Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9113

MGAFN: Multi-Granularity Alignment and Fusion Network Based on MultiModal Aspect-Based Sentiment Analysis

Fei Wei. Xiaoyu Sun. Haopeng Zhang and Yifang Chen
School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan 250357, Shandong, China
Email: wifi_sd@163.com

How to cite this paper: Wei, F., Sun, X., Zhang, H., & Chen, Y. (2026). MGAFN: Multi-granularity alignment and fusion network based on multimodal aspect-based sentiment analysis. Journal of Computer Science and Frontier Technologies, 5(1), 29–38. ISSN Print: 3104-4204, ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9113
Published: 2026-11-24

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Abstract

Multimodal aspect-based sentiment analysis (MABSA) aims to identify the sentiment polarity toward a target aspect using textual and visual information. Existing methods often perform cross-modal interaction at a single semantic granularity, making it difficult to capture both aspect-level details and global context. Moreover, visual content may be irrelevant to the target aspect, and static fusion can introduce noise into sentiment prediction. To address these challenges, we propose a Multi-Granularity Alignment and Fusion Network (MGAFN). MGAFN constructs fine- and coarse-grained representations of text and images, then aligns them at the token–patch and sentence–scene levels, respectively. An adaptive gating mechanism regulates the contribution of visual features according to their relevance to the target aspect. A coarse-grained semantic distribution constraint further encourages consistency between paired text and image representations. Experiments on the Twitter-2015 and Twitter-2017 datasets show that MGAFN achieves competitive performance against representative MABSA baselines. These results suggest that multi-granularity alignment and adaptive visual fusion can improve aspect-level sentiment prediction while reducing interference from irrelevant visual information.

Keywords

Multimodal Aspect-Based Sentiment Analysis; Multi-Granularity Cross-Modal Alignment and Fusion; Adaptive Gating Fusion; Visual Noise Suppression

1. Introduction

Aspect-based sentiment analysis (ABSA) identifies the sentiment expressed toward a particular target rather than assigning one label to an entire post. In text–image content, this distinction also determines how visual information should be interpreted. An image may clarify the entity being discussed, supply context for an otherwise ambiguous expression, or contribute little to the judgment of a given

aspect. Multimodal aspect-based sentiment analysis (MABSA) therefore requires a model to combine information across modalities while retaining the target-specific nature of the task 错误!不能识别的开关参数。. 错误!不能识别的开关参数。. illustrates this setting through posts involving aspects such as “Steph Curry” and “Kanye.” The central difficulty is deciding which visual evidence can usefully complement the text.

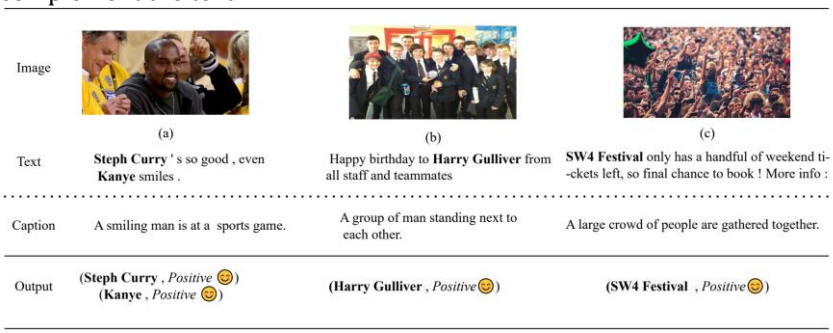

Figure 1. Examples of multimodal aspect-based sentiment analysis

Existing approaches provide several ways to connect textual and visual representations. Global image features extracted by pretrained encoders such as ResNet 错误!不能识别的开关参数。 offer a compact description of the scene, but can leave local details mixed with background content. Finer interactions seek to make these details accessible to the language representation. For example, AoM models aspect-oriented visual information 错误!不能识别的开关参数。, while CoolNet combines image descriptions with graph-based representations for cross-modal interaction 错误!不能识别的开关参数。. These approaches motivate closer examination of how local correspondence and broader context contribute to a sentiment decision, and how their contributions should be controlled when the image provides weak evidence.

Two connected challenges arise in this process. The first concerns the granularity of the evidence. A word or local image region may carry a decisive cue, whereas the meaning of a post can also depend on the sentence as a whole or the surrounding scene. Emphasizing only one scale can leave the other insufficiently represented. The second concerns the contribution of visual information. An image can be associated with a post without supporting the sentiment expressed toward every aspect in it. Consequently, retrieving a visually related feature does not by itself establish how strongly that feature should influence prediction. Combining multiple granularities and regulating their contributions are therefore separate requirements of the fusion process.

This perspective leads to three questions that guide our model design and evaluation. Do local and global cross-modal representations provide complementary information for aspect-level sentiment classification? Does adaptive gating improve prediction compared with fusion that omits this control? Does a global text–image contrastive objective provide useful supervision in addition to the sentiment labels?

These questions connect the proposed components to comparisons that can be examined through ablation experiments.

We address these questions with the Multi-Granularity Alignment and Fusion Network (MGAFN). The model constructs graph-based token features and patch-level image features for fine-grained interaction, together with pooled textual and visual representations for coarse-grained modeling. Image captions provide an additional description of visual content on the textual side. In the fine-grained branch, cross-modal attention retrieves patch information for the textual sequence. In the coarse-grained branch, sentence and scene representations supply global context. Separate scalar gates control visual contributions within the two branches, while a feature-wise gate combines their outputs. The resulting representation is used for sentiment classification. A bidirectional InfoNCE objective on the global features complements the classification loss by encouraging correspondence between paired text and images.

We evaluate MGAFN on Twitter-2015 and Twitter-2017. It achieves accuracies of 79.85% and 71.23% and macro-F1 scores of 75.35% and 69.44%, respectively. These results lead the compared methods in three of the four reported metrics; the Twitter-2017 macro-F1 remains below that of DAGG-Net. Ablation experiments assess the contributions of the two alignment branches, gating, captions, and contrastive learning. Together, these evaluations test whether coordinating evidence across granularities improves prediction on the two benchmarks.

The main contributions of this paper are as follows:

1. We propose MGAFN, a framework that combines token–patch interaction with sentence–scene modeling to represent local and global evidence for MABSA.
2. We develop a fusion scheme with scalar gates within each granularity and a feature-wise gate across granularities. A global contrastive objective supplies additional supervision for text–image correspondence.
3. We evaluate the framework on two public benchmarks and use component-wise ablations to examine the contributions of multi-granularity modeling, adaptive gating, caption features, and contrastive learning.

2. Method

Given a post T, its image I, and a target aspect A, MGAFN predicts a positive, neutral, or negative sentiment label. 错误!不能识别的开关参数。. organizes the computation into multimodal feature extraction, coarse-grained and fine-grained alignment, and multi-granularity fusion. Image captions enter the fine-grained textual sequence, while the projected global text and image features also supply the auxiliary InfoNCE objective.

2.1. Multimodal Feature Extraction

A fine-tuned RoBERTa 错误!不能识别的开关参数。 encoder produces token states

$H^t \in \mathbb{R}^{L_n \times d}$, where L_n is the post length and $d = 768$. We construct a semantic graph by retaining top-z links from aggregated multi-head attention and a syntactic graph using Perturbed Masking 错误!不能识别的开关参数。 followed by tree decoding 错误!不能识别的开关参数。. Their adjacency matrices, A^{sem} and A^{syn}, indicate retained attention links and dependency edges, respectively. Two DCGCN 错误!不能识别的开关参数。 paths share parameters and average their outputs:

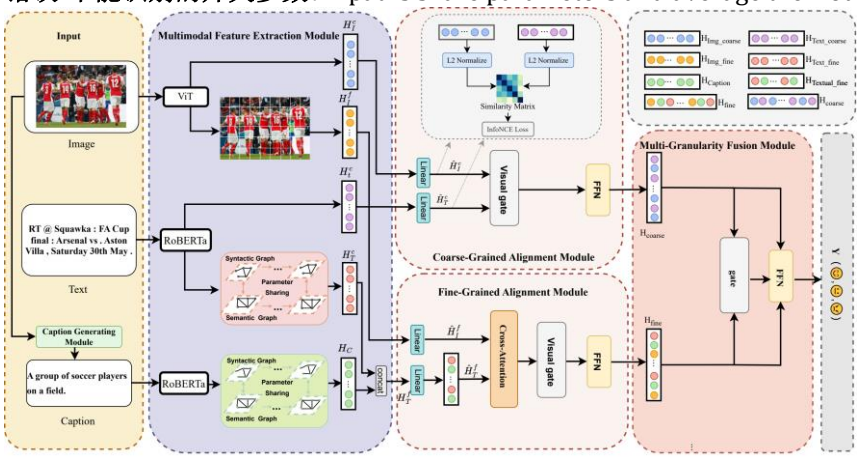

Figure 2. The overall architecture of MGAFN.

2.2. Multimodal Feature Extraction

A fine-tuned RoBERTa 错误!不能识别的开关参数。 encoder produces token states $H^t \in \mathbb{R}^{L_n \times d}$, where L_n is the post length and $d = 768$. We construct a semantic graph by retaining top-z links from aggregated multi-head attention and a syntactic graph using Perturbed Masking 错误!不能识别的开关参数。 followed by tree decoding 错误!不能识别的开关参数。. Their adjacency matrices, A^{sem} and A^{syn}, indicate retained attention links and dependency edges, respectively. Two DCGCN 错误!不能识别的开关参数。 paths share parameters and average their outputs:

$$H_t^f = 1/2\,[\text{DCGCN}(A^{\text{syn}}, H^t) + \text{DCGCN}(A^{\text{sem}}, H^t)]. \quad (1)$$

The image-caption branch generates a description of I, following the caption-based approach of CaTrBERT 错误!不能识别的开关参数。. RoBERTa and the same dual-path graph procedure encode its L_c tokens as $H_C \in \mathbb{R}^{L_c \times d}$. Concatenation along the sequence axis forms the fine-grained textual-side representation:

$$H_T^f = \text{Concat}_{\text{seq}}(H_t^f, H_C) \in \mathbb{R}^{(L_n + L_c) \times d}. \quad (2)$$

The frozen CLIP ViT-B/32 错误!不能识别的开关参数。 encoder supplies 49 patch states, H_I^f, under the 7×7 patch configuration, and a pooled image feature, H_I^c. For the global text feature, we apply an affine transformation to the post encoder's [CLS] state, $H_T^c = W_{tc} h_{[\text{CLS}]} + b_{tc}$.

Thus, graph-encoded post and caption tokens form the local text branch, while the post-level state forms the global text branch.The four feature sets pass through independent linear projections and layer normalization before alignment. For an

input X with feature width d_X, the corresponding parameters are $W_X \in \mathbb{R}^{d_X \times d_a}$ and $b_X \in \mathbb{R}^{d_a}$:

$$\hat{X} = \mathrm{LN}(XW_X + b_X), \qquad d_a = 256. \tag{3}$$

The projection preserves sequence length: $\hat{H}_T^f$ and $\hat{H}_I^f$ contain $L_n + L_c$ and 49 vectors, respectively, while $\hat{H}_T^c$ and $\hat{H}_I^c$ are global vectors. Every projected feature has width d_a.

2.3. Coarse- and Fine-Grained Alignment

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire journals, and not as an independent document. Please do not revise any of the current designations.

The upper alignment branch in Fig. 2 operates on the projected global features. Its textual and visual inputs are $u_c = \hat{H}_T^c$ and $v_c = \hat{H}_I^c$. A visual gate interpolates between these vectors, and an FFN produces the coarse-grained output. The same projected global features, before gating, are L2-normalized for the contrastive objective described in Section 2.4.

The lower branch first retrieves local visual information using $\hat{H}_T^f$ as the query sequence and $\hat{H}_I^f$ as the keys and values:

$$H_{T\leftarrow I}^f = \mathrm{CrossAttn}(\hat{H}_T^f, \hat{H}_I^f, \hat{H}_I^f). \tag{4}$$

Within each head, scaled dot-product attention is $\mathrm{softmax}(QK^\top/\sqrt{d_k})V$, where d_k is the key width. Four heads are used. Attention pooling then reduces the textual and cross-modal sequences to two vectors:

$$u_f = \mathrm{AttnPool}(\hat{H}_T^f), \qquad v_f = \mathrm{AttnPool}(H_{T\leftarrow I}^f). \tag{5}$$

Both branches use the same form of scalar visual gate and residual fusion. Writing $\ell \in \{c, f\}$ for the granularity, their operations are:

$$g_\ell = \sigma\left(W_g^\ell [u_\ell; v_\ell] + b_g^\ell\right) \in (0,1), \tag{6}$$

$$h_\ell = \mathrm{FFN}_\ell([u_\ell;\, u_\ell + g_\ell(v_\ell - u_\ell)]) \in \mathbb{R}^{d_a}. \tag{7}$$

Here, $[\,;\,]$ denotes feature concatenation, σ is the sigmoid, and FFN is a two-layer feed-forward network. The interpolation approaches u_ℓ as the gate approaches zero and v_ℓ as it approaches one; the separate u_ℓ path is retained in either case. In the fine-grained branch, this textual-side path includes caption information. The outputs h_c and h_f correspond to H_{coarse} and H_{fine} in Fig. 2.

2.4. Multi-Granularity Fusion

The right-hand module in Fig. 2 combines the two branch outputs. A feature-wise gate controls their contributions at each coordinate:

$$g = \sigma\big(W_f[h_f; h_c] + b_f\big) \in (0,1)^{d_a},$$
$$h_{\text{fused}} = g \odot h_f + (1-g) \odot h_c. \tag{8}$$

The fusion FFN receives the gated vector together with direct paths from both branch outputs, matching the three incoming paths in the diagram:

$$h_{\text{final}} = \text{FFN}\big([h_f; h_c; h_{\text{fused}}]\big) \in \mathbb{R}^{d_a}. \tag{9}$$

2.5. Sentiment Prediction and Optimization

The final representation is mapped to three class probabilities. For a minibatch of B samples, let y_{ik} be the one-hot label for class k and $\hat{y}_{ik}$ its predicted probability. Classification uses the batch-averaged cross-entropy:

$$\hat{y}_i = \text{softmax}(W_u h_{\text{final},i} + b_u),$$
$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{3} y_{ik} \log \hat{y}_{ik}. \tag{10}$$

For the auxiliary global objective, we L2-normalize $\hat{H}_{T,i}^{c}$ and $\hat{H}_{I,i}^{c}$ to obtain $\boldsymbol{t}_i$ and $\boldsymbol{v}_i$. Their similarities $s_{ij} = \boldsymbol{t}_i^{\top}\boldsymbol{v}_j$ form the matrix shown above the coarse-grained branch. Bidirectional InfoNCE [错误！不能识别的开关参数。] treats each paired text–image example as positive and the remaining batch positions as negatives:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2B}\sum_{i=1}^{B}\Big[\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ji})}\Big]. \tag{11}$$

The classification and global contrastive objectives are combined as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{con}}, \qquad \lambda = 0.1. \tag{12}$$

The contrastive term supervises global text–image correspondence; the scalar and feature-wise gates participate in the classification path. Caption features enter local alignment, whereas the contrastive branch uses only the global post and image features.

3. Experiments

3.1. Datasets and Evaluation Metrics

We evaluate MGAFN on Twitter-2015 and Twitter-2017 [错误！不能识别的开关参数。] for three-class aspect sentiment classification. lists the class distributions and dataset splits.

Table 1. Statistics of the Twitter-2015 and Twitter-2017 datasets.

Label	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	517	573
Negative	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234

We report accuracy (Acc) and macro-F1, which gives equal weight to each sentiment class.

3.2. Implementation Details

We use the encoders described in Section 2.1, with Perturbed Masking at layer 11 and semantic-graph top-z = 3.The alignment width is d_a = 256 with four attention heads; the contrastive weight is λ = 0.1 and the temperature is initialized to 0.07. AdamW uses an initial learning rate of 2×10^{-5} with cosine annealing. We train for 30 epochs with batch size 16 and dropout 0.1. Results average three runs with different random seeds, using PyTorch 1.13 on an NVIDIA RTX 4090 GPU.

3.3. Baseline Models

Unimodal baselines include image-based Res-Target 错误!不能识别的开关参数。 and text-based AE-LSTM 错误!不能识别的开关参数。, MGAN 错误!不能识别的开关参数。, BERT 错误!不能识别的开关参数。, ASGCN 错误!不能识别的开关参数。, Dual-GCN 错误!不能识别的开关参数。, and BERT4GCN 错误!不能识别的开关参数。.

Cross-modal baselines include MIMN 错误!不能识别的开关参数。, ESAFN 错误!不能识别的开关参数。, TomBERT 错误!不能识别的开关参数。, and CoolNet 错误!不能识别的开关参数。.

We also compare multi-granularity or adaptive fusion models: TMFN [21], MGFN-SD 错误!不能识别的开关参数。, AMIFN 错误!不能识别的开关参数。, DMIN 错误!不能识别的开关参数。, MG-VTFM 错误!不能识别的开关参数。, and DAGG-Net 错误!不能识别的开关参数。.

3.4. Comparative Results and Analysis

MGAFN ranks first in three of the four dataset–metric combinations in 错误!不能识别的开关参数。.

Table 2. Comparison of MGAFN with state-of-the-art MABSA methods (%).

Modality	Method	Twitter-2015		Twitter-2017	
		Acc	Macro-F1	Acc	Macro-F1
Image	Res-Target	59.88	46.48	58.59	53.98
Text	AE-LSTM	70.30	63.43	61.67	57.97
Text	MGAN	71.17	64.21	64.75	61.46
Text	BERT	74.15	68.86	68.15	65.23
Text+Image	MIMN	71.84	65.69	65.88	62.99
Text	ASGCN	72.15	70.40	–	–
Text+Image	ESAFN	73.38	67.37	67.83	64.22
Text	Dual-GCN	75.92	73.29	–	–
Text+Image	TomBERT	77.15	71.15	70.34	68.03
Text	BERT4GCN	74.77	72.02	70.15	68.40
Text+Image	CoolNet	79.07	74.62	70.50	68.58
Text+Image	TMFN	72.43	66.92	70.66	68.95
Text+Image	MGFN-SD	79.36	74.81	70.77	68.07
Text+Image	AMIFN	78.69	74.83	70.29	68.21
Text+Image	DMIN	78.72	73.12	70.55	68.45
Text+Image	MG-VTFM	79.00	74.40	70.69	68.58
Text+Image	DAGG-NET	79.42	74.63	70.68	69.52
Text+Image	MGAFN(ours)	79.85	75.35	71.23	69.44

Note: Best results are shown in bold, and second-best results are underlined.

On Twitter-2015, MGAFN achieves 79.85% accuracy and 75.35% macro-F1, exceeding DAGG-Net and AMIFN by 0.43 and 0.52 percentage points, respectively.

On Twitter-2017, accuracy reaches 71.23%, exceeding MGFN-SD by 0.46 points; macro-F1 ranks second at 69.44%, 0.08 points below DAGG-Net.

These gains are modest and metric-dependent; the reported point estimates do not establish statistical significance.

3.5. Ablation Study

错误!不能识别的开关参数。. reports ablations of fine-grained alignment (w/o Fine), coarse-grained alignment (w/o Coarse), gating (w/o Gate), the contrastive objective (w/o InfoNCE), and captions (w/o Caption).

Table 3. Ablation study of MGAFN (%).

Model	Twitter-2015		Twitter-2017	
	Acc	Macro-F1	Acc	Macro-F1
w/o Fine	73.39	69.41	62.70	58.95
w/o Coarse	77.39	72.42	69.28	67.19
w/o Gate	78.65	73.89	70.48	68.58
w/o InfoNCE	79.36	74.71	70.75	68.85
w/o Caption	78.33	73.57	70.47	68.39
MGAFN	79.85	75.35	71.23	69.44

Removing fine-grained alignment causes the largest losses across all metrics, including macro-F1 decreases of 5.94 and 10.49 points on Twitter-2015 and Twitter-2017, respectively. Removing coarse-grained alignment lowers macro-F1 by 2.93 and 2.25 points. Both branches contribute, with the larger observed effect from fine-grained alignment.

Removing gating reduces macro-F1 by 1.46 and 0.86 points on the two datasets; removing captions reduces it by 1.78 and 1.05 points, respectively.

Without InfoNCE, macro-F1 falls by 0.64 and 0.59 points, indicating a smaller auxiliary contribution. These ablations support the combined architecture but do not directly establish that gates identify aspect-relevant image regions.

References

[1] Zhao, H., Yang, M., Bai, X., Liu, H. (2024). A Survey on Multimodal Aspect-Based Sentiment Analysis. IEEE Access, 12, 12039–12052. doi: 10.1109/access.2024.3354844

[2] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. doi: 10.1109/cvpr.2016.90

[3] Zhou, R., Guo, W., Liu, X., Yu, S., Zhang, Y., Yuan, X. (2023). AoM: Detecting Aspect-oriented Information for Multimodal Aspect-Based Sentiment Analysis. Findings of the Association for Computational Linguistics: ACL 2023, 8184–8196. doi: 10.18653/v1/2023.findings-acl.519

[4] Xiao, L., Wu, X., Yang, S., Xu, J., Zhou, J., He, L. (2023). Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. Information Processing & Management, 60(6), Article 103508. doi: 10.1016/j.ipm.2023.103508

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692

[6] Wu, Z., Chen, Y., Kao, B., Liu, Q. (2020). Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4166–4176. doi:

10.18653/v1/2020.acl-main.383

[7] Chu, Y.J., Liu, T.H. (1965). On the shortest arborescence of a directed graph. Scientia Sinica, 14(10), 1396–1400. https://www.cs.cmu.edu/~15850/handouts/chu-liu_1965.pdf

[8] Guo, Z., Zhang, Y., Teng, Z., Lu, W. (2019). Densely Connected Graph Convolutional Networks for Graph-to-Sequence Learning. Transactions of the Association for Computational Linguistics, 7, 297–312. doi: 10.1162/tacl_a_00269

[9] Khan, Z., Fu, Y. (2021). Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation. Proceedings of the 29th ACM International Conference on Multimedia, 3034–3042. doi: 10.1145/3474085.3475692

[10] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning (PMLR), 139, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[11] van den Oord, A., Li, Y., Vinyals, O. (2018). Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748. doi: 10.48550/arXiv.1807.03748

[12] Yu, J., Jiang, J. (2019). Adapting BERT for Target-Oriented Multimodal Sentiment Classification. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 5408–5414. doi: 10.24963/ijcai.2019/751

[13] Wang, Y., Huang, M., Zhu, X., Zhao, L. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 606–615. doi: 10.18653/v1/d16-1058

[14] Fan, F., Feng, Y., Zhao, D. (2018). Multi-grained Attention Network for Aspect-Level Sentiment Classification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3433–3442. doi: 10.18653/v1/d18-1380

[15] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. doi: 10.18653/v1/n19-1423

[16] Zhang, C., Li, Q., Song, D. (2019). Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4567–4577. doi: 10.18653/v1/d19-1464

[17] Li, R., Chen, H., Feng, F., Ma, Z., Wang, X., Hovy, E. (2021). Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 6319–6329. doi: 10.18653/v1/2021.acl-long.494

[18] Xiao, Z., Wu, J., Chen, Q., Deng, C. (2021). BERT4GCN: Using BERT Intermediate Layers to Augment GCN for Aspect-based Sentiment Classification. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 9193–9200. doi: 10.18653/v1/2021.emnlp-main.724

[19] Xu, N., Mao, W., Chen, G. (2019). Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 371–378. doi: 10.1609/aaai.v33i01.3301371

[20] Yu, J., Jiang, J., Xia, R. (2020). Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 429–439. doi: 10.1109/taslp.2019.2957872

[21] Wang, D., He, Y., Liang, X., Tian, Y., Li, S., Zhao, L. (2024). TMFN: A Target-oriented Multi-grained Fusion Network for End-to-end Aspect-based Multimodal Sentiment Analysis. Proceedings of the 2024 Joint International Conference on Computational

Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 16187–16197. https://aclanthology.org/2024.lrec-main.1407/

[22] Yang, J., Xiao, Y., Du, X. (2024). Multi-grained fusion network with self-distillation for aspect-based multimodal sentiment analysis. Knowledge-Based Systems, 293, Article 111724. doi: 10.1016/j.knosys.2024.111724

[23] Yang, J., Xu, M., Xiao, Y., Du, X. (2024). AMIFN: Aspect-guided multi-view interactions and fusion network for multimodal aspect-based sentiment analysis. Neurocomputing, 573, Article 127222. doi: 10.1016/j.neucom.2023.127222

[24] Wang, Y., Zhao, X., Zhong, Z. (2025). Dual-stream multi-level interaction network for aspect-based multimodal sentiment analysis. Signal, Image and Video Processing, 19(10), Article 827. doi: 10.1007/s11760-025-04387-y

[25] Chen, Y., Shi, L., Lin, J., Chen, J., Zhong, J., Dong, C. (2025). Multi-granularity visual-textual jointly modeling for aspect-level multimodal sentiment analysis. The Journal of Supercomputing, 81(1), Article 46. doi: 10.1007/s11227-024-06567-y

[26] Han, H., Ding, H., Qiao, B., Wang, Y., Dang, L., Xie, Y., Wang, Y., Zou, X. (2026). DAGG-Net: Dual adaptive graph and gating network for multimodal aspect-based sentiment analysis. Expert Systems with Applications, 314, Article 131591. doi: 10.1016/j.eswa.2026.131591