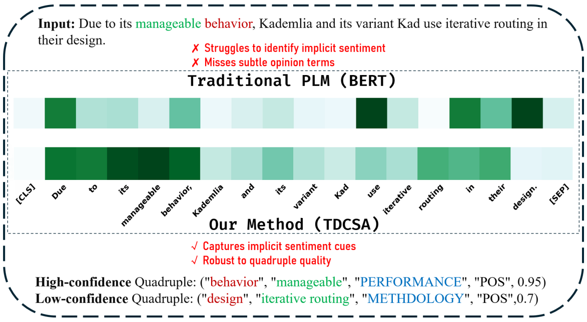
TDCSA: LLM-Guided Top-Down Approach for Robust Citation Sentiment Analysis

Fan Gao[1,2], **Jieyang Peng**[1], **Xiaoming Tao**[1], **WANG Youzheng**[1,2*]
[1]Department of Electronic Engineering, Tsinghua University
[2]Department of Electronic Engineering, BNRist, Tsinghua University

Abstract

Citation Sentiment Analysis (CSA) plays a crucial role in understanding academic influence and knowledge diffusion. While pre-trained language models (PLMs) and large language models (LLMs) showed remarkable success in general sentiment analysis, they encounter specialized challenges in CSA due to the less significant and implicit sentiment expressions in academic writing, as well as complex sentiment transitions. In order to address the challenges, We propose TDCSA, a Top-Down framework that leverages LLMs' semantic understanding capabilities to enhance PLM-based CSA, which transforms the traditional bottom-up feature engineering paradigm into a top-down architecture. Our framework consists of three key components: (1) a Dual LLM Feature Generation module for robust quadruple extraction, (2) a Multi-view Feature Representation mechanism for neutral citation processing, and (3) a Quad Feature Enhanced PLM. Experiments demonstrate that TDCSA[1] significantly outperforms existing methods, achieving state-of-the-art performance while maintaining robustness to quadruple quality variations.



Input: Due to its manageable behavior, Kademlia and its variant Kad use iterative routing in their design.

✗ Struggles to identify implicit sentiment
✗ Misses subtle opinion terms
Traditional PLM (BERT)
[CLS]
Due
to
its
manageable
behavior,
Kademlia
and
its
variant
Kad
use
iterative
routing
in
their
design.
[SEP]
Our Method (TDCSA)
✓ Captures implicit sentiment cues
✓ Robust to quadruple quality

High-confidence Quadruple: ("behavior", "manageable", "PERFORMANCE", "POS", 0.95)
Low-confidence Quadruple: ("design", "iterative routing", "METHDOLOGY", "POS",0.7)

Figure 1: The key limitations of traditional PLMs in CSA. While baseline models struggle to identify implicit sentiment expressions in academic writing, our method leverages LLM-extracted quadruples to enhance semantic understanding. Even with varying quadruple quality (e.g., the low-confidence quadruple), the model maintains robust performance.

1 Introduction

Citations play a fundamental role in academic contributions, reflecting complex relationships and attitudes between scholarly works (Saggion et al., 2016). While sentiment analysis has advanced significantly in domains like consumer reviews and social media posts, Citation Sentiment Analysis (CSA) presents unique challenges. Scientific citations employ formal and technical language, emphasizing objective descriptions over explicit evaluations, making sentiment expressions often implicit and requiring deeper semantic understanding (Yousif et al., 2019b). For instance in Figure 1, phrases "*manageable behavior*" in research contexts implicitly convey positive sentiment about a method's stability and usability, though appearing neutral on the surface.

Machine learning approaches to CSA have followed a bottom-up paradigm, where researchers carefully design and combine low-level linguistic features such as n-grams, dependency relations, or sentiment lexicons to capture sentiments (Athar, 2011; Sula and Miller, 2014; Athar and Teufel, 2012; Xu et al., 2015). These feature engineering methods have achieved considerable success through domain expertise and manual effort (Ihsan et al., 2023; Liu et al., 2024). The evolution of deep learning brought innovations through word embeddings (Mikolov et al., 2013; Pennington et al., 2014) and neural networks (Munkhdalai et al., 2016; Budi and Yaniasih, 2022), followed by pre-trained language models (PLMs) and large language models (LLMs). PLMs like BERT offer efficient fine-tuning capabilities for downstream tasks (Mercier et al., 2020; Yang et al., 2023), while LLMs benefit from massive training corpora that enable better semantic comprehension of scientific

[*] Corresponding author

[1] Codes and datasets are available at `https://github.com/goghvan1113/TDCSA`

Findings of the Association for Computational Linguistics: ACL 2025, pages 6467–6484
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

text (Wang et al., 2023). However, current solutions face two significant limitations.

1. **Suboptimal performance:** While traditional machine learning approaches with carefully engineered features remain competitive, particularly SVM using dependency relations, domain-specific PLMs still face challenges in capturing implicit sentiment expressions in academic writing. As shown in Figure 1, PLM struggles to capture subtle positive sentiment conveyed through phrases like "*manageable behavior*". While LLMs demonstrate better semantic comprehension, their direct application proves computationally intensive and potentially suboptimal for specific downstream tasks (Zhang et al., 2023), as demonstrated in Section C.2.

2. **Oversight of sentiment transitions:** Existing research predominantly overlooks the dynamic nature of sentiment expression in scientific citations, where authors often present opposing viewpoints before ultimately reaching a final evaluation. As shown in Figure 2, these transitions frequently determine the overall citation polarity, yet current approaches lack mechanisms to effectively capture these sequential sentiment dynamics.

In order to tackle these limitations, we propose a novel framework that leverages LLMs as robust feature generator to generate structured sentiment information, which then enhance smaller, task-specific PLMs for sentiment classification. Notably, our framework maintains robustness to quadruple quality variations, as it focuses on capturing general sentiment patterns rather than requiring perfect extraction accuracy. We design three components: a **Dual LLM Feature Generation (DFG)** framework that employs critical thinking (Betz and Richardson, 2020) and Chain-of-Thought (Wei et al., 2022) reasoning to extract robust sentiment quadruples. A **Multi-view Pattern Representation (MPR)** mechanism that handles the dominant neutral citations with complementary perspectives, avoiding feature redundancy. An **Quad Feature Enhanced PLM (QFE-PLM)** that integrates LLM-extracted features with PLM through adaptive fusion mechanisms. The main contributions of this paper can be summarized as follows:

- We propose a novel top-down framework that leverages LLMs' semantic understanding to enhance PLM-based citation sentiment classification.

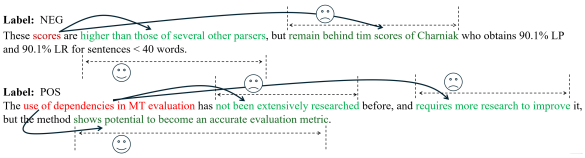

Label: NEG
These scores are higher than those of several other parsers, but remain behind tim scores of Charniak who obtains 90.1% LP and 90.1% LR for sentences < 40 words.
Label: POS
The use of dependencies in MT evaluation has not been extensively researched before, and requires more research to improve it, but the method shows potential to become an accurate evaluation metric.

Figure 2: Sentiment transition patterns in scientific citations. Solid arrows indicate aspect-opinion relationships, while dashed boxes highlight the scope of opinion expressions. The intensity of the different opinions is highlighted with light and dark green.

- We present an effective method DFG that provide complementary two-stage thinking process for sentiment quadruples generation.

- We design the MPR mechanism and an adaptive fusion strategy, which not only effectively integrates the sentiment quadruples information for sentiment classification but also introduces adversarial distribution perturbations to prevent the model from overfitting.

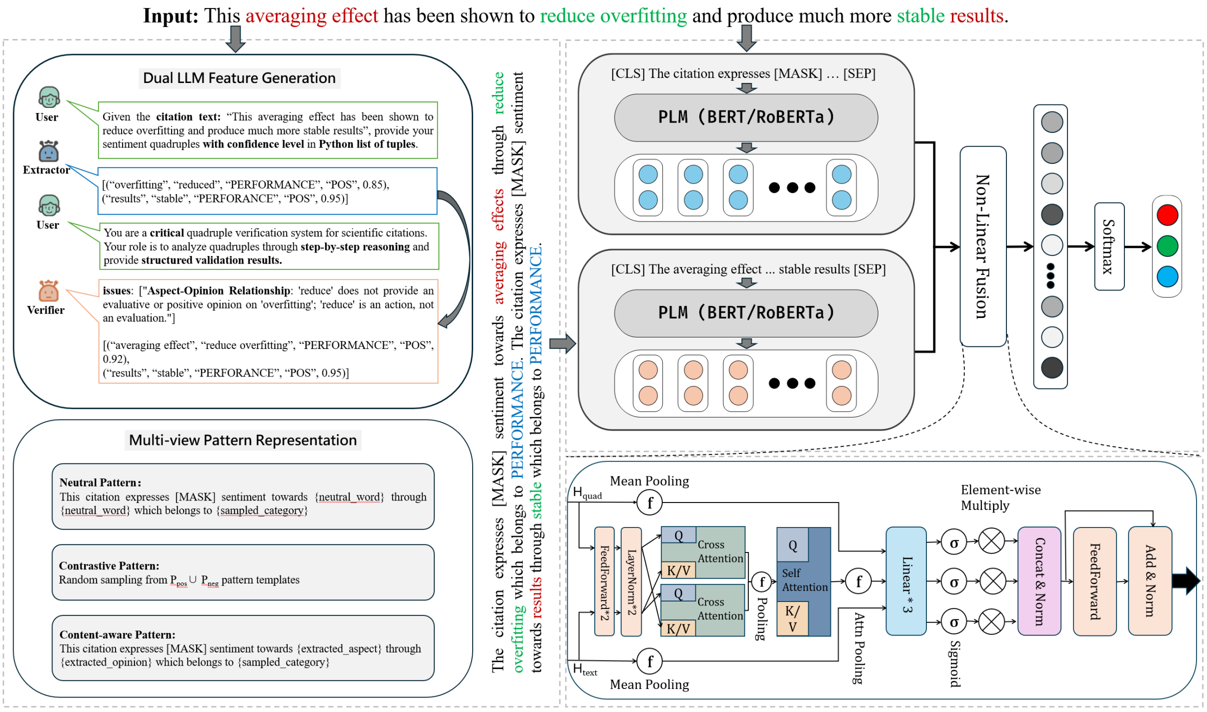
- We conduct extensive experiments on benchmark datasets, demonstrating the effectiveness of our approach and providing detailed analysis.

2 Methodology

The overall architecture of TDCSA is shown in the Figure 3: it consists of three key components: (1)a Dual LLM Feature Generation framework using critical thinking and CoT reasoning to extract sentiment quadruples; (2)a Multi-view Pattern Representation mechanism for handling neutral citations with three different representations and improving model generalization performance; (3)a Quad Feature Enhanced PLM-based model augmenting PLMs with structured sentiment information.

2.1 Critical Thinking DFG

Citation ASQP Problem Statement Following the Aspect Sentiment Quad Prediction (ASQP) task definition proposed by (Zhang et al., 2021) and (Cai et al., 2021), we define the CASQP (Citation Aspect Sentiment Quad Prediction) task for scientific citations as follows. Given a scientific citation text X, the CASQP task aims to extract a set of quadruples $Q = q_1, q_2, \ldots, q_n$, where each quadruple $q_i = (a_i, o_i, c_i, p_i)$ comprises an aspect term, an opinion term, a predefined category $\mathcal{C} =$



Input: This averaging effect has been shown to reduce overfitting and produce much more stable results.

Given the **citation text:** “This averaging effect has been shown to reduce overfitting and produce much more stable results”, provide your sentiment quadruples **with confidence level** in **Python list of tuples**.

[(“overfitting”, “reduced”, “PERFORMANCE”, “POS”, 0.85),
(“results”, “stable”, “PERFORANCE”, “POS”, 0.95)]

You are a **critical** quadruple verification system for scientific citations. Your role is to analyze quadruples through **step-by-step reasoning** and provide **structured validation results.**

issues: ["**Aspect-Opinion Relationship**: 'reduce' does not provide an evaluative or positive opinion on 'overfitting'; 'reduce' is an action, not an evaluation."]

[(“averaging effect”, “reduce overfitting”, “PERFORMANCE”, “POS”, 0.92),
(“results”, “stable”, “PERFORANCE”, “POS”, 0.95)]

Neutral Pattern:
This citation expresses [MASK] sentiment towards {neutral_word} through {neutral_word} which belongs to {sampled_category}

Contrastive Pattern:
Random sampling from $P_{pos} \cup P_{neg}$ pattern templates

Content-aware Pattern:
This citation expresses [MASK] sentiment towards {extracted_aspect} through {extracted_opinion} which belongs to {sampled_category}

The citation expresses [MASK] sentiment towards averaging effects through reduce overfitting which belongs to PERFORMANCE. The citation expresses [MASK] sentiment towards results through stable which belongs to PERFORMANCE.

[CLS] The citation expresses [MASK] … [SEP]

[CLS] The averaging effect … stable results [SEP]

Figure 3: Overall architecture of our TDCSA framework. aspect term and opinion term are highlighted in original text. The left bottom module consists of Dual LLM Feature Generation for quadruples extraction and verification. The left top presents Multi-view Pattern Representation modules that process both citation texts and extracted features. These representations are then integrated through an Feature Fusion module, then fed into the final classification layer.

as {*METHODOLOGY, PERFORMANCE, INNOVATION, APPLICABILITY, LIMITATION*}, and a polarity $p_i \in \text{NEU}, \text{POS}, \text{NEG}$. Scientific citations contain domain-specific terminology and often express sentiment through technical evaluative phrases rather than conventional sentiment markers. These characteristics necessitate specialized prompt designs for accurate aspect-sentiment extraction. Complete definitions of each category with annotation criteria are provided in Appendix B.

Extractor and Verifier Recent studies have demonstrated that LLM leveraging In-Context Learning (ICL) (Min et al., 2022) capabilities, can effectively perform ASQP tasks (Zhou et al., 2024; Bai et al., 2024). However, through our systematic analysis of 500 LLM-generated quadruples, we identified two recurring hallucination patterns in the interpretation of citation content. First, LLMs frequently over-interpret objective statements in citations as strong sentiments. Second, the confidence level of LLM output quadruples correlates with the uncertainty of judgment due to its probabilistic modeling properties. Inspired by dual-process theory in cognitive psychology, we propose a two-stage framework that mimics human thinking processes: fast intuitive thinking and slow critical thinking. The Quadruple Extractor employs ICL with carefully selected demonstrations to perform rapid pattern recognition as shown in Table 1. Through well-designed examples covering various cases like implicit opinions and sentiment transitions, it generates initial quadruple candidates with confidence scores $\alpha_i = \text{conf}(q_i) \in [0, 1]$ at a higher temperature 0.7. The Validity Verifier implements Chain-of-Thought reasoning at a lower temperature 0.3. It enforces both hard constraints and soft constraints through step-by-step verification reasoning, producing a final verification score $\beta_i = \text{verify}(q_i, \alpha_i) \in [0, 1]$.

2.2 Multi-view Pattern Representation

We observe a unique challenge in citation sentiment analysis that neutral citations constitute the majority of citation texts. Direct extraction of sentiment features from these neutral citations would lead to redundant feature representations due to their massive quantity and similar semantic patterns. To address this issue, we propose a Multi-view Pattern Representation mechanism that generates diverse

Text:	*However, one of the major limitations of these advances is the structured syntactic knowledge, which is important to global reordering, has not been well exploited.*
Quadruples:	('structured syntactic knowledge', 'has not been well exploited', 'LIMITATION', 'NEG', 0.90)

Table 1: Examples in extractor prompt. The aspect term, opinion term, aspect category are highlighted in different colors. Confidence level is also provided as a float variant.

complementary views for neutral citations:

Pattern Generation Given a neutral citation text $X_{neutral} = \{x_1, x_2, ..., x_n\}$, we construct three different representations:

1. **Neutral Pattern**: We randomly select tokens from a predefined neutral scientific vocabulary $V_{neutral}$ to replace the aspect and opinion terms:

   P_1 = "This citation expresses [MASK] sentiment towards $\{n_1\}$ through $\{n_2\}$ which belong to $\{$*category*$\}$ category"

   where $n_i \in V_{neutral}$, $category \in \mathcal{C}$.

2. **Contrastive Pattern**: We randomly sample positive or negative patterns from the existing quadruple templates:

   $$P_2 \in \{P_{pos} \cup P_{neg}\}$$

   This introduces controlled sentiment noise that helps the model learn robust neutral representations.

3. **Content-aware Pattern**: We extract key phrases from the neutral citation to construct a pseudo-quadruple:

   P_3 = "This citation expresses [MASK] sentiment towards $\{$*aspect*$\}$ through $\{$*opinion*$\}$ which belongs to $\{$*category*$\}$ category"

   where, *aspect* and *opinion* are phrases randomly selected from the citation, and *category*$\in \mathcal{C}$.

Thus, the Multi-view Pattern representation for neutral samples is defined by randomly selecting from the three patterns described below. Similarly, given a positive or negative text $X_{pos\_or\_neg} = \{x_1, x_2, ..., x_n\}$ and its corresponding quadruple set $Q = \{q_1, q_2, ..., q_m\}$ where each $q_i = (aspect_i, opinion_i, category_i, polarity_i)$, we use the same pattern for each q_i then concatenate them in order:

P = "This citation expresses [MASK] sentiment towards $\{$*aspect*$_i\}$ through $\{$*opinion*$_i\}$ which belongs to $\{$*category*$_i\}$ category [SEP] This citation expresses [MASK] sentiment towards ... "

For positive and negative citations, implicit sentiment is mapped to specific contextual features via quadruples. In contrast, for neutral samples, the generated pseudo-quadruples disrupt the original feature distribution, which improves the model's robustness to random noise.

2.3 Quad Feature Enhanced PLM-based Model

Dual Encoding Given a citation text $X = \{x_1, x_2, ..., x_n\}$ and its corresponding feature pattern $P = \{p_1, p_2, ..., p_m\}$, we first encode both sources using the PLMs to to obtain the last hidden state of contextual representations:

$$H_{\text{text/quad}} = \text{Encoder}_{\text{text/quad}}(X) \quad (1)$$

where $H_{\text{text}} \in \mathbb{R}^{n \times d}$ and $H_{\text{quad}} \in \mathbb{R}^{m \times d}$ represent the hidden states of citation text and sentiment quadruples, respectively.

Cross-attention Interaction To capture the mutual influence between citation content and sentiment quadruples, we employ bidirectional cross-attention:

$$A_{i \to j} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (2)$$

where Q, K, V are projections of H_{text} and H_{quad} using learnable weight matrices.

Non-linear Gated Fusion We propose an adaptive fusion mechanism to integrate the three complementary representations:

$$G = \sigma(W \cdot \text{pool}(\{H_{\text{text}}, H_{\text{quad}}, A_{\text{cross}}\}) + b) \quad (3)$$

where A_{cross} is the mean of bidirectional cross-attention, pool$(\cdot)$ represents the corresponding pooling operation applied to each input representation.

$$Z = G_1 \odot \phi(H_{text}) + G_2 \odot \phi(H_{quad}) + G_3 \odot \phi(A_{cross}) \quad (4)$$

where $\sigma(\cdot)$ is the sigmoid function, $\phi(\cdot)$ is a non-linear activation, and $\odot$ denotes element-wise multiplication. Finally, Z is fed into a fully connected classification layer W_{FC} and softmax function to obtain sentiment polarity score y:

$$y = \text{softmax}(ZW_{FC}) \quad (5)$$

3 Experimental Results

3.1 Datasets and Experimental Settings

The widely adopted CSA dataset Citation Sentiment Corpus (Athar Dataset)(Athar, 2011) presents relatively unbalanced distribution, consisting of 7627 neutral, 829 positive and 280 negative samples. To address quality issues stemming from automatic extraction, we constructed an enhanced dataset (CSCE) containing 6,328 neutral, 1,237 positive, and 631 negative samples. Each sample has been re-annotated by two human experts, with any conflicting cases carefully reviewed and resolved. For elaborate dataset annotation guidelines and processing steps, see Appendix A.

For our experimental settings, we employ LLaMA3.1-8B (Dubey et al., 2024) as the extractor (temperature: 0.7) and LLaMA3.1-70B as the verifier (temperature: 0.3). The QFE-PLM leverages pre-trained RoBERTa-base and SCIBERT models from Huggingface[2], maintaining bf16 precision and flash-attention2[3] acceleration. We conduct 10-fold cross-validation, reporting both accuracy and macro-F1 scores. All ablation studies maintain the same random seeds, and experiments are executed in an identical virtual environment using a single NVIDIA 4090 GPU with 24G RAM.

3.2 Baselines

We evaluate TDCSA with state-of-the-art approaches, including PLMs, deep neural network and conventional machine learning methods that have demonstrated superior performance on the Athar Dataset and CSA tasks. **PLMs baselines:** XLNet (Yang, 2019), BART (Lewis et al., 2019), DistilBERT (Sanh et al., 2019), BERT-base and BERT-large (Xu et al., 2019), SCIBERT (Beltagy et al., 2019), DictSentiBERT (Yu and Hua, 2023), DistilRoBERTa (Sanh et al., 2019), RoBERTa-base and RoBERTa-large (Liu, 2019), RoBERTa-llama3.1405B-twitter-sentiment (RoBERTa-sentiment) (Adam Lucek, 2024), and ImpactCite (Mercier et al., 2020). **GloVe based DNNs:** TextCNN, BiLSTM and BiLSTM with attention (Kong et al., 2024) using 300-dimensional GloVe embeddings[4] (Pennington et al., 2014). **Conventional machine learning methods:** SVM (dependency relation) (Athar, 2014) that uses dependency relations extracted by parsers[5]. In addition, **LLMs ICL and SFT:** Qwen2.5 (Yang et al., 2024), LLaMA, DeepSeek (Shao et al., 2024), and ChatGPT (OpenAI, 2022), as results shown in Appendix C.2.

3.3 Overall Results

The experimental results presented in Table 2 demonstrate that QFE-BERT and QFE-RoBERTa consistently outperform all baselines on both benchmark datasets. SVM based on dependency relations demonstrates robust performance on the Athar Dataset, suggesting the effectiveness of carefully engineered syntactic features for CSA. Among baseline models, SCIBERT's domain-specific pre-training yields notable improvements on CSCE Dataset but shows limited advantages on the Athar Dataset. RoBERTa-Sentiment, despite extensive sentiment pre-training, fails to surpass base models, indicating challenges in transferring social media sentiment analysis capabilities to academic contexts. Our QFE-PLM addresses the limitations of previous approaches by effectively integrating multi-view quadruples pattern representations and fusion mechanisms. Notably, it achieves significant improvements in F1 scores, demonstrating robust performance in the presence of dataset imbalances.

4 Analysis

4.1 Ablation Studies

We conducted comprehensive ablation experiments to evaluate the contribution of each component, as shown in Table 3. The results revealed that removing cross-attention degraded performance, indicating that simple concatenation inadequately

[2] https://github.com/huggingface/transformers
[3] https://github.com/Dao-AILab/flash-attention
[4] https://pytorch.org/text/stable/index.html
[5] https://github.com/explosion/spaCy

Base PLM	Models	Athar Dataset		CSCE Dataset	
		Acc.	F1.	Acc.	F1.
—	Most Frequent Class	88.3	93.2	77.2	87.1
	SVM(dependency relations)	89.8	76.4	85.7	70.0
	TextCNN	88.2	47.1	89.0	77.9
	BiLSTM	88.0	43.0	88.3	75.3
	BiLSTM+Attention	87.8	46.2	88.1	72.6
BART	BART-base	89.3	59.1	91.4	84.9
XLNet	XLNet-base	90.0	65.9	94.0	89.1
	ImpactCite	77.7	77.7	—	—
BERT	DistilBERT	89.2	60.3	94.0	89.1
	BERT-base	89.7	65.1	93.5	87.7
	BERT-large	89.5	66.0	93.4	88.6
	SCIBERT	90.1	63.8	95.5	92.2
	DictSentiBERT	—	—	95.2	86.0
	Ours: QFE-BERT	**95.5**	**81.2**	**97.8**	**95.0**
RoBERTa	DistilRoBERTa	89.9	64.9	93.5	88.6
	RoBERTa-base	90.1	68.0	94.5	90.4
	RoBERTa-large	90.5	68.4	94.3	90.6
	RoBERTa-Sentiment	89.7	66.9	93.4	88.3
	Ours: QFE-RoBERTa	**95.7**	**81.7**	**98.1**	**95.7**

Table 2: The main experimental results on Athar and CSCE datasets. The best results are marked as **bold** .

captures aspect/opinion term interactions. Replacing non-linear fusion with concatenation led to substantial performance deterioration, demonstrating the effectiveness of non-linear gated fusion for feature integration. Single BERT encoder variants showed reduced performance, suggesting that compressing features into a unified representation space may cause feature interference and diminish discriminative power. Notably, isolated components (Cross-attention or Quad only) exhibited significant performance drops, indicating that the model's discriminative ability relies more on the integrated representations than individual components. When examining interaction mechanisms shown in Figure 4, we found the bidirectional approach significantly outperforms unidirectional and cascaded alternatives. Thus, the interaction between original text and structured sentiment information provides complementary perspectives for sentiment understanding.

Model Variant	CSCE Dataset	
	Acc.	F1.
Base Model (Text Only)	95.5	92.2
QFE-RoBERTa	**98.1**	**95.7**
w/o Cross-attention	97.9	94.9
w/o Non-linear Fusion	97.3	93.9
Single BERT	97.2	93.8
Cross-attention Only	95.8	89.0
Quad Only	95.6	88.4

Table 3: Ablation study results.

4.2 Analysis of Multi-view Representation

The effectiveness of our Multi-view Pattern approach is validated through three systematic experiments. First, as shown in Table 4, compared to single pattern baselines, our method demonstrates consistent performance improvements in both accuracy and F1 score, showing the advantage of diverse representations for sentiment classification. Second, we conducted two critical ablation studies

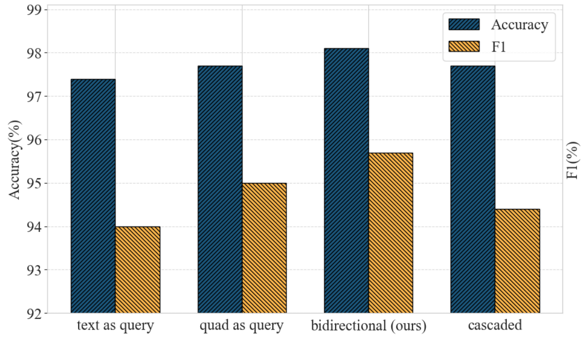


Figure 4: Analysis of different attention interaction mechanisms.

Multi-view Representation	CSCE Dataset	
	Acc.	F1.
Neutral Pattern	**98.4**	95.5
Contrastive Pattern	97.9	95.4
Content-aware Pattern	98.2	95.5
Ours: Multi-view Pattern	98.3	**95.7**
Random Contrastive Pattern[1]	93.5	88.2
Neutral Quad Pattern[2]	95.3	91.4

Table 4: Comparative Analysis of different pattern representations. Random Contrastive Pattern[1] represents an ablation study where sentiment quadruples are randomly assigned to positive and negative citation samples, rather than using their matched quadruples. Neutral Quad Pattern[2] indicates experiments where DFG extracts sentiment quadruples from neutral citations as an alternative to the multi-view pattern representation.

to verify our design choices. The Random Contrastive Pattern where sentiment quadruples were randomly assigned to citations, resulted in performance degradation, confirming the importance of maintaining semantic relationships in quadruple representations. Similarly, the Neutral Quad Pattern experiment, which attempted to extract sentiment quadruples from neutral citations, showed reduced effectiveness. As visualized in Figure 6 in Appendix, this approach introduces feature redundancy and blurs classification boundaries, whereas our Multi-view Pattern creates clearer sentiment separations.

To validate the statistical significance of the performance improvements achieved by our Multi-view Pattern approach, we conducted paired t-tests comparing our method against the Random Contrastive Pattern and Neutral Quad Pattern. For each comparison, we performed 10-fold cross-validation with 5 repetitions using different random seeds, collecting a total of 50 performance measurements for each method. The results confirm that our Multi-view Pattern representation achieves statistically significant improvements over both baseline approaches ($p < 0.05$). Specifically, when comparing to Random Contrastive Pattern, we obtained $t = 3.142$, $p = 0.007$, and when comparing to Neutral Quad Pattern, we obtained $t = 2.896$, $p = 0.012$.

Train Pattern	Test Pattern			
	NP	CP	CAP	Multi
NP	95.5	30.7	85.1	67.8
CP	95.8	**95.4**	95.8	95.7
CAP	95.5	57.7	95.5	80.1
Multi	**96.6**	93.4	**96.1**	**95.7**

Table 5: Pattern transfer results (F1-scores). Abbreviations: NP (Neutral Pattern), CP (Contrastive Pattern), CAP (Consistent Adversarial Pattern), Multi (Multi-view Pattern). The best and second best generalization performance is marked as **bold** and underlined respectively.

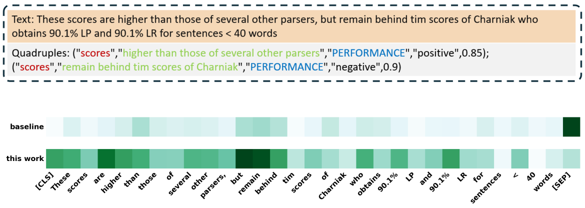
Finally, we evaluate pattern generalization through transfer experiments. As shown in Table 5, Models trained with our Multi-view Pattern demonstrate strong cross-pattern generalization, maintaining high F1 scores across different test patterns. This robust generalization is particularly evident compared to simpler approaches like Neutral Pattern and Content-aware Pattern, which suffer from performance drops when tested on different pattern types. These results suggest that our approach successfully learns transferable features that capture fundamental sentiment characteristics rather than pattern-specific features.

4.3 Analysis of Dual LLM Feature Generation

Through the analysis of the DFG framework, we identify two key findings. First, as shown in Table 6, The extractor tends to over-interpret neutral statements as subjective, often with relatively low confidence scores. However, smaller verifier models effectively mitigate such misinterpretations through validation. While the framework does not always perfectly delineate aspect term boundaries, it consistently captures essential terms, such as "method". We also performed a comparison of different LLM configurations and detailed results are presented in Appendix E.

Citation:	Hanks proposed using pointwise mutual information to identify collocations in lexicography; however, the method may result in unacceptable collocations for low-count pairs(Jian et al., 2004).
Dual LLM:	(1) ('method', 'may result in unacceptable collocations', 'LIMITATION', 'NEG', 0.90) ✓ (2) ('pointwise mutual information', 'proposed', 'METHODOLOGY', 'POS', 0.78) ✗
Human Annotation:	*aspect* in quad (1) should be 'method using pointwise mutual information'

Table 6: Quality analysis of DFG framework and human annotation. The ✓/✗ indicates the verifier's validation results.


Text: These scores are higher than those of several other parsers, but remain behind tim scores of Charniak who obtains 90.1% LP and 90.1% LR for sentences < 40 words

Quadruples: ("scores","higher than those of several other parsers","PERFORMANCE","positive",0.85); ("scores","remain behind tim scores of Charniak","PERFORMANCE","negative",0.9)

Figure 5: Sentiment transition attention weights visualization comparison of baseline and our method. Our method learns the aspect and contrasted opinion terms and predicts **NEG**.

4.4 Case Study

We analyze three representative cases demonstrating our model's capability in handling different sentiment expressions. In complex cases with sentiment transitions as shown in Figure 5, our model effectively captures the crucial comparative phrases that determine the final sentiment, whereas baseline approaches struggle with such mixed expressions. Cases with explicit and implicit sentiment expressions are elaborated in Appendix F.

5 Related Work

Scientific Citation Sentiment Analysis Early approaches employed conventional machine learning algorithms and sentiment lexicon-based methods (Yousif et al., 2019b). Athar (Athar, 2011) pioneered this field by developing an open-source dataset and demonstrated the effectiveness of combining 3-grams with dependency relations using SVM. Subsequently, approaches evolved through sentiment lexicons (Xu et al., 2015) and basic machine learning methods (Sula and Miller, 2014; Athar and Teufel, 2012). With the development of deep learning, methods progressed from CNN/RNN-based approaches (Munkhdalai et al., 2016; Yousif et al., 2019a) to Pre-trained Language Models (PLMs). Notable achievements include XLNet-based approaches (Mercier et al., 2020), DictSentiBERT (Yu and Hua, 2023), and SCIBERT (Beltagy et al., 2019). Recent work has focused on multi-task learning combining citation aspects and sentiments (Kong et al., 2024). However, these approaches still struggle with implicit sentiment expressions in academic writing.

Citation Aspect-Based Sentiment Analysis Aspect-Based Sentiment Analysis (ABSA) provides fine-grained analysis through aspect term extraction, opinion term identification, aspect category classification, and sentiment polarity determination (Zhang et al., 2022). While BERT-based models have shown promise in general ABSA tasks (Li et al., 2019), their application to citation analysis remains limited. Recent approaches like Multiview Prompting (Gou et al., 2023) and unified generative frameworks (Gao et al., 2022) have advanced the field but primarily focus on general domain tasks. Large Language Models (LLMs) have demonstrated capabilities in sentiment analysis (Lu et al., 2025), with recent studies exploring their potential in ABSA through instruction tuning (Varia et al., 2022) and few-shot learning (Šmíd et al., 2024). However, their application to citation sentiment analysis faces unique challenges due to the domain-specific nature of scientific citations. Previous work in citation ABSA (Hernández-Alvarez and Gomez, 2016; Ikram and Afzal, 2019) has primarily focused on basic pattern-based approaches, leaving room for more sophisticated methods that can capture the nuanced relationships in academic writing.

6 Conclusion

In this paper we introduced TDCSA, a novel Top-Down approach designed to address the challenges in Citation Sentiment Analysis. By leveraging LLMs for robust quadruple generation and enhancing pre-trained language models with adaptive fusion mechanisms, TDCSA achieved significant performance improvements over existing methods. Our analysis yields three key findings: First, structured sentiment quadruples effectively extraction provides interpretable features that enhance citation sentiment detection, particularly for implicit sentiment expressions. Second, Our Multi-view Pattern Representation effectively addresses the challenge of dominant neutral citations by generating complementary feature representations. Third, the cross-pattern generalization experiments in Section 4.2 verified that our model maintains consistent performance across different pattern types. Experiments on benchmark datasets demonstrate that TDCSA significantly outperforms existing methods, achieving state-of-the-art performance.

Limitations

We summarize the limitations of our framework that could be followed by future work. The main limitation points to a broader challenge in CSA. Our work, while advancing the state-of-the-art in general citation sentiment classification, highlights the necessity for aspect-based sentiment analysis in scientific citations. A more fine-grained ABSA approach would enable better understanding of how different aspects of scientific work are evaluated in the academic community, potentially offering deeper insights into research impact and knowledge evolution.

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 596–606.

Adam Lucek. 2024. roberta-llama3.1405b-twitter-sentiment(revision 75d7d53).

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session*, pages 81–87.

Awais Athar. 2014. Sentiment analysis of scientific citations. Technical report, University of Cambridge, Computer Laboratory.

Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 597–601.

Yinhao Bai, Zhixin Han, Yuhua Zhao, Hang Gao, Zhuowei Zhang, Xunzhi Wang, and Mengting Hu. 2024. Is compound aspect-based sentiment analysis addressed by llms? In *Conference on Empirical Methods in Natural Language Processing*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Gregor Betz and Kyle Richardson. 2020. Critical thinking for language models. *ArXiv*, abs/2009.07185.

Frédérique Bordignon and Philippe Gambette. 2024. A corpus of critical citations contexts. *Journal of Open Humanities Data*, 10.

Indra Budi and Yaniasih Yaniasih. 2022. Understanding the meanings of citations using sentiment, role, and citation function classifications. *Scientometrics*, 128:735–759.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *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)*, pages 340–350.

Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. Lego-absa: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th international conference on computational linguistics*, pages 7002–7012.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. *arXiv preprint arXiv:2305.12627*.

Myriam Hernández-Alvarez and José M Gomez. 2016. Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, 22(3):327–349.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Imran Ihsan, Hameedur Rahman, Asadullah Shaikh, Adel Sulaiman, Khairan Rajab, and Adel Rajab. 2023. Improving in-text citation reason extraction and classification using supervised machine learning techniques. *Computer Speech & Language*, 82:101526.

Muhammad Touseef Ikram and Muhammad Tanvir Afzal. 2019. Aspect based citation sentiment analysis using linguistic patterns for better comprehension of scientific knowledge. *Scientometrics*, 119:73–95.

Jia-Yan Jian, Yu-Chia Chang, and Jason S Chang. 2004. Tango: Bilingual collocational concordancer. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 166–169.

Ling Kong, Wei Zhang, Haotian Hu, Zhu Liang, Yonggang Han, Dongbo Wang, and Min Song. 2024. Transdisciplinary fine-grained citation content analysis: A multi-task learning perspective for citation aspect and sentiment classification. *Journal of Informetrics*, 18(3):101542.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting bert for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*.

Hanfeng Liu, Minping Chen, Zhenya Zheng, and Zeyi Wen. 2024. Exploiting careful design of svm solution for aspect-term sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5897–5906.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Heng-yang Lu, Tian-ci Liu, Rui Cong, Jun Yang, Qiang Gan, Wei Fang, and Xiao-jun Wu. 2025. Qaie: Llm-based quantity augmentation and information enhancement for few-shot aspect-based sentiment analysis. *Information Processing & Management*, 62(1):103917.

Zheng Ma, Jinseok Nam, and Karsten Weihe. 2016. Improve sentiment analysis of citations with author modelling. In *Proceedings of the 7th workshop on computational approaches to subjectivity, Sentiment and Social Media Analysis*, pages 122–127.

Dominique Mercier, Syed Tahseen Raza Rizvi, Vikas Rajashekar, Andreas Dengel, and Sheraz Ahmed. 2020. Impactcite: An xlnet-based method for citation impact analysis. *arXiv preprint arXiv:2005.06611*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *ArXiv*, abs/2202.12837.

Tsendsuren Munkhdalai, John P Lalor, and Hong Yu. 2016. Citation analysis with neural attention models. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 69–77.

OpenAI. 2022. Chatgpt: A language model-based dialogue system. *OpenAI Blog*. Accessed: 2024-11-17.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*.

Horacio Saggion, Ahmed Ghassan Tawfiq AbuRa'ed, and Francesco Ronzano. 2016. Trainable citation-enhanced summarization of scientific articles. In *Cabanac G, Chandrasekaran MK, Frommholz I, Jaidka K, Kan M, Mayr P, Wolfram D, editors. Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL); 2016 June 23; Newark, United States. CEUR Workshop Proceedings; 2016. p. 175-86.* CEUR Workshop Proceedings.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Zhihong Shao, Damai Dai, Daya Guo, Bo Liu (Benjamin Liu), Zihan Wang, and Huajian Xin. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *ArXiv*, abs/2405.04434.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2024. Llama-based models for aspect-based sentiment analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 63–70.

Chris Alen Sula and Matthew Miller. 2014. Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. *Literary and Linguistic Computing*, 29(3):452–464.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2022. Instruction tuning for few-shot aspect-based sentiment analysis. *arXiv preprint arXiv:2210.06629*.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *ArXiv*, abs/2304.04339.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Jun Xu, Yaoyun Zhang, Yonghui Wu, Jingqi Wang, Xiao Dong, and Hua Xu. 2015. Citation sentiment analysis in clinical trial papers. In *AMIA annual symposium proceedings*, volume 2015, page 1334. American Medical Informatics Association.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. 2024. Qwen2 technical report. *ArXiv*, abs/2407.10671.

Ning Yang, Zhiqiang Zhang, and Feihu Huang. 2023. A study of bert-based methods for formal citation identification of scientific data. *Scientometrics*, 128(11):5865–5881.

Zhilin Yang. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Abdallah Yousif, Zhendong Niu, James Chambua, and Zahid Younas Khan. 2019a. Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, 335:195–205.

Abdallah Yousif, Zhendong Niu, John K Tarus, and Arshad Ahmad. 2019b. A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review*, 52:1805–1838.

Dahai Yu and Bolin Hua. 2023. Sentiment classification of scientific citation based on modified bert attention by sentiment dictionary. In *EEKE/AII@ JCDL*, pages 59–64.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Conference on Empirical Methods in Natural Language Processing*.

Wenxuan Zhang, Yue Deng, Bing-Quan Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *ArXiv*, abs/2305.15005.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

Changzhi Zhou, Dandan Song, Yuhang Tian, Zhijing Wu, Hao Wang, Xinyu Zhang, Jun Yang, Ziyi Yang, and Shuhao Zhang. 2024. A comprehensive evaluation of large language models on aspect-based sentiment analysis. *arXiv preprint arXiv:2412.02279*.

A Details on Datasets

A.1 Annotation Guidelines

While numerous studies have explored Citation Sentiment Analysis (Munkhdalai et al., 2016; Abu-Jbara et al., 2013; Ma et al., 2016), publicly available datasets remain scarce. The widely used Citation Sentiment Corpus (Athar Dataset) (Athar, 2011) exhibits an imbalanced distribution, consisting of 7627 neutral, 829 positive, and 280 negative samples from the ACL Anthology Network. However, this dataset suffers from quality issues due to its automatic extraction and manual annotation process, as shown in Table 7.

To address these limitations, we re-annotate the dataset following comprehensive guidelines. We recognize that scientific literature contains many genuinely neutral citations by design, serving essential functions such as establishing context or referencing methodologies without evaluation. Our goal is not to force sentiment interpretations where none exist, but rather to accurately distinguish between truly neutral citations and those containing subtle implicit sentiment.

Our annotation guidelines define:

- **Neutral Citations (NEU)**: Citations that serve foundational functions without evaluative content, such as establishing background knowledge, referencing methodologies, or attributing data sources. These citations are intrinsically neutral by design.

  The minimum error training was used on the development data for parameter estimation.

- **Positive Citations (POS)**: Citations expressing approval, endorsement, or positive assessment through explicit praise, adoption/extension, positive comparison, or implicit endorsement.

Problem	Example / Details
(1)Too long sentence	It us widely acknowledged that word sense d samblguatmn (WSD) us a central problem m natural language processing... **(12518 words)**
(2)Too short sentence	(Cutting et al. , 1992)). **(no semantics)**
(3)Incomplete sentence	5http://www.statmt.org/wmt08 185 the BLEU score (Papineni et al., 2002), and tested on test2008. **(no semantics)**
(4)Wrong label	By segmenting words into morphemes, we can improve the performance of natural language systems including machine translation (Brown et al. 1993) and information retrieval (Franz, M. and McCarley, S. 2002). **(mislabel)**

Table 7: Problem and Example. Too long/short or incomplete sentences are removed and mislabel sentences are relabeled manually.

For the multilingual dependency parsing track, which was the other track of the shared task, Nilsson et al. achieved the best performance using an ensemble method.

- **Negative Citations (NEG)**: Citations indicating criticism, identification of limitations, or disagreement through explicit criticism, contrasting statements, limitation highlighting, or implicit critique.

  Other systems (Morinaga et al., 2002; Kushal et al., 2003) also look at Web product reviews but they do not extract opinions about particular product features.

Each citation was independently annotated by two computational linguistics experts with prior experience in sentiment analysis. Annotators achieved substantial agreement, with disagreements resolved through discussion to determine the final label. To mitigate potential subjective bias or over-interpretation, annotators were explicitly instructed to maintain the neutral classification unless clear linguistic evidence of sentiment was present.

Following these guidelines, we refined the Athar Dataset by removing problematic instances (1-3) and correcting questionable labels (4) as shown in Table 7.

A.2 CSCE Construction

We constructed an enhanced dataset through a systematic approach. First, we integrated the complementary dataset from (Yu and Hua, 2023), which utilizes the supported and unsupported labels from the SCICite dataset (Cohan et al., 2019) as positive and negative samples, respectively. Next, We adpoted the refined Athar Dataset and augmented the negative class with key citations from (Bordignon and Gambette, 2024). The resulting dataset, refered as CSCE, comprises 6328 neutral, 1237 positive, and 631 negative samples.

B Citation Aspect Sentiment Quad Prediction Task Definition

Due to the difference between scientific citations and general sentiment texts like product reviews or social media posts, where aspects are typically simple noun phrases and opinions are predominantly single adjectives, scientific citations exhibit more complex and domain-specific linguistic patterns. Similarly, rather than straightforward adjectives, an opinion term o_i in citations often appears as intricate verb phrases, compound expressions combining adverbs and adjectives, or technical evaluative phrases. An aspect term a_i in citations frequently manifests as sophisticated noun phrases encompassing technical terminology, methodological components, or theoretical constructs. Furthermore, we define the aspect category set $\mathcal{C}$ = as {*METHODOLOGY, PERFORMANCE, INNOVATION, APPLICABILITY, LIMITATION*} five categories as follows.

- *METHODOLOGY*: This category encompasses assessments of research methods, experimental designs, and technical approaches. It is essential as methodological rigor is a fundamental criterion in scientific evaluation (e.g., "Their robust experimental design validates the findings").

- *PERFORMANCE*: This category addresses quantitative and qualitative outcomes, including accuracy, efficiency, and effectiveness. It is crucial for comparing and benchmarking scientific contributions (e.g., "The model achieves superior accuracy on standard benchmarks").

- *INNOVATION*: This category captures the novelty and originality of the research contribution. It is vital in academic discourse as advancing the state-of-the-art is a key measure of research impact (e.g., "They propose a novel framework that breaks new ground").

- *APPLICABILITY*: This category reflects the practical utility and generalizability of the research. It is important as it bridges the gap between theoretical contributions and real-world applications (e.g., "Their approach can be effectively applied to various domains").

- *LIMITATION*: This category addresses constraints, drawbacks, and areas for improvement. It is critical for maintaining scientific rigor through balanced critique and identifying future research directions (e.g., "The method's computational complexity limits its practical deployment").

C Extended Experimental Analysis

C.1 Implementation Details

PLMs fine-tuning employs the following hyperparameters: batch size of 32, dropout rate of 0.1, weight decay of 0.05, and cosine annealing learning rate decay with an initial rate of 2.0e-5 using AdamW optimizer.

In zero-shot and few-shot learning scenarios, we employ various models from the Qwen and LLaMA family (Dubey et al., 2024), specifically the 3B version of LLaMA 3.2, the 7B and 14B versions of Qwen2.5 and the 8B, 70B, and 405B versions of LLaMA 3.1. Additionally, we incorporate ChatGPT and DeepSeek versions of DeepSeek V2 Chat, GPT-3.5-turbo, and GPT-4o-2024-0806.

When fine-tuning we use LLaMA-Factory framework[6] and QLoRA (Dettmers et al., 2023) with 4-bit NormalFloat (NF4) and double quantization using bitsandbytes[7], cosine annealing learning rate decay with an initial rate of 1.0e-4 using AdamW optimizer for 5 epochs, and batch size of 8. We set LoRA adapters (Hu et al., 2021) with $r = 8$ and $\alpha = 16$. Both ICL and SFT use bf16 training/inference precision and flash-attention2 acceleration. Temperature is set as 0.1 for certain responses. 6-shot ICL indicates two golden examples are selected from three categories as demonstrations.

C.2 LLMs In-Context Learning and Supervised Fin-Tuning results

The ICL and SFT experimental results presented in Table 8. 6-shot ICL generally shows better performance than 0-shot ICL, but worse than SFT. The gap between ICL and SFT is substantial, with SFT outperforming by 10-15 %. In zero-shot and few-shot scenarios, LLaMA3.1-405B and GPT-4o shows best performance in F1 while DeepSeek V2 achieves the highest accuracy in both ICL. Notably, Qwen models show strong performance across models of similar scale like LLaMA. However, the SFT of the 14B version of Qwen exhibits performance characteristics that are nearly identical to the native BERT model, which demonstrates that the performance of LLM in a specific domain such as CSA is suboptimal to the BERT-based model.

D Detailed Analysis of Multi-view Patterns

Furthermore, we employed t-SNE (Van der Maaten and Hinton, 2008) to reduce dimensionality of the pooler output of BERT encoder and quad-enhanced fusion features. Figure 6 presents that the degradation can be attributed to two primary factors: First, neutral citations in academic writing predominantly manifest as descriptive or declarative statements, making sentiment quadruple extraction potentially redundant. Second, the augmentation of numerous neutral samples with quadruple structures may introduce excessive feature redundancy into the representation space. This redundancy ultimately compromises the model's capacity to effectively discriminate sentiment features.

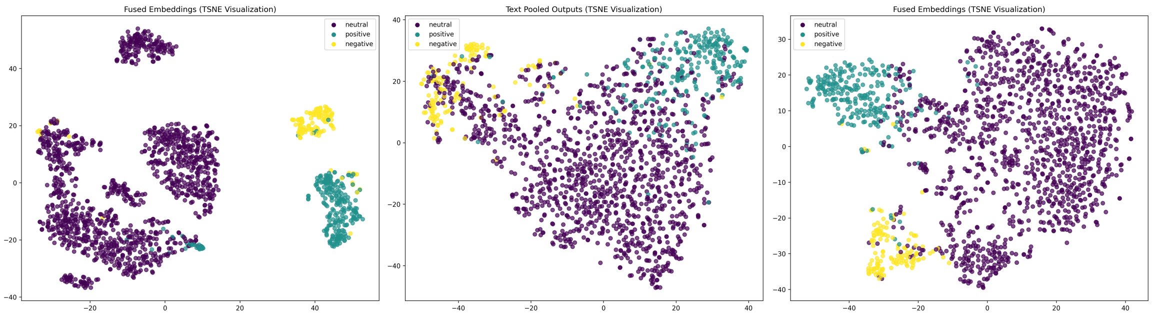
E Detailed Analysis of Dual LLM Feature Generation

We performed a comparison of different LLM configurations as presented in Table 9, various LLM configurations consistently enhance CSA performance, highlighting the robustness of our method to variations in quadruple quality.This suggests that basic semantic understanding capability is sufficient for our framework, rather than necessitating extremely large models.

We assign confidence scores to quadruple extraction and verification tasks across various LLM implementations, with results presented in Table 10. Our empirical analysis reveals that the dual LLM mechanism consistently achieves higher confidence levels across most models, with the notable

[6] https://github.com/hiyouga/LLaMA-Factory

[7] https://github.com/bitsandbytes-foundation/bitsandbytes



(a) Multi-view Pattern (b) Baseline (BERT) (c) Neutral Quad Pattern

Figure 6: t-SNE dimensionality reduction of our model embeddings. Colors represent different sentiment classes (purple: neutral, green: positive, yellow: negative). Ours Multi-view Pattern reduces the distance within classes and increases the distance between classes, resulting in a clearer classification boundary. However, Neutral Quad Pattern (neutral samples enhanced with their own quadruples) blurs classification boundary, making classification harder.

exception of LLaMA3.2-3B. The relatively lower performance of LLaMA3.2-3B can be attributed to potential overconfidence due to its limited parameter size.

F Detailed Case Study

As shown in Figure 7, in explicit sentiment cases, our model successfully focuses attention on critical evaluative phrases like "*reduce overfitting*" and "*stable results*", while baseline models only partially capture these signals. For implicit sentiment, our quadruple-enhanced model precisely identifies subtle evaluations like "*manageable*" as positive assessment of "behavior", maintaining accuracy even with imperfect quadruples.

G Error Analysis

Through comprehensive error analysis of challenging cases, we identify several patterns that our model struggles with in scientific citations. Table 11 depicts representative examples that highlight opinion terms in quadruples. Attention visualization of wrong predictions is shown in Figure 8. The fist challenge in cases (1), (3) and (5) is to predict overall sentiment in mixed sentiment expressions. Case (1) stems from the model's difficulty in recognizing the rhetorical structrure where initial praise ("best efforts") is overshadowed by subsequent criticism ("leaving unassessed"), while case (3) and (5) present a factual relationship between methods without negative evaluation. Another challenge in cases (2), (4), and (6) arises from the model's tendency to either over-reliance or under-reliance on the information provided by the quadruples. In figure 8(b), the model focuses more attention on "taking advantage of" and recognize the citation as POS, while in figure 8(d), 8(f), the model does not select the unconventional expression of sentiment provided by the quadruple such as "go beyond" and "compensates overfit", thus, predict the citation as NEG.

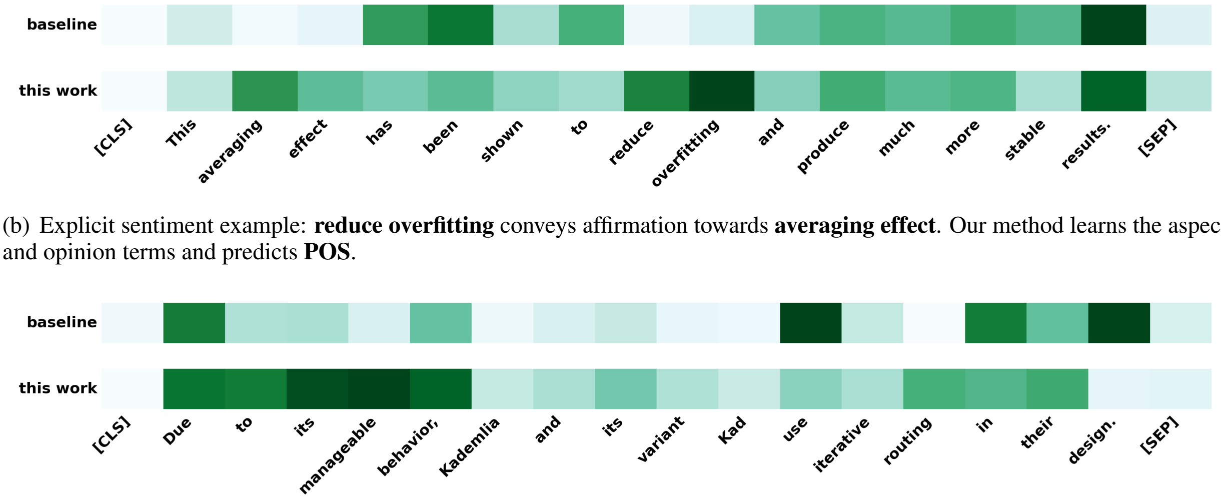
Text: This averaging effect has been shown to reduce overfitting and produce much more stable results.

Quadruples: ("averaging effect", "reduce overfitting", "PERFORMANCE", "POS", 0.92);
("results", "stable", "PERFORMANCE", "POS",0.95)

Text: Due to its manageable behavior, Kademlia and its variant Kad use iterative routing in their design.

Quadruples: ("behavior", "manageable", "PERFORMANCE", "POS", 0.85);
("design", "Iterative routing", "METHDOLOGY", "POS",0.8)

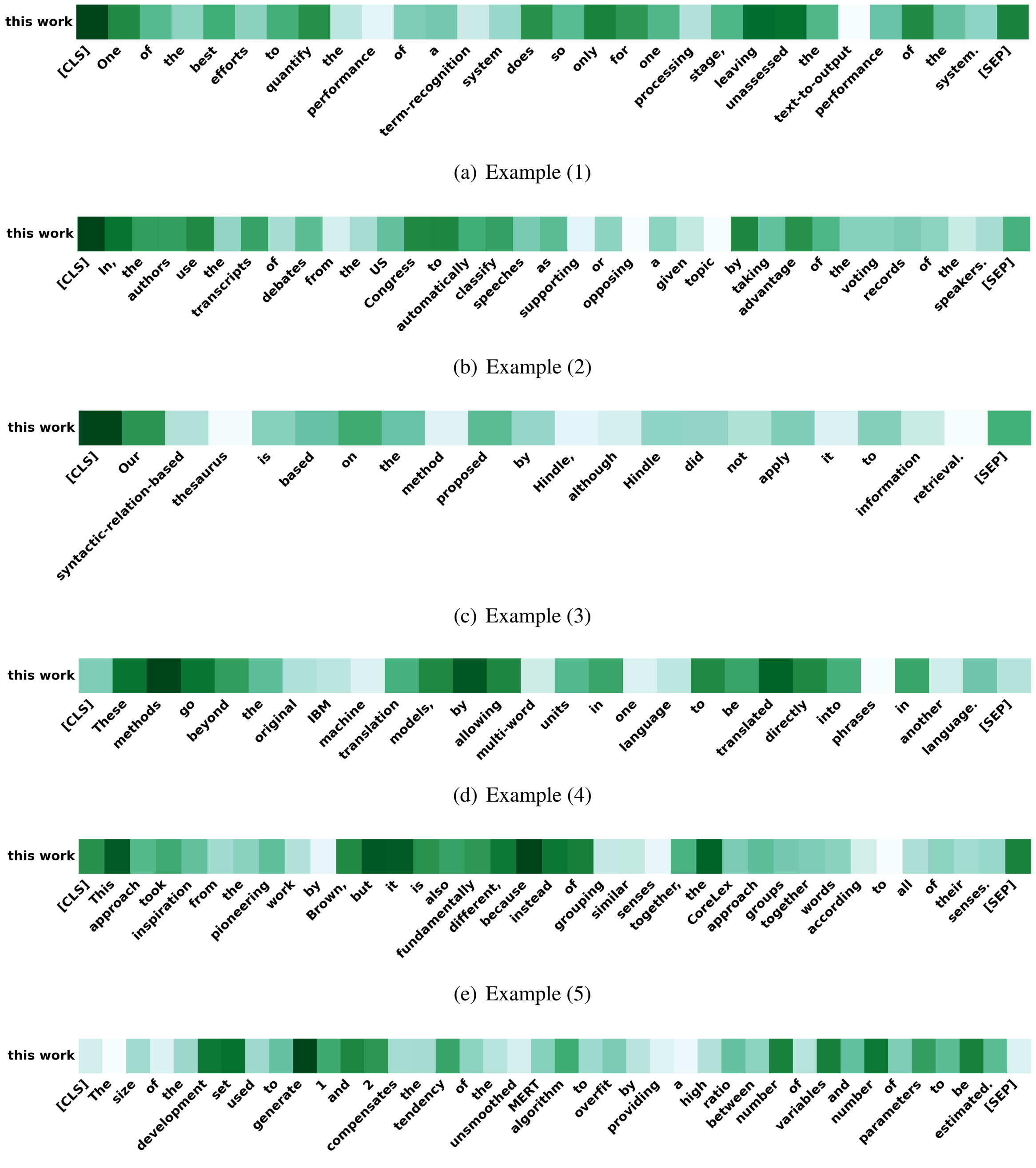
(a) Examples of the case study including explicit and implicit citations. Quadruples: (*aspect*, *opinion*, *category*, *polarity*, *confidence*).



(b) Explicit sentiment example: **reduce overfitting** conveys affirmation towards **averaging effect**. Our method learns the aspect and opinion terms and predicts **POS**.

(c) Implicit sentiment example: **manageable** shows advantages towards **behavior**. Our method learns the aspect and opinion terms and predicts **POS**.

Figure 7: Attention weights visualization comparison of baseline and our method.



(a) Example (1)

(b) Example (2)

(c) Example (3)

(d) Example (4)

(e) Example (5)

(f) Example (6)

Figure 8: Attention weights visualization of error samples in our method.

Method	Model	CSCE Dataset	
		Acc.	F1.
0-shot	Qwen2.5-7B	78.4	58.7
	Qwen2.5-14B	79.9	61.2
	LLaMA3.2-3B	64.0	52.6
	LLaMA3.1-8B	77.1	57.4
	LLaMA3.1-70B	80.4	62.3
	LLaMA3.1-405B	81.0	**64.9**
	DeepSeek V2	**82.7**	62.8
	GPT-3.5	80.3	63.5
	GPT-4o	80.8	64.7
6-shot	Qwen2.5-7B	82.1	63.7
	Qwen2.5-14B	82.4	64.8
	LLaMA3.2-3B	45.2	43.7
	LLaMA3.1-8B	80.1	61.1
	LLaMA3.1-70B	81.5	66.6
	LLaMA3.1-405B	83.1	**70.5**
	DeepSeek V2	**83.6**	66.5
	GPT-3.5	81.6	67.0
	GPT-4o	82.5	69.8
SFT	Qwen2.5-7B	93.9	90.7
	Qwen2.5-14B	**96.3**	**91.6**
	LLaMA3.2-3B	90.3	81.9
	LLaMA3.1-8B	94.5	89.5

Table 8: ICL and SFT results on the CSCE dataset, comparing various models across zero-shot, few-shot, and supervised fine-tuning methods. The highest-performing models are highlighted in **bold** (best) and underlined (second best).

Extractor/Verifier LLMs	CSCE Dataset
	F1.
LLaMA-3B/70B	93.8
LLaMA-8B/70B	94.9
LLaMA-405B/70B	**95.7**
LLaMA-70B/—	95.1
GPT-3.5/LLaMA-70B	95.3
GPT-4o/LLaMA-70B	95.4
LLaMA-8B/DeepSeek-V2	94.7
LLaMA-8B/GPT-4o	94.6

Table 9: Different Extractor+Verifier LLMs on QFE-RoBERTa performance. Our experiments demonstrate robustness of different Extractor/Verifier configurations to model's performance.

Extractor	Verifier	Extr Conf.	Veri Conf.	POS Num.	NEG Num.
LLaMA3.2-3B	LLaMA3.1-70B	0.932	0.902 ↓	2638	1177
LLaMA3.1-8B	LLaMA3.1-70B	0.872	0.903 ↑	2611	993
LLaMA3.1-70B	—	0.902	—	2688	902
LLaMA3.1-405B	LLaMA3.1-70B	0.896	0.901 ↑	2662	836
GPT-3.5[1]	LLaMA3.1-70B	0.907	0.909 ↑	2585	843
GPT-4o[2]	LLaMA3.1-70B	0.910	0.912 ↑	2615	837
DeepSeek V2[3]	LLaMA3.1-70B	0.920	0.921 ↑	2706	867

Table 10: Confidence Scores and Sample Distribution Across Language Models. Most models show increased confidence scores after verification, dual LLM framework demonstrates improved confidence scores compared to single-model approaches. [1]GPT-3.5-turbo, [2]GPT-4o-2024-08-06, [3]DeepSeek-V2-Chat.

Error examples	Label	Prediction
(1) One of the **best** efforts to quantify the performance of a term-recognition system does so only for one processing stage, **leaving unassessed** the text-to-output performance of the system.	NEG	NEU
(2) In, the authors use the transcripts of debates from the US Congress to automatically classify speeches as supporting or opposing a given topic by **taking advantage of** the voting records of the speakers.	NEU	POS
(3) Our syntactic-relation-based thesaurus is **based on** the method proposed by Hindle, although Hindle **did not apply** it to information retrieval.	NEU	NEG
(4) These methods **go beyond** the original IBM machine translation models, by allowing multi-word units in one language to be translated directly into phrases in another language.	POS	NEU
(5) This approach took inspiration from the **pioneering** work by Brown, but it is also fundamentally different, because instead of grouping similar senses together, the CoreLex approach groups together words according to all of their senses.	NEU	POS
(6) The size of the development set used to generate 1 and 2 **compensates the tendency** of the unsmoothed MERT algorithm to **overfit** by providing a high ratio between number of variables and number of parameters to be estimated.	POS	NEU

Table 11: Examples of model's prediction errors with *opinion* terms in quadruples highlighted in **bold**.