Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL33468

Exploration into the Optimization of a Lightweight Sentiment Perception and Hierarchical Response System for Small and Medium-sized E-commerce Platforms

Yan Li

International (European and American) Institute of Technology, Henan University, Zhengzhou, China

2410240637@henu.edu.cn

Abstract. Addressing the pain points of high computational costs and significant latency associated with deep learning models on small and medium-sized e-commerce platforms, this study proposes a lightweight sentiment perception and hierarchical response system based on Snow NLP optimization. By refactoring the inference logic to reduce instantiation overhead, the system constructs a multi-level response engine to enable automated interventions. Experimental results indicate that, while maintaining an accuracy of 82.2%, the system's operational efficiency improves by 34.33% compared to the baseline, achieving a response speed 24.5 times faster than BERT. This research demonstrates that lightweight models can expand business depth even under extremely low computing power, offering small and medium-sized enterprises an intelligent customer service solution that balances efficiency with real-time response capabilities. Future work will focus on integrating continuous learning mechanisms to seamlessly adapt to evolving e-commerce terminologies and exploring multi-lingual support. Additionally, expanding the system to handle multi-modal inputs, such as customer emojis and voice snippets, will further enhance interactive experiences while strictly preserving the model's lightweight architecture and low computational footprint.

Keywords: Snow NLP, Small and medium-sized e-commerce platforms, Operational efficiency optimization, Hierarchical response system, Efficiency balance

1. Introduction

In light of the significant interconnection between the digital economy and internet-based technologies, e-commerce and local service platforms create huge volumes of short text user reviews. Such texts contain the experiences of customers with regards to consumption and their central demands that can be used as important data sources by companies to mine demands and improve services [1,2]. Sentiment analysis technology is an essential branch of natural language processing, and it can help determine the emotional inclinations of users automatically, which can make a significant contribution to the process of making decisions within a business, and it has become one of the most popular areas of research.

DOI: 10.54254/2755-2721/2026.GL33468

Nevertheless, when dealing with large batches of instant reviews, small and middle-sized e-commerce companies are frequently confronted by both the challenge of restricted computing resources and that of intense real-time responsiveness. According to recent studies and applications, despite the fact that deep learning models such as BERT have a very high level of accuracy, they are also very expensive to deploy, complex to design and have a very high inference time which makes it hard to adjust them to lightweight settings where there are limited resource [2]. On the other hand, lightweight Chinese NLP solutions have numerous benefits when it comes to small-scale short-text analysis situations because of the ease of implementation and low resource usage.

Snow NLP is among other lightweight tools that are an example of a library built using Python particularly to work with short texts in Chinese. Sentiment analysis module is trained with machine learning; once text is preprocessed, it provides sentiment scores between 0 and 1 (with closer to 1 being more positive sentiment) and allows custom thresholds to classify emotion labels [3]. It has basic features and can be quickly deployed locally but, on implementation to real-world business use, it has two severe shortcomings. Firstly, it is functionally singular - it simply produces sentiment scores and labels and does not provide practical application-level guidance that could help with post-purchase communications and emotional calming. Secondly, its inference mechanism is repetitive. In its unoptimized form, it does not perform concurrent processing, so that processing large batches of reviews results in numerous object instances, which significantly increases execution time.

The current studies have been primarily aimed at confirming the situation-oriented use of Snow NLP, rather than carrying out meaningful code-level fixes that would aim at its functional and efficiency-related flaws. These pain points can be addressed by rule-based text generation technology as well as model lightweighting technologies which are practical solutions [3]. The rule-based text generation, which uses pre-defined rule libraries and keyword matching, is capable of fast generation rates with controlled logic, which makes it highly applicable in intelligent customer service situations. At the same time, the optimization of lightweight NLP models frequently involves techniques such as sharing the outputs of computations and eliminating redundant invocations. Reorganization of the code workflow alone without changing the fundamental algorithms can make great contributions to the efficiency of the operations.

Considering this, this paper takes native Snow NLP as the basis and modifies the inference process by programming in order to make it more efficient in operations, and integrates a multi-level strategic response engine. It builds on a better system which is not only efficient but also deep in business, which offers small and medium enterprises a very cost-efficient and "efficient-balanced" working solution to the problem of intelligent review processing of large volumes.

2. Model architecture optimization and improvement design

2.1. Dataset and baseline model settings

The test data used in this work was obtained through the Chinese e-commerce service review corpus offered by the open-source Github project Py-SnowNLP. As stated in the official documentation of the project, the source of data used to perform the sentiment analysis module pre-training mostly involves real-life feedback of buyers and sellers during e-commerce purchase situations, which is exactly what this research aims to focus on, i.e., small and medium-sized e-commerce sites. In order to test the model, this paper sampled 500 semantically full and feature-diverse short texts in this corpus to construct a test sample (TestModel.csv) with review texts and manually labeled emotions (1 indicating positive, 0 indicating negative) [4]. The basic model invokes the native Snow NLP to

DOI: 10.54254/2755-2721/2026.GL33468

process data and the sentiment classification threshold has been defined as 0.6. Experimental analysis reveals that the baseline model has 82.2% sentiment analysis accuracy on the given dataset.

2.2. Pipeline fusion and refactoring of the computing link

This study optimized inference link by doing code refactoring to resolve an issue with the native model where its computation steps were fragmented which caused repeated data traversal and reading of large text processing. The pipeline pattern separates sentiment calculation, classification judgment, and strategy generation, integrating them into a "single-pass, parallel settlement". During one iteration, the sentiment score that was extracted is saved in local memory and simultaneously employed in multi-level response generation as well as emotion label classification. It eliminates multiple I/O operations and considerably decreases the runtime cost of the Python interpreter.

2.3. Multi-level sentiment perception and hierarchical response engine

In order to make up for the absence of application output in the baseline model, this paper abandons straightforward binary classifications, and builds a multi-level response mechanism that is strongly connected to business situations, allowing automatic hierarchical actions depending on sentiment scores and keywords.

In case of extremely positive sentiments with scores of 0.85 and above, the system detects high-loyalty customers and activates high-value maintenance procedures to include motivating repurchasing; In case of positive sentiments that fall within the range of 0.6 and 0.85, the engine automatically produces traditional expressions of gratitude in order to maximize the interaction experience [4,5]; In case where the score is below or equal to 0.25, the system considers it to be an extreme negative warning with high level of public relations risk and flags it automatically so that advanced human support can be used; Moreover, in cases of texts with low scores and containing certain negative keywords such as quality or logistics, the system does a second semantic match to produce highly targeted core demand compensation and after-sales pacification recommendations [6,7].

3. Experiments and result analysis

3.1. Experimental design

The setup of the experimental environment consists of Python 3.13 (CPU Intel Core Ultra 7 155H, 32GB RAM) and core dependencies are pandas 2.1.0 and snow NLP 0.12.3. To show the benefit of the lightweight model in the form of an efficiency balance, the experiment explicitly adds a popular deep learning large model (BERT) as the high-precision inference standard. The industrial utility of the enhanced scheme is determined through comparison of the accuracy (Acc), overall processing time and average time taken per single item of the three groups of models on the same 500 models [8,9]. To guarantee that the data obtained is objective, operational efficiency is the arithmetic mean of 10 repeated measurements.

3.2. Performance comparison analysis

In order to evaluate the performance of the improved system intuitively, this paper has extensively reported on the operational performance of the three model groups in the same hardware environment, and comparison data are also represented in Table 1.

DOI: 10.54254/2755-2721/2026.GL33468

Table 1. Comprehensive comparison of operational efficiency

Model Metric	Accuracy (Acc)	Total Processing Time (500 items)	Average Single-item Response Speed	Applicable Scenarios
BERT (Benchmark)	91.00%	60.0000 s	0.1200 s	High computing power / Non-real-time needs
Native Snow NLP	82.20%	3.7300 s	0.0075 s	Offline general analys
Improved Lightweight System	82.20%	2.4500 s	0.0049 s	Small & medium e-commerce / Real-time response

3.3. Experimental results discussion

According to experimental findings, on the assumption of preserving 82.2% sentiment analysis accuracy, the total processing time of the improved system decreased by 34.33 percent to 2.45 seconds compared to the baseline of 3.73 seconds. This is a very good evidence of the efficacy of workflow refactoring in removing unnecessary computation and unlocking the computational capabilities of light tools. At the same time, in terms of efficiency balance, despite the fact that the BERT model has the highest possible accuracy advantage, the enhanced system in the current study had a rate of response time, which was 24.5 times higher. The millisecond-level real-time response ability and multi-level business intervention functionality, which is acquired at the cost of reduced accuracy by some 8.8 percent, are extremely cost-effective to server hardware and system maintenance requirements to resource-constrained small and medium-sized e-commerce platforms. Such an exact understanding of where the efficiency balance lies enables the lightweight system to satisfy the high-concurrency processing requirements of particular business contexts with greater effectiveness [10].

4. Discussion

4.1. Core advantages of the improved model

The present research departs the linear competition of NLP algorithm accuracy and aims at finding the Achilles heel of small and medium enterprises to obtain a very high commercial cost-performance. The code implementation is short and efficient; with just refactoring the logic and implementing hierarchical response engine, it jumps to become a simple analytical tool to an intelligent business system. The multi-stage early warning system allows companies to intervene in the negative public sentiment as soon as possible which has significant practical value.

4.2. Research limitations and future directions

As of now, the given responses of the present study are also largely based on fundamental keyword matching. The probabilistic model architecture behind Snow NLP restricts the capability of recognizing complex semantics, including irony and metaphor. The next stage of future research will be an attempt to incorporate a jieba word segmentation library along with custom domain dictionaries to refine the accuracy of feature extraction to enhance the quality of capturing e-commerce "jargon" and saracasm. Another plan is to add lightweight knowledge graphs to the strategy generation engine, which will increase the personalization and fit of suggestion generation, by using domain common-sense modeling, and thus, change the generation process to knowledge-

DOI: 10.54254/2755-2721/2026.GL33468

driven generation. In respect to architectural performance, multi-threading programming or distributed stream processing systems will be further presented to investigate the efficiency boundary of the system in ultra-large-scale real-time analytics, and adjust to transient traffic peaks when big promotion event takes place. Also, it will be extended to the combination of lightweight image recognition models to develop a multi-dimensional sentiment perception engine to overcome the growing problem of mixed text-and-image reviews in social e-commerce settings.

5. Conclusion

To resolve the intrinsic drawbacks of single functionality and relatively poor efficiency of native Snow NLP, this paper performed in-depth optimization and improvement studies on the system level that were specific to small and medium-sized e-commerce. The refactoring of the calculation link into a form of a "single-pass, parallel settlement" pipeline pattern minimized the redundant overhead of object instantiation and greatly boosted the concurrency support ability of the system. At the same time, the system integrated multi-level strategic response engine to gain full business closed-loop by providing precise sentiment perception and automatic post-sale intervention, which efficiently overcame the bottleneck of the traditional lightweight models without application outputs. The analysis demonstrates that, as compared with the baseline, the improved system has raised the operational efficiency by 34.33 percent at the cost of 82.2 percent accuracy, and the response rate has been multiplied by 24.5 compared with the high accuracy large model (BERT). The scheme of "efficiency balance" in this paper can offer a low-cost and very timely sentiment perception solution to SMEs with limited computing resources, which have great engineering practice values. Even though this system has achieved considerable engineering efficiency improvements, it is still struggling with processing sophisticated semantic depths and scaling to ultra-massive datasets, which will become the main focus of deepening the engineering application research of lightweight NLP models in the future.

References

[1] Yan, J., & Wei, J. (2026). Research on User Experience of Knowledge Service Platforms Based on Snow NLP Sentiment Analysis and BERTopic Model—A Case Study of the Dedao App. Bianji Xuekan, (01), 41-51.

[2] Yan, Z. (2023). Design and Application of Sentiment Analysis System Based on BERT and K-Means (Master's thesis, Southwest University). Master's. https: //doi.org/10.27684/d.cnki.gxndx.2023.000352.

[3] isnowfy. (2019). snownlp: Python library for processing Chinese text [Computer software]. GitHub. https: //github.com/isnowfy/snownlp

[4] Tang, X. (2026). Design and Implementation of Student Emotion Recognition Algorithm Based on NLP. Computer Programming Skills & Maintenance, (02), 119-121. https: //doi.org/10.16184/j.cnki.comprg.2026.02.017.

[5] Wang, X., & Qi, F. (2026). Deconstruction and Evaluation of Chinese Sentiment Analysis Capabilities of Large Language Models. Huawen Jiaoxue yu Yanjiu, (01), 27-35. https: //doi.org/10.16131/j.cnki.cn44-1669/g4.2026.01.015.

[6] Emexidis, C., Gkonis, P., & Liapakis, A. (2025). Analyzing Employee Job Satisfaction Through Sentiment Analysis for Enhanced Workplace Improvement and Business Success. Theoretical and Applied Ergonomics, 1(2), 10. https: //doi.org/10.3390/tae1020010

[7] Zhang, Y., & Zhai, N. (2023). Sentiment Analysis, Making Machines Understand You Better—Instructional Design of "Exploring Artificial Intelligence". China Information Technology Education, (19), 23-26.

[8] Tian, X. (2024). Predicting Test Case Error Detection Capability for Sentiment Analysis Tasks (Master's thesis, Southeast University). Master's. https: //doi.org/10.27014/d.cnki.gdnau.2024.004677.

[9] Yu, Y., Jia, S., & Gao, H. (2023). Research on the Application of Deep Learning in Sentiment Analysis. Modern Computer, 29(17), 45-48.

[10] Lorenzo MAMELONA. (2025). Enhancing Sentiment Analysis Using Hybrid Transformer Models and Adaptive Active Learning Strategies (Doctoral dissertation, Nanjing University of Information Science and Technology).

DOI: 10.54254/2755-2721/2026.GL33468

Doctoral. https: //doi.org/10.27248/d.cnki.gnjqc.2025.000078.