Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29617

Research and Analysis of Deep Learning Models for Emotion Analysis Tasks

Mingcheng Yang

School of Computer Science and Software, Southwest Petroleum University, Chengdu, China
202331061227@stu.swpu.edu.cn

Abstract. The field of Natural Language Processing (NLP) currently has a bright future, but it still faces challenges due to a series of issues such as language complexity, data resources and limitations. Therefore, this paper starts with the classic sentiment analysis problem. Based on the IMDB movie review dataset, by evaluating the original dataset, the semantic contradiction set obtained by filtering the original data, and the easily confused dataset obtained by training the large prediction model with prompt words, this paper systematically estimates the basic performance of a series of models including CNN, LSTM, BiLSTM, GRU, MLP, Attention, Multi-HeadAttention, Transformer, and PositionalEmbedding+Transformer. With this in-depth study, the basic generalization ability, anti-interference ability, and fine-grained semantic understanding ability of the model are studied. Comparison of the original test set and the semantically contradictory set shows that each model has excellent basic generalization capabilities. The GRU demonstrates the strongest interference resistance in the semantically contradictory set, while the LSTM demonstrates the best fine-grained semantic understanding in the easily confused set. Combining the scores of indicators across the test sets, the Attention model demonstrates the most comprehensive overall performance. This research reflects the potential for further development of RNN and its variants and suggests the possibility and potential of models such as the RNN+Attention model.

Keywords: Sentiment Analysis, IMDB Dataset, Deep Learning, Attention Mechanism

1. Introduction

Currently, the field of artificial intelligence is in the era of weak artificial intelligence and is developing in depth. This is mainly reflected in the breakthroughs in Natural Language Processing (NLP) technology, which is gradually approaching human levels in tasks such as translation, question answering, and sentiment analysis [1]. However, there are still some technical difficulties and pain points in the development of NLP that need to be solved, such as the need to improve the self-generated cognitive ability of large models, insufficient data supplementation, and the need to improve the evaluation system [2, 3].

Given this, re-examining classic tasks and datasets is crucial, as they can serve as a benchmark for technological advancement. Different models have strengths and limitations for different language problems. Sentiment analysis of IMDB movie reviews is a classic NLP project. This

© 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.TJ29617

dataset can be used to test whether a model can grasp certain aspects of stylized language, such as correctly handling irony and metaphors or accurately capturing connections between certain contexts.

This article explores and analyzes the ability of several deep learning models to identify semantically challenging examples based on a sentiment analysis project for IMDB movie reviews. The following sections describe the model selection and dataset evaluation methods used in this research.

In terms of model selection, the first thing to consider is that the current NLP field focuses on the development of models with attention mechanism as the core [4], so a series of model architectures such as Attention, MultiHeadAttention, Transformer, PositionalEmbedding+Transformer are selected [5, 6]; secondly, considering that the time series model has excellent understanding ability in sentence context, the variant models of RNN model LSTM, BiLSTM and GRU are selected [7]; considering that MLP is a basic model of deep learning and a pioneer in text sentiment analysis projects and has strong nonlinear fitting ability and CNN has the ability to capture local features [8], finally these two models are included in the selection.

In terms of the dataset measurement method, the study prepared the original test set, the semantically contradictory set screened from the original test set [9], and the easily confused set generated by the large prediction model [10]. It is hoped that the basic generalization ability, anti-interference ability, and fine-grained semantic understanding ability of each model can be measured and compared through the performance of each model on these three datasets.

2. Related work

Sentiment analysis is an important branch of NLP. Its main content is to identify and judge the emotional tendency in text data. It has gone through the early stages of completing tasks based on sentiment dictionaries and related matching rules [11], traditional machine learning such as TF-IDF training, and the stage of deep learning methods that are still being expanded [8]. At present, the research on text sentiment analysis using deep learning methods has become mature, and various language models such as LSTM and Transformer have performed well in this task.

Currently, many classic metrics are used to evaluate classification tasks. Each metric has different focuses, reflecting the capabilities of a model from various perspectives. Research is also moving beyond traditional metrics. However, in practice, some aspects require careful consideration, such as robustness to adversarial forces, semantic consistency, and good generalization.

3. Research methods

3.1. Dataset and preprocessing

The dataset used in this article is the IMDB movie review dataset, a classic sentiment analysis benchmark in the field of NLP. It was compiled and published in 2011 by Stanford University researchers Andrew Mass and others. This dataset contains 50,000 movie review texts, with 25,000 each for training and testing. The dataset uses binary sentiment labels (positive and negative), with a 50% positive and 50% negative sentiment each. All reviews are genuine reviews from IMDB users. These reviews are long, rich in content, and have complex semantic structures, making them highly valuable for research.

For the preprocessing step, the first step is to clean the text of the original dataset to remove special characters, numbers (such as punctuation marks, line breaks, etc.), and HTML tags; then

DOI: 10.54254/2755-2721/2026.TJ29617

perform label encoding conversion to convert the original text labels (negative/positive) into numerical labels (0/1); then perform text vectorization and serialization; finally, encode all labels.

3.2. Model selection and training

In terms of model selection, a variety of model architectures have been constructed, which are representative in various sub-fields of deep learning, including the basic deep learning models CNN and MLP, variant models of recurrent neural networks (RNN) LSTM, BiLSTM, GRU, and a series of models with attention mechanism as the core, such as Attention, MultiHeadAttention, Transformer, PositionalEmbedding+Transformer.

After selecting the model, define a list of model architecture names, train each model in a loop, evaluate each successfully trained model on the original test set, and save the evaluation results, model, and tokenizer.

3.3. Evaluation strategy

Accuracy is used as the primary evaluation metric. When accuracy rates are close, the F1 score is used to further differentiate performance. The following sections explain the meaning of the formula elements, the metric formula, and the reasons for choosing accuracy and F1 score as the basis for judging model performance.

Table 1 shows the various indicators and the corresponding formulas, where the parameters of the formulas are as follows.

TP (True Positive): Predicts a correct positive outcome.

FP (False Positive): Predicts an incorrect positive outcome (false positive).

FN (False Negative): Predicts an incorrect negative outcome (false negative).

TN (True Negative): Predicts a correct negative outcome.

Table 1. Indicators and their formulas

indicators	formulas
Accuracy	(TP+TN)/(TP+TN+FP+FN)
Precision	TP/(TP+FP)
Recall	TP/(TP+FN)
F1-Score	2*(Precision*Recall) /(Precision+Recall)

As can be seen from this, accuracy reflects the model's overall predictive ability and effectively provides an overview of performance when the data distribution is relatively uniform. The F1 score, as the harmonic mean of precision and recall, represents the model's overall performance in both precision and recall, reflecting its overall performance in scenarios with uneven data distribution. Therefore, this experiment selected these two metrics as the basis for measuring model performance.

After obtaining the evaluation results of the original test set, the model performance is further evaluated using the semantic contradiction set screened out from the original test set and the easily confused set generated by the large language model to analyze the model's basic generalization ability, anti-interference ability, and fine-grained semantic understanding ability.

DOI: 10.54254/2755-2721/2026.TJ29617

4. Research methods

Table 2 shows the evaluation results of each model on the original test set. As can be seen from the table, CNNs demonstrate strong performance, achieving the highest accuracy. The Transformer model also performs well by leveraging its own attention mechanism. The Transformer+PositionEmbedding, which incorporates position information, also has the ability to understand the structure of text sequences, performing slightly better than the Transformer. The performance of traditional recurrent neural networks (LSTMs), GRUs, and BiLSTMs is similar, with BiLSTM offering slightly better performance. MLP performance is relatively low, but has reached a generally acceptable level.

Table 2. Model evaluation on the original test set

Model	Accuracy	Precision	Recall	F1_score
LSTM	0.87280	0.87449	0.87280	0.87265
GRU	0.87032	0.87124	0.87032	0.87023
BiLSTM	0.87012	0.87479	0.87012	0.86970
CNN	0.88252	0.88253	0.88252	0.88252
MLP	0.86952	0.87106	0.86952	0.86939
Attention	0.88136	0.88138	0.88136	0.88136
MultiHeadAttention	0.87400	0.87534	0.87400	0.87390
Transformer	0.87708	0.87710	0.87708	0.87708
PositionalEmbedding+Transformer	0.88096	0.88097	0.88096	0.88096

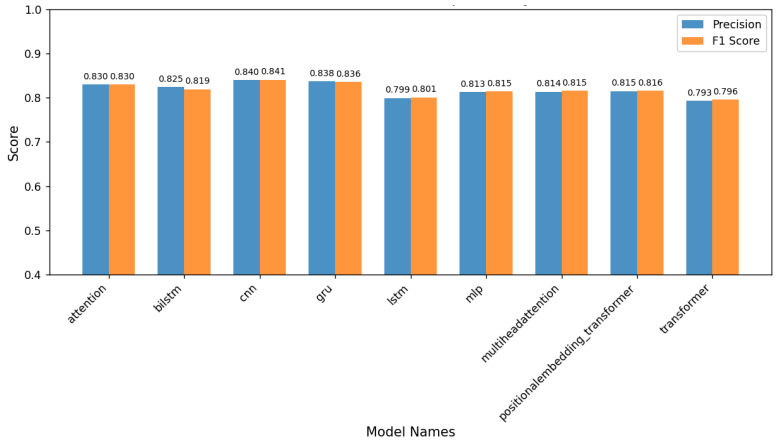


Figure 1. Comparison of accuracy and F1 score of each model under semantic contradiction set

As shown in figure 1, in the evaluation of the semantic contradiction set, the performance of each model showed a slight decline, among which the GRU performance decline was the smallest, and the BiLSTM and CNN models had relatively low declines. The performance of the Attention and MultiAttention models declined more than the above three models. The LSTM, Transformer, and PositionalEmbedding+Transformer models had the largest performance decline.

DOI: 10.54254/2755-2721/2026.TJ29617

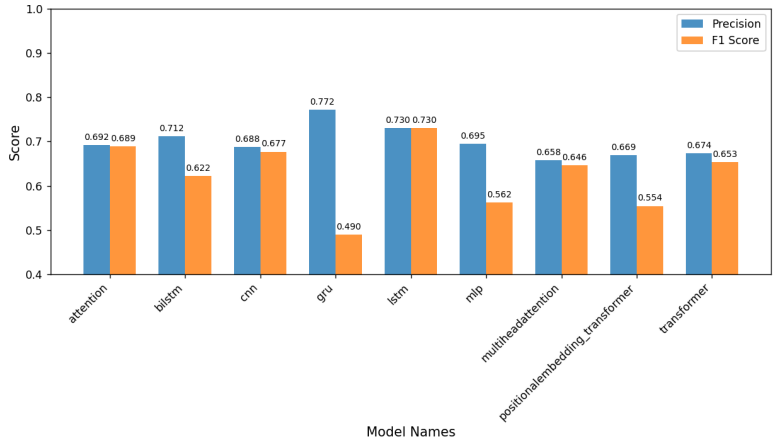


Figure 2. Comparison of accuracy and F1 score of each model under the easily confused set

As shown in figure 2, in the evaluation of the easily confused dataset, the performance of each model decreased more significantly than that of the semantically contradictory dataset. The LSTM model showed the smallest decrease, while the performance of the other models decreased more severely. Although the GRU model maintained some accuracy, its F1 score decreased particularly severely. The performance of the PositionalEmbedding+Transformer model decreased more severely, as can be seen from the accuracy and F1 score.

Based on the original test set, the study conclude that deep learning models have become relatively mature in sentiment analysis tasks, with CNNs, Transformers, and PositionalEmbedding + Transformers demonstrating strong generalization capabilities. While performance declined somewhat on semantically contradictory and easily confused datasets, the models maintained high accuracy, demonstrating their robustness to interference. The GRU model performed better than other models on the semantically contradictory dataset. Because it includes update and reset gates, it can more accurately determine when to retain or discard information, thus mitigating the semantic contradiction issue. The performance of each model on the easily confused dataset demonstrates that the LSTM model possesses strong fine-grained semantic understanding capabilities and stability, as its gating mechanism allows for more precise control over information transmission and storage. The Attention model demonstrated relatively stable performance in both overall interference resistance and fine-grained semantic understanding, as measured by both test sets.

5. Conclusion

This study trained multiple deep learning models on the IMDB e-sentiment analysis task, and evaluated and analyzed each model using the original test set, a semantically contradictory dataset processed from the original test set, and a confusing dataset generated by feeding prompt words to a large language model. The following results were obtained: CNN and Positional Embedding+Transformer performed best in the basic generalization capabilities of each model, the GRU model performed best on the semantically contradictory dataset, and the LSTM model performed best on the confusing dataset. It is worth noting that the Attention model maintained the top three performance in all three test sets and was the most stable.

The findings demonstrate that deep learning models possess strong generalization capabilities for sentiment analysis, strong resistance to interference, and the ability to discern semantic details. The Attention model achieved the best overall performance, demonstrating the most comprehensive performance in this experiment. These results provide insights for future research: CNNs excel at

DOI: 10.54254/2755-2721/2026.TJ29617

extracting local features, while models using the attention mechanism focus on capturing global dependencies and important information. Both models performed well in this study, suggesting the potential for exploring hybrid architectures, such as CNN+Attention, to demonstrate their performance in complex scenarios. Furthermore, the field of NLP still faces challenges, such as difficulties in understanding semantics and the need to consider computational resources and energy consumption in practical scenarios. The RNN variants observed in this study, such as LSTM and GRU, which do not utilize attention mechanisms, offer considerable potential for development in these areas due to their good performance, low parameter count, and relatively simple structure.

References

[1] Manning, C. D. (2022). Human language understand & reasoning. Daedalus.
[2] Kadavath, S., et al. (2022). Language models (mostly) know what they know. arXiv preprint arXiv: 2207.05221.
[3] Liang, P., Bommasani, R., Lee, T., Tsipras, D., et al. (2022). Holistic evaluation of language models. arXiv preprint arXiv: 2211.09110.
[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS).
[5] Lyu, B., Wang, D., & Zhu, Z. (2025). A solvable attention for neural scaling laws. University of Southampton Institutional Repository.
[6] Chen, P. C., Tsai, H., Bhojanapalli, H. W., Chung, H. W., et al. (2021). A simple and effective positional encoding for transformers. arXiv preprint arXiv: 2109.08677.
[7] Sachan, D. S., Zaheer, M., & Salakhutdinov, R. (2019). Revisiting lstm networks for semi-supervised text classification via mixed objective function. Proceedings of the AAAI Conference on Artificial Intelligence.
[8] Shaukat, Z., Zulfiqar, A. A., Xiao, C., Azeem, M., et al. (2020). Sentiment analysis on IMDB using lexicon and neural networks. SN Applied Sciences.
[9] Kiela, D., Bartolo, M., Nie, Y., Kaushik, A., Geiger, A., et al. (2021). Dynabench: Rethinking benchmarking in NLP. arXiv preprint arXiv: 2102.13249.
[10] Gardner, M., Artzi, Y., Basmova, V., Berant, J., et al. (2020). Evaluating models' local decision boundaries. arXiv preprint arXiv: 2004.03036.
[11] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining pp. 168