Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29486

Comparative Study and Analysis of the Performance of Locally Fine-Tuned DistilBERT and Mainstream Large-Scale Language Models in Text Sentiment Classification Tasks

Xueli Shi

School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China
2023090911011@std.uestc.edu.cn

Abstract. This research compares two commonly used Natural Language Processing (NLP) techniques, enabling each of them to identify the sentiment contained in text and categorize it into one of three categories. One is the DistilBERT model trained on local machines, which is built on the Transformers architecture. The other includes three currently mainstream large language model in China, which are Zhipu GLM-4, Doubao-seed-1.6-Flash and DeepSeek-V3.1. The data used in this research is sourced from a public dataset available on Kaggle. The four models were placed on the same set of experimental corpora, with accuracy values, weighted F1-scores and inference time collected respectively, and the errors they would make were observed. The results show that the lightweight model DistilBERT achieves the highest accuracy and the fastest inference speed. However, it relies heavily on keywords and struggles to handle complex contexts. GLM-4 demonstrates deeper context understanding, but it makes misjudgments in cases of emotional conflicts and implicit expression. The Doubao-seed-1.6-Flash and DeepSeek-V3.1 models exhibit a conservative tendency, which means that when faced with ambiguous information, the two models tend to lean toward neutral judgments. This tendency comes at the cost of recall. Therefore, the selection of models needs to be closely aligned with the specific requirements of application scenarios for speed, cost, precision, and recall, so as to achieve optimal performance.

Keywords: Sentiment Three-classification, DistilBERT, Mainstream Large Language Models, Model Performance Evaluation, Zero-shot Learning

1. Introduction

In recent years, with the continuous iteration of natural language processing technology, sentiment analysis—one of the core tasks within this field—has achieved breakthrough progress in both academic research and industrial application domains. Particularly driven by the technologies of pre-trained language models and large language models, the application boundaries and performance upper limits of sentiment analysis have been continuously expanded. From the perspective of technical application scenarios, sentiment classification tasks have gradually extended from early

© 2025 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.TJ29486

high-resource language scenarios to low-resource language and multilingual mixed scenarios. In some low-resource language environments where labeled data is scarce, research on sentiment classification is still in its initial stage. However, relying on the transfer learning capability of cross-lingual pre-trained models, relatively reliable classification performance can be achieved, breaking the technical barriers of sentiment analysis for low-resource languages [1]. Meanwhile, in scenarios involving multilingual mixed and symbol-mixed texts that are abundant on social media platforms, models based on the Transformer architecture—by virtue of their ability to accurately capture semantic correlations—exhibit classification performance far superior to that of traditional machine learning methods, effectively addressing the challenge of difficult sentiment recognition for unstandardized texts [2].

Large language models demonstrate strong generalization capabilities in text classification tasks through zero-shot learning and chain-of-thought prompting, making them particularly suitable for scenarios with limited labeled resources or requirements for rapid deployment [3,4]. The research indicates that appropriate property strategies can significantly improve the performance of models in sentiment recognition tasks [5]. Such models do not require task-specific training and can be directly applied to various natural language understanding tasks including sentiment classification, thereby providing a feasible solution for researchers and enterprises with limited resources.

However, despite the advantages of large language models in zero-shot scenarios, their performance in specific sentiment classification tasks still often lags behind that of medium and small-scale models that have undergone targeted fine-tuning [1]. These lightweight models not only perform better in metrics such as accuracy and F1-score but also feature lower inference latency and higher stability, making them particularly suitable for application scenarios with strict requirements on response time and computational cost [2].

Against this backdrop, this research aims to systematically compare the performance of fine-tuned models (represented by DistilBERT) and zero-shot large language models (three models, namely GLM-4, Doubao-seed-1.6-Flash, and DeepSeek-V3.1, are used in this research) in sentiment classification tasks, with evaluations conducted from dimensions including accuracy, F1-score and inference latency. The research is based on three types of sentiment datasets, through quantitative and qualitative analysis from an academic perspective, this research can fill the gap in current horizontal comparison studies between the two technical paths, provide supplements to the model performance evaluation system in the field of sentiment analysis, and especially offer a reference experimental paradigm for subsequent related studies and in the quantitative analysis of multi-dimensional evaluation indicators. From an application perspective, the research results can directly provide clear model selection basis for natural language processing practitioners. Whether it is small and medium-sized enterprises with limited resources that need rapid implementation, or large enterprises pursuing high-precision large-scale deployment, all can find suitable technical solutions according to their own business scenarios, effectively reducing model trial-and-error costs and improving the development efficiency and implementation effect of sentiment analysis systems. In addition, the model comparison framework and evaluation logic proposed in this research can also provide references for model selection in other natural language processing tasks, and have strong industry reference value.

DOI: 10.54254/2755-2721/2026.TJ29486

2. Methods

2.1. Dataset

This research adopts the available datasets named sentiment analysis dataset from Kaggle, which includes three sentiment orientations--positive negative and neutral. After data cleaning was conducted locally, the dataset was divided into training, validation and test sets. DistilBERT was fine-tuned on the training and validation sets, while all final metrics were calculated on the test set to ensure the consistency of evaluation.

2.2. The locally fine-tuned DistilBERT

This research refers to the twitter-emotion-classification project on Kaggle and implements the fine-tuning of DistilBERT by calling the Hugging Face Transformers library. The base weights used in the project are distilbert-base-uncased. During the training process on the data set, 3 training epochs are set, with a learning rate of 2e-5 and a batch size of 64. In the evaluation phase, the fine-tuned model is used to make predictions on each sample in the test set of the dataset.

2.3. The three large language models

In this research, three large language models—GLM-4, Doubao-seed-1.6-Flash, and DeepSeek-V3.1--are invoked via API for zero-short sentiment classification [5]. The system prompt remains unchanged, the large language models are instructed to act as experts in the field of sentiment analysis and required to classify the sentiment implied in the text into one of the three sentiment categories, with the answers strictly limited to a single word (i.e., the sentiment label). Texts in the test set are sent one by one, and the results returned by the models are received accordingly.

2.4. Evaluation metrics

Table 1. Model evaluation metrics

Accuracy	The proportion of correct predictions
Weighted F1-Score	A weighted comprehensive score that includes metrics such as accuracy and recall
Classification Report	The precision, recall and weighted F1-score for each category
Confusion Matrix	The confusion among the three categories
Average Inference Latency	The average processing time per text (ms/text)
Latency Standard Deviation	The standard deviation of processing time fluctuations (ms)

Table 1 clearly presents the key metrics used to evaluate model performance, including accuracy and weighted F1-score which reflect the overall prediction effect, classification-related metrics such as classification reports and confusion matrices, as well as average inference latency and latency standard deviation which reflect processing efficiency.

DOI: 10.54254/2755-2721/2026.TJ29486

3. Results

3.1. Overall performance comparison

Table 2. Overall performance comparison of the four models

Metric	DistilBERT	GLM-4	Doubao-seed-1.6-Flash	DeepSeek-V3.1
Accuracy	0.7735	0.6393	0.6920	0.6900
Weighted F1-Score	0.7736	0.6038	0.6790	0.6835

As shown in the data in Table 2, in terms of Accuracy, DistilBERT reaches 0.7735, which is significantly higher than GLM-4 (0.6393), Doubao-seed-1.6-Flash (0.6920), and DeepSeek-V3.1 (0.6900). Regarding the Weighted F1-Score, DistilBERT achieves 0.7736, which is also higher than GLM-4 (0.6038), Doubao-seed-1.6-Flash (0.6790), and DeepSeek-V3.1 (0.6835). Based on these two key metrics, DistilBERT outperforms the other three models—GLM-4, Doubao-seed-1.6-Flash, and DeepSeek-V3.1—in terms of overall performance in this research.

3.2. Efficiency comparison

Table 3. Efficiency comparison

Metric	DistilBERT	GLM-4	Doubao-seed-1.6-Flash	DeepSeel-V3.1
Average Inference Latency	6.17 ms/sample	356.61 ms/sample	2231.56 ms/sample	597.41 ms/sample
Latency Standard Deviation	±0.59 ms	±167.76 ms	±1518.09 ms	±350.96 ms

Table 3 compares the efficiency of the four models. In terms of average inference latency, DistilBERT is only 6.17 ms/text, which is much lower than GLM-4 (356.61 ms/ text), Doubao-seed-1.6-Flash (2231.56 ms/ text), and DeepSeek-V3.1 (597.41 ms/ text). Regarding latency standard deviation, DistilBERT has a value of ±0.59 ms. However, due to the fact that the other three models are invoked via API—subject to round-trip network delays and other factors—their standard deviations are at least nearly 300 times higher [6].

3.3. Classification report and confusion matrix

Table 4. DistilBERT classification report

	Precision	Recall	F1-Score	Support
Negative	0.79	0.77	0.78	566
Neutral	0.73	0.74	0.74	809
Positive	0.81	0.82	0.82	626
Accuracy	/	/	0.77	2000
Macro Avg	0.78	0.78	0.78	2000
Weighted Avg	0.77	0.77	0.77	2000

DOI: 10.54254/2755-2721/2026.TJ29486

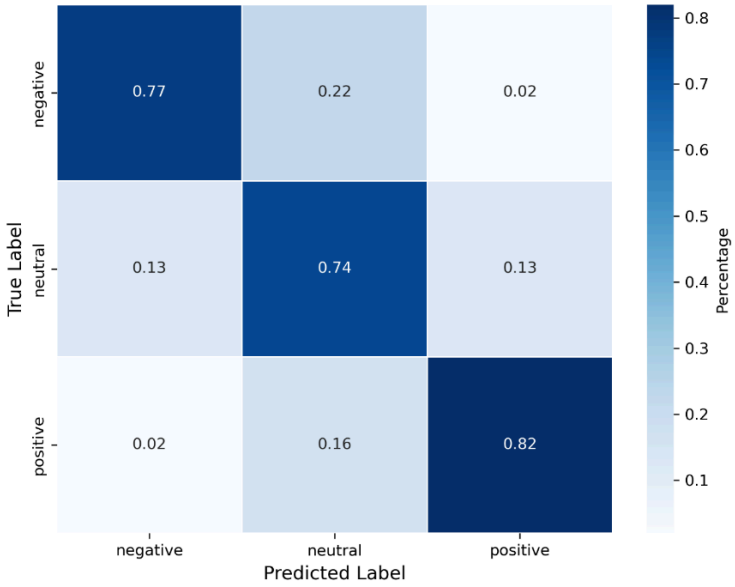

Figure 1. DistilBERT confusion matrix

As shown in Table 4 and Figure 1, the overall accuracy of the DistilBERT model is 0.77. Among the three sentiment categories, it achieves an F1-score of 0.82 for the positive category, demonstrating strong performance; the F1-score for the negative category is 0.78, and for the neutral category, it is 0.74. The confusion matrix indicates that the model correctly predicts 77% of samples with the true label of negative, 74% of samples with the true label of neutral, and 82% of samples with the true label of positive. Although there are cases of misclassification between different categories, the model's overall performance is relatively high and stable.

Table 5. GLM-4 classification report

	Precision	Recall	F1-Score	Support
Negative	0.61	0.89	0.73	566
Neutral	0.94	0.95	0.94	806
Positive	0.59	0.85	0.70	624
Accuracy	/	/	0.64	1996
Macro Avg	0.70	0.68	0.62	1996
Weighted Avg	0.72	0.64	0.60	1996

DOI: 10.54254/2755-2721/2026.TJ29486

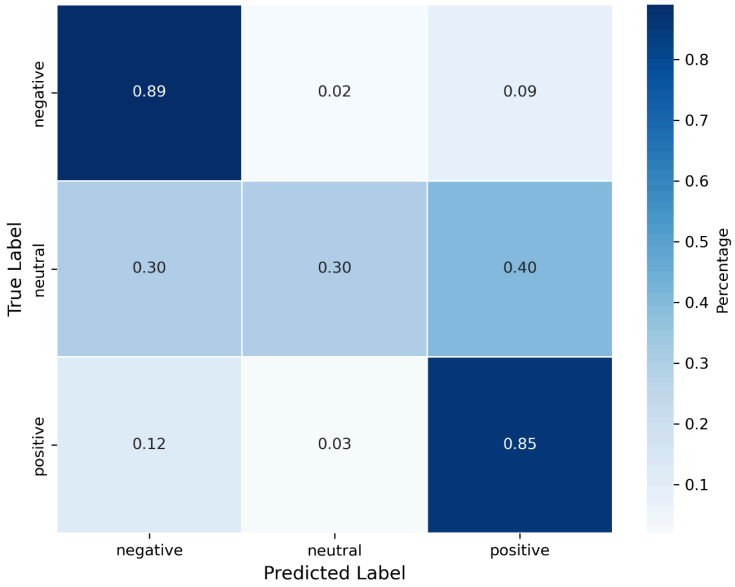

Figure 2. GLM-4 confusion matrix

As shown in Table 5 and Figure 2, the GLM-4 model has an overall accuracy of 0.64, indicating subpar performance. For the Neutral category, its precision, recall, and F1-score are all around 0.94, showing excellent classification results. For the Negative category, the recall rate is relatively high at 0.89, but the precision is low at 0.61, which means there are many cases where samples from the other two categories are misclassified as Negative. For the Positive category, the recall rate of 0.85 is acceptable, but the precision is also low at 0.59, which means that the misclassification issue is prominent.

Table 6. Doubao-seed-1.6-Flash classification report

	Precision	Recall	F1-Score	Support
Negative	0.65	0.85	0.74	566
Neutral	0.82	0.46	0.59	809
Positive	0.66	0.84	0.74	625
Accuracy	/	/	0.69	2000
Macro Avg	0.71	0.72	0.69	2000
Weighted Avg	0.72	0.69	0.68	2000

DOI: 10.54254/2755-2721/2026.TJ29486

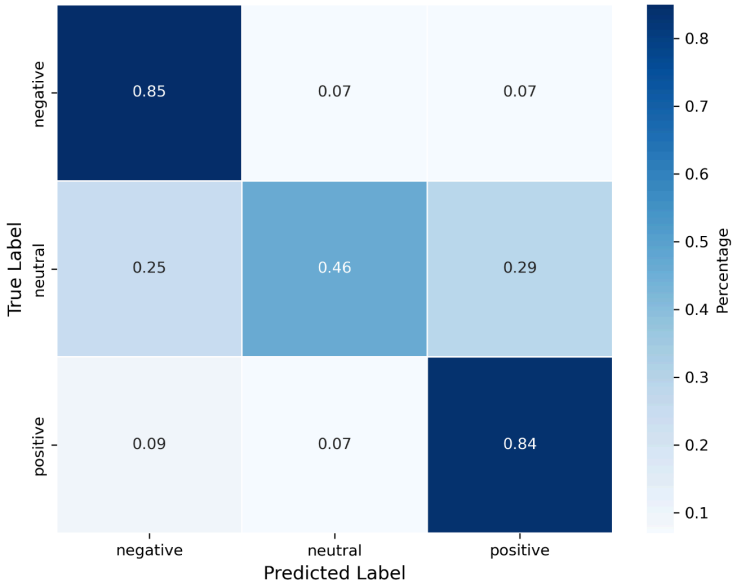

Figure 3. Doubao-seed-1.6-Flash confusion matrix

As shown in Table 6 and Figure 3, the Doubao-seed-1.6-Flash model has an overall accuracy of 0.69, placing it at a moderate level. For the Negative category, the precision is 0.65, the recall is 0.85, and the F1-score reaches 0.74; for the Positive category, the precision is 0.66, the recall is 0.84, and the F1-score is also 0.74—both categories demonstrate acceptable classification performance. However, the model performs poorly in classifying the Neutral category, although its precision is relatively high at 0.82, the recall is only 0.46, and the F1-score is merely 0.59. Additionally, it can be seen from the confusion matrix that a considerable number of samples with the true label of Neutral are misclassified as Negative or Positive.

Table 7. DeepSeek-V3.1 classification report

	Precision	Recall	F1-Score	Support
Negative	0.69	0.79	0.74	566
Neutral	0.76	0.52	0.62	809
Positive	0.64	0.82	0.72	625
Accuracy	/	/	0.69	2000
Macro Avg	0.70	0.71	0.69	2000
Weighted Avg	0.70	0.69	0.68	2000

DOI: 10.54254/2755-2721/2026.TJ29486

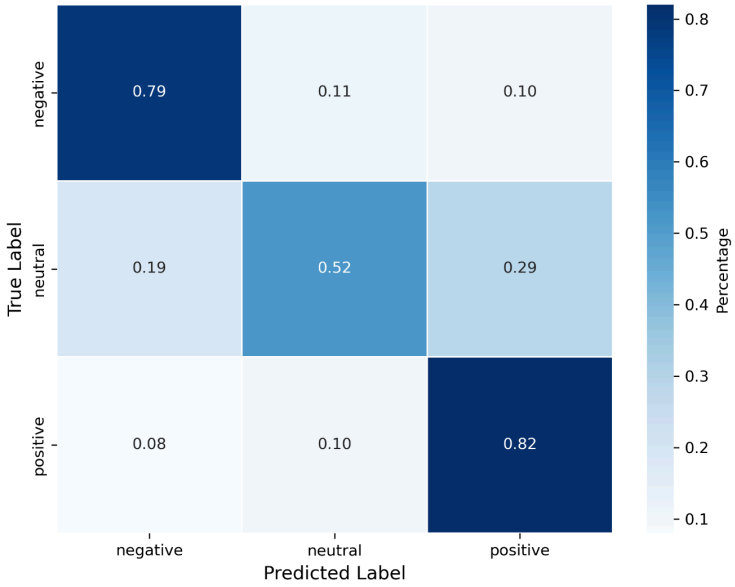

Figure 4. DeepSeek-V3.1 confusion matrix

As shown in Table 7 and Figure 4, the DeepSeek-V3.1 model has an overall accuracy of 0.69. For the Negative category, it achieves a relatively high recall rate of 0.79 and an F1-score of 0.74; for the Positive category, the recall rate is 0.82 and the F1-score is 0.72—both categories exhibit good classification performance. However, the model also performs poorly in classifying the Neutral category: its recall rate is only 0.52 and the F1-score is 0.62. Although this performance is slightly better than that of the previous two large language models, the recall rate is still only half of the total Neutral samples, which remains unsatisfactory.

3.4. Typical misclassification analysis

3.4.1. Negative but positive or neutral

The DistilBERT model and the DeepSeek-V3.1 model have the highest misclassification rates for the Negative category, at 24% and 21% respectively.

Among these, the DistilBERT model has limited ability to understand sarcasm [7], complex semantics and ambiguous emotions. It tends to assign strong weight to words that express emotions, making it prone to be misled by superficial positive words in negative contexts. For example, consider the following text, "no one told me the mutual admiration society was meeting this morning! LOL Hi boys!" (Label: negative; Prediction: positive). This is a typical sarcastic scenario. The speaker is complaining about being excluded, which carries obvious dissatisfaction. Moreover, the phrase "mutual admiration society" itself is sarcastic, referring to a small group of people who excessively praise and flatter each other without objectivity. The model assigned high weight to seemingly positive signals such as "admiration," "LOL," and "Hi boys!"—ultimately leading to misclassification.

A notable characteristic of the DeepSeek-V3.1 model is that when processing mixed texts containing both factual information and emotional content, it sometimes prioritizes the factual part, thereby ignoring the emotion expression.

For example, consider the text, "my monitor won’t turn on. This can only be the work of my mother" (Label: negative; Prediction: neutral). The first half of the text—stating that the monitor

DOI: 10.54254/2755-2721/2026.TJ29486

won’t turn on--undoubtedly carries negative emotions such as frustration and annoyance while the second half reflects helpless blame. However, the model interpreted the entire sentence as a statement of an objective problem and its possible cause, overlooking the implicit emotions like frustration and complaint within it. This ultimately led to the model given a neutral judgment.

3.4.2. Neutral but positive or negative

The three large language models all show unsatisfactory accuracy in judging the Neutral category. Among them GLM-4 has the lowest accuracy at 30%, while Doubao-seed-1.6-Flash and DeepSeek-V3.1 perform slightly better at 46% and 52% respectively.

GLM-4 exhibits a tendency to overinterpret, often assigning emotional connotations to neutral texts such as objective statements.

For example, take the following text, "meant to be doing hw but not going to happen!!! Ill pray for tia! Rose is in the shower so sh.." (Label: neutral; Prediction: negative). This text similar to a life update, contains three neutral events including failing to do homework, praying for Tia and Rose being in the shower. However the model may have detected resistance toward homework from the phrase "not going to happen!!!" or inferred that something bad might have happened from "pray for tia!". When faced with such fragmented information, GLM-4 tends to make negative associations.

The Doubao-seed-1.6-Flash and DeepSeek-V3.1 models exhibit similar characteristics to the DistilBERT when analyzing the Neutral category—specifically, they tend to let keywords dominate their judgments. Take the following text as an example, "hahaha this is super late, but thank you!!" (Label: neutral; Prediction: positive). Doubao-seed-1.6-Flash was drawn to the strongly positive words "hahaha" and "thank you!!" but inappropriately ignored the contextual detail "super late". This indicates that the model struggles to integrate emotional information from different parts of a sentence. Another example is the text, "half of me wants to go to sleep, half of me doesn't. shoot. my other half is dying of sleepiness" (Label: neutral; Prediction: negative). This text describes a conflicting state where one part of the speaker doesn’t want to sleep and the other is extremely sleepy, with no clear emotional tendency. However, DeepSeek-V3.1 was most likely influenced by the phrase "dying of sleepiness". While "dying" here is used as an exaggeration, its strong negative connotation caused the model to overlook the contradictory and balanced state presented earlier in the text.

3.4.3. Positive but neutral or negative

All four models achieve a recall rate of over 80% for the Positive sentiment, which is quite remarkable. However, regarding the misclassified cases, there is still room for analysis.

When faced with emotional conflicts or complex expressions, GLM-4 always overly focuses on the negative parts. For example, consider the following text, "Khan oh please, you don’t have to do that to me. Don't bother" (Label: positive; Prediction: negative). This sentence is clearly a polite refusal in a social context, expressing the speaker’s reluctance to trouble the other person—its emotional core should be gratitude and embarrassment. However, the imperative sentences "you don’t have to" and "Don't bother" carry negative and refusal connotations. This indicates that the model is not very adept at handling such implicit expressions in social etiquette.

As for the Doubao-seed-1.6-Flash model, when processing texts related to wishes or hopes, it tends to make a safer neutral judgment. Take the following text as an example, "Good stuff, Smiles back? Why won’t you go to his concert? :o. Wish I could instead of working on my music which is failing lol x" (Label: positive; Prediction: neutral). This text contains extremely complex and

DOI: 10.54254/2755-2721/2026.TJ29486

conflicting emotional signals, "Good stuff" and "Smiles" are positive; "Why won't you go" carries a tone of doubt and a hint of urging; "Wish I could" expresses a positive desire; while "my music which is failing" conveys negative self-mockery, and the final "lol x" attempts to lighten the mood. Faced with such chaotic information, the model refrained from judging the dominant emotion and opted for the safest choice—neutral. This reflects the model's limitation in handling texts with information overload and highly mixed emotions.

4. Discussion

Sarcasm and complex contexts pose challenges for all models [8]. Whether it is the lightweight DistilBERT or the large-scale models (GLM-4, Doubao-seed-1.6-Flash, and DeepSeek-V3.1), they all tend to make mistakes when processing contexts that require in-depth reasoning, such as sarcasm, humor, and metaphors. Instead, they are more adept at understanding literal meanings [9,10].

The ability of these models to handle mixed emotions needs improvement. When a text contains both positive and negative signals simultaneously, it is difficult for the models to accurately identify which one constitutes the dominant emotion. They either get biased by the signal word with the strongest intensity, or opt for a conservative judgment of neutral when the intensities of conflicting signals are comparable.

These models are overly sensitive to negative words and words with negative connotations. There is a common tendency among them, as long as a sentence contains negative words (e.g., no, not, don't) or words with negative meanings (e.g., bored, suck, fail), the models tend to skew the overall sentiment of the entire sentence toward negative or neutral—even if the sentence conveys a positive meaning on the whole.

When a sentence contains both an objective factual statement and a subjective emotional expression, the model sometimes focuses more on the objective facts of the former, thereby ignoring the emotional tone defined by the latter—ultimately leading to misclassification as neutral.

5. Conclusion

This research analyzes the performance of four models. The results show that the accuracy gap between the models is relatively small, DistilBERT leads with an accuracy rate of 77%, and there is still room for improvement due to issues related to the adaptability of its code and dataset; the accuracy rates of the other three large language models fall within the range of 63% to 69%. In terms of inference efficiency, DistilBERT has an average latency of only 6.17 ms per sample, while the latency of the other three models is at least 50 times that of DistilBERT. In particular, Doubao-seed-1.6-Flash has a latency exceeding 2000 ms per sample, making it very time-consuming to use.

From the perspective of characteristics and scenario adaptation, DistilBERT, as a high-speed keyword catcher, is suitable for large-scale, low-risk, and high-real-time scenarios, such as preliminary screening of e-commerce reviews and public opinion monitoring. GLM-4 excels in context understanding and is adapted to high-precision scenarios that can tolerate latency, such as brand analysis and feedback interpretation. Doubao-seed-1.6-Flash and DeepSeek-V3.1, due to their cautious and neutral characteristics, are suitable for scenarios where false positives need to be avoided, such as auxiliary content moderation.

In addition, a cascaded system combining DistilBERT with other large models can balance efficiency, cost, and performance, making it an effective solution for large-scale sentiment analysis.

DOI: 10.54254/2755-2721/2026.TJ29486

This research provides a clear basis for model selection in industrial sentiment analysis tasks. It can help enterprises select models accurately according to their business requirements—such as real-time performance, accuracy, and false positive control—thereby reducing trial-and-error costs. Meanwhile, the cascaded system solution offers an efficient and feasible path for large-scale data processing, which is conducive to improving the implementation efficiency and economic benefits of sentiment analysis tasks.

In the future, further optimizations can be made in the following aspects. Enhance DistilBERT's ability to understand deep context, narrowing its shortcomings in recognizing sarcasm and implicit emotions. Address GLM-4's overinterpretation issue by training it with more data from diverse cultural scenarios, thereby improving its ability to judge complex social intentions. For Doubao-seed-1.6-Flash and DeepSeek-V3.1, while retaining their advantage of low false positives, optimize the emotional signal capture algorithm to increase their sensitivity to samples with non-direct emotions and reduce sample omissions caused by low recall rates. Additionally, explore a dynamic multi-model collaboration mechanism, adjusting model combinations in real time based on data characteristics to further improve the comprehensive performance of sentiment analysis.

References

[1] Das, A., Sharif, O., Hoque, M. M., et al. (2021). Emotion classification in a resource constrained language using transformer-based approach. arXiv preprint arXiv: 2104.08613.
[2] Li, C., Wang, J., Zhang, Y., et al. (2023). Large language models understand and can be enhanced by emotional stimuli. arXiv preprint arXiv: 2307.11760.
[3] Wang, Z., Pang, Y., Lin, Y. (2023). Large language models are zero-shot text classifiers. arXiv preprint arXiv: 2312.01044.
[4] Li, Z., Zhu, H., Lu, Z., et al. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. arXiv preprint arXiv: 2310.07849.
[5] Almalki, S. S. (2025). Sentiment Analysis and Emotion Detection Using Transformer Models in Multilingual Social Media Data. International Journal of Advanced Computer Science & Applications, 16(3).
[6] Chen, L., Zaharia, M., Zou, J. (2023). Frugalgpt: How to use large language models while reducing cost and improving performance. arXiv preprint arXiv: 2305.05176.
[7] Zhu, L., Zhu, Z., Zhang, C., et al. (2023). Multimodal sentiment analysis based on fusion methods: A survey. Information Fusion, 95: 306-325.
[8] Wankhade, M., Rao, A. C. S., Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7): 5731-5780.
[9] Schaeffer, R., Miranda, B., Koyejo, S. (2023). Are emergent abilities of large language models a mirage?. Advances in neural information processing systems, 36: 55565-55581.
[10] Bender, E. M., Gebru, T., McMillan-Major, A., et al. (2021). On the dangers of stochastic parrots: Can language models be too big?//Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 610-623.