

# Holistic Visual-Textual Sentiment Analysis with Prior Models

Junyu Chen

*Department of Computer Science  
University of Rochester  
Rochester, USA  
jchen175@ur.rochester.edu*

Jie An

*Department of Computer Science  
University of Rochester  
Rochester, USA  
jan6@cs.rochester.edu*

Hanjia Lyu

*Department of Computer Science  
University of Rochester  
Rochester, USA  
hlyu5@ur.rochester.edu*

Christopher Kanan

*Department of Computer Science  
University of Rochester  
Rochester, USA  
ckanan@cs.rochester.edu*

Jiebo Luo

*Department of Computer Science  
University of Rochester  
Rochester, USA  
jluo@cs.rochester.edu*

**Abstract**—Visual-textual sentiment analysis aims to predict sentiment with the input of a pair of image and text, which poses a challenge in learning effective features for diverse input images. To address this, we propose a holistic method that achieves robust visual-textual sentiment analysis by exploiting a rich set of powerful pre-trained visual and textual prior models. The proposed method consists of four parts: (1) a visual-textual branch to learn features directly from data for sentiment analysis, (2) a visual expert branch with a set of pre-trained “expert” encoders to extract selected semantic visual features, (3) a CLIP branch to implicitly model visual-textual correspondence, and (4) a multimodal feature fusion network based on BERT to fuse multimodal features and make sentiment predictions. Extensive experiments on three datasets show that our method produces better visual-textual sentiment analysis performance than existing methods.

**Index Terms**—visual-textual sentiment analysis, multimodal fusion, deep neural network.

## I. INTRODUCTION

Visual-textual sentiment analysis [1] lies at the intersection of computer vision and natural language processing. Given a pair of image and text, the visual-textual sentiment analysis task aims at recognizing polarized sentiments (i.e., positive, negative, and neutral) or fine-grained emotions (e.g., happy, angry, calm, etc.). A typical application is to predict sentiments in social media posts where users post text and images to jointly express their feelings (Fig. 1).

The key to visual-textual sentiment analysis is extracting useful visual/textual features for sentiment inference. Existing studies [1]–[3] typically employ two encoders for visual and textual information extraction, with the visual encoder based on CNN and the text encoder based on either RNN or Transformer. A feature fusion module then combines visual and textual features for sentiment inference. Although this framework has made significant progress, a main challenge remains: it is difficult to effectively extract useful visual features and recognize the *abstract* sentiment of images. Sentiment

analysis requires a relatively higher level of information, which may be represented by scene, object, facial expression, or their combinations [4], [5], which is **difficult to learn with only one network without any prior knowledge**. Furthermore, images in visual-textual sentiment datasets can come from **highly diverse domains**. For example, as shown in Fig. 1, some images are screenshots of text. Without an understanding of the embedded image text, it is unlikely to infer the sentiment accurately. Therefore, a powerful visual feature extractor that can capture various effective information from images and handle the heterogeneity of various domains of images is desired for visual-textual sentiment analysis.

In this study, we propose a holistic approach for visual-textual sentiment analysis, aiming for robust performance by leveraging the power of pre-trained feature extractors. The proposed algorithm, named **VSA-PF** (Visual-textual Sentiment Analysis with Pre-trained Features; Fig. 2), consists of four parts. First, a visual-textual branch based on the Swin Transformer [6] and BERTweet [7] is used to learn visual and textual features for sentiment prediction. It captures the important information for sentiment prediction directly from the dataset. These learnable encoders are pre-trained on large-scale datasets for generalizability.

Second, a visual expert branch with a rich set of pre-trained “expert” visual encoders is used. These encoders extract semantic visual features, such as face [8], object [9], scene [10], and OCR, which are useful for sentiment analysis of **different visual contents** [4], [11], [12] but hard to learn directly from the visual-textual sentiment analysis dataset.

Third, we adopt a CLIP branch. Since CLIP is trained with a large number of image-text pairs, on one hand, it improves the generalizability of the proposed method to handle *unseen or weakly* sentiment images. On the other hand, because CLIP is trained to minimize the cosine distance between paired images and textual captions, image-text correspondence can be implicitly induced with the extracted visual and textual

Image			
Text	Centre Mall where pedestrians do not have the right of way ! #HamOnt #OnlyinHamilton	You only allow imprisonment... Because you are #weak...and full of #fear.... Control can only work if you #Believe	@USER pitches Green vision in Halifax #elxn42 #elxn2015 #ndpoli HTTPURL
Sentiment	Negative	Negative	Positive

Fig. 1. Examples from a visual-textual sentiment dataset, where image and text are used to jointly express the sentiment. Clearly, it is challenging to predict the sentiment only from the image in some cases.

features from the CLIP encoder.

Finally, a multimodal feature fusion network based on BERT [13] (i.e., Transformer encoder) is used to fuse multi-domain multimodal features and make sentiment predictions, which facilitates effective modeling of the complex interactions between different features. Extensive experiments on three datasets demonstrate that the proposed method outperforms state-of-the-art frameworks on visual-textual sentiment analysis. Furthermore, we conduct an ablation study to demonstrate the necessity of the visual-textual branch, show the contribution of each pre-trained expert feature, and compare the different architecture choices of the fusion network.

**Our contributions are three folds:** (1) Utilizing the pre-trained CLIP model, we extract *aligned visual and textual features*, enhancing sentiment prediction by implicitly modeling visual-textual correspondence. (2) Incorporating a set of “expert” visual encoders pre-trained for various visual tasks, our method benefits from a **strong visual prior**, capturing *complicated, diverse, and nuanced* visual information for sentiment analysis. (3) Our holistic framework, composed of CLIP, visual-expert, and trained visual-textual branches, achieves better performance than the state-of-the-art methods on three existing datasets.

## II. RELATED WORK

**Textual Sentiment Analysis.** Textual sentiment analysis is a well-established task in NLP. Traditional methods such as VADER [14] typically rely on hand-crafted and lexicon-based features. In recent years, deep learning techniques have been widely adopted. For instance, [15] proposed using a combination of CNN and LSTM to capture both local affective information and long-distance sentence dependencies. More recently, transformer-based models [13] such as BERTweet [7] have set new benchmarks in a variety of NLP tasks, including sentiment analysis.

**Visual Sentiment Analysis.** Visual sentiment analysis aims to explore the sentiment evoked or expressed by images. Traditional methods often rely on low- to mid-level features such as color distribution [4] and adjective-noun-pair detector [16]. Many recent approaches are based on CNNs to take advantage of the expressive power of pre-training. For example, [17] designed a progressive fine-tuning procedure for enhanced

transfer learning performance. [18] adopted attention modules with CNNs to highlight sentiment-revealing image regions. Compared with textual data, visual sentiment analysis poses **a unique challenge due to the high level of visual diversity, semantic abstraction, and viewer subjectivity**. Some approaches attempt to address these by incorporating extra features [19], such as utilizing facial expression recognition to improve sentiment prediction in images with faces [4], [20].

**Visual-textual Sentiment Analysis.** Visual-textual sentiment analysis integrates information from both visual and textual modalities to predict sentiment. Some approaches in this field focus on fusing multimodal information effectively [2], [3]. Other works concentrate on extracting effective features from each modality [21]–[23]. For example, [21] used pre-text tasks such as masked language/region modeling to help the model capture sentiment-related features. [11] and [24] adopted object and scene features of images. The most relevant work to ours is Se-MLNN [12], which adopted pre-trained encoders to extract visual features including objects, scenes, facial expressions, and CLIP embeddings. However, **our work differs in several major ways** from [12]. First, we incorporate a **learnable** branch to directly learn important multimodal features from the training dataset, in contrast to Se-MLNN’s reliance on pre-trained and fixed feature extractors. Second, we employ a pre-trained OCR encoder to explicitly handle texts in images, which is very prevalent in all the studied datasets (but largely overlooked by existing methods). This OCR feature proves important for accuracy in our experiments. Additionally, we use BERT to fuse different features, which is more effective at learning the interactions between multimodal multi-domain features. Overall, our experiments demonstrate that our holistic method outperforms Se-MLNN in all three public datasets.

## III. METHOD

Our framework, illustrated in Fig. 2, integrates four primary components: (1) a trainable visual-textual branch to directly extract sentiment features from the dataset, (2) a visual expert branch consisting of four visual feature extractors to capture semantic visual information from open-domain images, (3) a CLIP branch for both image and text inputs to implicitly model the cross-modal correspondence, and (4) a multimodal fusion network to fuse all information and make holistic sentiment prediction. This section details each component.

### A. Visual-textual Branch

The visual-textual branch aims at extracting sentiment-related features directly from the datasets. Two networks are adopted to extract features from text and image, respectively.

**Text Encoder.** We employ BERTweet-large [7] as the text encoder. BERTweet is based on BERT architecture [13] and was pre-trained using 873M English tweet. It achieves state-of-the-art performance in tweets classification tasks. Since the text of visual-textual sentiment datasets originated from social media which is similar to tweets, the text encoder can acquire a strong prior from pre-training. We fine-tune the model with the

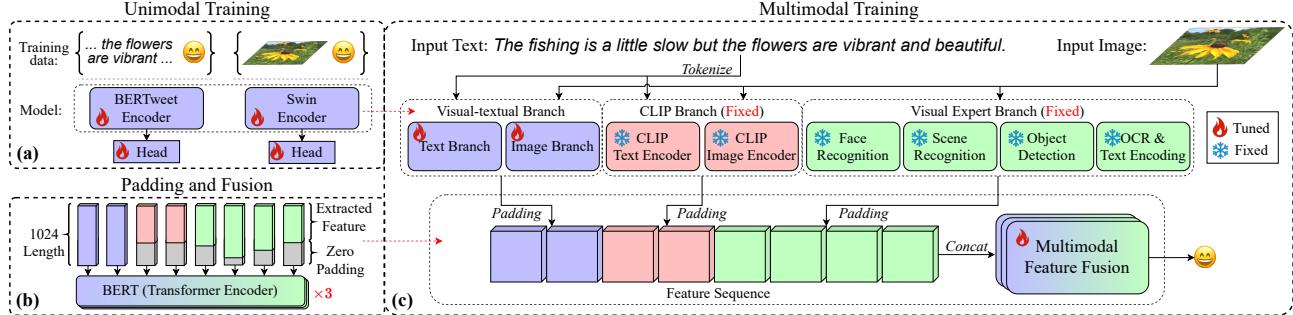


Fig. 2. Overview of the proposed method’s architecture and training procedure. Our framework (c) consists of four parts: (1) A visual-textual branch to learn visual and textual features for sentiment prediction, (2) A visual expert branch to equip the method with a strong visual prior, (3) A CLIP branch to implicitly model the visual-textual correspondence with aligned embeddings, and (4) A multimodal feature fusion module to integrate all information and make holistic sentiment predictions. Initial training (a) fine-tunes the visual-textual branch separately on unimodal data to capture sentiment features before proceeding to multimodal training (b, c). The dataset splits remain consistent throughout the two training phases.

text of the training set to adapt it to sentiment analysis. Next, we remove the classification head of the BERTweet model, and utilize the last layer’s hidden state of the classification token as the text representation, which is an embedding of 1024 dimensions.

**Image Encoder.** We use Swin Transformer [6] as the image encoder, specifically a Swin-base network pre-trained on ImageNet-21k at a resolution of  $384 \times 384$ . This pre-training enables the model to more robustly handle diverse and noisy open-domain images in the datasets. Next, similar to the text encoder, we fine-tune the Swin-base model with only images of the datasets, remove the classification head, and use the last hidden-state layer’s average pooling to generate a 1024-dimensional image embedding.

### B. Visual Expert Branch

The visual expert branch employs four pre-trained neural networks to selectively extract high-valued face, scene, object, and optical character recognition (OCR) features in images.

**Face Recognition.** Face images are of pivotal importance in sentiment analysis [4], [12]. As shown in Table I, portrait photos with faces account for over 40% of images in the datasets. To extract effective face features, we first use the MTCNN face detector [8] to filter out images without faces, then identify and retain the largest face in each image. We then adopt a pre-trained Facenet [25] based on Inception-ResNet architecture as the face feature extractor, which was pre-trained on VGGFace2. The resulting 512-dimensional embedding represents the face features.

**Scene and Object Detection.** Scenes and objects are pivotal in conveying sentiment in images [16], [26]. For example, a photo of a delicious meal in a fancy restaurant reveals enjoyment, while a solitary tree in a graveyard implies melancholy. To capture these elements, our visual expert branch employs YOLOv5 [9] pre-trained on the MS COCO dataset for object detection. The output, i.e., the predicted logit for each category, is summed up into an 80-dimensional vector to represent the number and types of objects in each image. For scene detection, we use a DenseNet161 pre-trained on the Place365 dataset [10]. This model can recognize 365 types of scenes, such as “airfield”, “campus” and “highway”. We use

the output logit of each image as the scene features, which are 365-dimensional.

**OCR.** Text within images, such as internet memes, event posters, and screenshots of messages, is prevalent on social media platforms like Twitter. To extract OCR features, we employ a pre-trained Sentence-Transformer based on MPNet [27] as our OCR encoder. The process begins with extracting English words from images using Tesseract-OCR Engine [28]. These words are then concatenated into a sequence and processed through the Sentence-Transformer. The output 768-dimensional sentence embeddings serve as the OCR features. To ensure relevance, we only extract OCR features from images containing at least five words, minimizing the impact of trivial text.

### C. CLIP Branch

The CLIP branch leverages CLIP’s pre-trained image and text encoders to extract corresponding features. The image encoder is a 24-layer, 16-head vision transformer, while the text encoder is a 12-layer, 12-head transformer, both with an embedding dimension of 768. CLIP [29] was pre-trained on a massive dataset of 400 million image-text pairs using a contrastive loss to minimize the distance between the paired text and image. This training enables CLIP’s encoders to discern high-level correlations between images and text that unimodal feature extractors cannot capture.

### D. Multimodal Feature Fusion

We employ a 3-layer, 512-width BERT model (i.e., Transformer encoder) with 4 attention heads to integrate all visual and textual features for sentiment prediction. As illustrated in Fig. 2b, we first zero pad all the extracted features to the same size of 1024 to prevent distortion before fusion. These features are arranged into a sequence and processed by the BERT model. This process effectively fuses the information from the visual-textual, visual expert, and CLIP branches to make the final sentiment prediction.

## IV. EXPERIMENTS

To demonstrate the performance of the proposed method, we conduct experiments on three public datasets. In this section, we introduce the experimental setup and show the

TABLE I

THE SUMMARY STATISTICS OF THE DATASETS. WE SHOW THE NUMBER OF SAMPLES BELONGING TO EACH SENTIMENT/EMOTION CATEGORY AND THE RATIO (%) OF SAMPLES THAT CONTAIN FACES, OBJECTS, OR IMAGE TEXT IN THREE DATASETS. NOTE THAT **THE PERCENTAGE OF IMAGE TEXT IS QUITE HIGH** ACROSS ALL THREE DATASETS.

Dataset	Sentiment	Samples	Face	Object	OCR
MVSA-Single	Positive	2,683	55.65	86.92	14.39
	Neutral	470	22.98	73.40	18.30
	Negative	1,358	46.32	82.11	17.89
	All	4,511	49.43	84.06	15.85
MVSA-Multiple	Positive	11,318	50.14	84.94	20.81
	Neutral	4,408	45.30	79.81	31.22
	Negative	1,299	47.04	79.37	28.41
	All	17,025	48.65	83.19	24.08
TumEmo	Angry	14,554	44.32	73.01	21.82
	Bored	32,283	51.86	83.79	12.32
	Calm	18,109	18.56	60.82	18.82
	Fearful	20,264	24.37	59.64	42.42
	Happy	50,267	49.12	80.57	19.81
	Loving	34,511	42.31	74.71	26.42
	Sad	25,277	35.18	67.10	40.54
	All	195,265	40.81	73.76	24.83

experiment results along with key findings. The source code and the dataset splits used in our experiments are available at <https://github.com/jchen175/VSA-PF>.

#### A. Datasets and Preprocessing

**MVSA-Single and MVSA-Multiple.** These datasets contain 4,869 and 19,600 image-text pairs from Twitter. Each pair is assigned a human-annotated sentiment label (positive, negative, neutral). We follow the preprocessing procedure of [11] to remove pairs with conflicting sentiments in two modalities or disagreement among annotators. This results in 4,511 and 17,025 pairs for MVSA-Single and MVSA-Multiple, respectively.

**TumEmo.** TumEmo [24] is a large visual-textual emotion dataset containing 195,265 image-text pairs from Tumblr, classified into seven emotion classes (i.e., angry, bored, calm, fearful, happy, loving, and sad) if the hashtags in the text contain certain keywords related to that emotion class. We adhere to the original study’s methodology, removing all hashtags that could reveal the ground truth label.

We follow BERTweet’s normalization procedure for text preprocessing, where emojis, user mentions, and URLs are replaced with special tokens, and the use of punctuation and abbreviations is standardized across sentences. Table I shows the detailed statistics of the datasets after preprocessing. For each dataset, we adopt an 8:1:1 split ratio for the training, validation, and test sets in a stratified manner, where the distributions of annotation labels are consistent across these sets. Pre-trained features are extracted once and saved for reuse.

#### B. Model Training

In model training, the CLIP and visual expert branches are fixed. As shown in Fig. 2, our training procedure consists of two steps.

**Unimodal Training.** We first train the visual-textual branch separately with either text or images in the datasets. This

TABLE II

QUANTITATIVE RESULTS ON THREE DATASETS. WE SHOW REPORTED BASELINE RESULTS IN THEIR PAPERS AND COMPARE OUR METHOD WITH THEIRS USING OUR DATASET SPLITS (REP). THE RESULTS OF TWO MVSA DATASETS ARE AVERAGED OVER 10 FOLDS. THE BEST PERFORMANCE IS IN **BOLD** AND THE SECOND BEST IS UNDERLINED.

Model	MVSA-Single		MVSA-Multiple		TumEmo	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
MGNNS [26]	73.77	72.70	72.49	69.34	66.72	66.69
Se-MLNN [12]	75.33	73.76	66.35	61.89	-	-
CLMLF [21]	75.33	73.46	72.00	69.83	-	-
MGNNS (rep)	71.04	69.68	69.43	65.88	65.33	65.02
Se-MLNN (rep)	73.09	71.31	69.22	64.57	58.16	57.99
CLMLF (rep)	72.73	<u>71.62</u>	70.08	<u>66.33</u>	71.12	71.01
VSA-PF (ours)	<b>75.58</b>	<b>74.48</b>	<u>71.26</u>	<b>68.72</b>	<b>76.58</b>	<b>76.57</b>

procedure enables the encoders to learn enough knowledge for sentiment analysis from unimodal data. For the visual encoder (Swin Transformer), we train exclusively with images for 15 epochs, with a learning rate of  $5e-5$  for MVSA and  $1e-4$  for TumEmo, and batch sizes of 16 for MVSA and 32 for TumEmo. For the textual encoder (BERTweet), we train with only text for 20 epochs, setting the learning rate to  $8e-6$  for MVSA and  $5e-5$  for TumEmo, and batch sizes of 16 for MVSA and 64 for TumEmo.

**Multimodal Training.** After unimodal training, we train the entire model with both images and text. Note the dataset splits remain the same in the two stages. We initialize the visual-textual branch with the best-performing parameters (those with the lowest loss on the validation set) in unimodal training. Next, we train the visual-textual branch and the final multimodal feature fusion module together end-to-end. We use the AdamW optimizer with learning rates of  $5e-6$  for MVSA and  $1e-5$  for TumEmo, a batch size of 16 for all datasets, and a dropout rate of 0.5 to mitigate over-fitting. The training lasts 30 epochs, and the parameters that yield the highest accuracy on the validation set are then used for testing.

#### C. Baselines

We benchmark our method against three publicly available, state-of-the-art methods in visual-textual sentiment analysis:

**MGNNS** [26] introduces a multi-channel graph neural network to model object, scene, and text representations based on the global co-occurrence patterns across the dataset.

**Se-MLNN** [12] integrates several pre-trained image features with contextual text features extracted via RoBERTa to comprehensively represent multimodal contents.

**CLMLF** [21] introduces a contrastive learning framework to enable effective learning of generalized representations for both image and text in visual-textual sentiment analysis.

Among the compared methods, Se-MLNN relies only on pre-trained features and is lightweight with 1M parameters, while our method and compared CLMLF, MGNNS have comparable sizes (500M, 200M, 100M respectively).

#### D. Quantitative Comparison

Table II shows the quantitative comparison between our proposed method and established baselines. While baseline results from their respective papers vary due to different

TABLE III

ABLATION STUDY RESULTS. WE SHOW THE PERFORMANCE CHANGES ON ACCURACY AND WEIGHTED-F1 WHEN: (1) REMOVING AN ENCODER IN OUR FRAMEWORK, (2) REMOVING THE FINE-TUNING PROCEDURE OF THE VISUAL-TEXTUAL BRANCH, OR (3) REPLACING THE FUSION NETWORK WITH MLP.

Model	MVSA-Single		MVSA-Multiple		TumEmo	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
VSA-PF (ours)	75.58	74.48	71.26	68.72	76.58	76.57
VSA-PF (using MLP)	-0.52	-0.77	-0.22	-0.76	+0.09	+0.09
w/o Swin	-2.91	-3.39	-0.62	-1.09	-0.34	-0.31
w/o BERTweet	-4.70	-5.31	-2.41	-4.02	-6.68	-6.64
w/o Fine-tuning	-1.23	-1.21	-1.08	-1.53	-0.20	-0.19
w/o CLIP-Image	-0.42	+0.04	-0.31	-0.67	-0.70	-0.72
w/o CLIP-Text	-0.46	-0.42	-0.33	-0.74	-0.18	-0.19
w/o Face	-0.82	-0.94	-0.38	-0.73	-0.19	-0.21
w/o OCR	-0.93	-1.42	-0.42	-0.80	-0.05	-0.06
w/o Scene	-0.34	-0.41	-0.27	-0.61	-0.02	-0.04
w/o Object	-0.68	-0.69	-0.40	-0.66	-0.14	-0.17

dataset splits, we ensure a fair comparison by training each model with our dataset splits and reporting the metrics. Results for the MVSA datasets are averaged over 10 folds.

On the MVSA and TumEmo datasets, the proposed method outperforms other strong baselines, highlighting its effectiveness. The proposed visual-textual branch is based on pre-trained Swin Transformer and BERTweet, which allow our framework to employ the power of advanced neural network architectures. In contrast, Se-MLNN relies solely on pre-trained features **without learning useful sentiment-related features directly from the dataset**, which may lead to its compromised performance on the more complex TumEmo dataset with fine-grained emotion annotations. Regarding the visual expert branch, our method has four different visual features while OCR and face features are not used by MGNNS [26]. It is hard for a trainable visual encoder to learn these missing visual features directly.

#### E. Ablation Study

In our ablation study, we assess the contribution of each pre-trained and learned feature, the significance of multimodal input, and compare different fusion network architectures.

**The Importance of Different Branches.** We analyze performance changes when individual encoders are removed. Additionally, we provide a qualitative illustration (Fig. 4), showcasing how each branch contributes to accurate sentiment predictions. As detailed in Table III, the most substantial decline occurs with the removal of BERTweet, underscoring the critical role of textual information in visual-textual sentiment analysis. We hypothesize the reason is that learning sentiment from text is easier than from images, since adjectives in text can convey sentiment more directly. In the MVSA datasets, the removal of Swin Transformer, or the fine-tuning of the visual-textual branch (i.e., using fixed pre-trained Swin and BERTweet) leads to a notable performance decrease. This indicates the significance of directly learning sentiment-related features from the datasets. Among visual expert features, removing OCR causes the largest performance drop. This is in line with our observation about the prevalence and importance of image text in social media, which has been ignored in previous studies. We show some example cases in Fig. 3.

TABLE IV

ABLATION STUDY RESULTS. WE SHOW THE PERFORMANCE UNDER UNIMODAL SETTINGS BY KEEPING ONLY THE IMAGE OR TEXT ENCODERS AND COMPARE IT WITH THE MULTIMODAL MODEL.

Model	MVSA-Single		MVSA-Multiple		TumEmo	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
BERTweet	72.73	70.87	69.84	64.18	67.90	67.58
+ Pre-trained textual feature	72.49	71.28	70.18	66.90	68.99	68.75
Swin	68.46	66.40	67.94	60.18	54.04	53.89
+ Pre-trained visual features	70.15	68.96	68.28	63.45	61.40	61.51
BERTweet + Swin	75.52	74.60	70.83	68.23	76.18	76.17
VSA-PF (ours)	75.58	74.48	71.26	68.72	76.58	76.57

When the input image contains text to deliver sentiment-related information, the OCR feature can indeed help make correct sentiment predictions. In general, removing any of the pre-trained features diminishes performance, highlighting the necessity of each encoder. We present examples in Fig. 4 to illustrate the importance of these encoders, where their removal leads to incorrect results. Removing the CLIP image encoder results in a decrease in accuracy and a marginal increase in weighted F1 on MVSA-Single. This could probably be caused by the class imbalance present in the dataset, as shown in Table I. For TumEmo, the removal of the CLIP image encoder significantly reduces performance. This suggests that the CLIP encoder captures unique image features distinct from the learnable Swin Transformer and is particularly effective in handling the noisy, open-domain images in this complex dataset.

In the ablation study on TumEmo, we find that the OCR and scene features only marginally help sentiment analysis. This observation aligns with [26]’s ablation study, where scene features had a lesser impact on TumEmo compared to MVSA. A possible explanation is that the OCR and scene encoders might not be sufficiently robust for the complex task of seven-class emotion classification: (1) OCR may not detect all the text present in the images (last column in Fig. 3), (2) the text information in images may be inconsistent with the input text, and (3) it is more difficult to infer seven-class emotional predictions from scene features.

**The Necessity of Multimodal Input.** Table IV shows that the pre-trained visual (CLIP image and visual expert features) and textual (CLIP text feature) elements enhance unimodal sentiment analysis over using Swin Transformer and BERTweet alone. This supports the effectiveness of pre-trained features. When using unimodal data only, text-based analysis generally outperforms image-based analysis. This is *in line with* previous findings [21], [26], and could be due to a higher level of abstraction and subjectivity in the visual modality. Multimodal models show consistently superior performance than unimodal models, particularly on the challenging TumEmo dataset, which demonstrates the necessity of multimodal input for comprehensive sentiment analysis.

**Fusion Module Architecture.** We compare our BERT-based fusion module to a lightweight 3-layer MLP with a 768-size hidden layer, where MLP is used by [12] as the feature fusion module. In this setup, visual and textual features are concatenated and fed into the MLP with ReLU activation for sentiment prediction. Table III shows that MLP is comparable

<b>Input Image</b>					
<b>Input Text</b>	@USER RT #Calgary #alberta #Canada #AMBERALERT Share at : HTTPURL	When your prom date from a year ago still calls you prom date ???		if you're here make sure to put them up during you & i #OTRAToronto ??	Do this more
<b>Detected Image Text</b>	... father has been found <b>murdered</b> and his two-year-old daughter has been <b>abducted</b> from Alberta ...	Yay ! I'll see ya soon Mr. Sigler @USER Sounds <b>good</b> miss prom date & )		<b>Thank</b> you for Staying with us	' 2 SPOT COME WERE GOING o3 Bz Ze3 Ge

Fig. 3. Examples from visual-textual sentiment datasets, where the sentiment is mainly revealed by image text. We give three instances where the OCR engine accurately detected the image text and helped improve the accuracy, along with one failure case in which the OCR engine failed to detect hand-written image text.

<b>Input Image</b>										
<b>Input Text</b>	Centre Mall where pedestrians do not have the right of way ! #HamOnt #OnlyinHamOnt #OnlyinHamilton	INTERVIEW : We chat with @USER and @USER about the upcoming @USER #TIFF15 HTTPURL	How I feel	Eating donut @USER bought me last night as my cat sleeps on me . The joys of no movies til 3pm . #TIFF15						
<b>Sentiment</b>	VSA-PF w/o BERTweet	w/o Swin GT	VSA-PF w/o Swin GT	VSA-PF w/o CLIP-Image GT	VSA-PF w/o CLIP-Text GT					
	Negative	Positive	Negative	Positive	Angry	Love	Angry	Positive	Neutral	Positive
<b>Input Image</b>										
<b>Input Text</b>	White roses ? my favourites !	@USER pitches Green vision in Halifax #elxn42 #elxn2015 #cndpoli HTTPURL	When I Close My Eyes When it comes you It's very different...	Off to a #foggy start this #Thursday #morning #yeg #yegdt #yegwx #Edmonton						
<b>Sentiment</b>	VSA-PF Love	w/o Object Happy	GT Love	VSA-PF w/o Face Positive	w/o OCR Neutral	GT Positive	VSA-PF Bored	w/o Scene Sad	GT Positive	Negative

Fig. 4. Examples from visual-textual sentiment datasets. Each column represents a branch and a corresponding data pair, where our VSA-PF model accurately predicts sentiments for these examples, but the removal of the branch results in incorrect prediction for the associated sample.

to BERT on the TumEmo but worse on two MVSA datasets, demonstrating that BERT is more effective in fusing multi-modal multi-domain features than simple networks like MLP.

## V. CONCLUSION AND FUTURE WORK

In this paper, we present a holistic framework for visual-textual sentiment analysis, which consists of trainable visual-textual, visual-expert, and CLIP branches. The visual-textual branch exploits the power of state-of-the-art network architectures. The visual expert branch allows our framework to benefit from rich pre-trained visual encoders. The CLIP branch equips the proposed method with the ability to model visual-textual correspondence explicitly. The experimental results on three public datasets demonstrate the effectiveness of our

proposed framework across the board. The ablation study reveals the necessity of each designed branch. In the future, we plan to further investigate why some encoders are ineffective on certain datasets and address this limitation. Additionally, there is currently no systematic evaluation of foundational LLM/LVM models in multimodal sentiment analysis benchmarks. Our preliminary tests with GPT-4V on 30 samples from each dataset showed lower accuracies (76%/66%/43%) compared to our framework, particularly on complex datasets. An interesting direction for future research is to integrate our method with LLMs and LVMs. For example, we could explore using embeddings extracted from these models or employ special adapters to integrate our pre-trained features into these models.

## REFERENCES

- [1] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo, “Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks,” in *ACM MM*, 2016.
- [2] Zenan Xu, Qinliang Su, and Junxi Xiao, “Multimodal aspect-based sentiment classification with knowledge-injected transformer,” in *ICME*, 2023.
- [3] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, “Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia,” in *WSDM*, 2016.
- [4] Jianbo Yuan, Sean McDonough, Quanzeng You, and Jiebo Luo, “Sentiattribute: image sentiment analysis from a mid-level perspective,” in *WISDOM*, 2013.
- [5] Honglin Zheng, Tianlang Chen, Quanzeng You, and Jiebo Luo, “When saliency meets sentiment: Understanding how image content invokes emotion and sentiment,” in *ICIP*, 2017.
- [6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *ICCV*, 2021.
- [7] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen, “Bertweet: A pre-trained language model for english tweets,” in *EMNLP*, 2020.
- [8] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE SPL*, 2016.
- [9] Glenn Jocher, “YOLOv5 by Ultralytics,” May 2020.
- [10] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, “Places: A 10 million image database for scene recognition,” *TPAMI*, 2018.
- [11] Nan Xu and Wenji Mao, “Multisentinet: A deep semantic network for multimodal sentiment analysis,” in *CIKM*, 2017.
- [12] Gullal S Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth, “A fair and comprehensive comparison of multimodal tweet sentiment analysis methods,” in *MMPT*, 2021.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, 2019.
- [14] Clayton Hutto and Eric Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *ICWSM*, 2014.
- [15] Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang, “Dimensional sentiment analysis using a regional cnn-lstm model,” in *ACL*, 2016.
- [16] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang, “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” in *ACM MM*, 2013.
- [17] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, “Robust image sentiment analysis using progressively trained and domain transferred deep networks,” in *AAAI*, 2015.
- [18] Quanzeng You, Hailin Jin, and Jiebo Luo, “Visual sentiment analysis by attending on local image regions,” in *AAAI*, 2017.
- [19] Masaki Saito and Kyoko Sudo, “Visual sentiment analysis by distillation of instance-based sentiment vectors,” in *MIPR*, 2023.
- [20] Ge Ma and Jiebo Luo, “Is a picture worth 1000 votes? analyzing the sentiment of election related social photos,” in *ICME*, 2013.
- [21] Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao, “Clmlf: A contrastive learning and multi-layer fusion method for multimodal sentiment detection,” *arXiv preprint arXiv:2204.05515*, 2022.
- [22] Luwei Xiao, Ejian Zhou, Xingjiao Wu, Shuwen Yang, Tianlong Ma, and Liang He, “Adaptive multi-feature extraction graph convolutional networks for multimodal target sentiment analysis,” in *ICME*, 2022.
- [23] Shuanglu Dai and Hong Man, “Integrating visual and textual affective descriptors for sentiment analysis of social media posts,” in *MIPR*, 2018.
- [24] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang, “Image-text multimodal emotion classification via multi-view attentional network,” *IEEE TMM*, 2020.
- [25] Florian Schroff, Dmitry Kalenichenko, and James Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *CVPR*, 2015.
- [26] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang, “Multimodal sentiment detection based on multi-channel graph neural networks,” in *ACL*, 2021.
- [27] Nils Reimers and Iryna Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *EMNLP*, 2019.
- [28] Matthias A Lee, “Python tesseract,” <https://github.com/madmaze/pytesseract>.
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.