






AdaptiSent: Context-Aware Adaptive Attention for Multimodal Aspect-Based Sentiment Analysis

S M Rafiuddin¹^{*}, Sadia Kamal¹, Mohammed Rakib¹, Arunkumar Bagavathi¹, and Atriya Sen¹

Oklahoma State University, Stillwater, OK 74078, USA
{srafiud, sadia.kamal, mohammed.rakib, atriya.sen}@okstate.edu
b.arun410@gmail.com

Abstract. We introduce AdaptiSent, a new framework for Multimodal Aspect-Based Sentiment Analysis (MABSA) that uses adaptive cross-modal attention mechanisms to improve sentiment classification and aspect term extraction from both text and images. Our model integrates dynamic modality weighting and context-adaptive attention, enhancing the extraction of sentiment and aspect-related information by focusing on how textual cues and visual context interact. We tested our approach against several baselines, including traditional text-based models and other multimodal methods. Results from standard Twitter datasets show that AdaptiSent surpasses existing models in precision, recall, and F1 score, and is particularly effective in identifying nuanced inter-modal relationships that are crucial for accurate sentiment and aspect term extraction. This effectiveness comes from the model’s ability to adjust its focus dynamically based on the context’s relevance, improving the depth and accuracy of sentiment analysis across various multimodal data sets. AdaptiSent sets a new standard for MABSA, significantly outperforming current methods, especially in understanding complex multimodal information.

Keywords: Multimodal Sentiment Analysis, Adaptive Cross-Modal Attention, Context-Aware Modeling

1 Introduction

The rise of social media has led to an abundance of multimodal content blending text, images, and other media, which enriches expression but complicates sentiment understanding, particularly when sentiments are tied to specific aspects, motivating Multimodal Aspect-Based Sentiment Analysis (MABSA) that jointly analyzes textual and visual signals to infer aspect-specific sentiment. Historically, sentiment analysis focused on text only, but the proliferation of multimodal data has spurred methods capable of interpreting complex text–image relationships: early works like Yang *et al.* (2022) integrate visual data into text analysis via a Cross-Modal Multitask Transformer [1], while Zhu *et al.* (2015)

^{*} Corresponding author

emphasize leveraging linguistic structures in joint models [2]. Recent advances leverage large pre-trained transformers and cross-modal attention to fuse text and image features for MABSA [5,6], yet most apply direct fusion without addressing the modality gap, the differing ways text and images encode sentiment, which can lead to semantic inconsistencies and reduced performance [7,8]. While text often expresses opinions explicitly, images offer implicit emotional cues that may reinforce or contradict sentiment [3], but many models assume equal visual importance or ignore uncertain visual signals [10], and even selective fusion or semantic-bridging strategies [4] often fail to capture fine-grained aspect alignment or adaptively weight multimodal signals. This paper presents *AdaptiSent*, a new MABSA framework featuring (1) dynamic importance scoring, (2) context-aware modality weighting, (3) aspect-specific adaptive masking, (4) visual-guided textual augmentation with custom balancing, and (5) multimodal semantic alignment, which together enable adaptive, per-aspect attention modulation and cross-modal regularization to improve sentiment classification and aspect extraction without added architectural complexity.

2 Method

2.1 Problem Formulation

Multimodal Aspect-Based Sentiment Analysis (MABSA) processes a text sequence $T^0 \in \mathbb{R}^{L \times d_t}$ and image features $V_I \in \mathbb{R}^{K \times d_v}$ to extract a subset of aspect terms $\mathcal{A}_{\text{ext}} \subseteq \mathcal{A}$ and predict their sentiments in $\mathcal{S} = \{\text{positive}, \text{negative}, \text{neutral}\}$. Formally, we learn

$$f : \mathcal{A} \times \mathbb{R}^{L \times d_t} \times \mathbb{R}^{K \times d_v} \rightarrow \mathcal{S} \quad (1)$$

and output the set

$$D = \{(a_i, s_i) \mid a_i \in \mathcal{A}_{\text{ext}}, s_i = f(a_i, T^0, V_I)\} \quad (2)$$

Here $f : \mathcal{A} \times \mathbb{R}^{L \times d_t} \times \mathbb{R}^{K \times d_v} \rightarrow \mathcal{S}$ is a multimodal sentiment classification function, and $\mathbf{D} \subseteq \mathcal{A} \times \mathcal{S}$.

2.2 Multimodal Representation

Textual Representation: RoBERTa BPE tokenizes text into L tokens (including $t_{\text{cls}}, t_{\text{sep}}$); embeddings $E(t_i) \in \mathbb{R}^{d_t}$ plus positional P_i yield $T^0 \in \mathbb{R}^{(L+2) \times d_t}$.

Visual Representation: Image I is split into K patches embedded as $E(p_i) \in \mathbb{R}^{d_v}$ with prepended p_{cls} and positional P_i , giving $V_I \in \mathbb{R}^{(K+1) \times d_v}$.

2.3 Method for Multimodal Aspect Term Extraction:

Importance Score Computation For each token t_i we compute the visual-to-text relevance score

$$R_{\text{vis}}(t_i) = \text{softmax}(\text{att}(E[t_i], V_I) + \text{att}(E[t_i], C^0)) \quad (3)$$

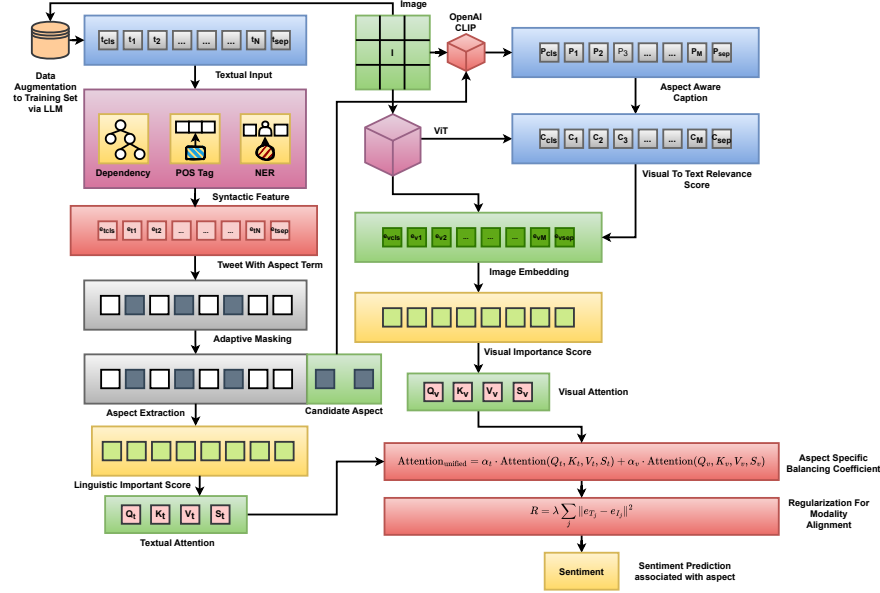


Fig. 1: AdaptiSent: an MABSA framework using LLM-augmented aspects, linguistic-masked RoBERTa text, CLIP captions and ViT visuals fused by relevance-weighted cross-modal attention for per-aspect sentiment.

the linguistic importance score

$$R_{\text{ling}}(t_i) = \text{sigmoid}(W_d d_i + W_p p_i + W_n n_i + b) \quad (4)$$

an adaptive threshold

$$\theta = \mu_S + \alpha_m \sigma_S \quad (5)$$

and finally feed the masked sequence into RoBERTa to get

$$\mathcal{A}_{\text{ext}} = \text{RoBERTa}_{\text{masked}}(m(T^0), V_I, C^0) \quad (6)$$

Here, C_0 denotes the caption embedding and m is the masking operator.

2.4 Method for Multimodal Aspect based Sentiment Classification:

Visual-Guided Textual Data Augmentation We encode image I via ViT to obtain \mathbf{e}_I , use an LLM to generate augmented text $\mathbf{T}' = \text{LLM}_{\text{aug}}(\mathbf{T}, \mathbf{e}_I, \mathcal{A}_{\text{ext}})$ and embed it with RoBERTa to $\mathbf{e}_{T'}$, including it in training based on $\cos(\mathbf{e}_{T'}, \mathbf{e}_I)$; concurrently, for each aspect a_j we extract \mathbf{e}_{T_j} (RoBERTa) and \mathbf{e}_{I_j} (ViT+C), fusing them as $\alpha_j \mathbf{e}_{T_j} + (1 - \alpha_j) \mathbf{e}_{I_j}$ with learnable α_j .

Context-Adaptive Cross-Modal Attention Mechanism We propose a cross-modal attention mechanism that dynamically integrates visual-to-text relevance and linguistic importance scores to enhance aspect-based sentiment analysis: given token-level linguistic $R_{\text{ling}}(t_i)$ and visual $R_{\text{vis}}(t_i)$ importance scores, we compute a combined importance score

$$\mathbf{S}(t_i) = \gamma R_{\text{ling}}(t_i) + (1 - \gamma) R_{\text{vis}}(t_i) \quad (7)$$

and modify the standard scaled dot-product attention to incorporate \mathbf{S} as an adaptive bias

$$\text{Attention}(Q, K, V, \mathbf{S}) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \beta \mathbf{S}\right)V \quad (8)$$

where β is a trainable scaling factor; we then compute modality weighting coefficients

$$\alpha_t = \frac{\sum_i R_{\text{ling}}(t_i)}{\sum_i R_{\text{ling}}(t_i) + \sum_i R_{\text{vis}}(t_i)}, \quad \alpha_v = 1 - \alpha_t \quad (9)$$

and form the unified attention output

$$\text{Attention}_{\text{unified}} = \alpha_t \text{Attention}(Q_t, K_t, V_t, \mathbf{S}_t) + \alpha_v \text{Attention}(Q_v, K_v, V_v, \mathbf{S}_v) \quad (10)$$

allowing dynamic focus on the most relevant cross-modal features while maintaining efficiency by operating solely over token-level importance scores without increasing model complexity.

Regularization for Modality Alignment To align textual and visual aspect embeddings, we map $\mathbf{e}_{T_j} \in \mathbb{R}^{d_t}$ and $\mathbf{e}_{I_j} \in \mathbb{R}^{d_v}$ into a shared space via $\mathbf{e}'_{T_j} = \mathbf{W}_T \mathbf{e}_{T_j} + b_T$, $\mathbf{e}'_{I_j} = \mathbf{W}_I \mathbf{e}_{I_j} + b_I$ (with $\mathbf{W}_T \in \mathbb{R}^{d \times d_t}$, $\mathbf{W}_I \in \mathbb{R}^{d \times d_v}$), then penalize their squared Euclidean distance:

$$R = \lambda \sum_{j=1}^m \|\mathbf{e}'_{T_j} - \mathbf{e}'_{I_j}\|^2 \quad (11)$$

where λ balances the alignment strength.

Key parameters: α_m (masking threshold scaling; trainable), α_j (modality balancing coefficient; trainable), β (attention scaling factor; trainable), $\gamma \in [0, 1]$ (linguistic–visual balance; hyperparameter, 0.3), and λ (modality alignment strength; hyperparameter, 0.1).

2.5 Training Procedure

Loss Function for MABSA The overall loss jointly optimizes aspect term extraction, sentiment classification, and modality alignment:

$$L = \sum_{i=1}^n w_i \text{CE}(p_i, y_i) + \lambda \sum_{j=1}^m \|\mathbf{e}'_{T_j} - \mathbf{e}'_{I_j}\|^2 \quad (12)$$

where w_i weights token-level cross-entropy by importance and λ controls modality-alignment.

3 Results

3.1 Compared Baseline Models

We evaluate AdaptiSent on *Twitter-15* and *Twitter-17*, marking an extraction correct only if both aspect term and sentiment match. Backbones are initialized with RoBERTa-base and ViT-base-patch16 (768-d hidden, 8 cross-modal heads). Models are trained with AdamW at 2×10^{-5} LR (warmup), a 60-token limit, and batch size 16 on NVIDIA A100 GPUs (<3 h per run). Results (precision, recall, F1) are averaged over three seeds.

Table 1: Performance comparison on MABSA datasets (**Twitter15** and **Twitter17**) with Precision (Prec), Recall (Rec), and F1 scores. Values in parentheses indicate standard deviation over 3 runs with different random seeds.

Model	Twitter15			Twitter17		
	Prec	Rec	F1	Prec	Rec	F1
Text-Only Models						
SPAN [11]	53.7	53.9	53.8	59.6	61.7	60.6
D-GCN [12]	58.3	58.8	58.6	64.2	64.1	64.1
BART [13]	62.9	65.0	63.9	65.2	65.6	65.4
RoBERTa	62.9	63.7	63.3	65.1	66.2	65.7
Multimodal Models						
UMT [14]	58.4	61.4	59.9	62.3	62.4	62.4
OSCGA [15]	61.7	63.4	62.5	63.4	64.0	63.7
JML [16]	65.0	63.2	64.1	66.5	65.5	66.0
VLP [17]	68.3	66.6	67.4	69.2	68.0	68.6
CMMT [1]	64.6	68.7	66.6	67.6	69.4	68.5
M2DF [9]	67.0	68.3	67.6	67.9	68.8	68.4
DTCA [18]	67.3	69.5	68.4	69.6	71.2	70.4
AoM [19]	67.9	69.3	68.6	68.4	71.0	69.7
TMFN [4]	68.4	69.6	69.0	70.7	71.2	71.0
DQPSA [10]	71.7	72.0	71.9	71.1	70.2	70.6
Large Language Models (Text Only)						
Llama2	53.6	55.0	54.3	57.6	58.8	58.2
Llama3	56.4	57.2	56.8	61.8	62.5	62.2
GPT-2.0	47.8	49.2	48.5	52.0	53.9	52.9
GPT-3.5	50.9	51.9	51.4	55.6	56.1	55.9
AdaptiSent	70.9 (± 0.27)	72.8 (± 0.39)	71.9 (± 0.18)	71.4 (± 0.52)	71.8 (± 0.31)	71.6 (± 0.24)

SPAN introduces span-based extraction to resolve sentiment inconsistencies, outperforming sequence-tagging methods [11]. **D-GCN** incorporates syntactic dependencies via directional graph convolutions [12]. **BART** uses denoising seq-to-seq pre-training for robust text understanding [13], while **RoBERTa** refines BERT’s objectives and data scale. Among multimodal methods, **UMT** unifies textual and visual encoders [14], **OSCGA** employs dense co-attention [15], **JML**, **VLP**, and **CMMT** leverage vision-language pre-training [16,17,1], **M2DF** and **DTCA** exploit advanced transformers and denoising channels [9,18], **AoM** selectively aligns image regions [19], **TMFN** uses multi-grained fusion [4], and **DQPSA** refines cross-modal gating [10]. General LLMs (**Llama2**, **Llama3**,

GPT-2.0, GPT-3.5) excel in language understanding but lack multimodal training. **AdaptiSent** (Ours) combines LLM-augmented aspect insertion, syntactic masking, and learnable cross-modal self-attention with modality-alignment regularization to isolate genuine sentiment signals and achieve state-of-the-art performance.

3.2 Ablation Studies

Table 2: Ablation study for MABSA with different feature combinations, evaluated on **Twitter15** and **Twitter17**. Results are averaged over 3 runs with random seeds.

Model	Twitter15			Twitter17		
	Prec	Rec	F1	Prec	Rec	F1
w/o Aspect-Aware Captions	67.13	63.51	65.27	68.37	65.53	66.92
w/o Regularization for Modality Alignment	67.89	64.44	66.12	70.22	66.26	68.18
w/o Aspect-Specific Balancing Coefficients	65.11	64.30	64.70	67.08	64.41	65.72
w/o Data Augmentation	74.56	66.64	70.38	74.50	67.68	70.93
w/o Context-Based Masking	70.11	64.56	67.22	72.34	67.77	69.98
AdaptiSent (Full Model)	70.95	72.85	71.89	71.42	71.83	71.62

Table 2 presents ablations: removing modality weights yields the largest F1 drop (71.89→64.70 on Twitter-15; 71.62→65.72 on Twitter-17), followed by captions (-6.62, -4.70), alignment regularizer (-5.77, -3.44), context masking (-4.67, -1.64), and data augmentation (-1.51, -0.69). Figure 2 shows $\gamma = 0.3$, $\lambda = 0.1$ near optimum. The full model attains 71.89 and 71.62 F1.

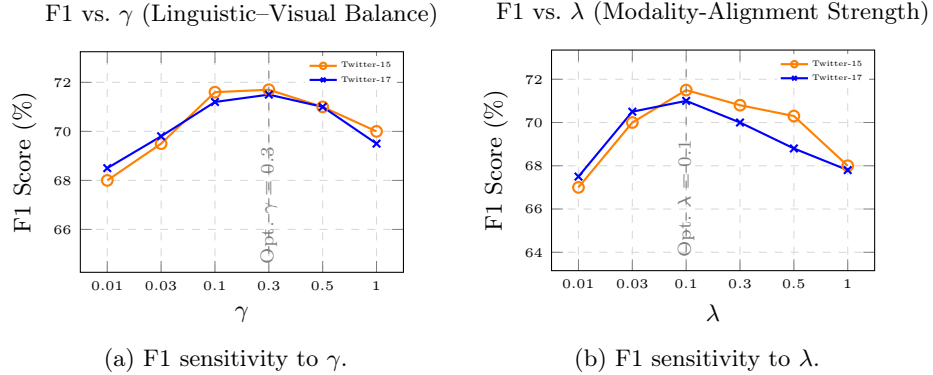


Fig. 2: Hyperparameter sensitivity: (a) variation with γ , peaking at 0.3; (b) variation with λ , peaking at 0.1.

3.3 Case Studies

Table 3: Comparison of sentiment analysis models (three case studies).




Image	Text	Ground Truth	TMFN Model	AoM Model	DPQSA Model	Ours
	First day of school in Chicago and at Cameron Elementary. This kindergartener wasn't impressed by the mayoral visit	(Chicago, Neutral) (Cameron Elementary, Negative)	✓ (Chicago, Neutral) × (Cameron Elementary, Positive)	✓ (Chicago, Neutral) × (Cameron Elementary, Neutral)	× (Chicago, Positive) ✓ (Cameron Elementary, Negative)	✓ (Chicago, Neutral) ✓ (Cameron Elementary, Negative)
	RT @ ItsChuckBass : Chuck Bass is everything #MCM	(Chuck Bass, Positive) (#MCM, Neutral)	× (Chuck Bass, Negative) ✓ (#MCM, Neutral)	× (Chuck Bass, Neutral) ✓ (#MCM, Neutral)	✓ (Chuck Bass, Positive) × (#MCM, Positive)	✓ (Chuck Bass, Positive) ✓ (#MCM, Neutral)
	Why Chris Brown and Beyonce look like they tryna lead Praise and Worship?	(Chris Brown, Negative) (Beyonce, Negative)	✓ (Chris Brown, Negative) × (Beyonce, Positive)	× (Chris Brown, Positive) ✓ (Beyonce, Negative)	× (Chris Brown, Neutral) ✓ (Beyonce, Negative)	✓ (Chris Brown, Negative) ✓ (Beyonce, Negative)

Table 3 contrasts ground truth with predictions from **TMFN** [4], **AoM** [19], **DPQSA** [10], and **AdaptiSent** across three case studies, where AdaptiSent correctly matches all aspect-sentiment pairs.

4 Conclusion & Future Work

AdaptiSent employs adaptive cross-modal attention with dynamic modality weights and a squared-distance regularizer for robust embedding alignment and improved out-of-distribution generalization. Future work will explore lightweight attention, misaligned input handling, scaling to noisier data, and neuro-symbolic sentiment reasoning using commonsense knowledge and contrastive/counterfactual methods.

References

1. Yang L, Na JC, Yu J. Cross-modal multitask transformer for end-to-end multi-modal aspect-based sentiment analysis. *Information Processing & Management*. 2022;59(5):103038.
2. Zhu L, Sun H, Gao Q, Yi T, He L. Joint multimodal aspect sentiment analysis with aspect enhancement and syntactic adaptive learning. In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*; 2015.
3. Yang H, Zhao Y, Qin B. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2022. p. 3324–3335.
4. Wang D, He Y, Liang X, Tian Y, Li S, Zhao L. TMFN: A target-oriented multi-grained fusion network for end-to-end aspect-based multimodal sentiment analysis. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*; 2024. p. 16187–16197.

5. Hu H. A vision-language pre-training model based on cross attention for multi-modal aspect-based sentiment analysis. In: 2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL); 2024. p. 370–375. IEEE.
6. Fan H, Chen J. Position perceptive multi-hop fusion network for multimodal aspect-based sentiment analysis. IEEE Access. 2024.
7. Xiang Y, Cai Y, Guo J. MSFNet: Modality smoothing fusion network for multi-modal aspect-based sentiment analysis. *Frontiers in Physics*. 2023;11:1187503.
8. Xu Z, Su Q, Xiao J. Multimodal aspect-based sentiment classification with knowledge-injected transformer. In: 2023 IEEE International Conference on Multimedia and Expo (ICME); 2023. p. 1379–1384. IEEE.
9. Zhao F, Li C, Wu Z, Ouyang Y, Zhang J, Dai X. 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 (EMNLP); 2023. p. 9057–9070.
10. Peng T, Li Z, Wang P, Zhang L, Zhao H. A novel energy based model mechanism for multi-modal aspect-based sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2024. 38, p. 18869–18878.
11. Hu M, Peng Y, Huang Z, Li D, Lv Y. Open-domain targeted sentiment analysis via span-based extraction and classification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 537–546.
12. Chen G, Tian Y, Song Y. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In: Proceedings of the 28th International Conference on Computational Linguistics; 2020. p. 272–279.
13. Lewis M. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. 2019.
14. Yu J, Jiang J, Yang L, Xia R. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. p. 3342–3352.
15. Wu Z, Zheng C, Cai Y, Chen J, Leung H, Li Q. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020. p. 1038–1046.
16. Ju X, Zhang D, Xiao R, Li J, Li S, Zhang M, Zhou G. 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 (EMNLP); 2021. p. 4395–4405.
17. Ling Y, Yu J, Xia R. Vision-language pre-training for multimodal aspect-based sentiment analysis. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022. p. 2149–2159.
18. Yu Z, Wang J, Yu LC, Zhang X. Dual-encoder transformers with cross-modal alignment for multimodal aspect-based sentiment analysis. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2022. p. 414–423.
19. Zhou R, Guo W, Liu X, Yu S, Zhang Y, Yuan X. AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023. p. 8184–8196.