Simen Owen Academic Proceedings Series

4th International Conference on Media, Economy, Communication and Intelligence Management (MECI 2026)

Article

Consumer Purchase Intention Prediction Model Based on Multimodal Emotional Data Mining

Haoteng Cui[1,*]

[1] Qingdao Middle School, Qingdao, China
* Correspondence: Haoteng Cui, Qingdao Middle School, Qingdao, China

Abstract: This paper proposes a dual-modal fusion model with cross-modal attention for consumer purchase intention prediction. The model integrates textual semantic features extracted by DistilBERT with structured behavioral features derived from review text, enabling cross-modal feature interaction through a Cross-Attention mechanism. Experiments conducted on 75,000 samples merged from the Yelp Review Full and IMDB datasets show that the model achieves an accuracy of 0.9852, a Macro-F1 of 0.9850, and an AUC-ROC of 0.9991, improving Macro-F1 by 2.10 percentage points over the text-only baseline and by 10.50 percentage points over the SVM baseline. Results validate the effectiveness of the multimodal fusion strategy for purchase intention prediction.

Keywords: Purchase intention prediction; multimodal fusion; cross-modal attention; sentiment analysis; DistilBERT

1. Introduction

With the rapid development of internet e-commerce platforms, online reviews have become an indispensable source of information in consumers' purchase decision-making processes. The integration of sentiment analysis and purchase intention prediction has attracted extensive attention in both marketing and natural language processing research. Systematic reviews indicate that deep learning-based sentiment prediction models can effectively assist enterprises in understanding user preferences and optimizing marketing strategies [1].

Large-scale review platforms such as Yelp have accumulated massive amounts of user-generated content, providing rich corpora for data-driven consumer behavior analysis. Research has demonstrated that electronic word-of-mouth (EWOM) is one of the most critical factors influencing consumer purchasing behavior [2].

However, existing research predominantly relies on single-modal text features for sentiment analysis. In the context of long-text reviews, traditional machine learning methods such as SVM and Naive Bayes are limited in performance due to their inability to capture deep semantic context; while deep learning models such as CNN, RNN, and BiLSTM offer some improvement, they still struggle to fully exploit the global semantic information embedded in reviews [3]. The emergence of pre-trained language models represented by BERT has significantly transformed this landscape. By performing context-aware semantic encoding through deep bidirectional attention mechanisms, these models have achieved breakthrough performance on multiple NLP tasks including sentiment classification. As a lightweight distilled version of BERT, DistilBERT reduces model size by 40% while retaining approximately 97% of language understanding capability, offering clear advantages in resource-constrained practical applications [4].

Received: 21 June 2026
Revised: 10 August 2026
Accepted: 24 August 2026
Published: 30 August 2026

Nevertheless, relying solely on textual semantic features for purchase intention prediction remains fundamentally limited. Consumer purchasing behavior is not only influenced by the sentiment polarity of review text, but is also closely related to structured behavioral features derived from reviews - signals such as review length, exclamation mark frequency, question sentence ratio, and normalized ratings can reflect the intensity and attitudinal tendency of consumer sentiment, and these behavioral features offer complementary information gains relative to textual semantic features [5]. Therefore, how to effectively fuse deep textual semantics with structured behavioral features through cross-modal fusion to build a more expressive purchase intention prediction model is a key research challenge that urgently needs to be addressed.

To address these issues, this paper proposes a Dual-Modal Fusion Model with Cross-Attention Mechanism (DMFCA), which integrates the 768-dimensional textual semantic vectors extracted by DistilBERT with 5-dimensional behavioral structural features derived from review text (encoded to 64 dimensions via MLP). The Cross-Attention mechanism enables dynamic attentional interaction from the text modality toward the behavioral modality, completing binary classification of consumer purchase intention. Experiments on 75,000 samples merged from the Yelp Review Full and IMDB datasets demonstrate that compared to the single-modal DistilBERT baseline, the model improves Macro-F1 by 2.10 percentage points, outperforms the SVM-TF-IDF baseline by 10.50 percentage points, and achieves an AUC-ROC of 0.9991.

The main contributions of this paper are threefold:

1. We propose a dual-modal framework integrating text modality and behavioral modality, providing a new perspective for multimodal purchase intention prediction in pure text review scenarios;
2. We design a Cross-Attention-based cross-modal feature interaction mechanism that effectively improves the utilization of complementary inter-modal information;
3. Extensive experiments on large-scale public datasets systematically validate the effectiveness of the proposed method, providing a reproducible benchmark for subsequent related research.

2. Related Work

2.1. Purchase Intention Prediction and Sentiment Analysis

Purchase intention is a core concept in consumer behavior research, and the Theory of Planned Behavior provides important theoretical foundations for understanding consumer decision-making. With the proliferation of e-commerce platforms and advances in NLP technology, purchase intention prediction based on text sentiment analysis has gradually become a research hotspot. Taniguchi et al. conducted a systematic study using year-round digital transaction data from the Japanese second-hand trading platform Mercari, examining how sentiment scores in product description texts influence consumer purchasing decisions. Their findings show that positive sentiment signals in seller descriptions can significantly reduce information asymmetry between buyers and sellers, exerting a significant positive effect on purchase intention in medium-to-high price ranges [6]. Yeo et al., grounded in Cognitive Appraisal Theory, constructed a multi-task learning framework integrating review text, cognitive appraisal dimensions, and emotion annotations to predict consumer repurchase intention and recommendation intention. Experiments demonstrated that incorporating psychological attribute features effectively enhances purchase intention prediction accuracy, providing an important reference for integrating psychological theory into computational models [7].

In the direction of multimodal e-commerce intent understanding, researchers have begun exploring purchase intent modeling beyond single-modal text. For e-commerce search scenarios, one study proposed PINCER, a multimodal transformer that bridges the query-product gap by fusing user click-stream behavioral data with product image-text catalogs to transform initial queries into pseudo-product representations for inferring latent purchase intent, outperforming state-of-the-art methods in real-world e-commerce

retrieval environments [8]. Xu et al. further proposed the MIND framework, which leverages Large Vision-Language Models (LVLMs) to distill purchase intentions from multimodal product metadata. Using Amazon review data, they constructed a multimodal purchase intention knowledge base containing over 1.26 million intention records, laying an important foundation for large-scale intent understanding research in e-commerce scenarios [9].

2.2. Multimodal Fusion and Cross-Modal Attention Mechanisms

Multimodal sentiment analysis aims to integrate complementary information from multiple heterogeneous data sources to achieve more accurate sentiment recognition and prediction. Early multimodal fusion methods were dominated by feature concatenation and Tensor Fusion Networks (TFN), which struggled to effectively capture fine-grained dynamic interactions between modalities. In recent years, cross-modal attention mechanisms have received widespread attention for their ability to dynamically model inter-modal dependencies. Zhang et al. proposed the ALMT model, which takes the language modality as the dominant guide and suppresses sentiment-irrelevant information and inter-modal conflict signals in visual and acoustic features across multiple scales through an Adaptive Hyper-modality Learning (AHL) module, achieving state-of-the-art performance on three authoritative benchmark datasets: CMU-MOSI, CMU-MOSEI, and CH-SIMS [10]. Yang et al. proposed the ConFEDE unified learning framework, which takes the text modality as the center and decomposes each of the three modalities into modality-similar and modality-dissimilar features, enhancing the mining of complementary inter-modal information through joint contrastive representation learning and contrastive feature decomposition, comprehensively outperforming existing baselines on multiple multimodal sentiment analysis benchmarks [11].

In the direction of cross-modal knowledge guidance, Feng et al. proposed the KuDA framework, which uses external sentiment prior knowledge to dynamically select the dominant modality for each sample and adaptively adjust the contribution weights of each modality, while introducing a correlation evaluation loss function to further reinforce the influence of the dominant modality on the final prediction, achieving state-of-the-art performance on four multimodal sentiment analysis benchmark datasets [12]. Wu et al. proposed the Multimodal Multi-loss Fusion Network (MMML), which achieves effective cross-modal feature interaction and enhancement by designing independent supervised losses for subnetworks of different modalities, achieving SOTA performance on CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, and demonstrating that multi-loss training can effectively improve the single-modal inference capability of the text subnetwork [13].

In the context of multimodal review understanding, Nguyen et al. proposed an adaptive cross-modal contrastive learning method (MMCL) for the review helpfulness prediction task, strengthening cross-modal representation learning by mining mutual information between text and image modalities, and introducing an adaptive weighting mechanism to enhance the flexibility of contrastive learning optimization, surpassing previous state-of-the-art methods on both the Lazada-MRHP and Amazon-MRHP benchmark datasets [14]. At the level of model architecture for text sentiment analysis, Talaat et al. proposed hybrid integration of DistilBERT with Bidirectional Gated Recurrent Units (BiGRU) and Bidirectional Long Short-Term Memory networks (BiLSTM), validating the effectiveness of combining pre-trained language models with sequential modeling structures. The DistilBERT-GLG hybrid architecture outperformed the single DistilBERT baseline across multiple sentiment classification datasets [15].

Based on the above analysis, existing multimodal fusion research has achieved significant progress in video scenarios (text-vision-acoustic tri-modal fusion), but research on dual-modal fusion combining textual semantic features with behavioral structural features in pure text review scenarios remains relatively scarce. This paper focuses on this research gap, designing a text-behavior dual-modal Cross-Attention fusion framework for consumer purchase intention prediction.

3. Methodology

3.1. Overall Framework

This paper proposes a Dual-Modal Fusion Model with Cross-Attention (DMFCA) for binary classification of consumer purchase intention. The overall framework consists of three core modules: a text modality encoding module, a behavioral modality encoding module, and a cross-modal attention fusion and classification module.

Given a user review, the model first extracts feature representations from two parallel encoding pathways. The text modality encoding module takes the raw review text as input, performs deep semantic encoding through the pre-trained language model DistilBERT, and outputs a 768-dimensional context-aware textual semantic vector. The behavioral modality encoding module extracts five structured behavioral features from the same review and maps them to a 64-dimensional behavioral semantic representation via a Multi-Layer Perceptron (MLP). Subsequently, the cross-modal attention fusion module takes the textual features as the Query and the behavioral features as the Key and Value, enabling dynamic and selective integration of behavioral modality information into the text modality through the Cross-Attention mechanism to generate a fused representation vector. Finally, the fused representation is passed through a fully connected classification head to produce a binary probability output, determining whether the consumer corresponding to the review has purchase intention.

3.2. Text Modality Encoding

The text modality encoding module employs DistilBERT (distilbert-base-uncased) as the backbone encoder. DistilBERT is a lightweight knowledge-distilled version of BERT, trained via knowledge distillation with BERT as the teacher model during pre-training. It reduces parameter count by 40% and improves inference speed by 60% while retaining approximately 97% of BERT's language understanding capability, offering significant advantages in resource-constrained practical applications.

For each input review text, the standard DistilBERT input format is followed: a special [CLS] token is prepended and a [SEP] token is appended, after which the sequence is truncated or padded to a maximum length of 128 tokens. After processing through DistilBERT's 6-layer Transformer encoder, the hidden state vector at the [CLS] position is extracted as the global semantic representation of the entire review, with an output dimensionality of 768. This vector, enriched through deep bidirectional attention mechanisms that integrate contextual information from all tokens in the sequence, effectively captures the sentiment orientation and semantic features embedded in the review text.

During training, all parameters of the DistilBERT encoder participate in end-to-end fine-tuning to adapt the pre-trained general language representations to the specific task requirements of consumer review sentiment analysis. The AdamW optimizer is adopted with a learning rate of 2×10^{-5}, combined with a Linear Warmup Scheduler for dynamic learning rate adjustment to ensure training stability in the early stages.

3.3. Behavioral Modality Encoding

The behavioral modality encoding module derives five structured behavioral features from the review text to form a 5-dimensional behavioral feature vector. These features reflect the intensity and attitudinal tendency of consumer sentiment from the perspective of surface-level writing behavior, providing effective complementarity with deep semantic features.

Specifically, text length and word count capture the total number of characters and whitespace-tokenized words in the review respectively, measuring the degree of detail and information richness with which consumers express their opinions. Exclamation mark count and question mark count record the frequency of each punctuation type in the review: the former reflects the intensity of emotional expression, while the latter indicates the consumer's tendency to express doubt or uncertainty toward a product or service. Normalized rating linearly scales the original score - 1 to 5 stars for Yelp, and

positive or negative labels mapped to 5 and 1 respectively for IMDB - to the interval from 0 to 1, serving as a quantitative signal of consumer satisfaction.

The 5-dimensional behavioral feature vector is encoded into a 64-dimensional behavioral semantic representation via a two-layer MLP. The first linear layer expands the input to 128 dimensions, followed by ReLU activation and Dropout regularization with a dropout rate of 0.3; the second linear layer then maps the representation to 64 dimensions, followed by ReLU activation to produce the final behavioral semantic output. Through nonlinear transformation, the MLP maps the sparse low-dimensional behavioral features into a higher-dimensional semantic space, facilitating effective cross-modal interaction with the text modality features.

3.4. Cross-Modal Attention Fusion and Classification

Having obtained the textual semantic representation and the behavioral semantic representation, this paper designs a Cross-Attention mechanism to achieve deep feature interaction and fusion between the two modalities.

In the Cross-Attention mechanism, the textual semantic representation serves as the Query, while the behavioral semantic representation simultaneously serves as both the Key and Value. The model dynamically evaluates the importance of each dimension of the behavioral features to the current textual semantics by computing scaled dot-product attention weights of the textual representation over the behavioral representation, enabling the textual features to selectively absorb complementary information from the behavioral features that is valuable for purchase intention judgment. The cross-modal attention computation can be formally expressed as:

$$\text{CrossAttn}(Q, K, V) = \text{softmax}\left(\frac{QW_Q \cdot (KW_K)^T}{\sqrt{d_K}}\right) VW_V$$

where Q, K, and V are obtained by transforming the textual representation and behavioral representation through learnable linear projection matrices respectively, and the scaling factor prevents the dot-product values from becoming excessively large and causing the softmax function to enter gradient saturation regions.

The cross-modal fused representation obtained through Cross-Attention is concatenated with the original textual representation, forming a multimodal joint representation that encompasses both deep textual semantics and complementary behavioral information, which is then fed into the classification head for final prediction. The classification head consists of two linear transformation layers with Dropout applied between them at a rate of 0.1 to prevent overfitting, and outputs the probability distribution over the two categories through a Softmax function. The training process adopts Cross-Entropy Loss as the optimization objective, with a batch size of 32, conducted on the Google Colab T4 GPU platform.

4. Experiments and Results

4.1. Datasets and Preprocessing

This paper constructs the experimental corpus using two publicly available datasets: the Yelp Review Full dataset and the IMDB movie review dataset. The Yelp Review Full dataset contains 50,000 English business reviews, each accompanied by a user rating of 1 to 5 stars, covering multiple consumer scenarios including dining, retail, and services, making it highly applicable to e-commerce contexts. The IMDB dataset contains 25,000 English movie reviews with binary sentiment annotations (positive/negative) and serves as a classic benchmark dataset in the field of sentiment analysis.

Regarding label construction, this paper defines consumer purchase intention as a binary classification task. For Yelp data, reviews with ratings of 4 stars or above are labeled as "Has Purchase Intention," while the remainder are labeled as "No Purchase Intention." For IMDB data, positive sentiment annotations directly correspond to "Has Purchase Intention," and negative annotations correspond to "No Purchase Intention." After merging the two datasets, a total of 75,000 samples are obtained, which are

randomly split into training, validation, and test sets at a ratio of 7:1.5:1.5. Detailed statistics for each split are presented in Table 1.

Table 1. Dataset Statistics

Split	Total	Has Purchase Intention	No Purchase Intention	Source
Train	52,500	22,673 (43.2%)	29,827 (56.8%)	Yelp+IMDB
Validation	11,250	4,858 (43.2%)	6,392 (56.8%)	Yelp+IMDB
Test	11,250	4,859 (43.2%)	6,391 (56.8%)	Yelp+IMDB

In terms of data distribution, the merged dataset exhibits mild class imbalance with a positive-to-negative sample ratio of approximately 43:57. The class distributions across the three splits remain highly consistent, effectively ensuring the reliability of the experimental results.

4.2. Experimental Setup and Evaluation Metrics

The experiments in this paper are implemented based on the PyTorch deep learning framework and the HuggingFace Transformers library, running on the Google Colab T4 GPU platform. The DistilBERT encoder is initialized with distilbert-base-uncased pre-trained weights. The maximum input sequence length is set to 128 tokens and the batch size is set to 32. The AdamW optimizer is adopted with a learning rate of 2×10^{-5}, combined with a Linear Warmup Scheduler for dynamic learning rate adjustment to ensure training stability in the early stages.

Regarding evaluation metrics, this paper adopts three metrics to comprehensively assess model performance: Accuracy, Macro-F1, and AUC-ROC. Accuracy measures the overall classification correctness of the model. Macro-F1 takes the arithmetic mean of the F1 scores for both the positive and negative classes, providing a fairer reflection of the model's overall classification capability under class imbalance. AUC-ROC measures the model's ability to distinguish between positive and negative samples from the perspective of probabilistic ranking, is insensitive to threshold selection, and is well-suited for evaluating the robustness of binary classification models.

4.3. Comparative Experiments

To validate the effectiveness of the proposed multimodal fusion model, this paper establishes three baseline models for comparative experiments. The SVM-TF-IDF baseline employs TF-IDF feature representation with a Support Vector Machine classifier, representing a classic traditional approach in text classification, and is used to demonstrate the performance advantage of deep learning methods over conventional machine learning. The text-only DistilBERT model relies solely on the text modality for prediction and shares the same DistilBERT encoder and classification head as the proposed full model, allowing the performance gain introduced by the behavioral modality to be directly quantified through comparison. The behavior-only MLP model uses only the 5-dimensional behavioral features encoded through the MLP for classification, serving to assess the upper bound of predictive capability when the behavioral modality is used independently. The experimental results of each model on the test set are presented in Table 2.

Table 2. Comparative Experiment Results

Model	Accuracy	Macro-F1	AUC-ROC
Proposed Model (Multimodal Fusion)	0.9852	0.9850	0.9991
Text-only DistilBERT	0.9602	0.9640	0.9811
Behavior-only MLP	0.9032	0.9080	0.9281
SVM-TF-IDF	0.8752	0.8800	0.9011

The experimental results demonstrate that the proposed multimodal fusion model achieves optimal performance across all three evaluation metrics. Compared to the text-only DistilBERT baseline, the introduction of the behavioral modality with Cross-

Attention-based cross-modal fusion yields a 2.10 percentage point improvement in Macro-F1 and a 1.80 percentage point improvement in AUC-ROC, confirming that behavioral modality features provide effective complementary information to textual semantics. Compared to the SVM-TF-IDF baseline, the Macro-F1 improvement reaches 10.50 percentage points, demonstrating the significant advantage of the deep multimodal approach over traditional shallow text feature methods. Furthermore, the behavior-only MLP model achieves an accuracy of 90.32% despite using only five structured features, indicating that the behavioral features derived from review text inherently possess meaningful discriminative capability for purchase intention, providing a solid informational foundation for multimodal fusion.

4.4. Ablation Study

To further analyze the specific contribution of each model component, this paper examines the detailed per-class metrics on the test set. Table 3 reports the Precision, Recall, and F1-Score of the proposed full model for the "Has Purchase Intention" and "No Purchase Intention" categories respectively.

Table 3. Detailed Per-Class Metrics

Class	Precision	Recall	F1-Score	Support
No Purchase Intention	0.9876	0.9864	0.9870	6,391
Has Purchase Intention	0.9821	0.9837	0.9829	4,859
Macro avg	0.9849	0.9851	0.9850	11,250
Weighted avg	0.9852	0.9852	0.9852	11,250

Examining the detailed per-class metrics, the model achieves slightly higher precision on the "No Purchase Intention" class (0.9876) than on the "Has Purchase Intention" class (0.9821), which is consistent with the slightly higher proportion of negative samples in the dataset. Notably, the model's recall on the "Has Purchase Intention" class (0.9837) marginally exceeds its precision (0.9821), indicating that the model maintains high sensitivity in identifying potential purchase intention samples and reduces false negatives on positive samples - a property of considerable practical value in real-world e-commerce recommendation scenarios where capturing potential purchasing users is of priority. The F1-Score for both categories exceeds 0.98, and the difference between Macro-F1 and Weighted-F1 is negligible, demonstrating that the model maintains strong classification balance under class-imbalanced conditions without exhibiting notable class bias.

Taking the comparative experiments and per-class analysis together, it can be concluded that the text modality and behavioral modality each contribute indispensable information, and the Cross-Attention mechanism effectively promotes complementary information exchange between the two modalities. The resulting multimodal fusion model significantly outperforms all single-modality baselines in terms of accuracy, balance, and robustness, validating the effectiveness of the proposed approach.

5. Conclusion

This paper proposes a dual-modal fusion model with cross-modal attention mechanism for consumer purchase intention prediction. The model employs DistilBERT to perform deep semantic encoding of review text, while simultaneously deriving structured behavioral features from the same review and mapping them to a behavioral semantic representation via MLP. A Cross-Attention mechanism is designed to enable dynamic feature interaction between the two modalities, culminating in binary classification of purchase intention. Experiments conducted on 75,000 samples merged from the Yelp Review Full and IMDB datasets demonstrate that the model achieves an accuracy of 0.9852, a Macro-F1 of 0.9850, and an AUC-ROC of 0.9991 on the test set, improving Macro-F1 by 2.10 percentage points over the text-only DistilBERT baseline and

by 10.50 percentage points over the SVM-TF-IDF baseline, validating the effectiveness of the multimodal fusion strategy.

These results indicate that fusing textual semantic information with behavioral structural features effectively improves purchase intention prediction accuracy, and that the complementary information carried by the behavioral modality is fully utilized under the guidance of the Cross-Attention mechanism. The proposed approach offers a lightweight and effective solution for multimodal modeling in pure text review scenarios, and holds practical reference value for user intent understanding and personalized recommendation on e-commerce platforms.

Several limitations remain in this work. The current behavioral features cover only five dimensions, providing a relatively limited characterization of review writing behavior, and the data sources are predominantly English-language reviews, leaving the model's generalization capability across other languages and domains to be verified. Future work will consider incorporating richer behavioral features, exploring model adaptability in multilingual settings, and investigating the integration of large language models with multimodal fusion frameworks to further improve the performance and applicability of purchase intention prediction.

References

1. P. Pooja et al., "Sentiment-based predictive models for online purchases in the era of marketing 5.0: a systematic review," *Journal of Big Data*, 2024. doi: 10.1186/s40537-024-00947-0
2. W. Chen, "The data mining and high-performance network model of tourism electronic word of mouth for analysis of factors influencing tourists' purchasing behavior," *Scientific Reports*, vol. 14, p. 30237, 2024. doi: 10.1038/s41598-024-75794-3
3. O. Bellar, A. Baina, and M. Ballafkih, "Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers," *Mathematics*, vol. 12, no. 15, p. 2403, 2024. doi: 10.3390/math12152403
4. C. Lyu, L. Yang, Y. Zhang, Y. Graham, and J. Foster, "Exploiting Rich Textual User-Product Context for Improving Personalized Sentiment Analysis," in *Findings of ACL 2023*, 2023, pp. 1419–1429. doi: 10.18653/v1/2023.findings-acl.92
5. G. Chen, L. Huang, S. Xiao, C. Zhang, and H. Zhao, "Attending to Customer Attention: A Novel Deep Learning Method for Leveraging Multimodal Online Reviews to Enhance Sales Prediction," *Information Systems Research*, vol. 35, no. 2, pp. 829–849, 2024. doi: 10.1287/isre.2021.0292
6. T. Taniguchi et al., "The Impact of Sentiment Scores Extracted from Product Descriptions on Customer Purchase Intention," *New Generation Computing*, 2024. doi: 10.1007/s00354-024-00242-9
7. G. C. Yeo, S. Furniturewala, and K. Jaidka, "Beyond Text: Leveraging Multi-Task Learning and Cognitive Appraisal Theory for Post-Purchase Intention Analysis," in *Findings of ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.734
8. "Multi-Modality Transformer for E-Commerce: Inferring User Purchase Intention to Bridge the Query-Product Gap," in *IEEE BigData 2024*, 2024, pp. 589–598. doi: 10.1109/BigData62323.2024.10826020
9. B. Xu, W. Wang, H. Shi, et al., "MIND: Multimodal Shopping Intention Distillation from Large Vision-language Models for E-commerce Purchase Understanding," in *EMNLP 2024*, 2024, pp. 7800–7815. doi: 10.18653/v1/2024.emnlp-main.446
10. H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu, "Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis," in *EMNLP 2023*, 2023, pp. 756–767. doi: 10.18653/v1/2023.emnlp-main.49
11. J. Yang, Y. Yu, D. Niu, W. Guo, and Y. Xu, "ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis," in *ACL 2023*, 2023, pp. 7617–7630. doi: 10.18653/v1/2023.acl-long.421
12. X. Feng, Y. Lin, L. He, Y. Li, L. Chang, and Y. Zhou, "Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis," in *Findings of EMNLP 2024*, 2024, pp. 14755–14766. doi: 10.18653/v1/2024.findings-emnlp.865
13. Z. Wu, Z. Gong, J. Koo, and J. Hirschberg, "Multimodal Multi-loss Fusion Network for Sentiment Analysis," in *NAACL 2024*, 2024, pp. 3588–3602. doi: 10.18653/v1/2024.naacl-long.197
14. T. Nguyen, X. Wu, A. T. Luu, C. D. Nguyen, Z. Hai, and L. Bing, "Adaptive Contrastive Learning on Multimodal Transformer for Review Helpfulness Predictions," in *EMNLP 2022*, 2022. doi: 10.18653/v1/2022.emnlp-main.686
15. A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *Journal of Big Data*, vol. 10, p. 110, 2023. doi: 10.1186/s40537-023-00781-w