MeViTSA: Multimodal Ensemble Approach for Sentiment Analysis of Visually Integrated Text Data

Soham Bhattacharya[1,4][0009−0004−0658−7049], Ali Reza Alaei[2][0000−0003−1669−2606], Sukalpa Chanda[3][0000−0002−9068−5845], and Umapada Pal[4][0000−0002−5426−2618]

[1] Ramakrishna Mission Vivekananda Educational and Research Institute Belur, India bhattacharyasoham026@gmail.com
[2] Southern Cross University, Australia ali.alaei@scu.edu.au
[3] Østfold University of Applied Science, Norway sukalpa@ieee.org
[4] Indian Statistical Institute Kolkata, India umapada@isical.ac.in

Abstract. Existing Multimodal Sentiment Analysis (MSA) systems often have difficulty in understanding images with embedded text, like memes, ads and infographics. Existing methods considered any embedded text on an image to be some visual noise. This study presents MeViTSA, a dual-stream multimodal ensemble framework created to distinctly separate and analyze visual and linguistic meanings. The system employs a CLIP encoder for processing the input image and a T5 based classifier for handling the textual data extracted using Google Cloud Vision OCR from the input image, and combined through a fixed late fusion approach. A notable advancement is the addition of a text-extraction fallback feature, that is if OCR fails to find any text, a BLIP powered component creates a caption acting as a substitute textual cue guaranteeing that it improves architectural robustness and maintains a continuous semantic stream for text-deficient images. To rigorously evaluate this approach, we present DocImSentv1, a manually curated and balanced dataset of 2,054 images refined from the DocImSent benchmark. Extensive experiments demonstrate that MeViTSA establishes a new state-of-the-art, achieving 96.59% accuracy on DocImSentv1 and outperforming existing baselines on the widely used MVSA-Single and MVSA-Multiple datasets. Our dataset, DocImSentv1, will be available after the acceptance of this paper. The corresponding code repository can be accessed here.

Keywords: Multimodal sentiment analysis · Ensemble model · Late fusion

1 Introduction

While sentiment analysis (SA) has matured significantly within the Natural Language Processing domain [1,19], traditional text-centric models are increasingly inadequate for the contemporary digital landscape. The proliferation of user-generated content across social media has shifted the focus toward Multimodal Sentiment Analysis (MSA), which aims to bridge the discrepancy between unimodal signals and the complex, multi-layered human emotions expressed through interleaved text, audio, and visual streams [25]. Despite the development of large-scale datasets like CMU-MOSEI [25] and MVSA [15], a fundamental limitation still persists. Most of the existing frameworks

rely on a single, isolated modality or assume the presence of all modalities during inference [9,15]. This reliance fails to capture the intricate interplay in visually-integrated textual data, where the sentiment is often contingent on the synergy between the image and the text embedded within it [1].

So despite significant advancements in multimodal sentiment analysis, a persistent gap still remains in developing generalizable models. Much of the existing literature has focused on domain-specific datasets, limiting the real-world applicability of proposed frameworks. For example, there is limited MSA research on visually integrated textual data in images [1].

Inspired by Ahuja's [1] work, in this paper, we propose a framework designed primarily to address the complex task of analyzing images with visually-integrated text. To ensure broader applicability, the architecture includes a fallback stream that allows the system to process text-deficient images (images having no visually-integrated text) by generating synthetic textual proxies, thereby maintaining a multimodal workflow regardless of the input type.

Our primary contributions are threefold: **(i)** the development of an enhanced MSA framework that demonstrates superior performance over the existing baselines, **(ii)** the integration of a novel Text-Extraction Fallback (TEF) mechanism that ensures architectural robustness against image instances where textual content is unreadable or absent, allowing the system to process any image instance uniformly, and **(iii)** extending and refining the dataset DocImSent introduced by [1] through comprehensive curation to improve data quality and volume of the dataset.

The primary novelty of our work lies in the TEF mechanism, which transforms a missing-modality problem into a proxy-modality solution using BLIP [13], an image captioning model.

The remainder of the paper is organized as follows: Section 2 provides an overview of the literature. Section 3 presents the proposed approach, followed by Section 4, which outlines the experimental setup and results. Finally, Section 5 concludes the paper mentioning the key findings and suggests directions for future research in sentiment analysis.

2 Related Work

The landscape of sentiment analysis has transitioned from isolated unimodal assessments to integrated systems that mirror human perception [9]. Modern research is currently defined by three converging streams: textual sentiment analysis, visual sentiment analysis, and the complex synthesis of these signals within multimodal frameworks [9]. Early foundations [16] established the necessity of polarity detection, which has since evolved into systematic reviews of multimodal fusion techniques that prioritize intermodality interactions. While current research often emphasizes large-scale language understanding, a primary challenge remains in developing models that can contextually bridge the gap between linguistic intent and its physical manifestation in digital media.

In parallel, Visual Sentiment Analysis (VSA) emerged to bridge the "affective gap" between raw pixel data and human emotional response. Foundational semantic frameworks, such as the Large-scale Visual Sentiment Ontology [2], pioneered the use of

Adjective-Noun Pairs (ANPs) to move beyond simple global image statistics. This inquiry has expanded into specialized domains, including the analysis of sentiment in disaster-related social media imagery [10] and general deep learning applications for diverse social datasets [4]. However, even state-of-the-art transformers like SentiFormer [8] and contemporary models focusing on semantic correlation enhancement [26] often struggle when textual information is embedded within the visual field, frequently misidentifying overlaid characters as non-semantic visual noise.

Multimodal Sentiment Analysis (MSA) serves as the critical junction where linguistic and visual signals are integrated to approximate human understanding. Early milestones in this synthesis utilized convolutional neural networks for multimedia sentiment analysis [3], which evolved through the study of multi-view social data [15]. Subsequent advancements introduced deep semantic networks like MultiSentiNet [23] and hierarchical attentional frameworks [22] to capture fine-grained cues. This progression has culminated in sophisticated co-memory networks [24] and transfer learning schemes employing joint fine-tuning [6] to align disparate modalities. While these methods have improved performance on standard benchmarks like CMU-MOSEI [25], they predominantly operate on datasets where text and images are treated as clean, independent inputs.

Despite this progress, a substantial research gap persists regarding visually integrated text, which is characteristic of modern formats like memes, infographics, and digital posters. Recent studies, such as the framework for images containing textual information [1], have begun to address this by acknowledging the interdependence of embedded words and their visual surroundings. Further innovations in novelty fusion [11] have combined deep neural networks with transformers to enhance classification robustness. However, many systems still lack the architectural synergy to treat integrated text as a primary, distinct modality. This necessitates a more robust multimodal ensemble approach that preserves the syntactic integrity of the language while simultaneously interpreting the visual semantics of the surrounding scene, ensuring the contextual interplay between overlaid words and background imagery is fully captured.

Our proposed framework addresses these specific deficiencies by implementing a dual-stream framework that explicitly handles the extraction and analysis of visually integrated text. By bridging the gap between unimodal textual analysis and global visual classification, we ensure that neither the semantic meaning of the words nor the emotional resonance of the imagery is sacrificed. This methodology establishes a new benchmark for processing complex, text-rich visual data, moving beyond the limitations of existing siloed datasets and standard fusion strategies.

3 The proposed framework

We introduce our proposed framework, the Multimodal Ensemble for Visually Integrated Text data for Sentiment Analysis ("`MeViTSA`"), designed to enhance sentiment analysis in this domain. It is important to note that we refer to images containing embedded text (visually integrated text) inside it as *Visually Integrated Text data*. The primary objective of `MeViTSA` is to accurately classify sentiment from single, static images that contain visually-integrated text (e.g., memes, advertisements, or product screen-

shots). This problem definition notably diverges from standard multi modal tasks that rely on separate, paired image-text data or tagged data and addresses sentiment analysis of images in a much more generalized context.

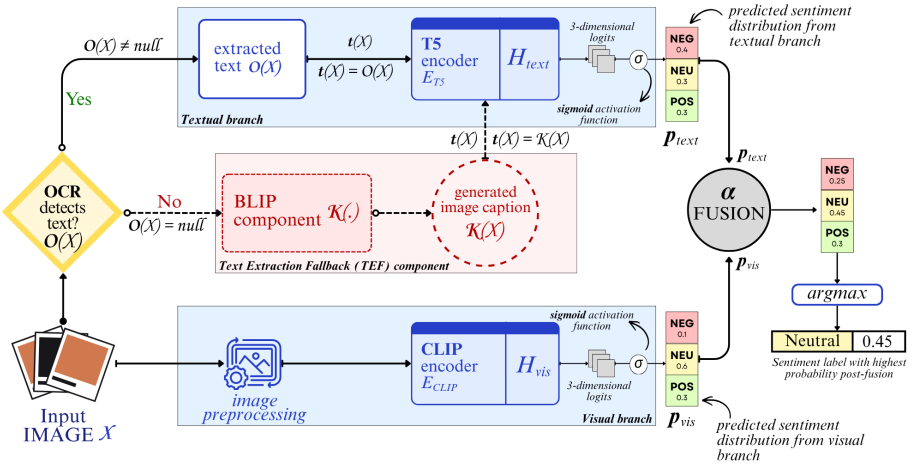

Fig. 1: MeViTSA framework design (Use 200% zoom for better visibility)

The proposed framework is depicted in Figure 1. "MeViTSA" is a dual-branch framework, where an input image is processed in parallel through two specialized pathways: a textual branch and a visual branch. The outputs of these branches are then integrated at a final fusion stage to produce a single sentiment classification.

3.1 Mathematical formulation of the proposed framework

To rigorously define the framework of MeViTSA, we formulate the sentiment analysis task as learning a mapping function $\mathcal{F} : \mathcal{I} \rightarrow \mathcal{Y}$, where $\mathcal{I}$ represents the input space as the set of all images (e.g., memes, infographics), and $\mathcal{Y} = \{0, 1, 2\}$ denotes the discrete label space corresponding to negative, neutral, and positive sentiments respectively. While our focus is on images containing visually integrated text, the system handles the domain $\mathcal{I}$ uniformly through a conditional mapping. Also we denote the softmax activation function as $\sigma(\cdot)$.

The framework processes an input instance $x \in \mathcal{I}$ through two three-dimensional probability distribution vectors, $\mathbf{p}_{text}$ and $\mathbf{p}_{vis}$, before aggregating them into a final decision. The vectors $\mathbf{p}_{text}$ and $\mathbf{p}_{vis}$ denote the predicted probability distributions over the three sentiment classes for the textual and visual branches respectively. Hence, $\mathbf{p}_{text}, \mathbf{p}_{vis} \in [0, 1]^3$.

The textual branch is designed to handle input variability through a conditional extraction mechanism. Let $\mathcal{O}(\cdot)$ represent the OCR API function and $\mathcal{K}(\cdot)$ represent the

image captioning function (Bootstrapping Language-Image Pre-training (BLIP) [13]). We define the intermediate textual representation t_x as given in Equation 1:

$$t_x = \begin{cases} \mathcal{O}(x) & \text{if } \mathcal{O}(x) \neq \emptyset \\ \mathcal{K}(x) & \text{if } \mathcal{O}(x) = \emptyset \end{cases} \tag{1}$$

Formally, $O(x) = \emptyset$ occurs when the OCR API returns an empty text-annotations object, at which point t_x takes the value of the synthetic caption $K(x)$ produced by the BLIP component.

This conditional logic ensures domain continuity in cases of OCR failure or text-deficient input image. The extracted token sequence t_x is then processed and mapped to a high-dimensional semantic embedding space by the T5 encoder [18], denoted as E_{T5}, followed by a transformation to a 3-dimensional vector space via a classification head H_{text}. The textual probability distribution vector $\mathbf{p}_{text} \in [0, 1]^3$ is computed by applying $\sigma(\cdot)$ on the output of the H_{text} as given in Equation 2:

$$\mathbf{p}_{text} = \sigma(H_{text}(E_{T5}(t_x))) \tag{2}$$

Parallel to the textual branch, the visual branch operates directly on the raw image x. We utilize the pre-trained CLIP image encoder [17], denoted as E_{CLIP}, to project the image into a high-dimensional semiotic embedding space. This embedding is subsequently mapped to a 3-dimensional vector via a dedicated visual classification head H_{vis}. This 3-dimensional vector is then mapped to the visual probability distribution vector $\mathbf{p}_{vis}$ via the $\sigma(\cdot)$ as expressed in Equation 3:

$$\mathbf{p}_{vis} = \sigma(H_{vis}(E_{CLIP}(x))) \tag{3}$$

The final predictive layer aggregates the unimodal distributions through a linear convex combination. The fusion function is governed by the scalar hyperparameter α that denotes the weight assigned to the textual branch probability predictions. This yields the final probability vector $\mathbf{p}_{final}$, as given in Equation 4:

$$\mathbf{p}_{final} = \alpha \cdot \mathbf{p}_{text} + (1 - \alpha) \cdot \mathbf{p}_{vis} \tag{4}$$

The predicted sentiment class $\hat{y}$ is determined by selecting the index with the maximum probability mass, as given in Equation 5:

$$\mathcal{F}(x) = \hat{y} = \underset{c \in \mathcal{Y}}{\text{argmax}}(\mathbf{p}_{final}^{(c)}) \tag{5}$$

In the above Equation 5, $\mathbf{p}_{final}^{(c)}$ refers the probability associated with class c. The above formulation ensures that the final decision boundary is a weighted consensus of the semantic (textual) and semiotic (visual) signals extracted from the input.

3.2 Textual branch with text-extraction fallback

The textual branch of our framework is designed with a specialized dual-pathway setup to ensure high reliability across varied visual inputs. The process begins with an OCR

checkpoint utilizing the Google Cloud Vision API (link available here), which was selected over traditional tools like Tesseract due to its superior performance in scene-text environments, achieving an accuracy of 84.43% compared to Tesseract's significantly lower accuracy of 47.02% [21]. To ensure framework reproducibility and out-of-the-box utility, the API was utilized with its default internal thresholding settings. This checkpoint acts as a logistical switch: if the OCR API returns a non-empty text-annotations object, then the system tokenizes the text into PyTorch tensors for direct analysis. If the OCR API returns a null text-annotations object, the system activates the Text-Extraction Fallback (TEF) pathway (the TEF component), utilizing the BLIP (Bootstrapping Language-Image Pre-training) model [13]. BLIP serves as a sophisticated semantic proxy, that translates the visual scene into a linguistic caption, effectively converting a purely visual signal into a textual format that the subsequent textual sentiment analysis branch can interpret. This ensures that every image produces a valid textual feature vector, regardless of whether it contains explicit characters.

The core analytical engine of this branch is the T5 (Text-to-Text Transfer Transformer) model [18]. Unlike standard encoder-only models such as BERT or RoBERTa [14], T5 utilizes a unified text-to-text objective that facilitates a deeper contextual understanding of the input. We adapted the T5-base architecture by adding a dedicated classification head and a softmax layer to project inputs into a 3-class probability distribution (negative, neutral, and positive). The model processes sequences up to 128 tokens, generating a 768-dimensional embedding vector for prediction. For training, we utilized the DocImSentv1 dataset, an enhanced version of the original DocImSent [1] repository.

3.3 Visual branch

For the visual branch, the framework incorporates Contrastive Language-Image Pre-training (CLIP) [17]. We selected CLIP over traditional Convolutional Neural Networks (e.g., ResNet, DenseNet) or standard Vision Transformers [7] (ViT) because of its unique pre-training objective. Unlike standard visual encoders trained solely on fixed object labels (e.g., ImageNet), CLIP is trained on massive image-text pairs to align visual and textual representations in a shared semantic space. This architectural choice is supported by comparative benchmarks conducted by Radford et al. [17], which demonstrated that CLIP matches the performance of a fully supervised ResNet-50 on ImageNet while significantly outperforming it on robustness tests and zero-shot transfer capabilities. This alignment is critical for our task, as the sentiment in our dataset is derived from the interaction between embedded text and visual content. CLIP's semantically rich embeddings are better suited to bridge this modality gap than unimodal visual encoders.

So, we utilized the CLIP visual encoder based on the ViT-B/32 backbone [7], which accepts images resized to 224×224 pixels and outputs a 512-dimensional embedding vector. The input image is first processed using OpenAI's CLIP preprocessing pipeline [17], ensuring consistency with the model's expected normalization and resizing, and is then passed through the pretrained CLIP image encoder to obtain a semantic-rich visual representation. This embedding is fed into a custom classification head, followed by a softmax layer to produce a three-class sentiment probability distribution based

solely on visual cues. The visual model was fine-tuned on the DocImSentv1 dataset, and the choice of this architecture was further validated by the experimental results on our dataset (Section 4).

3.4 Late fusion

Multi-modal fusion strategies generally follow three paths: early, intermediate, and late fusion. While early fusion risks modality dominance and intermediate approaches like cross-modal attention add heavy computational costs [9], we utilized a late fusion (decision-level) strategy for the MeViTSA framework. This choice was driven by the need for architectural independence; by processing the visual and textual branches separately, we prevent noise in one modality from destabilizing the feature extraction of the other. This modularity ensures that the final sentiment classification remains robust even if one branch encounters missing or ambiguous data. The integration of these independent signals is achieved through a weighted sum of their probability distributions, as given in Equation 4.

Our experimental results (refer to Section 4.5) indicated that a choosing a fixed scalar weight of $\alpha = 0.5$ over adaptive weighted strategy provided the most stable performance. This equal weighting acts as a strong regularization prior, preventing any modality collapse which is a common failure where a model over-relies on a single, easier modality at the expense of the other.

4 Experimental results and discussion

4.1 Dataset and evaluation metrics

The landscape of multimodal sentiment analysis is frequently hindered by a shortage of domain-agnostic datasets that specifically address hybrid, text-rich images. While existing resources, such as, image dataset MVSA [15] and video dataset CMU-MOSEI [25] provide valuable associations between separate modalities or acoustic signals, they often lack focus on static images where text is an inseparable, visually integrated component. To address these drawbacks, we utilized the DocImSent dataset introduced by Ahuja et al. (2025) [1] as a baseline. Originally containing 1,717 images across memes, infographics, and advertisements, this dataset provided a foundational 3-class sentiment distribution (positive, neutral, and negative). However, our initial audit revealed data redundancy issues, including 30 duplicate pairs identified through a cryptographic deduplication protocol using MD5 hashing-based analysis, which necessitated a rigorous refinement process.

One of the contributions of our work is the creation of DocImSentv1, a manually curated and expanded corpus designed to serve as a high-quality benchmark for the MeViTSA framework. We discarded the identified duplicates and integrated 367 new images annotated manually. To ensure label reliability, we employed a consensus-based review where any ambiguous instances were removed, resulting in a finalized corpus of 2,054 images. The final distribution remains nearly balanced, with 31.5% negative, 31.9% neutral, and 36.6% positive instances. For evaluation, we adopted a standard

5-fold cross validation strategy and utilized accuracy and macro-averaged precision, recall, and F1-score as metrics [9]. These were calculated both per-class and as overall averages to ensure a comprehensive assessment of the framework's performance and to avoid biases inherent in single-metric reporting.

4.2 MeViTSA training and parameter settings

The MeViTSA framework was evaluated using a stratified 5-fold cross-validation strategy to ensure a rigorous and fair representation of sentiment classes. All experiments were conducted on an NVIDIA T4 GPU. The framework maintains two independent branches: a visual branch leveraging the CLIP ViT-B/32 [17] encoder and a textual branch centered on the T5-base [18] transformer. Data preprocessing was tailored to each modality's specific requirements. The text branch utilized a pretrained Hugging Face tokenizer with a maximum sequence length of 128 tokens, while the image branch employed OpenAI's CLIP preprocessing pipeline for standardized normalization and resizing. The CLIP visual encoder projected the preprocessed image into a embedding of 512 dimensions followed by a custom classification head, while the T5 encoder generated 768-dimensional textual embeddings. Both branches were trained and optimized independently on our dataset DocImSentv1 using the AdamW optimizer with a learning rate of 2e-4 and a batch size of 16 over 25 epochs. Following independent training, the final sentiment prediction was derived through a late fusion mechanism. This process involved a simple averaging of the predicted probability distributions from both branches, using a fixed scalar weight of $\alpha = 0.5$.

The fallback $K(x)$ operator utilized the BLIP [13] model. To ensure the determinism of the generated captions and eliminate stochastic variance, the model was configured with greedy decoding (do_sample=False, num_beams=1), ensuring the most probable literal description was consistently selected as the semantic proxy.

Final performance was assessed using a comprehensive suite of metrics, including accuracy and macro-averaged precision, recall, and F1-score, to capture the framework's effectiveness across the positive, neutral, and negative sentiment classes.

4.3 Experimental results based on component combinations

As mentioned, our proposed framework contains textual and visual sentiment analysis branches. Based on state-of-the-art models and their corresponding performance, we tried a handful of recent and advanced textual (DistilRoBERTa [20], BART [12], RoBERTa [14] and T5 [18]) and visual (CLIP [17], ViT-L/16 [7] and Xception [5]) sentiment analysis models for the purpose of selecting the best model for each of the branches of our proposed framework. Each of the model was fine-tuned on dataset, DocImSentv1, independently on the extracted text data from the image for our textual branch, and the entire image for our visual branch. Likewise, the accuracies and necessary evaluation metrics, such as, precision, recall and F1-score for each of the models for each branch were noted down. The performance results for each of the fine-tuned models for the textual (T) and visual branch (V) of our framework have been noted down in Table 1.

Table 1: Textual and visual sentiment analysis results (in %) of various textual and visual models fine-tuned for the *textual* (T) and *visual* (V) sentiment branch of our framework and on our dataset DocImSentv1

Model-name	**Branch**	**Accuracy**	**Precision**	**Recall**	**F1-score**
BART [12]	T	90.1	90.2	90.1	90.0
DistilRoBERTa [20]	**T**	**92.0**	**92.1**	**91.9**	**91.9**
RoBERTa [14]	T	89.8	90.8	89.5	89.5
T5 [18]	T	90.5	90.5	90.4	90.4
CLIP [17]	**V**	**91.3**	**92.0**	**92.0**	**91.0**
ViT-L/16 [7]	V	82.9	83.0	83.0	83.0
Xception [5]	V	74.3	75.0	74.0	74.0

A critical aspect of our architectural design is the independence of these branches prior to fusion. Since the textual and visual data streams arise from the fission of a single source image, we initially hypothesized that combining the highest-performing individual models would naturally yield the superior ensemble. To validate this, we conducted a comprehensive inference experiment testing all possible pairings of the fine-tuned models (results tabulated in Table 2). Contrary to our hypothesis, the combination of the top-performing individual models (CLIP + DistilRoBERTa) ranked second in the fused setting. The CLIP + T5 combination emerged as the most effective ensemble, achieving the highest overall accuracy. Consequently, we selected CLIP + T5 for the final MeViTSA framework. This decision is justified by the empirical evidence that ensemble performance depends on complementarity rather than just individual strength; T5 and CLIP demonstrated a superior ability to correct each other's mis-classifications compared to other pairings.

Table 2: Classification results (in %) of paired fine-tuned models on our dataset DocImSentv1

Visual branch	**Textual branch**	**Accuracy**	**Precision**	**Recall**	**F1-score**
CLIP	BART	94.30	94.26	94.30	94.30
	DistilRoBERTa	**94.93**	**94.91**	**94.94**	**94.91**
	RoBERTa	94.35	94.32	94.35	94.32
	T5	**96.35**	**96.32**	**96.57**	**96.39**
ViT-L/16	BART	93.96	93.96	93.92	93.94
	DistilRoBERTa	**94.59**	**94.61**	94.52	**94.56**
	RoBERTa	94.06	94.05	94.01	94.03
	T5	**94.59**	**94.61**	**94.53**	**94.57**
Xception	BART	88.60	88.87	88.39	88.52
	DistilRoBERTa	89.48	89.70	89.26	89.40
	RoBERTa	89.89	89.33	88.88	89.02
	T5	89.48	89.71	89.25	89.39

Detailed inspection of the confusion matrices based on the performances (refer to Table 2) of the component combination revealed that T5 provided stronger and more consistent textual sentiment predictions than DistilRoBERTa, particularly for the negative (positive) class, where T5 achieved a high correct classification rate of 90% (95%). When combined with CLIP, which demonstrated strong visual discrimination, especially for the neutral class (98%), T5 complemented CLIP by reducing textual ambiguity and improving class separation. In contrast, DistilRoBERTa exhibited slightly higher confusion across classes, which limited the overall benefit when paired with CLIP. Consequently, the CLIP–T5 combination offered a more balanced and reliable multimodal representation, providing higher accuracy than CLIP combined with DistilRoBERTa.

4.4 Comparative analysis

We conducted comparative analysis to compare our method with other benchmarks. Detailed comparative results are given in Tables 4 and 5. To ensure the validity of these comparisons, we reimplemented the baseline architectures following the original authors' specifications and performed extensive hyperparameter tuning for each model on the DocImSentv1 corpus. This approach guarantees that the baselines were evaluated at their optimal capacity for this specific dataset.

From Table 4, it can be seen that the proposed MeViTSA framework achieves state-of-the-art performance on the DocImSentv1 dataset, reaching an accuracy of 96.59%. This result significantly exceeds the performance of its closest competitor, Ahuja et al. [1], whose method achieved 89.89% on our dataset. The primary distinction lies in MeViTSA's Text-Extraction Fallback component; whereas previous models fail when embedded text is stylized or undetectable, our integration of the BLIP captioning module ensures a semantic textual description is always generated. This architectural safeguard effectively bridges the gap left by failed OCR extractions, resulting in a performance boost of over 6% and high statistical reliability, as evidenced by our 5-fold cross-validation results with standard deviations below 0.005 across all metrics (refer to Table 3).

Table 3: Statistical reliability analysis of our method using five-fold cross-validation on our dataset DocImSentv1

Metric	*Fold 1*	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean* (μ)	Standard deviation (σ)
Accuracy	0.9635	0.9732	0.9635	0.9635	0.9659	**0.9659**	±0.0042
Precision	0.9631	0.9729	0.9636	0.9633	0.9659	**0.9658**	±0.0041
Recall	0.9655	0.9743	0.9655	0.9658	0.9680	**0.9678**	±0.0038
F1-score	0.9637	0.9735	0.9641	0.9639	0.9665	**0.9663**	±0.0041

Further comparative analysis (refer to Table 4) underscores the limitations of "tag-based" and purely visual architectures. For instance, VSCNet [26] approach by Zhang et al. achieved only 81.55% accuracy, largely because its method of converting affective

regions into simple object tags, such as “person” or “computer”, results in the loss of complex linguistic signals like sarcasm. Similarly, works such as [4], involving standard CNNs like DenseNet and ResNet struggle with the DocImSentv1 corpus as they lack a dedicated branch for visually integrated text in images.

Table 4: Comparative results (in %) of the proposed and the existing methods on our dataset DocImSentv1

Method name	**Accuracy**	**Precision**	**Recall**	**F1-score**
Zhang *et al.* [26]	81.55	81.75	81.37	81.46
Chandrasekaran *et al.* [4]	77.35	77.06	77.20	76.92
Ahuja *et al.* [1]	89.89	89.33	88.88	89.02
Our proposed method (MeViTSA)	**96.59**	**96.58**	**96.78**	**96.63**

We evaluated MeViTSA’s performance through direct inference on the MVSA-Single and MVSA-Multiple [15] benchmarks (refer to Table 5), without any data specific fine-tuning, utilizing the weights previously optimized on DocImSentv1. This zero-shot transfer demonstrated that MeViTSA continued to outperform recent models, including the late fusion method by Hung et al. [11] and joint optimization strategies like TL-JFT [6]. By utilizing a fixed-weight late fusion strategy, the framework effectively mitigated modality dominance, allowing for balanced representation without overfitting to the inconsistencies inherent in paired image-text social media data.

Table 5: Comparative results (in %) of the proposed and existing methods on **MVSA-Single** & **MVSA-Multiple** datasets

Methods	MVSA-Single		MVSA-Multiple	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**
TL-JFT [6]	69.2	67.0	68.4	62.2
Hung *et al.* [11]	69.9	69.8	67.4	62.5
Ahuja *et al.* [1]	70.9	68.3	70.7	64.1
Our proposed method (MeViTSA)	**79.1**	**71.6**	**73.9**	**65.5**

4.5 Ablation studies

Text-Extraction Fallback (TEF) component The Text-Extraction Fallback (TEF) component was designed to serve as a fallback mechanism for handling images where the OCR engine fails to detect text or a text-deficient image, which does not contain any visually integrated text in it. By utilizing the BLIP [13] model, the framework generates synthetic captions that act as linguistic proxies for the textual branch.

Quantitative results (refer to Table 6) demonstrate that TEF provides a measurable performance gain, increasing overall accuracy from 95.62% to 96.59%. These metrics were calculated as the mean performance across the held-out validation folds of our 5-fold cross-validation regimen to ensure the findings are generalizable and free from training-bias. A detailed inspection of the confusion matrices reveals that rather than introducing bias, the TEF module improves classification precision across categories. When TEF was enabled, negative instances misclassified as neutral dropped significantly from 32 to 21. Similarly, misclassifications of positive instances as neutral decreased from 37 to 35. These findings suggest that the descriptive captions generated by BLIP successfully preserve semantic context, allowing the textual branch to contribute meaningfully to the final sentiment prediction even in the absence of explicit overlaid characters.

Table 6: Classification results (in %) for two cases of our framework, with TEF enabled and disabled

Evalutation metric	**TEF enabled**	**TEF disabled**
Macro precision	96.58	95.75
Macro recall	96.78	95.66
Macro f1-score	96.63	95.61
Accuracy	**96.59**	**95.62**

Scalar hyperparameter α The scalar hyperparameter α serves as the decision-level fusion weight, where α is assigned to the textual branch probability distribution p_{text} and $(1 - \alpha)$ is assigned to the visual branch p_{vis}. It plays a critical role in balancing the influence of textual and visual modalities during the fusion process. Initially, we hypothesized that an adaptive heuristic, setting α based on the text-to-image area ratio, would enhance the framework's sensitivity to varying content densities. To validate this, we conducted an empirical grid search on our dataset, DocImSentv1, varying α from 0.1 to 0.9 with a constant increment of 0.1. To ensure a rigorous tuning protocol, each α configuration was evaluated using the same stratified 5-fold cross-validation regimen described in Section 4.2, with the reported metrics representing the mean performance across all validation folds. These performance metrics for each configuration were noted down (refer to Table 7). The results demonstrate that an α value of 0.5 yields to be the optimal choice, achieving a peak overall accuracy of 96.59%. This finding suggests that a balanced contribution from both modalities is essential for interpreting visually integrated text data accurately. Furthermore, these experimental outcomes indicate that a fixed weighted strategy provides superior stability and performance compared to the area-based adaptive weighting heuristic.

5 Conclusion and future work

In this paper, we introduced a new dual-branch multimodal framework designed primarily for the challenging task of sentiment analysis on images containing embedded text,

Table 7: Performance results (in %) of our proposed framework under different fusion weight (α) values on the our dataset DocImSentv1

α	**Accuracy**	**Precision**	**Recall**	**F1-Score**
0.1	92.50	92.62	92.44	92.44
0.2	93.43	93.54	93.38	93.38
0.3	94.40	94.47	94.37	94.37
0.4	95.67	95.66	95.66	95.64
0.5	**96.59**	**96.58**	**96.78**	**96.63**
0.6	96.54	96.49	96.58	96.52
0.7	96.45	96.39	96.49	96.43
0.8	96.35	96.28	96.38	96.32
0.9	96.34	96.26	96.39	96.32

along with handling text-deficient images to ensure wide applicability of the system. Our proposed approach is designed based on two distinct processing streams for visual and textual signals to successfully leverage this untapped information. By uniformly fusing the predictions from both branches, MeViTSA sets a new state-of-the-art on public benchmarks like MVSA-Single and MVSA-Multiple datasets. Across all experimental scenarios on our DocImSentv1 dataset, our proposed framework proved highly effective, achieving a maximum accuracy of **96.59%**.

For future work, the static alpha parameter could be replaced by exploring more sophisticated, attention-based fusion mechanisms to dynamically weigh the visual and textual branches on a per-sample basis. End-to-end architectures that learn to read and see simultaneously could also be investigated, which would mitigate the reliance on a discrete OCR step. Finally, the MeViTSA framework can be further enhanced by integrating a caption-generation model capable of producing more reliable and sentiment-aware captions.

References

1. Ahuja, G., Alaei, A., Pal, U.: A new multimodal sentiment analysis for images containing textual information. Multimedia Tools and Applications **84** (2025)
2. Borth, D., Ji, R., Chen, T., et al.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM International Conference on Multimedia (2013)
3. Cai, G., Xia, B.: Convolutional neural networks for multimedia sentiment analysis. NLPCC 2015 (2011)
4. Chandrasekaran, G., Antoanela, N., Andrei, G.: Visual sentiment analysis using deep learning models with social media data. Applied Sciences **12** (2022)
5. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. CoRR **abs/1610.02357** (2016)
6. De Toledo, G.L., Marcacini, R.M.: Transfer learning with joint fine-tuning for multimodal sentiment analysis. arXiv preprint arXiv:2210.05790 (2022)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020)

8. Feng, B., Ruan, S., Yang, M., et al.: Sentiformer: Metadata enhanced transformer for image sentiment analysis. In: Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (2025)
9. Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A.: Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion **91** (2023)
10. Hicks, H.A.: Visual sentiment analysis from disaster images in social media. Sensors **22** (2022)
11. Hung, B.T., Thu, N.H.M.: Novelty fused image and text models based on deep neural network and transformer for multimodal sentiment analysis. Multimedia Tools and Applications **83** (2024)
12. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
13. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the 39th International Conference on Machine Learning. vol. 162 (2022)
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. CoRR **abs/1907.11692** (2019)
15. Niu, T., Zhu, S., Pang, L., El Saddik, A.: Sentiment analysis on multi-view social data. In: MultiMedia Modeling. MMM 2016. LNCS, vol 9517 (2016)
16. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2** (2008)
17. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. CoRR **abs/2103.00020** (2021)
18. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: T5: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21** (2020)
19. Ranaware, G.: Multimodal sentiment analysis on cmu-mosei dataset using transformer based models. arXiv preprint arXiv:2507.06110 (2025)
20. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. CoRR **abs/1910.01108** (2019)
21. Thammarak, K., Kongkla, P., Sirisathitkul, Y., Intakosum, S.: Comparative analysis of tesseract and google cloud vision for thai vehicle registration certificate. International Journal of Electrical and Computer Engineering **12** (2021)
22. Xu, N.: Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In: 2017 IEEE International Conference on Intelligence and Security Informatics (2017)
23. Xu, N., Mao, W.: Multisentinet: A deep semantic network for multimodal sentiment analysis. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (2017)
24. Xu, N., Mao, W., Chen, G.: A co-memory network for multimodal sentiment analysis. In: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (2018)
25. Zadeh, A.B., Liang, P.P., Poria, S., et al.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018)
26. Zhang, H., Liu, Y., Xiong, Z., et al.: Visual sentiment analysis with semantic correlation enhancement. Complex and Intelligent Systems **10** (2024)