TRE: Mitigating Label Noise in Multimodal Aspect-Based Sentiment Analysis via LLM-Guided Dataset Reformation

Anonymous ACL submission

Abstract

With the rapid development of social media, Multimodal Aspect-based Sentiment Analysis (MABSA) has garnered significant attention. The integration of diverse modalities in MABSA presents a unique set of challenges. Among the most commonly used datasets in MABSA are Twitter-2015 and Twitter-2017. During our research, however, we identified labeling errors in these datasets, which we believe contribute to the difficulty in improving MABSA model accuracy. To address this issue, we introduced an expert system based on Large Language Models (LLMs) to assist in filtering abnormal samples and relabeling them manually. This process led to the creation of the **T**witter-**RE**vised datasets, namely TRE-2015 and TRE-2017. Experimental results indicate that our proposed TER dataset provides more accurate sentiment annotations while preserving well-defined and learnable sentiment features. The dataset exhibits sentiment consistency, making it more effective in enhancing the sentiment analysis capabilities of models. Our complete code and datasets will be made publicly available.

1 Introduction

Multimodal Aspect-based Sentiment Analysis (MABSA) has emerged as a popular research area in recent years since it incorporates multiple modalities beyond traditional text-only sentiment analysis tasks. As shown in Fig.1, MABSA aims to extract aspects in given text and identify sentiment towards them. This field gained significant attention since Yu and Jiang first introduced the MABSA task by labeling two datasets: Twitter-2015 and Twitter-2017. However, during our research, we discovered numerous samples in the Twitter-2015 and Twitter-2017 datasets with apparent sentiment annotation errors.

Fig.2 presents two examples of incorrect sentiment labeling. In sample (a), the text "RT @ wgkantai: See this lady? Her name is Joyce Njuguna. She's 'disabled' (I hate that word). She's won us a bronze in powerlifting" conveys a clearly positive and appreciative sentiment toward the aspect "Joyce Njuguna," yet it was labeled as negative in the original dataset. In sample (b), the text "Usually sceptical of celebrities: AngelinaJolie is different with consistent commitment – in tradition of. @miafarrow" expresses a negative sentiment toward "celebrities," while the sentiment toward "Angelina Jolie" should be positive. These inconsistencies highlight the need for a more robust approach to sentiment annotation in multimodal datasets.

	Text:	Image:
Input:	*Tim Tebow* is good for football and the *NFL* via @ buffa82 # Eagles *# NFL*	
Output:	(Tim Tebow, Positive), (NFL, Neutral), (# NFL, Neutral)	

Figure 1: Example of multimodal aspect-based sentiment analysis.

Leveraging Large Language Models (LLMs) for data enhancement has been widely used and proven effective (Choi et al., 2024). LLMs, with the rich knowledge acquired during their training process, are capable of identifying phrases that express emotional tendencies in text. These models excel at detecting implicit sentiment cues that might be missed by traditional methods, such as subtle expressions of praise, irony, or contextual references that can dramatically affect sentiment interpretation. For instance, LLMs can recognize that phrases like "is different" in the context of generally negative sentiment toward celebrities actually indicate positive sentiment toward the specific entity mentioned.

Based on these capabilities, we designed a data verification system using state-of-the-art LLMs.

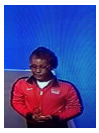
(a)		*RT @ wgkantai: See this lady? Her name is* **[Joyce Njuguna]**$_{Negative}$*. She's 'disabled' (I hate that word). She's won us a bronze in powerlifting.*
(b)		*Usually sceptical of celebrities:* **[AngelinaJolie]**$_{Negative}$ *is different with consistent commitment - in tradition of. @ miafarrow*

Figure 2: Two examples of incorrect sentiment labels from the Twitter-2015 dataset: (a) In this sample, the sentiment towards 'Joyce Njuguna' should be positive but was incorrectly labeled as negative; (b) In this sample, the sentiment towards 'AngelinaJolie' should also be positive but was similarly mislabeled as negative.

Our proposed expert system comprises five advanced models: Qwen2.5-VL (Bai et al., 2025), GPT-4o (Hurst et al., 2024), DeepSeek-V3 (Liu et al., 2024a), DeepSeek-R1 (Guo et al., 2025), and QwQ (Team Qwen, 2024). Among these five models, Qwen2.5-VL and GPT-4o are multimodal models capable of processing both text and images from the original datasets as input, allowing them to capture sentiment cues across modalities. DeepSeek-V3, DeepSeek-R1, and QwQ are text-only models. To ensure these text-only models can also obtain additional information from images, we concatenate image captions with the text as input. For each sample, these five expert models analyze the sentiment independently. When two or more models make judgments that differ from the original dataset labels for the same sample, our system flags that data for subsequent human verification.

Sentiment analysis is highly subjective, and LLMs cannot achieve complete accuracy in sentiment classification (Wang et al., 2023). We believe this is because the rich internal knowledge of large models influences their sentiment judgments, sometimes causing them to overthink or overinterpret neutral statements. For example, given the text "Former Bridgecorp boss Rod Petricevic will be released from jail next month," LLMs tend to interpret this tweet as describing the imprisonment of Bridgecorp's former boss, thus inferring negative sentiment toward "Bridgecorp," as shown in the reasoning process in Fig.3. However, this tweet merely presents a factual news statement without expressing any sentiment toward the entities mentioned in the text. Our goal is not only for models to identify what sentiment is present in the text but also to determine whether sentiment exists in the text at all.

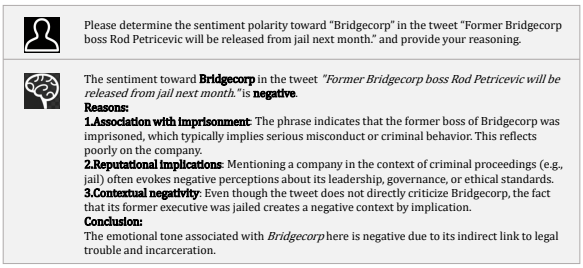


Figure 3: Analysis process demonstrating how Large Language Models incorrectly attribute sentiment polarity to inherently neutral entities, illustrating the "overthinking" phenomenon.

Therefore, we did not fully adopt the opinions of LLMs nor use mainstream majority voting algorithms. Instead, we filtered out all potentially problematic labels for manual annotation, significantly increasing the accuracy of sentiment labeling. We ultimately examined all 11,310 data samples in Twitter-2015 and Twitter-2017, identifying 2,638 and 2,627 abnormal samples respectively, conducted manual checks, and relabeled 1,071 and 1,263 data points accordingly. Subsequently, we selected LLMs and state-of-the-art models in the MABSA field and tested them on our TRE-2015/2017 datasets. The experimental results demonstrated that our TRE datasets have more accurate sentiment annotations and well-defined sentiment features, providing a more reliable benchmark for future MABSA research.

2 Related Works

Multimodal Aspect-based Sentiment Analysis (MABSA) has emerged as a significant research area in affective computing. Extending beyond text-only sentiment analysis, MABSA incorporates multiple modalities and focuses on identifying sentiment toward specific aspects mentioned in text. Yu and Jiang pioneered this field by adapting BERT (Devlin et al., 2019) for target-oriented multimodal sentiment classification and contributed two valuable datasets: Twitter-15 and Twitter-17. Their work introduced a Target-Image (TI) Matching layer that employed attention mechanisms to establish connections between targets and images, thus generating target-sensitive visual representations. This approach inspired subsequent researchers and became the dominant methodology in the field for a considerable period. Building

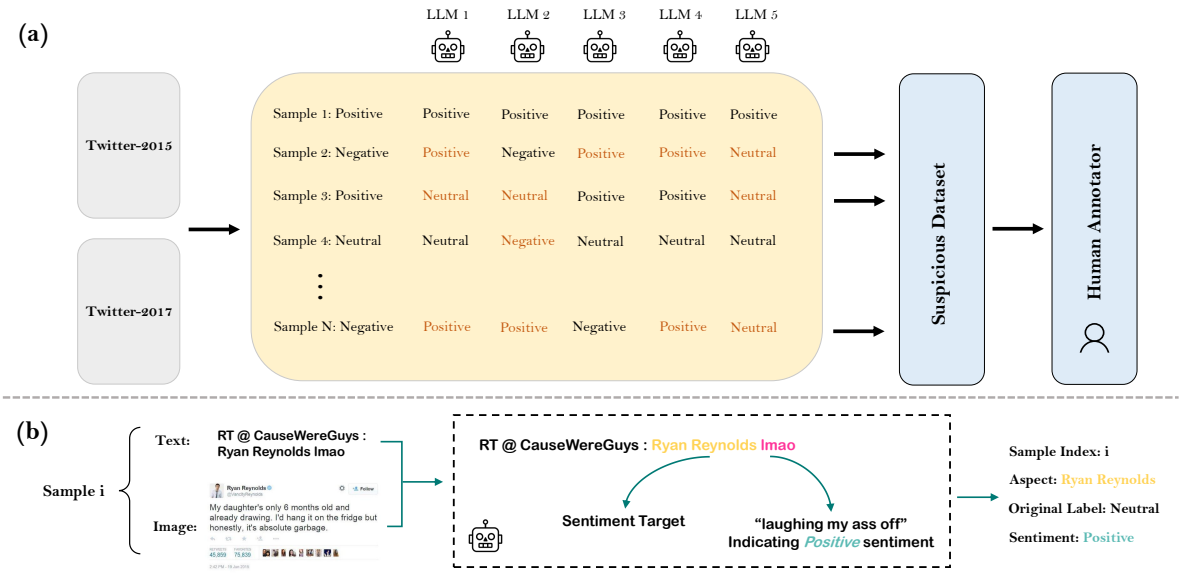


Figure 4: Our proposed framework. (a) illustrates our proposed LLM-based expert annotation system, where samples are added to the suspicious dataset for human annotator review when two or more LLMs predict results different from the original label. (b) demonstrates the judgment process of an LLM for a single sample, showing how LLMs can detect emotional cues that might be overlooked by humans.

on the original BERT architecture, Khan and Fu introduced a Caption Transformer layer that generated image descriptions to facilitate cross-modal information fusion. Similarly, Yang et al. proposed a comparable framework that transformed images into textual descriptions. The difference is that their approach, FITE, focused specifically on facial information in images to generate detailed facial descriptions and capture additional sentiment cues.

However, these methods share a common limitation: the inability to process images and text simultaneously. Some researchers argue that translating images to text inevitably results in information loss. To address this issue, a more effective solution involves mapping both images and text to the same vector space. Ling et al. were the first to employ encoder-decoder models, specifically BART (Lewis et al., 2020), to jointly process textual and visual features. They proposed three pre-training tasks: Textual Aspect-Opinion Extraction, Visual Aspect-Opinion Generation, and Multimodal Sentiment Prediction. Their approach inspired Zhou et al., who developed two novel modules: the Aspect-oriented Method (AoM) and the Aspect-Aware Attention Module (A^3M). These modules enable models to extract fine-grained information and capture interactions between different modalities.

With the advancement of large language models, researchers have begun exploring their application to multimodal aspect-based sentiment analysis. Yang et al. proposed a method leveraging In-Context Learning (ICL) to enhance LLMs' MABSA capabilities. However, despite their impressive performance on other NLP tasks, LLMs consistently underperform compared to traditional pre-trained models on MABSA tasks. Our investigation revealed a potential explanation: inconsistent and incorrect sentiment labels in the Twitter-2015 and Twitter-2017 datasets. We believe these annotation errors significantly contribute to the limitations in model accuracy improvement. In this paper, we address this issue by employing a combination of large language models and manual annotation to verify these datasets, resulting in a newly revised dataset that supports more accurate sentiment analysis.

3 Twitter-Enhanced

Twitter-2015 and Twitter-2017 play an important role in MABSA area. However, as shown in picture 1, there are several sentiment label mistakes in this two datasets. Hence, we develop a hybrid dataset verification methodology integrating a LLMs-based Expert System with manual annotation protocols, through which comprehensive inspection of the original dataset is conducted. In this section, we formally present the architectural framework of the LLMs-based Expert System and the standardized criteria governing the manual an-

notation process. Framework of our proposed method is shown in Fig.4.

3.1 Expert System for Data Verification

Model Selection and System Design. We designed an expert system consisting of five state-of-the-art LLMs to conduct preliminary verification of the datasets. The key reason for choosing LLMs for initial screening is that the rich internal knowledge these models acquired during training stage enables them to better capture subtle emotional cues that human annotators might overlook. For instance, as shown in Fig.4 (a), LLMs can accurately understand the emotional intensity expressed in internet abbreviations like "lmao" (laughing my ass off), and can identify other elements such as emoji combinations, specific patterns of punctuation repetition ("!!!"), sarcastic tones, subtle cultural references, and memes with multiple layers of meaning that human annotators might misinterpret or miss entirely.

To ensure model diversity, we carefully selected both multimodal and text-only models. For multimodal analysis, we employed ChatGPT-4o and Qwen2.5-VL-Chat, which can process both textual and visual information simultaneously, effectively capturing emotional consistency or contradictions between images and text. For text-only processing, we selected DeepSeekV3, DeepSeekR1, and QwQ. These models use different training corpora, parameter scales, and optimization strategies, representing diverse training methodologies and architectural designs, helping to mitigate biases and limitations of individual models and provide more comprehensive and objective data evaluation results.

Input Processing Protocol. Our input processing protocol is designed to maximize information availability across model types. Multimodal models received the original data unmodified, enabling them to directly process both text and associated images. Text-only models, however, received the textual content concatenated with detailed image descriptions to compensate for their inability to process visual information directly. This approach of providing image captions to text-only models has been shown to effectively bridge modality gaps in previous research (Li et al., 2021).We chose to use Qwen2.5-VL-Chat to generate image captions. We instructed the large model to produce detailed image descriptions while considering the accompanying text, thereby ensuring that sufficient information was preserved to enhance the accuracy of subsequent text models. The prompt we used is: '*This is an image accompanying a tweet with the text **T**. Please generate a detailed description of this image.*'

Verification Methodology. Each model independently evaluated every aspect in each data sample, producing sentiment judgments without knowledge of the original labels or other models' outputs. This blind evaluation protocol reduces confirmation bias. We established a threshold whereby samples were flagged as potentially erroneous when K or more models disagreed with the original sentiment label. To fully leverage the rich internal knowledge of large language models in identifying potentially erroneous data, we set the sensitivity parameter K to 2, rather than recording a data sample only when the majority of models (three or more) make an incorrect prediction. This approach effectively ensures that as many suspicious samples as possible are flagged from the original dataset, while avoiding the need for a complete manual inspection of the entire dataset—striking a balance between efficiency and annotation quality. Furthermore, recognizing the limitations of LLMs in sentiment analysis tasks (Wang et al., 2023), we deliberately avoided employing majority voting algorithms that would automatically relabel data based solely on model consensus.

Human-in-the-Loop Verification. After the automated screening phase, all flagged samples are compiled into a Suspicious Dataset and subsequently undergo comprehensive human review. To ensure that the annotation process remained unbiased, human annotators are not exposed to the predictions made by the LLMs; instead, they relied solely on their own judgment. We enlist three well-educated graduate students with strong reading comprehension skills to perform the annotations, and the final label for each sample is determined by majority vote.

By adopting this two-stage verification strategy: model-driven flagging and manual validation, we aim to systematically correct errors in the original datasets while making efficient use of LLMs' inherent knowledge.

3.2 Human Examination

In our manual annotation process, we employed a rigorous methodology to accurately represent the original attitude of tweet authors (Mohammad, 2016). To maintain impartiality throughout the an-

notation procedure, human annotators were deliberately isolated from LLMs predictions, enabling them to exercise independent judgment. We recruited three graduate students with advanced reading comprehension capabilities to conduct the annotations. The definitive sentiment label for each sample was subsequently determined through a majority voting mechanism.

When analyzing sentiment in social media content, we adhered to consistent methodological principles that carefully distinguished between objective factual reporting and subjective opinion expression. For instance, when a tweet merely reported factual information about a criminal incident occurring in a specific location, we did not interpret this as conveying negative sentiment toward that geographical entity. However, when a tweet explicitly characterized a location as having an elevated crime rate, we classified this as expressing negative sentiment toward the place, as such assertions transcend mere factual reporting by incorporating evaluative conclusions.

Similarly, tweets that exclusively conveyed informational content about artistic works and their creators without employing evaluative language were systematically classified as neutral. Conversely, when authors employed affective terminology such as "beautiful" or "stunning" to describe artistic creations, we identified these instances as expressing positive sentiment toward both the artwork and its creator. Our systematic analysis revealed that conclusive statements typically contain explicit sentiment orientations, whereas faithful descriptions of factual information generally maintain neutrality. This fundamental distinction between factual reporting and evaluative commentary served as a guiding principle throughout our annotation methodology. Through consistent application of these principles, we aimed to establish a more precise sentiment annotation framework that accurately reflects authentic sentiment expression patterns in social media contexts.

Quality Control. To ensure annotation reliability and consistency, we implemented a comprehensive quality control protocol featuring double-check verification mechanisms. Specifically, we randomly selected 20% of the annotated samples for independent re-annotation by different annotators who had not previously evaluated these particular instances. This secondary annotation process was conducted under identical guidelines and conditions as the primary annotation effort. We subsequently calculated the inter-annotator agreement rate between the primary and secondary annotations, achieving a Cohen's Kappa (McHugh, 2012) coefficient of 0.83, which indicates substantial agreement according to established interpretative standards in computational linguistics. Discrepancies identified during this verification process were systematically documented and resolved through collaborative discussion sessions involving all annotators and project supervisors. These sessions facilitated the clarification of annotation guidelines where necessary and ensured consistent application of sentiment classification principles across the entire dataset. This methodologically rigorous approach aligns with our broader research objective of enhancing sentiment label quality in multimodal datasets. By meticulously differentiating between neutral factual reporting and genuinely sentiment-laden expressions, we addressed a significant limitation in existing datasets where such distinctions were frequently overlooked, resulting in inconsistent annotations that adversely affected model performance and evaluation metrics.

4 Experiment

This section presents our experimental design and results aimed at addressing three key research questions:

Q1 : The annotation of the TRE dataset was conducted with the assistance of large language models. Could this lead to unfair model evaluation accuracy?

Q2 : Does the TRE dataset maintain sentiment consistency during the annotation process, providing good learnable sentiment features?

Q3 : Are the sentiment labels in the TRE dataset more accurate than those in the original dataset?

4.1 Experimental Setup

4.1.1 Datasets

We conducted experiments on the original Twitter-2015 and Twitter-2017 datasets (Yu and Jiang, 2019) along with our enhanced versions, Twitter-2015-Enhanced and Twitter-2017-Enhanced. The statistics of the size of the Suspicious Dataset and the amount of manually relabeled data are shown in Table2.

Table 1: A comparison of basic statistics between the original datasets and the TRE-2015/2017 datasets.

	TWITTER-15				TWITTER-17				TRE-2015				TRE-2017			
	#POS	#NEG	#Neutral	Total	#POS	#NEG	#Neutral	Total	#POS	#NEG	#Neutral	Total	#POS	#NEG	#Neutral	Total
Train	928	368	1,883	3,179	1,508	416	1,638	3,562	1,068	336	1,775	3,179	1,655	437	1,470	3,562
Dev.	303	149	670	1,122	515	144	517	1,176	412	142	568	1,122	525	150	503	1,176
Test	317	113	607	1,037	493	168	573	1,234	399	108	530	1,037	588	196	450	1,234

Table 2: Statistics of Suspicious Dataset and Relabeled Data.

	Original Dataset	Suspicious Dataset	Relabeled Data
Twitter-2015-Train	3,179	1,585	577
Twitter-2015-Dev.	1,122	552	261
Twitter-2015-Test	1,037	501	233
Twitter-2017-Train	3,562	1,551	605
Twitter-2017-Dev.	1,176	513	298
Twitter-2017-Test	1,234	563	360

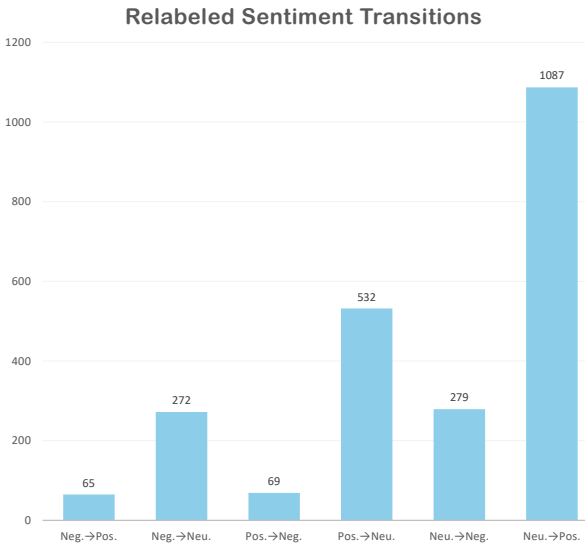


Figure 5: Statistics of sentiment label transitions after TRE relabeling: The bar chart shows the total number of each sentiment transition type across all subsets of the Twitter-2015 and Twitter-2017 datasets (Train, Dev, and Test). Neg. is short for *Negative*, Pos. is short for *Positive*, Neu. is short for *Neutral*.

For statistical validity, we maintained the same train/validation/test splits as the original datasets. The detailed statistics of the Twitter-2015/2017 and TRe-2015/2017 datasets are shown in Table1. To better illustrate the relabeling results of the TRE dataset on the original Twitter corpus, we aggregated the number of sentiment label transitions across all subsets and visualized them as shown in Fig.5. The most frequent transitions are from neutral to positive (Neu.→Pos.) and from positive to neutral (Pos.→Neu.), suggesting that distinguishing between these two sentiment categories is particularly challenging. Additionally, a notable number of transitions occurred between neutral and negative sentiments, indicating a degree of ambiguity and subjectivity in the original labels. These relabeling patterns highlight the inherent difficulties in annotating nuanced sentiment expressions and underscore the importance of high-quality labeling for MABSA.

4.1.2 Baseline Models

To comprehensively evaluate our enhanced datasets, we selected the following models:

Large Language Models:

LLaMA (Grattafiori et al., 2024) - A completely different model from those used in our annotation process.

LLaVA (Liu et al., 2024b) - An end-to-end trained large-scale multimodal model that connects a visual encoder with a large language model to enable general-purpose vision and language understanding.

Qwen-VL (Yang et al., 2024a) - A smaller multimodal model with open-source implementation.

We deliberately selected these smaller open-source models, distinct from those used in our data verification process, to ensure reproducibility and to test whether our enhancement genuinely improves sentiment representation rather than simply aligning with specific model biases.

MABSA Pre-trained Models:

BERT (Devlin et al., 2019) - A bidirectional Transformer-based model pre-trained to generate contextualized word embeddings for various NLP tasks

BART (Lewis et al., 2020) - A Transformer-based model employing an encoder-decoder architectural framework.

TomBERT (Yu and Jiang, 2019) - The first MABSA model to incorporate BERT.

FITE (Yang et al., 2022) - Introduces face-sensitive image descriptions in addition to the textual modality and employs a gating mechanism to control information fusion.

VLP-MABSA (Ling et al., 2022) - The first model to use a Vision-Language Pretraining (VLP) framework to map textual and visual features into a unified vector space.

Table 3: The performance of different LLMs comparison across different datasets

Method	Twitter-2015		Twitter-2017		TRE-2015		TRE-2017	
	ACC	Mac-F_1	ACC	Mac-F_1	ACC	Mac-F_1	ACC	Mac-F_1
LLaMA3.1-8B (Grattafiori et al., 2024)	54.87	**53.23**	49.43	48.82	59.21	57.49	64.26	63.10
LLaMA3.2-1B (Grattafiori et al., 2024)	48.60	32.30	38.57	31.30	52.65	32.86	45.71	33.86
LLaMA3.2-3B (Grattafiori et al., 2024)	44.55	37.19	41.98	38.60	45.52	40.11	43.84	38.13
LLaVA-1.5-7b (Liu et al., 2024b)	30.57	15.61	39.95	19.03	38.48	18.52	47.65	21.51
Qwen-VL-Chat (Bai et al., 2023)	**57.86**	52.95	**54.62**	**51.14**	**65.77**	**62.09**	**67.42**	**63.79**

Table 4: The performance of different MABSA pre-trained models comparison across different datasets.

Method	Twitter-2015		Twitter-2017		TRE-2015		TRE-2017	
	ACC	Mac-F_1	ACC	Mac-F_1	ACC	Mac-F_1	ACC	Mac-F_1
BERT (Devlin et al., 2019)	72.32	67.36	60.05	58.77	71.46	65.12	72.20	68.64
BART (Lewis et al., 2020)	72.20	62.80	68.87	66.53	71.92	66.43	71.34	69.35
TomBERT (Yu and Jiang, 2019)	75.12	68.72	69.43	66.97	74.12	66.74	69.57	68.41
FITE (Yang et al., 2022)	77.34	71.63	72.23	69.30	73.58	68.31	70.91	70.04
VLP-MABSA (Ling et al., 2022)	78.27	72.31	72.96	70.75	75.83	70.85	70.89	71.52
AoM (Zhou et al., 2023)	**79.17**	**74.65**	**73.26**	**71.65**	**76.66**	**71.51**	**72.58**	**70.72**

AoM (Zhou et al., 2023) - Builds upon the VLP-MABSA model by introducing an attention module to enhance the model's ability to capture cross-modal interactions.

4.2 Main Results

Comparison of Large Model Results on Two Datasets (Q1): We compared the accuracy of the LLMs on the two datasets and the results are shown in Table 3. By analyzing the experimental results, we observe that both multimodal and text-based large models achieve higher accuracy on our TRE dataset. To ensure the robustness of our evaluation, we selected completely unrelated open-source models for comparison—models with significantly fewer parameters, distinct architectures, and different pre-training strategies. This careful experimental design eliminates confounding factors such as data leakage and model-specific biases. The results demonstrate that our dataset better aligns with human emotional preferences and is more suitable for capturing sentiment-related features. Moreover, they confirm that our annotation approach effectively encodes key domain characteristics rather than merely adapting to the decision boundaries of a specific model. By ruling out the possibility of dataset contamination, our findings offer more reliable scientific evidence for the proposed methodology and establish a more rigorous validation paradigm for LLMs evaluation.

Comparison of MABSA Pre-trained Model Results on Two Datasets (Q2): We compared the performance of four MABSA pre-trained models across two datasets, as shown in Table 4. To ensure the comparability of the results and reduce potential bias, we do not directly adopt models' results from their original paper but instead retrained the model under the same experimental settings.

The results demonstrate that our dataset exhibits consistent sentiment labeling and enables the model to effectively learn sentiment analysis capabilities. Moreover, prior MABSA models still performed well on our dataset, suggesting that the correction of sentiment labels did not compromise the models' ability to capture multimodal sentiment features. In other words, we successfully revised the sentiment annotations while preserving consistent sentiment characteristics.

4.3 Case Study (Q3)

To demonstrate that our newly constructed dataset provides more accurate sentiment annotations than the original dataset, we present a case study in this section by selecting several representative samples. We compare the sentiment labels from the original dataset with our revised annotations and explain why our labels better reflect the true sentiment expressed in the multimodal content. In sample a from Figure 6, the original dataset fails to detect the sarcastic tone in the text and labels the sentiment toward Nintendo as neutral. However, we argue that the correct sentiment should be negative, as the text clearly conveys irony and criticism. In sample b, the phrase "killed it" is a

Image				
Text	(a) 10 seconds into the game and already an ad. Smooth, Nintendo.	(b) RT @ DanceGoals : Chris Brown killed it !	(c) RT @ TheNatsBlogJoe : Seriously , how do you not like Bryce Harper ? #Nats	(d) RT @ JustAGirlThing : How can anyone hate Charlie sheen
Sentiment	~~Neutral~~ Negative	~~Negative~~ Positive	~~Negative~~ Positive	~~Negative~~ Positive

Figure 6: Case analysis of original Twitter-2015/2017 datasets and our newly proposed TRE-2015/2017 datasets. Aspects are highlighted in red within the text. In the **Sentiment** row, the strikethrough text represents the label from original dataset, while the label below it is the one we re-annotated.

slang expression meaning someone performed exceptionally well. Therefore, the sentiment toward Chris Brown should be positive, contrary to the original label. Sample c uses rhetorical questioning to express admiration rather than dislike for Bryce Harper, indicating that the correct sentiment label should be positive. In sample d, the original annotation makes a similar mistake. When considering the accompanying image, it becomes clear that the sentiment toward Charlie Sheen is positive, not negative as originally labeled. These examples highlight the limitations of the original dataset in capturing nuanced expressions such as sarcasm, idiomatic phrases, and multimodal context. Our re-annotations address these issues by aligning sentiment labels more closely with the actual intent and tone of the content, thereby providing a more reliable resource for MABSA.

5 Conclusion

In this paper, we developed a novel data inspection platform that combines LLMs-based expert-system and manual annotation. Using this platform, we systematically examined and re-annotated two widely used datasets in the MABSA field: Twitter-2015 and Twitter-2017. Our analysis revealed that the low sentiment prediction accuracy of LLMs on MABSA tasks was primarily caused by incorrect sentiment annotations in the original datasets. Based on this observation, we introduced two revised datasets: TRE-2015 and TRE-2017. Experimental results show that, on our newly proposed datasets, LLMs achieve sentiment prediction accuracies at a reasonable level—unlike before, where their performance was significantly below that of conventional pre-trained models, which was clearly inconsistent with their general capabilities. Furthermore, the TRE datasets maintain well-defined and learnable sentiment patterns, allowing previous MABSA pre-trained models to retain their performance, thereby confirming that the revisions did not impair the models' ability to learn multimodal sentiment features. The TRE dataset we proposed possesses consistent and learnable sentiment features, enabling the model to acquire accurate sentiment analysis capabilities. We hope that our work can offer valuable insights and inspiration for future research in the field.

Limitations

Despite our efforts to improve the quality of sentiment annotations, it is important to acknowledge that sentiment remains an inherently subjective and context-dependent construct. The same expression may convey different emotions depending on the speaker's intent, cultural background, or audience interpretation. As such, it is difficult to guarantee complete correctness or universal agreement on all sentiment labels, even after careful re-annotation. Moreover, while our work focuses on correcting sentiment labels in MABSA datasets, these datasets also include aspect annotations, which we did not examine or modify in this study. The omission of aspect-level validation limits the overall impact and completeness of our dataset refinement, as sentiment and aspect information are often intertwined in multimodal sentiment analysis tasks. In future work, we plan to conduct a comprehensive analysis and refinement of aspect annotations to further enhance the reliability of MABSA datasets.

Ethics Statement

All data in our proposed TRE dataset originates from Twitter-2015/2017, two well-established open-source datasets widely used in academic research. The annotators involved in our labeling process were graduate students from Chinese universities, who were compensated at local wage standards throughout the annotation procedure.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Juhwan Choi, Jungmin Yun, Kyohoon Jin, and Youngbin Kim. 2024. Multi-news+: Cost-efficient dataset cleansing via llm-based data annotation. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL-HLT 2019*, pages 4171–4186.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *ACM Multimedia 2021*, pages 3034–3042.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2476–2483.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *ACL 2022*, pages 2149–2159.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.

Team Qwen. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *EMNLP 2022*, pages 3324–3335.

Li Yang, Zengzhi Wang, Ziyan Li, Jin-Cheon Na, and Jianfei Yu. 2024b. An empirical study of multimodal entity-based sentiment analysis with chatgpt: Improving in-context learning via entity-aware contrastive learning. *Information Processing & Management*, 61(4):103724.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. IJCAI.

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 *ACL 2023 Findings*, pages 8184–8196.