0.49	0.51	91.66	0.920	0.916	0.916
	Efficient-NetV2l	0.09	0.91	89.16	0.900	0.891	0.891
	Xception	0.5	0.5	92.08	0.925	0.920	0.920
	ResNet152V2	0.44	0.56	90	0.910	0.900	0.900
	InceptionV3	0.51	0.49	90.83	0.914	0.908	0.908
	Inception-ResNetV2	0.39	0.61	90.41	0.910	0.904	0.904
DistilBERT	MobileNetV2	0.44	0.56	90.83	0.915	0.908	0.908
	VGG19	0.53	0.47	86.66	0.867	0.866	0.866
	DenseNet201	0.55	0.44996	87.08	0.870	0.870	0.870
	Efficient-NetV2l	0.51	0.49	87.08	0.875	0.870	0.870
	Xception	0.51	0.49	88.33	0.883	0.883	0.883
	ResNet152V2	0.42	0.58	87.91	0.880	0.879	0.879
	InceptionV3	0.51	0.49	86.66	0.870	0.866	0.866
AIBERT	Inception-ResNetV2	0.51	0.49	87.91	0.881	0.879	0.878
	MobileNetV2	0.46	0.54	87.91	0.881	0.879	0.879
	VGG19	0.51	0.49	91.25	0.913	0.912	0.912
	DenseNet201	0.34	0.65999	90.83	0.908	0.908	0.908
	Efficient-NetV2l	0.03	0.97	90.41	0.904	0.904	0.903
	Xception	0.45	0.55	91.25	0.913	0.912	0.912
	ResNet152V2	0.43	0.57	91.66	0.917	0.916	0.916
XLNet	InceptionV3	0.37	0.63	91.66	0.917	0.916	0.916
	Inception-ResNetV2	0.34	0.65999	91.66	0.917	0.916	0.916
	MobileNetV2	0.35	0.65	91.25	0.913	0.912	0.912
	VGG19	0.53	0.47	83.33	0.837	0.833	0.833
	DenseNet201	0.5	0.5	82.91	0.828	0.829	0.828
	Efficient-NetV2l	0.26	0.74	77.91	0.779	0.779	0.777
	Xception	0.67	0.33	83.75	0.837	0.837	0.836
XLNet	ResNet152V2	0.39	0.61	82.91	0.829	0.829	0.828
	InceptionV3	0.39	0.61	82.08	0.822	0.820	0.819
	Inception-ResNetV2	0.49	0.51	83.75	0.837	0.837	0.837
	MobileNetV2	0.38	0.62	83.75	0.838	0.837	0.837

91.66% was obtained from the combination of RoBERTa and DenseNet201, and AIBERT and ResNet152V2, which is lower than the 92.08% accuracy obtained from our proposed framework.

5.5 Erroronous analysis

As demonstrated in Table 2, our proposed framework failed to detect sentiments correctly in some cases. For instance, the image in the second row of Table 2 was associated with the positive true label but was predicted as neutral by our framework. This discrepancy may be attributed to the image's dark and gloomy nature, which some models associate with negative or neutral sentiments. Interestingly, this highlights the lack of a semantic connection (correlation) between the visual and textual aspects of the image. The image in the ninth row of Table 2 was also labeled as negative but predicted as positive. This discrepancy arose from the contextual interpretation of the text 'Pray for Kerala,' which, when taken out of context (image), can be perceived as a positive sentiment. Our proposed framework computed a strong positive sentiment for textual elements compared to visual parts, leading to an overall positive sentiment despite the flood shown in the image.

5.6 Comparative analysis

We initially employed three recent state-of-the-art visual sentiment analysis methods [17, 18, 47] on our DocImSent dataset. The method presented in [17] was based on the DenseNet121 pretrained model to detect sentiment from images. The second method [18] was designed to detect sentiment from images based on a VGG19 network. The third method [47], the Visual Semantic Correlation Network (VSCNet), was also based on a transformer model to capture global visual features and filter out redundant and noisy visual proposals, thereby enhancing sentiment prediction. The results obtained from our proposed method and these three methods [17, 18, 47] using our dataset (DocImSent) are provided in Table 11.

Notably, the reported results for visual sentiment classification by Chandrasekaran et al. [17] and Hassan et al. [18] indicated an accuracy of 89% and 92.88% in their experimental analysis, respectively. However, their accuracy decreased substantially (77.91% and 78%) when applied to our dataset, which involves images with embedded text. Similarly, the method proposed by Zhang et al. [47] achieved an accuracy of 91.46% in their experiments on a binary sentiment classification dataset. However, the result dropped to 53.75% on our dataset containing three classes. This finding highlighted the superiority of our proposed sentiment analysis method when applied to images containing textual information.

Furthermore, methods [18, 19, and 49] are single-modality methods that classify sentiment from images only. Our proposed method combines both image and text to predict sentiment, resulting in significantly higher accuracy compared to single-modality methods, as shown in Table 11. We also implemented various approaches using only images and only text, with results presented in Tables 4 and 6. The maximum accuracy achieved using only images was

Table 11 Comparative results of the proposed and the existing methods on our DocImSent dataset

Methods	Accuracy (%)	Precision	Recall	F1-Score
Chandrasekaran et al. [17]	77.91	0.781	0.779	0.778
Hassan et al. [18]	78.00	0.783	0.783	0.78
Zhang et al. [47]	53.75	0.565	0.537	0.527
Our proposed method	92.08	0.925	0.920	0.920

around 78%, given by Xception, a pretrained CNN. For only text, the maximum accuracy was 90%, achieved by ALBERT, a pretrained text transformer. Our proposed method achieves 92%, demonstrating the effectiveness of a multi-modal approach over a single modality.

Moreover, we considered two publicly available datasets, the MVSA single and MVSA multiple, for a further comparative study. We employed our proposed framework on the MVSA single and MVSA multiple datasets and obtained results regarding accuracy and the F1-score. Table 12 demonstrates the results obtained from our proposed framework and eight state-of-the-art models reported in [27]. From Table 12, it can be observed that our proposed framework provided excellent results regarding accuracy compared to the state-of-the-art methods. Notably, as our proposed framework used a late fusion strategy (weighted average) in aggregating the results, it performed well compared to the late fusion method presented in [27]. In addition, the results in terms of F-score were comparable to the state-of-the-art methods. Our proposed framework provided lower F1-score results than some of the methods listed in Table 12 because, in our framework, the primary focus was to train the model with a high fine-grained correlation between image-text pairs. This potentially limited its performance in scenarios with less cross-modal correlation.

6 Conclusion and future work

In this research, we proposed a multimodal sentiment analysis framework based on two pipelines and the weighted fusion strategy fed by only image data. The VSA pipeline employed a modified version of the Xception model, while the TSA pipeline utilized a RoBERTa-based model. The TSA pipeline used OCR for text extraction, followed by extensive text preprocessing methods to detect sentiment from textual elements of the input image. Both models showed competitive performance in their respective domains. The weighted fusion strategy, combining predictions from the VSA and TSA pipelines, enhanced the overall sentiment analysis performance. The proposed multimodal sentiment analysis framework, combining visual and textual sentiment analysis pipelines with a weighted fusion strategy, provided promising results across various datasets, including our custom dataset (DocImSent) and MVSA datasets, demonstrated the effectiveness of our framework compared to the state-of-the-art methods in multimodal sentiment analysis in cases where images incorporate text elements.

Table 12 Comparative analysis of results on MVSA single and MVSA multiple datasets according to [27]

Dataset Method	MVSA single		MVSA multiple	
	Accuracy (%)	F1-Score	Accuracy (%)	F1-Score
SentiBank & SentiStrength [29]	52.1	0.501	65.6	0.554
CNN-Multi [30]	61.2	0.584	66.4	0.642
MultiSentiNet [31]	69.8	0.696	68.9	0.681
HSAN [32]	66.8	0.669	68.2	0.678
CoMN [33]	70.5	0.700	68.9	0.688
SFNN [28]	68.3	0.662	68.8	0.657
TL-JFT [34]	69.2	0.670	68.4	0.622
Late Fusion [27]	69.9	0.698	67.4	0.625
Proposed Method	70.9	0.683	70.7	0.641

Looking forward, enhancing visual sentiment analysis models and combining them with text sentiment analysis holds the potential for further accuracy improvements, particularly in images featuring text. In addition, further exploration should be conducted to enhance the robustness of sentiment detection models in scenarios with less cross-modal correlation between images and text. Moreover, finding a meaningful semantic relation between image and text would be a promising avenue for future research.

Data availability Comparative data is included from referenced papers, and researchers can request the custom dataset and source code after the paper is published.

Declarations

Conflict of interest The authors have no Funding and/or Conflicts of interest/Competing interests.

References

1. Yadollahi A, Shahraki AG, Zaiane OR (2017) Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)* 50(2):1–33
2. Ortis A, Farinella GM, Battiato S (2020) Survey on visual sentiment analysis. *IET Image Proc* 14(8):1440–1456
3. Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A (2023) Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges, and future directions. *Inf Fusion* 91:424–444
4. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Stoyanov V. (2019). Roberta: A robustly optimized Bert pretraining approach. arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692)
5. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI pp 1800–1807. <https://doi.org/10.1109/CVPR.2017.195>
6. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
7. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI pp 2261–2269. <https://doi.org/10.1109/CVPR.2017.243>
8. Tan M, Le QV (2019) EfficientNet: rethinking model scaling for convolutional neural networks. [ArXiv.org/abs/1905.11946](https://arxiv.org/abs/1905.11946)
9. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV pp 770–778. <https://doi.org/10.1109/CVPR.2016.90>
10. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV pp 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
11. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17). AAAI Press 4278–4284
12. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L (2018) MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT pp 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>
13. Sanh V, Debut L, Chaumond J, Wolf T (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint [arXiv:1910.01108](https://arxiv.org/abs/1910.01108)
14. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. (2019). Albert: A lite BERT for self-supervised learning of language representations. arXiv preprint [arXiv:1909.11942](https://arxiv.org/abs/1909.11942)
15. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY 517:5753–5763


16. Hussain M, Bird JJ, Faria DR (2019) A study on CNN transfer learning for image classification. In: Lotfi A, Bouchachia H, Gegov A, Langensiepen C, McGinnity M (eds) *Advances in Computational Intelligence Systems*. UKCI 2018. *Advances in Intelligent Systems and Computing*, vol 840. Springer, Cham. https://doi.org/10.1007/978-3-319-97982-3_16
17. Chandrasekaran G, Antoanela N, Andrei G, Monica C, Hemanth J (2022) Visual sentiment analysis using deep learning models with social media data. *Appl Sci* 12(3):1030
18. Hassan SZ, Ahmad K, Hicks S, Halvorsen P, Al-Fuqaha A, Conci N, Riegler M (2022) Visual sentiment analysis from disaster images in social media. *Sensors* 22(10):3628
19. Devlin J, Chang MW, Lee K, Toutanova K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
20. Arief R, Mutiara AB, Kusuma TM, Hustinawaty (2018) Automated extraction of large scale scanned document images using google vision OCR in apache hadoop environment. *Int J Adv Comput Sci Appl(IJACSA)* 9(11). <https://doi.org/10.14569/IJACSA.2018.091117>
21. Salim F, Saeed F, Basurra S, Qasem SN, Al-Hadhrani T (2023) DenseNet-201 and Xception pre-trained deep learning models for fruit recognition. *Electronics* 12(14):3132
22. Haddi E, Liu X, Shi Y (2013) The role of text pre-processing in sentiment analysis. *Procedia Computer Science* 17:26–32
23. Adoma AF, Henry N-M, Chen W (2020) Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In: 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China pp 117–121. <https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379>
24. Cortiz, D. (2021). Exploring transformers in emotion recognition: a comparison of Bert, distillbert, Roberta, Xlnet and Electra. arXiv preprint [arXiv:2104.02041](https://arxiv.org/abs/2104.02041)
25. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G Ravi S. (2020). GoEmotions: A dataset of fine-grained emotions. arXiv preprint [arXiv:2005.00547](https://arxiv.org/abs/2005.00547)
26. Kuncheva LI (2002) A theoretical study on six classifier fusion strategies. *IEEE Trans Pattern Anal Mach Intell* 24(2):281–286
27. Hung BT, Thu NHM (2024) Novelty fused image and text models based on deep neural network and transformer for multimodal sentiment analysis. *Multimed Tools Appl* 83:66263–66281. <https://doi.org/10.1007/s11042-023-18105-8>
28. Niu T, Zhu S, Pang L, El Saddik A (2016) Sentiment analysis on multi-view social data. In: Tian Q, Sebe N, Qi GJ, Huet B, Hong R, Liu X (eds) *MultiMedia Modeling, MMM 2016. Lecture Notes in Computer Science*, vol 9517. Springer, Cham. https://doi.org/10.1007/978-3-319-27674-8_2
29. Borth D, Ji R, Chen T, Breuel T, Chang S (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference* pp 223–232. <https://doi.org/10.1145/2502081.2502282>
30. Cai G, Xia B (2015) Convolutional neural networks for multimedia sentiment analysis. In: Li J, Ji H, Zhao D, Feng Y (eds) *Natural Language Processing and Chinese Computing, NLPCC 2015. Lecture Notes in Computer Science*, vol 9362. Springer, Cham. https://doi.org/10.1007/978-3-319-25207-0_14
31. Xu N, Mao W (2017) MultiSentiNet: a deep semantic network for multimodal sentiment analysis. In: *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM '17)*. Association for Computing Machinery, New York, NY pp 2399–2402. <https://doi.org/10.1145/3132847.3133142>
32. Xu N (2017) Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China pp 152–154. <https://doi.org/10.1109/ISI.2017.8004895>
33. Xu N, Mao W, Chen G (2018) A co-memory network for multimodal sentiment analysis. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. Association for Computing Machinery, New York pp 929–932. <https://doi.org/10.1145/3209978.3210093>
34. De Toledo GL, Marcacini RM. (2022). Transfer learning with joint fine-tuning for multimodal sentiment analysis. arXiv preprint [arXiv:2210.05790](https://arxiv.org/abs/2210.05790)
35. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. *J Am Soc Inform Sci Technol* 61(12):2544–2558
36. Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B (2011) Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HPLaboratories, Technical Report HPL-2011 89*
37. Rehman AU, Malik AK, Raza B, Ali W (2019) A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications* 78:26597–26613
38. Nandwani P, Verma R (2021) A review on sentiment analysis and emotion detection from text. *Soc Netw Anal Min* 11(1):81

39. You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15). AAAI Press 381–388
40. Ribeiro FN, Araújo M, Gonçalves P et al (2016) SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPI Data Sci* 5:23. <https://doi.org/10.1140/epjds/s13688-016-0085-1>
41. Feldman R (2013) Techniques and applications for sentiment analysis. *Commun ACM* 56(4):82–89. <https://doi.org/10.1145/2436256.2436274>
42. Jindal S, Singh S (2015). Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning. In *Proceedings of the International Conference on Information Processing (ICIP)*, Pune, India, , pp. 447–451. <https://doi.org/10.1109/INFOP.2015.7489424>
43. Zhang H, Liu Y, Xiong Z et al (2023) Visual sentiment analysis with semantic correlation enhancement. *Complex Intell Syst*. <https://doi.org/10.1007/s40747-023-01296-w>
44. Jiang Z, Zaheer W, Wali A et al (2024) Visual sentiment analysis using data-augmented deep transfer learning techniques. *Multimed Tools Appl* 83:17233–17249. <https://doi.org/10.1007/s11042-023-16262-4>
45. Ganesh Kumar P, S, A.A.V., V, J.P. et al (2023) A context-sensitive multi-tier deep learning framework for multimodal sentiment analysis. *Multimed Tools Appl*. <https://doi.org/10.1007/s11042-023-17601-1>
46. Zhao T, Peng J, Huang Y et al (2023) A graph convolution-based heterogeneous fusion network for multimodal sentiment analysis. *Appl Intell* 53:30455–30468. <https://doi.org/10.1007/s10489-023-05151-w>
47. Zhang H, Liu Y, Xiong Z, Wu Z, Xu D (2024) Visual sentiment analysis with semantic correlation enhancement. *Complex & Intelligent Systems* 10(2):2869–2881
48. Tiruwa A, Yadav R, Suri PK (2020) Sentiment analysis: an effective way of interpreting consumer's inclinations towards a brand. In: Suri P, Yadav R (eds) *Transforming Organizations Through Flexible Systems Management*. Flexible Systems Management. Springer, Singapore. https://doi.org/10.1007/978-981-13-9640-3_12
49. Kauffmann E, Peral J, Gil D, Ferrández A, Sellers R, Mora H (2019) Managing marketing decision-making with sentiment analysis: An evaluation of the main product features using text data mining. *Sustainability* 11(15):4235
50. Ansari MZ, Aziz MB, Siddiqui MO, Mehra H, Singh KP (2020) Analysis of political sentiment orientations on Twitter. *Procedia computer science* 167:1821–1828
51. Wang W, Han C, Zhou T, Liu D (2022) Visual recognition with deep nearest centroids. <https://doi.org/10.48550/arXiv.2209.07383>
52. Yan L, Ma S, Wang Q, Chen Y, Zhang X, Savakis A, Liu D (2022) Video Captioning Using Global-Local Representation. *IEEE Trans Circuits Syst Video Technol* 32(10):6642–6656. <https://doi.org/10.1109/TCSVT.2022.3177320>
53. Yan L, Han C, Xu Z, Liu D, Wang Q (2023) Prompt learns prompt: exploring knowledge-aware generative prompt collaboration for video captioning pp 1622–1630. <https://doi.org/10.24963/ijcai.2023/180>
54. Yan L, Wang Q, Ma S, Wang J, Yu C (2022) Solve the puzzle of instance segmentation in videos: A weakly supervised framework with spatio-temporal collaboration. *IEEE Trans Circuits Syst Video Technol* 33(1):393–406
55. Shao Z, Han J, Marnerides D, Debattista K (2022) Region-object relation-aware dense captioning via transformer. In: *IEEE Transactions on Neural Networks and Learning Systems*. <https://doi.org/10.1109/TNNLS.2022.3152990>
56. Shao Z, Han J, Debattista K, Pang Y (2023) Textual context-aware dense captioning with diverse words. *IEEE Trans Multimedia* 25:8753–8766
57. Shao Z, Han J, Debattista K, Pang Y (2024) DCMSTRD: end-to-end dense captioning via multi-scale transformer decoding. In: *IEEE Transactions on Multimedia* 26:7581–7593. <https://doi.org/10.1109/TMM.2024.3369863>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Garvit Ahuja¹  · Alireza Alaei² · Umapada Pal³

✉ Garvit Ahuja
garvit.ahuja02@gmail.com; garvit.219310157@muj.manipal.edu

Alireza Alaei
ali.alaei@scu.edu.au

Umapada Pal
umapada@isical.ac.in

¹ School of Computer Science and Engineering, Manipal University Jaipur, Jaipur, India

² Faculty of Science and Engineering, Southern Cross University, Gold Coast, Australia

³ Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India