A Dual-Module Denoising Approach with Curriculum Learning for Enhancing Multimodal Aspect-Based Sentiment Analysis

Nguyen Van Doan, Dat Tran Nguyen, Cam-Van Thi Nguyen*
Faculty of Information Technology
VNU University of Engineering and Technology
{21020111, 21020011, vanntc}@vnu.edu.vn

Abstract

Multimodal Aspect-Based Sentiment Analysis (MABSA) combines text and images to perform sentiment analysis but often struggles with irrelevant or misleading visual information. Existing methodologies typically address either sentence-image denoising or aspect-image denoising but fail to comprehensively tackle both types of noise. To address these limitations, we propose **DualDe**, a novel approach comprising two distinct components: the *Hybrid Curriculum Denoising Module* (HCD) and the *Aspect-Enhance Denoising Module* (AED). The HCD module enhances sentence-image denoising by incorporating a flexible curriculum learning strategy that prioritizes training on clean data. Concurrently, the AED module mitigates aspect-image noise through an aspect-guided attention mechanism that filters out noisy visual regions which unrelated to the specific aspects of interest. Our approach demonstrates effectiveness in addressing both sentence-image and aspect-image noise, as evidenced by experimental evaluations on benchmark datasets.

1 Introduction

Sentiment analysis is a fundamental task in natural language processing (NLP) (Zhang and Liu, 2012), which seeks to uncover and interpret the opinions, attitudes, and emotions embedded in user-generated content. Multimodal Aspect-Based Sentiment Analysis (MABSA) extends this analysis by combining textual and visual modalities to achieve a deeper and more comprehensive understanding of sentiment. MABSA is typically organized into three principal subtasks: Multimodal Aspect Term Extraction (MATE), which focuses on the identification and extraction of aspect-specific terms from text (Wu et al., 2020a); Multimodal Aspect-Oriented Sentiment Classification (MASC), which involves classifying the sentiment associated with each aspect term into categories such as positive, neutral, or negative (Yu and Jiang, 2019); and Joint Multimodal Aspect-Sentiment Analysis (JMASA), which concurrently addresses aspect extraction and sentiment classification to provide a unified analysis of both aspects and sentiments (Ju et al., 2021).

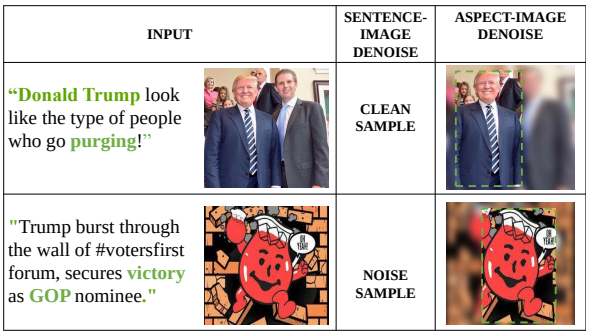

Figure 1: Illustration of Sentence-Image Denoising and Aspect-Image Denoising. Sentence-Image Denoising classifies an image as clean if it is relevant to the overall sentence meaning. Aspect-Image Denoising identifies regions as noise (e.g., blurred areas) when they lack strong relevance to any specific aspect.

In real-world scenarios, not all images are relevant to the accompanying text; some even mislead the contextual and emotional understanding of the sentence. For images that are related to the text, not all visual blocks in the image are closely tied to the aspect; in fact, there are often blocks that introduce noise. In real-world scenarios, images accompanying text may not always be relevant and can sometimes mislead the interpretation of the sentence's context and emotion. Even when images are relevant, not all visual elements are tied to the aspect of interest, often introducing noise. To address these challenges, existing methods focus on either sentence-image or aspect-image denoising. Approaches such as those by (Ju et al., 2021)

*Corresponding author. Cam-Van Thi Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.TS147.

and (Sun et al., 2021) utilize text-image relation detection to filter out non-contributory visual information but may miss significant details in images deemed irrelevant. (Zhao et al., 2023) address this with Curriculum Learning, progressively exposing the model to noisy images; however, their fixed noise metric limits flexibility. On the other hand, methods like those by (Zhang et al., 2021) and (Yu et al., 2022) concentrate on the interaction between visual objects and specific words, while (Zhou et al., 2023) use an aspect-aware attention module for fine-grained alignment. Despite their advantages, these methods often neglect the importance of sentence-image denoising, as illustrated in Figure 1.

In this paper, we propose **DualDe**, an advanced approach designed to comprehensively address both sentence-image and aspect-image noise. DualDe integrates two principal components: the Hybrid Curriculum Denoising Module (HCD) and the Aspect-Enhance Denoising Module (AED). The Hybrid Curriculum Denoising Module advances sentence-image denoising by implementing a flexible curriculum learning approach that dynamically adjusts noise metrics based on both model performance and pre-defined standards, thereby enhancing adaptability. The Aspect-Enhance Denoising Module (AED) utilizes an aspect-guided attention mechanism to selectively filter out irrelevant visual regions and textual tokens related to each specific aspect, thereby improving image-text alignment. Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to present a model, DualDe, that concurrently addresses both sentence-image and aspect-image noise.
- We introduce the Hybrid Curriculum Denoising Module(HCD), which effectively balances generalization and adaptability within the training framework.
- We demonstrate the effectiveness of our approach through extensive experiments on the Twitter-15 and Twitter-17 datasets.

2 Related Work

2.1 Multimodal Aspect-based Sentiment Analysis

With the proliferation of social media, where posts frequently encompass multiple modalities such as text and images, there has been considerable interest in utilizing multimodal approaches to analyze aspects and sentiments in user-generated content (Cai et al., 2019). The Multimodal Aspect-Based Sentiment Analysis (MABSA) task is typically segmented into three core subtasks: Multimodal Aspect Term Extraction (MATE) (Wu et al., 2020a), which focuses on identifying aspect terms within text; Multimodal Aspect-Oriented Sentiment Classification (MASC) (Yu and Jiang, 2019), which classifies the sentiment associated with each aspect term; and Joint Multimodal Aspect-Sentiment Analysis (JMASA) (Ju et al., 2021), which integrates MATE and MASC by concurrently extracting aspect terms and predicting their associated sentiments.

With the prevalence of noisy images in multimodal data, several methods have been proposed to address this issue. Ju et al. (2021) and Sun et al. (2021) address the issue of noisy images by incorporating an auxiliary cross-modal relation detection module that filters and retains only those images that genuinely contribute to the text's meaning. Ling et al. (2022) propose a Vision-Language Pre-training architecture specifically for MABSA, which enhances cross-modal alignment between text and visual elements, thereby mitigating the impact of noisy visual blocks. Meanwhile, Zhang et al. (2021) and Yu et al. (2022) focus on eliminating noise by disregarding image regions without visual objects and concentrating solely on regions containing relevant visual elements and their interaction with text. Zhou et al. (2023) propose an aspect-aware attention module that enhances image-text alignment by weighting tokens according to their relevance to the aspect, thereby effectively reducing aspect-image noise.

2.2 Curriculum Learning

Curriculum Learning (CL), introduced by Bengio et al. (2009), is a machine learning strategy that mimics human learning by starting with simpler concepts and progressively tackling more complex ones. CL has shown benefits across various tasks (Wang et al., 2019; Lu and Zhang, 2021; Platanios et al., 2019; Nguyen et al., 2024) and has been effective in mitigating noisy images in the Multimodal Aspect-Based Sentiment Analysis (MABSA) task (Zhao et al., 2023). While Zhao et al. (2023) utilize CL to progressively expose the model to noisy images, starting from cleaner data to address sentence-image noise, they do not account

for aspect-image noise. In this paper, we extend this concept by proposing the Hybrid Curriculum Denoising Module (HCD), specifically designed to reduce sentence-image noise and enhance overall performance.

3 Methodology

Our model comprises two main modules: (1) the Hybrid Curriculum Denoising Module (HCD) and (2) the Aspect-Enhanced Denoising Module (AED). The Aspect-Enhanced Denoising Module (AED) is constructed on a BART-based architecture and incorporates two sub-components situated between the encoder and decoder: Aspect-Based Enhanced Sentic Attention (AESA) and Graph Convolutional Network (GCN). An overview of the architecture is illustrated in Figure 2.

Task Definition. In this task, given a tweet with an image I and a sentence T consisting of m words $T = \{t_1, t_2, \ldots, t_m\}$, the objective is to generate an output sequence $Z = [b^1_{begin}, b^1_{end}, p_1, \ldots, b^m_{begin}, b^m_{end}, p_m]$. Each tuple $[b^i_{begin}, b^i_{end}, p_i]$ represents the i-th aspect, where b^i_{begin} and b^i_{end} denote the starting and ending positions of the aspect, and p_i indicates its sentiment polarity (Positive, Negative, or Neutral). Aspects can span multiple words, and a single sentence may include multiple aspects, each with different sentiment polarities.

Feature Extractor. We pre-trained BART (Lewis et al., 2019) model for embeddings word and ResNet (Chen et al., 2014) for embeddings image. The formatted output is $I = \{\texttt{<img>}i_1\texttt{</img>}, \ldots, \texttt{<img>}i_m\texttt{</img>}\}$ and $T = \{\texttt{<bos>}t_1\texttt{<eos>}, \ldots, \texttt{<bos>}t_n\texttt{<eos>}\}$ where m is the number of image features extracted by Resnet (surround by <img>...</img>), n is the number of text features(surround by <bos>...<eos>). These features are combined into a sequence X, which is then used as the input for the BART encoder.

The encoder generates multimodal hidden states $H = \{h^I_0, h^I_1, \ldots, h^I_m, h^T_0, h^T_1, \ldots, h^T_n\}$, where h^I_i represents the feature of the i-th visual block from the image I, and h^T_j represents the feature of the j-th word from the sentence T, with m visual blocks and n words in total.

3.1 Hybrid Curriculum Denoising Module (HCD)

This HCD module employs a flexible training strategy that adapts to varying levels of image noise, starting with cleaner data and progressively incorporating noisier examples. By integrating dynamic noise metrics from both model predictions and predefined standards, this module enhances the model's ability to mitigate sentence-image noise effectively.

3.1.1 Similarity Difficulty Metric

As depicted in Figure 1, when a sentence is paired with images that closely align with its content, it enhances the comprehension of the sentence's meaning and sentiment. Consequently, the degree of similarity between the text and accompanying images can be considered an indicator of learning difficulty: greater similarity suggests an easier learning process, whereas lower similarity indicates increased difficulty. The similarity score is computed as follows:

$$S_{(X^T_i, Y^I_i)} = \cos(X^T_i, Y^I_i) \tag{1}$$

where S is the similarity score calculated by the cosine function $\cos(\cdot)$, X^T_i and Y^I_i represent the textual and visual features, respectively, obtained through the text and image encoders of the pre-trained CLIP model (Radford et al., 2021).

Subsequently, we define and normalize the difficulty at the sentence level of i-th sample as follows:

$$d^s_i = 1.0 - \frac{\mathrm{S}_{(X^T_i, Y^I_i)}}{\max\limits_{1 \le k \le N} \mathrm{S}_{(X^T_k, Y^I_k)}}, \tag{2}$$

where N is length of train dataset, d^s_i is normalized within the range $[0.0, 1.0]$. A lower value of d^s_i indicates that the data is likely to be easier to learn or predict accurately and will therefore be prioritized in the learning process.

3.1.2 Model loss Diffculty Metric

The individual loss function for each data sample in a sequential model can be expressed as:

$$L_i = -\sum_{t=1}^{O} \log P(y_t \mid Y_{<t}, X_i) \tag{3}$$

where L_i represents the loss for the i-th data sample, X_i is the input for that sample, and O is the sequence length. y_t denotes the word or character at time step t, and $Y_{<t}$ represents all preceding words or characters. $P(y_t \mid Y_{<t}, X_i)$ is the probability predicted by the model for the word y_t given the context $Y_{<t}$ and input X_i.

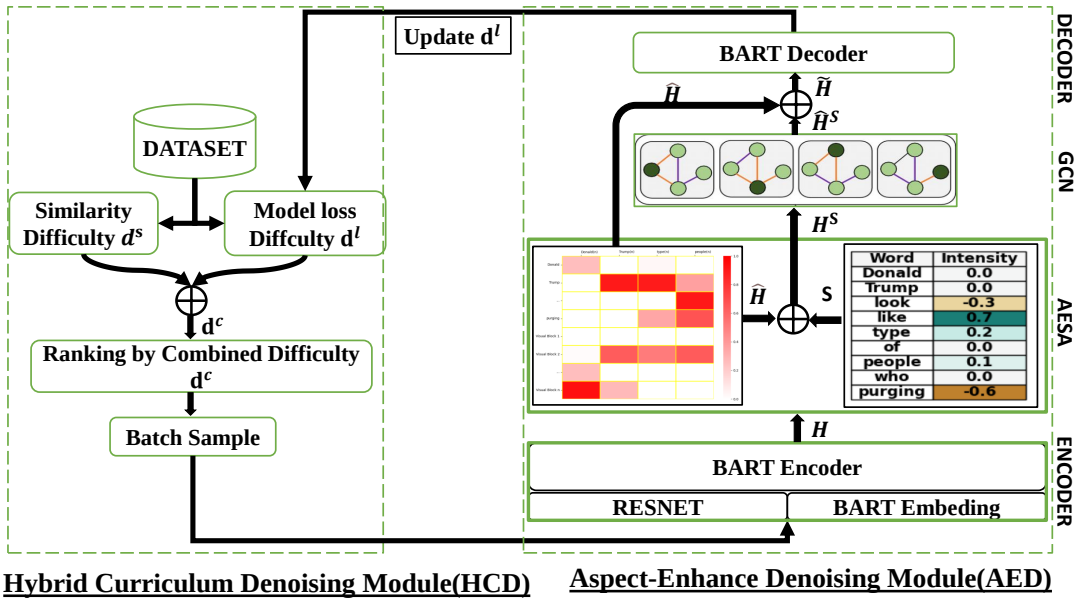

Figure 2: Model Overview

After that, we normalized this difficulty score of i-th sample to $[0.0, 1.0]$ by following formula:

$$d_i^l = \frac{L_i}{\max\limits_{1 \le j \le N} L_j} \tag{4}$$

where N is length of train dataset.

Since the difficulty a batch sample (based on the loss metric) is entirely dependent on the model's state, we update the loss metric at each epoch to ensure accurate evaluation.

3.1.3 Comprehensive Difficulty Metric

The difficulty metric d_i^s is a predefined metric that remains constant throughout the training process. Conversely, the difficulty metric d_i^l is based on the model's current learning state and changes at each epoch. To balance the generalization of d_i^s and the adaptability of d_i^l in training schedules, we propose a new composite difficulty metric d_i^c for i-th sample, defined as:

$$d_i^c = \alpha \cdot d_i^l + (1 - \alpha) \cdot d_i^s \tag{5}$$

where α is a weighting factor that balances the contribution of d_i^l and d_i^s. Empirical results indicate that setting $\alpha = 0.8$ yields optimal performance.

3.1.4 Curriculum Training

Platanios et al. (2019) introduced the concept of "Competence-Based Curriculum Learning", highlighting that competence reflects the model's learning ability, which progressively increases from an

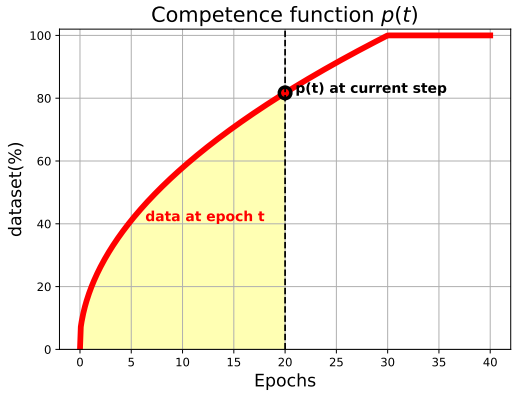

Figure 3: Illustrate the curve of the competence function $p(t)$ and the corresponding amount of selected data at epoch t.

initial value λ_{init} to 1 over a duration T. At the t-th epoch, the model selects only those training data that are well-aligned with its current capabilities, defined by the condition $d^c < p(t)$, where $p(t)$ is the model's learning competence. The curve $p(t)$ is depicted as the red curve in Figure 3 and is computed using the following formula:

$$p(t) = \begin{cases} \sqrt{\frac{t}{T}\left(1 - \lambda_{\text{init}}^2\right) + \lambda_{\text{init}}^2} & \text{if } t \le T, \\ 1.0 & \text{otherwise.} \end{cases} \tag{6}$$

When $p(t) \ge 1.0$, the model selects 100% of the training dataset.

3.2 Aspect-Enhance Denoising Module (AED)

This module enhances text-image alignment for sentiment analysis by using an aspect-guided attention mechanism to filter out irrelevant visual data and focus on extracting meaningful features tied to each aspect.

3.2.1 Aspect-Based Enhance Sentic Attention(AESA)

We leverages the Aspect-Aware Attention(A3M) Module from (Zhou et al., 2023) to filter out visual block noise—visual blocks that are very few or nearly irrelevant to the aspect. A3M uses an aspect-guided attention mechanism as described by the following formula:

$$Z_t = \tanh((W_{CA}H^{CA} + b_{CA}) \oplus (W_H h_t + b_H)), \tag{7}$$

$$\alpha_t = \text{softmax}(W_\alpha Z_t + b_\alpha), \tag{8}$$

where $H^{CA} = \{h_1^{CA}, h_2^{CA}, \ldots, h_n^{CA}\}$ is the list of all n noun in sentence, Z_t is the comprehensive feature extracted from both the noun list H^{CA} and the hidden states h_t. $W_{CA}, W_H, W_\alpha, b_{CA}, b_H$, and b_α are the learned parameters. $\oplus$ is an concatenate operator. We then get the aspect-related hidden feature h_t^A by calculating the weighted sum of all candidate aspects following the equation:

$$h_t^A = \sum_{i=1}^{k} \alpha_{t,i} h_i^{CA}. \tag{9}$$

To mitigate noisy visual blocks, the parameter β_t is learned to aggregate the atomic feature h_t with its aspect-related hidden feature h_A^t.

$$\beta_t = \text{sigmoid}(W_\beta[W_1 h_t; W_2 h_A^t] + b_\beta), \tag{10}$$

$$\hat{h}_t = \beta_t h_t + (1 - \beta_t) h_A^t, \tag{11}$$

where W_β, W_1, W_2, and b_β are parameters, and $[;]$ denotes the concatenation operator for vectors. $\hat{h}_t$ is the final output of A3M after the semantic alignment and noise reduction procedure.

We utilize SenticNet (Cambria et al., 2016), an external affective commonsense knowledge base, to enhance sentiment feature representations for each concept. The affective values in SenticNet range from $[-1, 1]$, where values closer to 1 indicate a stronger positive sentiment. The attention output $\hat{h}_t$ is further refined by incorporating these affective values from SenticNet as follows:

$$s_i = W_S \cdot \text{SenticNet}(w_i) + b_S, \tag{12}$$

$$h_i^S = \hat{h}_i + s_i \tag{13}$$

where w_i is the word in the sentence, and W_S and b_S are learned parameters.

3.2.2 Weighted Association Matrix

First, we use the Spacy library to create matrix D representing the dependency tree, where D_{ij} is the distance between the i-th word and the j-th word in the tree.

Next, we initialize a zero-weighted association matrix A, $A \in \mathbb{R}^{(m+n)\times(m+n)}$, where the image features range from 1 to m, and the text features range from $m+1$ to $m+n$. We divide matrix A into 3 regions: $A_{\textit{image-image}}$ contains all A_{ij} with $(i, j \leq m)$, $A_{\textit{text-image}}$ contains all A_{ij} with $(i < m < j)$ or $(j < m < i)$, and $A_{\textit{text-text}}$ contains all A_{ij} with $(i, j > m)$. We fill the values for A as follows:

- For $A_{\textit{image-image}}$, we initialize the main diagonal with 1. **(I)**
- For $A_{\textit{text-image}}$, to ensure aspect-oriented directionality:
  - If the i-th feature is an aspect, we set $A_{ik} = \cos(\hat{h}_i, \hat{h}_k)$ for $0 \leq k \leq m+n$.
  - Similarly, if the j-th feature is an aspect, we set $A_{kj} = \cos(\hat{h}_k, \hat{h}_j)$ for $0 \leq k \leq m+n$. **(II)**
- For $A_{\textit{text-text}}$, we set $A_{ij} = \cos(\hat{h}_i, \hat{h}_j)$ if $\text{D}_{ij} \leq \textit{threshold}$. In this paper, we set the *threshold* to 2. **(III)**

The above conditions can be rewritten as follows:

$$A_{ij} = \begin{cases} 1 & \text{(I)}, \\ \cos(\hat{h}_i, \hat{h}_j) & \text{(II) and (III)}, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

where $\cos(\cdot)$ is the cosine function.

3.2.3 Graph Convolutional Network (GCN)

Based on the weighted association matrix A above and the enhanced sentiment feature h_i^S, we feed the graph into the GCN layers to learn the affective dependencies for the given aspect. Then each node in the l-th GCN layer is updated according to the following equation:

$$h_{i,0}^S = h_i^S, \tag{15}$$

$$h_{i,l}^S = \text{ReLU}\left(\sum_{j=1}^{n} A_{ij} W_l h_{i,l-1}^S + b_l\right). \tag{16}$$

Table 1: Statistics of two benchmark datasets

Datasets	Positive	Neutral	Negative
Twit15 Train	928	1883	368
Twit15 Dev	303	670	149
Twit15 Test	317	607	113
Twit17 Train	1508	1638	416
Twit17 Dev	515	517	144
Twit17 Test	493	573	168

where $h_{i,l}^S$ is the hidden state of i-th node at l-th GCN layer, W_l, b_l are learned parameters.

3.2.4 Prediction and Loss Function

Based on (Lewis et al., 2019), the BART decoder predicts the token probability distribution using the following approach:

$$\tilde{H} = \alpha_1 \hat{H} + \alpha_2 \hat{H}^S, \tag{17}$$

$$h_t^d = \text{Decoder}(\tilde{H}; Y_{<t}), \tag{18}$$

$$\overline{H}_T = \frac{W + \tilde{H}_T}{2}, \tag{19}$$

$$P(y_t) = \text{softmax}([\overline{H}_T; C^d] h_t^d), \tag{20}$$

$$L = -\mathbb{E}_{X \sim D} \left[\sum_{t=1}^{O} \log P(y_t \mid Y_{<t}, X) \right], \tag{21}$$

where $\hat{H}$ denote the output from the AESA module, and $\hat{H}^S$ represent the output from the GCN. The parameters α_1 and α_2 indicate the respective contributions of $\hat{H}$ and $\hat{H}^S$. The hidden state of the decoder at time step t is denoted by h_t^d. The term $\tilde{H}_T$ refers to the textual portion of $\tilde{H}$. The matrix W represents the embeddings for input tokens, and C^d denotes the embeddings for sentiment categories, L is the loss function, $O = 2M + 2N + 2$ is the length of Y, and X denotes the multimodal input.

4 Experiment

4.1 Experimental Settings

Datasets:In this study, we utilize two primary benchmark datasets: Twitter2015 and Twitter2017, as detailed by (Yu and Jiang, 2019). The statistics of these two datasets are presented in Table 1.

Evaluation Metrics: The performance of our model is assessed across different tasks using various metrics. For the MABSA and MATE tasks, we utilize the F1 score, Precision (P), and Recall (R) to evaluate the performance, and in the MASC task, we only adopt Accuracy (ACC) and F1 score.

4.2 Comparison models

We compare our model with all competitive baseline models list below:

For JMASA Task: SpanABSA (Hu et al., 2019), D-GCN (Chen et al., 2020), GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019), BART (Yan et al., 2021), UMT-collapsed (Yu et al., 2020b), OSCGA-collapsed (Wu et al., 2020b), and RpBERT-collapsed (Sun et al., 2021), CLIP (Radford et al., 2021), RDS (Xu et al., 2022), JML (Ju et al., 2021), VLP-MABSA (Ling et al., 2022), AoM (Zhou et al., 2023). **For MASC Task:** ESAFN (Yu et al., 2020a), TomBERT (Yu and Jiang, 2019), CapTrBERT (Khan and Fu, 2021). **For MATE Task:** RAN (Wu et al., 2020a), UMT (Yu et al., 2020b), OSCGA (Wu et al., 2020b)

4.3 Main Results

Table 2 summarizes the results for the JMASA task. Our model achieves the highest scores across Precision, Recall, and F1 metrics on both the Twitter2015 and Twitter2017 datasets, with notable improvements of 0.95%, 0.08%, and 0.75% in Precision, Recall, and F1 on Twitter2015, and 0.41%, 0.20%, and 0.24% on Twitter2017 compared to the second-best results. This consistent performance across datasets demonstrates robust generalizability. For the MASC task as shown in Table 3, our model shows F1 score increases of 0.63 and 0.34 on the Twitter2015 and Twitter2017 datasets, respectively, though accuracy metrics vary slightly. In the MATE task in Table 4, F1 scores increase by 0.08 and 0.15 on Twitter2015 and Twitter2017, respectively, but there are inconsistencies in precision and recall metrics across datasets.

4.4 Ablation Study

4.4.1 Module Effectiveness

In this section, we evaluate the impact of each module on the model's performance, as detailed in Table 5. Removing the Aspect-based Emotion Sentiment Analysis (AESA) module results in the most significant drop in performance, highlighting its crucial role in aspect alignment and the integration of external affective commonsense knowledge. The removal of the Hybrid Curriculum Denoising (HCD) module also leads to a substantial performance decrease, underscoring its importance in enhancing overall model effectiveness. On the other hand, omitting the Graph Convolutional Network (GCN) causes only a modest reduction in

Table 2: Results of different approaches for JMASA task, Italic value denote for second-best result and bold-typed value for best result. The Δ values show the difference between our model and the previous state-of-the-art.

Modality	Approaches	2015_P	2015_R	2015_F1	2017_P	2017_R	2017_F1
TEXT	SpanABSA (Hu et al., 2019)	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN (Chen et al., 2020)	58.3	58.8	59.4	64.2	64.1	64.1
	GPT-2 (Radford et al., 2019)	66.6	60.9	63.6	55.3	59.6	57.4
	RoBERTa (Liu et al., 2019)	62.4	64.5	63.4	65.3	66.6	65.9
	BART (Yan et al., 2021)	62.9	65.0	63.9	65.2	65.6	65.4
MULTIMODAL	UMT-collapse (Yu et al., 2020b)	60.4	61.6	61.0	60.0	61.7	60.8
	OSCGA-collapse (Wu et al., 2020b)	63.1	63.7	63.2	63.5	63.5	63.5
	RpBERT-collapse (Sun et al., 2021)	49.3	46.9	48.0	57.0	55.4	56.2
	CLIP (Radford et al., 2021)	44.9	47.1	45.9	51.8	54.2	53.0
	RDS (Xu et al., 2022)	60.8	61.7	61.2	61.8	62.9	62.3
	JML* (Ju et al., 2021)	64.8	63.6	64.2	65.6	66.1	65.9
	VLP-MABSA* (Ling et al., 2022)	64.1	*68.1*	66.1	65.8	67.9	66.9
	AoM* (Zhou et al., 2023)	*65.15*	67.6	*66.35*	*65.94*	*68.0*	*67.06*
	DualDe (Ours)	**66.1**	**68.18**	**67.1**	**66.35**	**68.2**	**67.3**
	Δ	**0.95**	**0.08**	**0.75**	**0.41**	**0.2**	**0.24**

Table 3: Results of the MASC Task. Italicized values represent the second-best results, while bolded values indicate the best results. The Δ values denote the difference between our model and the previous SOTA model.

Methods	2015_ACC	2015_F1	2017_ACC	2017_F1
ESAFN	73.4	67.4	67.8	64.2
TomBERT	77.2	71.8	70.5	68.0
CapTrBERT	78.0	73.2	72.3	70.2
JML	**78.7**		72.7	
VLP-MABSA	78.6	73.8	*73.8*	71.8
AoM*	78.2	*73.81*	73.6	*72.05*
DualDe (Ours)	*78.62*	**74.44**	**74.14**	**72.39**
Δ	**-0.08**	**0.63**	**0.34**	**0.34**

Table 4: Results of different approaches for MATE task, Italic value denote for second-best result and bold-typed value for best result. The Δ values denote the difference between our model and the previous SOTA model.

Methods	2015_P	2015_R	2015_F1	2017_P	2017_R	2017_F1
RAN*	80.5	81.5	81.0	90.7	90.7	90.0
UMT*	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA*	81.7	82.1	81.9	90.2	90.7	90.4
JML*	83.6	81.2	82.4	**92.0**	90.7	91.4
VLP-MABSA*	83.6	**87.9**	*85.7*	90.8	92.6	*91.7*
AoM*	*83.72*	86.79	85.23	89.58	**92.71**	91.12
DualDe (Ours)	**84.34**	*87.27*	**85.78**	*91.01*	**92.71**	**91.85**
Δ	**0.62**	**-0.63**	**0.08**	**-0.99**	**0.0**	**0.15**

the F1-score compared to HCD, suggesting that while GCN is important for handling semantic and structural aspects of the data, its impact is less pronounced than that of HCD.

Table 5: Ablation Modules Performance

Methods	2015_P	2015_R	2015_F1
w/o AESA	62.5	62.7	62.6
w/o HCD	65.4	66.96	66.17
w/o GCN	65.2	67.5	66.33
DualDe (Ours)	**66.1**	**68.18**	**67.1**

4.4.2 Ratio Contribution Test

Figure 4 provides a detailed examination of the fine-tuning process for the contribution ratio of d_l - d_s at Hybrid Curriculum Denoising module (HCD), aiming to determine the optimal ratio for the model. Based on Figure 4, the ratio of (0.8 - 0.2) achieves the highest F1-score of 67.1. This indicates that the (0.8 - 0.2) ratio is the most effective configuration for optimizing model performance between d_l and d_s. Therefore, we select this ratio as the optimal setting for the model.

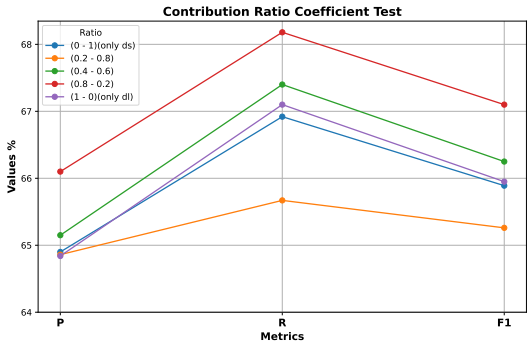

Figure 4: Illustration Contribution Ratio Coefficient Test

4.5 Case Study

Figure 5 illustrates how each module in our model processes data samples with two different levels of difficulty: "easy" (sample 1) and "hard" (sample 2). In the *Sentence-Image Denoise* step, sample 1 is considered "clean" because the image is strongly related to the text, whereas sample 2 is not. In the *Aspect-Image Denoise* step, the most impor-

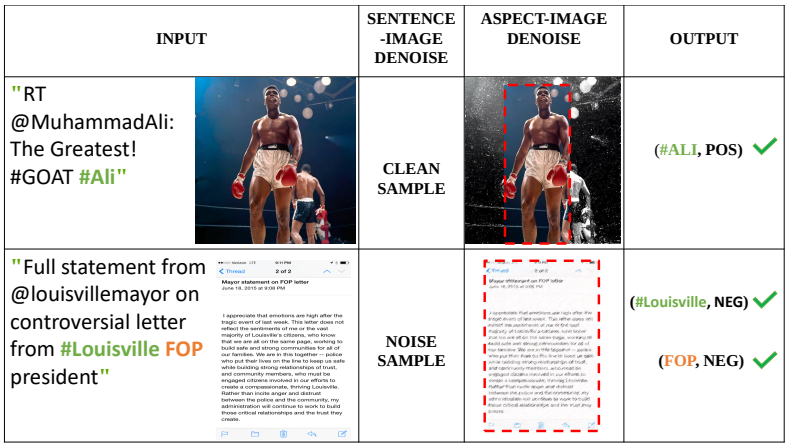

Figure 5: The figure illustrates instances where sentence-image noise and aspect-image noise impact the effectiveness of sentiment analysis. The easy sample features a clear alignment between the sentence and image, enhancing sentiment detection, while the hard sample involves a blurry image with minimal relevance to the sentence's aspects, complicating accurate sentiment evaluation.

tant image regions related to the specific aspect are highlighted, while the blurred parts are considered noise and are not emphasized during training. The output represents the model's predictions for each sample, demonstrating the effectiveness of our model.

5 Conclusion

This paper introduced DualDe, a novel framework for enhancing Multimodal Aspect-Based Sentiment Analysis (MABSA) by addressing both sentence-image and aspect-image noise. The framework comprises the Hybrid Curriculum Denoising Module(HCD), which utilizes Curriculum Learning to incrementally manage noisy data, and the Aspect-Enhanced Denoising Module(AED), which employs aspect-guided attention to filter irrelevant visual information. Empirical evaluations on the Twitter2015 and Twitter2017 datasets demonstrate that DualDenoise significantly improves Precision, Recall, and F1 scores compared to existing methods. These results affirm the model's efficacy in managing multimodal noise and its robust performance across diverse datasets. Future research may focus on refining the curriculum learning strategy and exploring broader applications of the proposed methodology.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *International Conference on Machine Learning*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Annual Meeting of the Association for Computational Linguistics*.

E. Cambria, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *International Conference on Computational Linguistics*.

Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *International Conference on Computational Linguistics*.

Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *ArXiv*, abs/1410.8586.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. *ArXiv*, abs/1906.03820.

Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4395–4405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zaid Khan and Yun Raymond Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. *Proceedings of the 29th ACM International Conference on Multimedia*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart:

Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Jinliang Lu and Jiajun Zhang. 2021. Exploiting curriculum learning in unsupervised neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

Cam-Van Thi Nguyen, Cao-Bach Nguyen, Duc-Trong Le, and Quang-Thuy Ha. 2024. Curriculum learning meets directed acyclic graph for multimodal emotion recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4259–4265, Torino, Italia. ELRA and ICCL.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom Michael Mitchell. 2019. Competence-based curriculum learning for neural machine translation. *ArXiv*, abs/1903.09848.

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. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based bert model for multimodal ner. *ArXiv*, abs/2102.02967.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. *ArXiv*, abs/1906.01130.

Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020a. Multimodal aspect extraction with region-aware alignment network. In *Natural Language Processing and Chinese Computing*.

Zhiwei Wu, Changmeng Zheng, Y. Cai, Junying Chen, Ho fung Leung, and Qing Li. 2020b. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. *Proceedings of the 28th ACM International Conference on Multimedia*.

Bo Xu, Shizhou Huang, Ming Du, Hongya Wang, Hui Song, Chaofeng Sha, and Yanghua Xiao. 2022. Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts. In *International Conference on Computational Linguistics*.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. *ArXiv*, abs/2106.04300.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. In *International Joint Conference on Artificial Intelligence*.

Jianfei Yu, Jing Jiang, and Rui Xia. 2020a. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020b. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Annual Meeting of the Association for Computational Linguistics*.

Jianfei Yu, Jieming Wang, Rui Xia, and Junjie Li. 2022. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4482–4488. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *AAAI Conference on Artificial Intelligence*.

Lei Zhang and B. Liu. 2012. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies*.

Fei Zhao, Chunhui Li, Zhen Wu, Yawen Ouyang, Jianbing Zhang, and Xinyu Dai. 2023. M2DF: Multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9057–9070, Singapore. Association for Computational Linguistics.

Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. 2023. Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics*.