Comparative Analysis of Language Models for Sentiment Classification

Shangjiafeng Guo

Xiamen University Malaysia, Kuala Lumpur, 43900, Malaysia
Email: CST2309143@xmu.edu.my
Manuscript received June 11, 2026; accepted July 11, 2026; published July 30, 2026

***Abstract*—By comparing and analysing the performance of several machine learning algorithms on fine-grained sentiment classification problems to examine their suitability and shortcomings for use as models in sentiment analysis. Sentiment analysis remains a prominent research area in Natural Language Processing (NLP). However, systematic comparisons of whether these methods demonstrate superiority in fine-grained classification settings have yet to be conducted. Based on the GoEmotions dataset that includes 28 sentiment classes as our experiment's reference for evaluating four kinds of models: logistic regression, BiLSTM, BERT, and the large-scale language model Claude (claude-haiku-4-5). The experiments employ TF-IDF feature extraction, sequence encoding, pre-training with fine-tuning, and zero-shot and few-shot prompting strategies. Based on this experiment, the model's generalisation capabilities improve as architectural depth increases. The accuracy of logistic regression was only 55.11%; BERT obtained the highest F1 score and performed the best overall, and bidirectional pre-trained representations were most valuable among other methods. However, large language models perform significantly worse on the 28-class classification task in zero-shot settings, suggesting that they are better suited for generative and open-ended emotional interaction than for standardized classification benchmarks.**

***Keywords*—sentiment analysis, fine-grained emotion classification, BERT, BiLSTM, GoEmotions dataset**

I. Introduction

Many applications based on various fields utilize the functions of natural language processing (NLP), such as patients' records monitoring, users' feedback handling systems human-machine interactive Ssystem. Many methods have been used in this study for the introduction of classical statistics to high-end technologies such as deep learning and large language models; however, most of these comparisons primarily examine generalised sentiment classification. Currently, comprehensive comparisons among various models in fine-grained, multi-class settings are lacking. Furthermore, standardized evaluations have yet to determine whether LLMs outperform smaller, specialized models.

Based on the deficiency of existing studies, this paper uses GoEmotions to build a fine-graded sentiment classification model by combining four types of models: Logistic regression, Bi-directional LSTMs (Bidirectional LSTM), BERT, and Claude. The evaluation encompasses a range of approaches, including TF-IDF feature extraction, sequence modeling, fine-tuning of pre-trained models, and both zero-shot and few-shot inference. The above-mentioned results serve as a reference for model selection in sentiment-sensitive NLP applications, help define the appropriate application scope of large language models, and provide a theoretical foundation for further research on emotions, emotion recognition, and emotion-driven intelligent assistive technologies.

II. Comparative Analysis of Language Models

According to LaValley, logistic regression solves binary classification problems by analysing the probability of outcome occurrence based on dichotomous variables, and can be extended to continuous, categorical, or multivariate predictors [1]. Ramadhan *et al.* applied multinomial logistic regression to social media sentiment classification, achieving 74% accuracy with TF-IDF and K-Fold cross-validation, demonstrating its viability as an NLP baseline [2]. However, logistic regression has several limitations: the relationship between predictors and the outcome may be misspecified; the model lacks a closed-form solution and thus requires iterative optimization; and poor model fit can produce results lacking practical significance [1].

A. Neural Network Models

Recurrent Neural Networks (RNNs) enabled the modeling of variable-length text and advanced sentiment analysis through distributed representations (Wang *et al.*, 2016), yet they often struggle to capture subtle sentiment shifts [3]. Hochreiter and Schmidhuber (1997) showed that ordinary RNNs fail to acquire long-term memory because errors diminish across iterations, with BPTT and RTRL suffering exponential amplification or attenuation during backward propagation [4]. These limitations motivated LSTM, which maintains accurate error signals over longer periods and retains initial information through forget, input, and output gates alongside a memory cell (Hochreiter & Schmidhuber, 1997; Wang *et al.*, 2016) [3, 4].

LSTMs have shown solid baseline performance: Wang *et al.* (2016) reported 82% accuracy on three-class restaurant reviews, while Mousa and Schuller (2017) achieved 89.58% on IMDB, also finding bidirectional LSTMs outperformed unidirectional ones [3-5]. Nevertheless, ordinary LSTMs capture only coarse-grained sentiment, as they lack a mechanism for selecting task-relevant keywords (Wang *et al.*, 2016) [3].

B. Transformer-based Models

Unlike sequential LSTMs, the Transformer (Vaswani *et al.*, 2017) relies solely on attention, enabling parallel processing [6]. Self-attention reduces the path length between any two tokens to a constant, facilitating the capture of long-range dependencies, whereas multi-head attention learns representations across multiple subspaces simultaneously (Vaswani *et al.*, 2017) [6]. Attention weights

carry semantic meaning, assigning higher weight to emotionally charged words (Wu *et al.*, 2020) [7].

Building on this, BERT uses bidirectional pre-training and a [CLS] token for classification (Devlin *et al.*, 2019) [8]. BERT_BASE achieved 93.5% and BERT_LARGE 94.9% accuracy on SST-2 (Devlin *et al.*, 2019) [8]. RoBERTa further improved performance via more data and longer training (Liu *et al.*, 2019) [9], with both BERT and RoBERTa reaching F1 scores around 0.93 (El Azzouzy *et al.*, 2025) [10]. However, Transformer-based models are subject to several limitations, including unstable fine-tuning on small datasets, high computational costs (Devlin *et al.*, 2019) [8], sensitivity to hyperparameters, and limited interpretability of attention mechanisms (Liu *et al.*, 2019; Wu *et al.*, 2020) [7-9].

C. LLMs

Large language models pre-trained on massive datasets exhibit emergent capabilities, including in-context learning (Zhao *et al.*, 2023) [11]. GPT-3, with 175 billion parameters, performs diverse tasks without fine-tuning (Brown *et al.*, 2020) [12]. LLMs excel in few-shot learning, approaching fine-tuned performance with minimal examples (Brown *et al.*, 2020) [12], and operate effectively in zero-shot settings with appropriate instructions (Kojima *et al.*, 2022) [13].

Zero-shot LLMs approach fully-supervised small-model performance on basic classification, though they lag on aspect-based sentiment analysis; they outperform small models in low-resource settings and can generate emotional dialogue (Zhang *et al.*, 2024) [14]. Limitations include imitative rather than genuine empathy; repetitive or overly long responses, which raise mental health concerns (Sorin *et al.*, 2024) [15]; sensitivity to prompt formulation and instability on emotion-related tasks (Zhang *et al.*, 2024) [14]; and reduced coherence when processing longer texts (Brown *et al.*, 2020) [12].

III. Research Design

A. Dataset Description

This experiment utilized the GoEmotions dataset, which was released by Google Research in 2020. The data come from real English comments posted by Reddit users and include 58,009 samples. Each sample belongs to one of the following types: positive, negative, uncertain or ambiguous (unclear) or neutral. There are unevenly divided categories in the data. Neutral is the most frequent emotion; Grief and Pride have relatively few occurrences among all expressions. This is similar to the distribution of emotions in real social media data.

B. Data Preprocessing

Basic preprocessing was first carried out on the original data. URL links, HTML tags, and extra whitespace were removed during text cleaning. Original capitalisation and punctuation were kept because some models are sensitive to these features. Texts shorter than 10 characters were then excluded, leaving 44,640 valid entries.

Regarding label processing, the original annotations were provided in list format; the first label of each entry was selected as the final category, yielding 28 emotion categories. The dataset was then split into a training set and a test set in an 80:20 ratio, with 35,703 entries in the training set and 8,926 in the test set. The stratify parameter was used to keep the distribution of emotion categories similar across the two sets.

The above steps formed the standardised processing procedures of this dataset. Specific pre-processing of the models individually is carried out furtherly.

IV. Experiment and Analysis

A. Statistical Model Experiments

Logistic regression is a statistical classification model. In this experiment, after basic preprocessing of the dataset, TF-IDF was used for feature extraction. The maximum number of training iterations was set to 1000. The experimental results are as follows Table 1.

Table 1. Experimental results

Metric	Score
Accuracy	0.5511
Precision	0.5680
Recall	0.5511
F1-score	0.4867

In this experimental dataset with 28 emotion classification categories, the logistic regression model achieved an accuracy of 55.11%, which is reasonable as a baseline model. The macro F1 score was lower than accuracy, primarily due to the uneven distribution of emotion categories in the dataset and the model's low recall for infrequent emotions such as grief and pride.

Furthermore, TF-IDF processes each iteration separately, causing the model to fail to capture contextual information in the text data, which has significant limitations in scenarios involving hidden expressions. These limitations motivate the introduction of more sophisticated models in subsequent sections.

B. Neural Network Models Experiments

LSTM is a neural network model capable of capturing text sequence information. In this experiment, the pre-processed text data is further segmented and sequence-encoded to build a vocabulary of 20,000 words for the model, with the maximum sequence length set to 100. A BiLSTM architecture was employed with the following hyperparameters: embedding dimension of 128, hidden layer dimension of 256, a dropout rate of 0.3 for regularization, and training for 5 epochs with a batch size of 64 using the Adam optimizer. The experiments were run on an NVIDIA GeForce RTX 4070 GPU as shown in Table 2.

Table 2. Experimental results

Metric	Score
Accuracy	0.5544
Precision	0.5224
Recall	0.5544
F1-score	0.5206

Compared to logistic regression, LSTM shows an improvement in F1-score, indicating that sequence modeling helps capture contextual dependencies in textual sentiment expression. However, BiLSTM's accuracy remains comparable to that of logistic regression, indicating that sequence modeling alone may not adequately capture

complex emotional expressions—particularly those involving subtle or long-range dependencies.

C. Transformer-based Models Experiments

A pre-trained model of BERT, which utilises the Transformer framework. The selected fine-tune model in this paper is bert-base-uncased. The text data were encoded using the BERT tokeniser and a maximum sequence length of 128 was used. A linear classification layer was added on top of the pre-trained model to map the final hidden states to the sentiment categories in the dataset. The model was trained for 3 epochs with a batch size of 32 using the AdamW optimizer as shown in Table 3.

Table 3. Experimental results

Metric	Score
Accuracy	0.6052
Precision	0.5941
Recall	0.6052
F1-score	0.5956

BERT outperforms logistic regression and LSTM on all metrics. This improvement stems from BERT's bidirectional self-attention mechanism, which takes into account both the left and right contexts of words concurrently, thereby more effectively capturing emotional expression.

Compared to LSTM, BERT improves the F1-score by 0.075, indicating that the language knowledge gained from pre-training has some value for sentiment analysis tasks. However, BERT has high computational cost and training time is significantly longer than statistical models and LSTM, which limits its effectiveness in situations with limited resources.

D. large Language Models Experiments

This experiment uses Claude (claude-haiku-4-5) as a large-scale language model, called via the Anthropic API, to test both zero-shot and few-shot inference methods. 200 samples were randomly selected from the test set for the experiment.

In the zero-shot setting, the model was provided solely with a list of 28 emotion categories and instructed to classify the input text accordingly, without any example demonstrations.

In the few-shot setting, four additional labeled examples (covering love, anger, excitement, and remorse) are provided in the prompt to guide the model in understanding the task format.

Due to API output parsing failures (zero-shot: 3.0%, few-shot: 1.5%), these samples were excluded from evaluation.

LLMs performed poorly compared with the other three methods on a 28 category emotion classification problem as shown in Table 4.

Table 4. Experimental results

Metric	Zero-shot	Few-shot
Accuracy	0.2835	0.2843
Precision	0.4536	0.4651
Recall	0.2835	0.2843
F1-score	0.2987	0.3152

The few-shot setting was slightly better than the zero-shot setting, with an F1 score of 0.3152 compared with 0.2987. This indicates that providing the model with a small number of examples offers a modest benefit, though the gains remain limited.

Another reason might be that there may not be distinct boundaries among the twenty-eight emotions in GoEmotions. Without task-specific fine-tuning, the model cannot precisely differentiate among these categories. Also, there might be differences in how the model interprets the label names compared to what they are defined as by humans when labelling manually.

LLMs have a stronger ability to produce text than standardised classification tasks and more empathy. They will be more conducive to the generation of emotion-based dialogues and providing various emotional types than other kinds.

V. Comparative Analysis

A. Performance Comparison

Overall, BERT performed best among all fine-tuning models, while LLM performed the worst in unsupervised settings as shown in Table 5.

Table 5. Experimental results

Model	Accuracy	Precision	Recall	F1-score
Logistic regression	0.5511	0.5680	0.5511	0.4867
LSTM	0.5544	0.5224	0.5544	0.5206
BERT	0.6052	0.5941	0.6052	0.5956
LLM Zero-shot	0.2835	0.4536	0.2835	0.2987
LLM Few-shot	0.2843	0.4651	0.2843	0.3152

B. Error Analysis

Most models exhibited limited performance in detecting low-frequency emotions, such as sadness, anger, and shame. The most significant cause is a disparity between classes in the dataset, and there are not enough training samples for them. Emotions with similar meanings, such as anger and annoyance or joy and excitement, were also often confused. This kind of mistake occurred among all the systems.

C. Feature Analysis

Different models have shown different results because they have extracted features differently. Logistic Regression primarily employs the TF-IDF word frequency features. Therefore, it cannot derive much contextual information from them. LSTM performs well on sequences, learning the dependence between them to some extent. Because BERT is more inclined to understand the context from both sides when performing sentiment analysis. Large language models excel at open-ended natural language understanding; however, when applied to standardized classification tasks with strict label definitions, they perform poorly.

D. Summary

According to the experimental results, most of these model's accuracy improves with the increased complexity in structure. Among these models, the pre-trained model of BERT achieved a better result among them. The precise value got was 0.6052. In standardized classification tasks, the large language model achieved an accuracy of approximately 0.2835 in the zero-shot setting and roughly 0.2843 in the few-shot setting. Their performance on the task of low-resources is still weak.

VI. Conclusion

This study systematically compare the precise sentiment classification effects of four different model systems in this paper: logistic regression, Bidirectional LSTM, BERT, and Claude. From the experimental data shows that with an increase in scale of the structure, this model has enhanced its generalisation performance. Claude performs significantly worse than the fine-tuned models under both zero-shot and few-shot conditions. This suggests that large language models may be less suited for standardized classification tasks, whereas their greater potential lies in open-ended emotional communication.

This study still has several limitations. First, the sample size for evaluating large language models was relatively small. Second, only text data were used, which means the study focused only on a unimodal setting. Third, class imbalance issues and effects on prediction results for low-frequency sentiment classes in a model. Future research could improve upon this work by fine-tuning large language models on domain-specific features, integrating multiple modalities to enhance recognition accuracy, and exploring the development of practical AI-driven sentiment support systems that incorporate clinical expertise.

Conflict of Interest

The author declares no conflict of interest.

References

[1] M. P. LaValley, "Logistic regression," Circulation, vol. 117, no. 18, pp. 2395–2399, 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.

[2] W. P. Ramadhan, A. Novianty, and C. Setianingsih, "Sentiment analysis using multinomial logistic regression," in 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), 2017, doi: 10.1109/ICCEREC.2017.8226700.

[3] Y. Wang, M. Huang, L. Zhao, and X. Zhu, "Attention-based LSTM for aspect-level sentiment classification," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 606–615.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] A. E.-D. Mousa and B. Schuller, "Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 1023–1032.

[6] A. Vaswani *et al.*, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998–6008.

[7] Z. Wu, T.-S. Nguyen, and D. C. Ong, "Structured self-attention weights encode semantics in sentiment analysis," arXiv preprint arXiv:2010.04922, 2020. [Online]. Available: https://arxiv.org/abs/2010.04922

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423. [Online]. Available: https://arxiv.org/abs/1810.04805

[9] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[10] O. El Azzouzy, T. Chanyour, and S. Jai Andaloussi, "Transformer-based models for sentiment analysis of YouTube video comments," Scientific African, vol. 29, e02836, 2025, doi: 10.1016/j.sciaf.2025.e02836.

[11] W. X. Zhao *et al.*, "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023. [Online]. Available: https://arxiv.org/abs/2303.18223

[12] T. B. Brown *et al.*, "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[13] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in Advances in Neural Information Processing Systems (NeurIPS), 2022.

[14] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," in Findings of the Association for Computational Linguistics: NAACL 2024, 2024, pp. 3881–3906.

[15] V. Sorin *et al.*, "Large language models and empathy: Systematic review," Journal of Medical Internet Research, vol. 26, e52597, 2024, doi: 10.2196/52597.