Proceedings of CONF-MLA 2026 Symposium: Explainable Computing, Modeling & Data Science in Complex Systems
DOI: 10.54254/2755-2721/2026.GF35112

Aspect-Dimension Sentiment Analysis for Short Chinese E-Commerce Reviews: A Comparative Study of Lightweight CNN, LSTM and Transformer Models

Jiaye Huang

College of Artificial Intelligence, Dongguan City University, Dongguan, China
2658517653@qq.com

Abstract. The rapid growth of online shopping platforms has significantly increased the analytical value of product review data. Short reviews on Chinese e-commerce platforms, typically containing fewer than 50 characters, simultaneously address multiple dimensions, including product quality, logistics, customer service, and value for money, rendering conventional sentence-level sentiment classification insufficient for fine-grained analysis. This study adopts aspect-based sentiment analysis (ABSA) as the core framework and constructs a labeling schema comprising four sentiment dimensions and three polarity levels, yielding 12 classification categories. Rule-based automated annotation was applied to 62,770 publicly available Chinese consumer reviews; manual verification on a 500-sample subset confirmed an annotation accuracy of 87.2%. Three lightweight models—TextCNN, BiLSTM with attention (BiLSTM+Attn), and ERNIE 3.0-Nano—were trained and evaluated under identical experimental conditions on an NVIDIA RTX 4060 laptop GPU. ERNIE 3.0-Nano proved to be the most accurate, with an accuracy of 0.6363 and a macro-average F1 score of 0.5798. BiLSTM+Attn was the fastest on GPU: 0.004 ms per sample. TextCNN had the smallest parameter count: 0.69 million. These findings provide quantitative evidence for model selection under diverse deployment constraints.

Keywords: Aspect-based sentiment analysis, Chinese e-commerce reviews, TextCNN, BiLSTM, ERNIE 3.0-Nano

1. Introduction

E-commerce platforms such as JD.com and Taobao host billions of product reviews daily. For merchants, this data constitutes a direct window into genuine consumer feedback; for platforms, it serves as an essential signal for improving recommendation algorithms and optimizing supply chains. Short Chinese reviews present unique analytical challenges: most reviews contain fewer than 50 characters yet simultaneously address multiple product aspects. Furthermore, Chinese users frequently convey negative sentiment through factual statements rather than explicit evaluative terms. The phrase "waited a whole week," for instance, expresses a logistics complaint without containing any overtly negative words. Conventional sentence-level sentiment classification is

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GF35112

inadequate for such data, making aspect-based sentiment analysis (ABSA) a more appropriate technical framework [1, 2].

Although BERT-based large models achieve strong performance on ABSA tasks [3], parameter counts exceeding 100 million generate prohibitive inference costs in high-throughput production settings. Lightweight alternatives—TextCNN [4], BiLSTM with attention [5], and knowledge-distilled models [6]—offer clear advantages in computational efficiency. However, systematic comparative studies across these three architecture families targeting Chinese e-commerce short reviews specifically remain limited in the existing literature.

This study focuses on three goals. First, it designs a reproducible, rule-based annotation framework covering four sentiment dimensions for Chinese e-commerce short reviews. Second, using that dataset, this study compares the classification performance and inference efficiency of TextCNN, BiLSTM+Attn, and ERNIE 3.0-Nano under identical experimental settings. Third, it analyzes how each model fits different real-world deployment conditions. The aim is to fill a clear gap—fine-grained sentiment analysis for short Chinese e-commerce texts has been underexplored in the lightweight model literature—and to give practitioners concrete numbers to guide model choices.

2. Related work

2.1. Aspect-based sentiment analysis

Sentiment analysis research has progressively advanced from document-level to sentence-level and subsequently to aspect-level granularity. The ABSA evaluation benchmark established by SemEval 2014 [1] standardized progress in this area and catalyzed numerous attention-based and pre-trained language model approaches. BERT-based models [3] subsequently pushed performance further through auxiliary sentence construction strategies [7]. However, the continued scaling of these models has raised the practical barrier for deployment, particularly in scenarios requiring real-time processing of large-scale review streams where inference latency is a binding constraint.

2.2. Lightweight model research

Exploration of lightweight approaches has proceeded along two primary lines. The first involves models designed from the ground up for efficiency: TextCNN, introduced by Kim [4], extracts local character n-gram features using multi-scale convolutional kernels; Wang et al. [5] incorporated additive attention into BiLSTM to improve sequential modeling of aspect-sentiment relationships. The second line employs knowledge distillation to compress large pre-trained models: TinyBERT [6] and ERNIE 3.0-Nano [8] exemplify this approach, targeting English and Chinese, respectively. Despite the practical relevance of both lines, direct horizontal comparisons across these architecture families on Chinese e-commerce short reviews remain sparse.

2.3. Linguistic characteristics of Chinese short reviews

Processing challenges in Chinese short reviews extend beyond their brevity. Chinese lacks natural word boundaries, meaning tokenization errors propagate downstream and affect model inputs. Reviews frequently contain internet abbreviations and platform-specific vocabulary that exceeds the coverage of general sentiment lexicons. Most critically, Chinese users tend to express sentiment through factual descriptions rather than direct evaluative language [9, 10]—a pattern that poses

DOI: 10.54254/2755-2721/2026.GF35112

considerable difficulty for keyword-matching approaches. These distinctive characteristics motivate the present study's focus on Chinese e-commerce short reviews as a specialized and underexplored domain in the ABSA literature.

3. Methodology

3.1. Dataset construction and annotation

The corpus originates from a publicly available Chinese consumer review dataset spanning ten product categories, including books, tablets, smartphones, apparel, and food. The original data provides only coarse-grained positive and negative labels. After removing empty and duplicate entries, 62,770 valid samples remain.

To construct fine-grained labels, four sentiment dimensions are defined—product quality, logistics, customer service, and value for money—with positive, neutral, and negative polarity labels assigned independently for each, yielding a "4×3=12 category" aspect-sentiment schema. Reviews that do not explicitly address a particular dimension are categorized as "no mention" and excluded from the learning objective for that dimension during training.

Annotation follows a four-step rule-based pipeline. First, keyword dictionaries are compiled per dimension; terms such as "shipping", "courier" and "delivery" map to logistics, while "workmanship", "material" and "durability" map to product quality. Second, HowNet [11] and the BosonNLP sentiment lexicon assign baseline polarity scores to sentiment words within each aspect term context window. Third, contextual rules for negation words and degree adverbs handle polarity reversal and intensity modulation. Fourth, conflict-resolution rules address reviews containing contradictory signals across multiple dimensions. Manual verification on a random sample of 500 reviews yields an annotation accuracy of 87.2%, consistent with benchmarks reported in comparable studies [12].

After expanding multi-aspect reviews into independent training instances, the total dataset comprises 86,551 samples, partitioned in a 7:1:2 ratio into training (60,585), validation (8,655), and test (17,311) sets. Given the substantial class imbalance—the quality dimension accounts for approximately 43% of all samples—inverse class-frequency weighting is applied to the loss function during training.

3.2. Model design

All three models adopt individual characters as the basic input unit, bypassing Chinese word-segmentation errors and more appropriately handling the ambiguous word boundaries common in short review text.

TextCNN [4] follows the classic architecture of Kim, employing convolutional kernel sizes of 2, 3 and 4 to capture local character n-gram features across varying spans. After max pooling and concatenation, dropout and a fully connected classification layer produce the output. The model contains approximately 0.69 million parameters with minimal inference overhead.

BiLSTM+Attn [5] appends an additive attention layer to the bidirectional LSTM output, enabling adaptive token weighting according to the current classification objective. Compared with dependence on the final hidden state alone, this mechanism provides greater advantages when processing contrastive constructions. The model contains approximately 1.2 million parameters; gradient clipping (maximum norm 1.0) is applied during training to stabilize convergence.

DOI: 10.54254/2755-2721/2026.GF35112

ERNIE 3.0-Nano [8] is the ultra-lightweight variant of Baidu's ERNIE 3.0 series, containing approximately 17.92 million parameters and pre-trained on Chinese knowledge-enhanced corpora. During preliminary experiments, the English-trained TinyBERT was evaluated as an alternative but yielded a Macro-F1 of only 0.22 on Chinese classification—near-random performance—and was consequently replaced by ERNIE 3.0-Nano as the lightweight Transformer representative. Fine-tuning uses AdamW with a 10% linear warm-up followed by linear decay, training over 8 epochs at a batch size of 64.

3.3. Experimental setup

The primary evaluation metrics are accuracy and macro-average F1 (Macro-F1). Macro-F1 is preferred over weighted F1 because the latter is dominated by majority classes under imbalanced conditions; with the largest class exceeding the smallest by more than a factor of 12, Macro-F1 assigns equal weight to each category and more objectively reflects performance on minority classes. Inference latency—averaged across 100 batches of 100 samples, measured in milliseconds per sample—and total parameter count are additionally recorded for efficiency comparison. All experiments are conducted on an NVIDIA RTX 4060 Laptop GPU (8 GB VRAM).

4. Results and analysis

4.1. Overall performance

Table 1 presents the overall test-set results for the three models. ERNIE 3.0-Nano achieves the highest accuracy (0.6363) and Macro-F1 (0.5798), confirming that Chinese pre-trained representations deliver substantial gains even when compressed to approximately 18 million parameters. BiLSTM+Attn outperforms TextCNN on both metrics, consistent with expectations—sequential context modeling captures negation constructions more effectively than local convolutional features.

Table 1. Overall performance comparison of three lightweight models

Model	Accuracy	Macro-F1	Parameters (M)	Latency (ms/sample)
TextCNN	0.5570	0.4741	0.69	0.049
BiLSTM+Attn	0.6124	0.5516	1.20	0.004
ERNIE 3.0-Nano	0.6363	0.5798	17.92	0.340

4.2. Per-dimension analysis

Table 2 presents average F1 scores per sentiment dimension—computed as the mean F1 across three polarity classes for each dimension. The product quality dimension yields the highest scores across all three models, attributable to its status as the largest sample class (approximately 43% of the total) and to the relative directness of quality-related expressions, which are more amenable to lexicon-based annotation. Logistics and value-for-money rank in the intermediate range, while customer service consistently records the lowest scores across all architectures.

The persistently low scores on the customer service dimension merit closer examination. Negative sentiment in customer service descriptions is frequently conveyed through narrative statements, for example, "asked three times and received no response", rather than explicit negative lexical items. This implicit sentiment pattern falls outside the coverage of rule-based lexicons,

DOI: 10.54254/2755-2721/2026.GF35112

introducing substantial annotation noise that subsequently impairs model learning. This bottleneck is shared across all three model types, indicating that the limitation originates at the data annotation level rather than in any architecture-specific weakness.

Table 2. Per-dimension average F1 score comparison

Sentiment Dimension	TextCNN	BiLSTM+Attn	ERNIE 3.0-Nano
Product Quality	0.5933	0.6800	0.7067
Logistics	0.4700	0.5333	0.5633
Customer Service	0.4000	0.4733	0.5100
Value for Money	0.4367	0.5200	0.5367

4.3. Efficiency analysis

From an efficiency perspective, BiLSTM+Attn records the lowest inference latency under GPU batch processing at 0.004 ms per sample, followed by TextCNN at 0.049 ms per sample, with ERNIE 3.0-Nano at 0.340 ms per sample. Surprisingly, BiLSTM+Attn ran faster on GPU than TextCNN—0.004 ms versus 0.049 ms per sample. This is likely because LSTM matrix operations parallelize well on GPU hardware, and our BiLSTM has only 1.2 million parameters. However, this ranking could reverse on CPU, where TextCNN's simpler structure might be faster.

For practical deployment, ERNIE 3.0-Nano is the appropriate choice when accuracy is the primary criterion—such as in brand sentiment monitoring or after-sales alerting systems—given its Macro-F1 advantage of approximately 10.6% over TextCNN. BiLSTM+Attn achieves a favorable balance between accuracy and efficiency for GPU-equipped environments. Where hardware resources are severely constrained, such as mobile or edge deployment, TextCNN's 0.69 million parameter footprint provides a clear advantage.

5. Conclusion

This study investigates fine-grained sentiment analysis for short Chinese e-commerce reviews through three main contributions: the design of a "4-dimension × 3-polarity" 12-category annotation framework, the construction of a domain-specific corpus via rule-based automated annotation applied to 62,770 samples, and a systematic comparative evaluation of TextCNN, BiLSTM+Attn, and ERNIE 3.0-Nano under unified experimental conditions on the same hardware platform.

From these results, three main conclusions are drawn. (1) ERNIE 3.0-Nano is the most accurate. Even at roughly 18 million parameters, its Macro-F1 is about 10.6 percentage points higher than TextCNN's. This shows that pre-training on Chinese knowledge-enhanced corpora still pays off for aspect-level classification. (2) BiLSTM+Attn is the fastest on GPU. It runs at 0.004 ms per sample with only 1.2 million parameters, offering a good trade-off between accuracy and efficiency. (3) TextCNN is the smallest. Its 0.69 million parameters make it a natural choice when memory or compute is tightly limited, for example, on mobile devices or embedded edge hardware.

The experiments also reveal several limitations that warrant attention in future research. The customer service dimension consistently yields the lowest F1 scores across all three models; the core cause is the prevalence of implicit sentiment expressions that rule-based lexicons cannot capture effectively. This limitation manifests at both the annotation stage and in subsequent model learning, representing a shared bottleneck rather than an architecture-specific deficit. The approximately 12.8% annotation noise introduced by rule-based labeling affects absolute metric values; however,

DOI: 10.54254/2755-2721/2026.GF35112

the relative ranking across the three models remains stable, lending reliability to the overall conclusions drawn from this comparative study.

Several directions for future research merit consideration. The dataset in this study originates from a single publicly available source, and cross-platform generalization remains unverified; incorporating reviews from multiple e-commerce platforms in future work would assess transfer capability and domain robustness. The rule-based annotation pipeline also has limited capacity for handling complex negation patterns and irony; integration of large language models as annotation assistants represents a promising avenue for improving label quality at scale. Additionally, extending this framework to cross-platform transfer learning or multimodal settings that incorporate both textual reviews and product images could substantially expand the applicability of ABSA methods in real-world Chinese e-commerce contexts.

References

[1] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I. (2014) SemEval-2014 Task 4: Aspect based sentiment analysis. Proceedings of SemEval 2014, 27–35.
[2] Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., Si, L. (2020) Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. Proceedings of AAAI 2020, 8600–8607.
[3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.
[4] Kim, Y. (2014) Convolutional neural networks for sentence classification. Proceedings of EMNLP 2014, 1746–1751.
[5] Wang, Y., Huang, M., Zhao, L., Zhu, X. (2016) Attention-based LSTM for aspect-level sentiment classification. Proceedings of EMNLP 2016, 606–615.
[6] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q. (2020) TinyBERT: Distilling BERT for natural language understanding. Proceedings of EMNLP 2020, 4163–4174.
[7] Sun, C., Huang, L., Qiu, X. (2019) Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. Proceedings of NAACL-HLT 2019, 380–385.
[8] Wang, S., Sun, Y., Li, X., Feng, S., Tian, H., Wu, H., Wang, H. (2021) ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv: 2107.02137.
[9] Zhang, C., Li, Q., Song, D. (2021) Sentiment analysis of Chinese product reviews via a novel hybrid deep learning model. IEEE Access, 9: 23613–23625.
[10] Tang, D., Qin, B., Liu, T. (2016) Aspect level sentiment classification with deep memory network. Proceedings of EMNLP 2016, 214–224.
[11] Dong, Z., Dong, Q. (2010) HowNet and the Computation of Meaning. World Scientific Publishing, Singapore.
[12] Xu, H., Liu, B., Shu, L., Yu, P.S. (2019) BERT post-training for review reading comprehension and aspect-based sentiment analysis. Proceedings of NAACL-HLT 2019, 2324–2335.