Proceedings of CONF-MSS 2025 Symposium: Automation and Smart Technologies in Petroleum Engineering
DOI: 10.54254/2755-2721/2025.KL22625

Enhancing Sentiment Analysis with Fine-Tuned DeBERTa in a Cross-Domain Study on IMDb and Amazon Polarity Datasets

Zijian Cao

Dundee International Institute, Central South University, Changsha, China
7802230128@csu.edu.cn

Abstract: Sentiment analysis has become crucial for understanding consumer behavior in e-commerce and social media. While models based on transformers like Bidirectional Encoder Representations from Transformers (BERT) and A Robustly Optimized BERT Pretraining Approach (RoBERTa) have great performance in Natural Language Processing (NLP) tasks, challenges persist in domain-specific language nuances and context sensitivity. This study addresses these gaps by utilizing DeBERTa, an advanced model with disentangled attention mechanisms, to improve sentiment classification. The methodology involves fine-tuning the DeBERTa-base on the IMDb dataset and evaluating its generalization on the Amazon Polarity dataset. The results achieved 93.21% accuracy, 91.28% precision, 94.98% recall, and 93.32% F1 score, achieving a good balance between precision and recall while maintaining high accuracy. The result of this study demonstrates that Decoding-enhanced BERT with Disentangled Attention (DeBERTa) has strong performance in handling long, structured text and short, diverse comments. This illustrates DeBERTa's superiority in capturing complex semantic relationships and context-dependent sentiments.

Keywords: Sentiment classification, Disentangled attention mechanisms, DeBERTa

1. Introduction

With the surge in user reviews on social media and e-commerce platforms, accurately identifying positive and negative sentiments has become a key requirement for optimizing products and services. For customers, when booking a ticket for a movie, or buying an appliance from Amazon, they also need to browse bad reviews to assess product quality. For example, according to Muzumdar, the role of review valence is critical in influencing consumer attitudes during time-sensitive purchase scenarios [1]. Based on these demands, analysis of sentiment provides a deep insight into the emotional and psychological thoughts of individuals. However, it is not a simple task for analysis of sentiment, for text is ambiguous, context-based, and genuine intention is related to specific domains. Therefore, devising an AI model to analyze the true thoughts beyond the complex information and obtaining the real meaning is inevitable.

In recent years, transformers has brought revolutionary changes in many domains, one of which is natural language processing, offering a strong tool to play a role in the analysis of sentiment. Among these, models such as Bidirectional Encoder Representations from Transformers (BERT) which can capture deep relationships between different elements become more prevalent. It surpasses earlier

© 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: 10.54254/2755-2721/2025.KL22625

deep learning architectures like Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs).

Devlin et al. introduced the BERT model, which elevated the state-of-the-art results on the General Language Understanding Evaluation (GLUE) benchmark to new heights, while also surpassing the previous best models on reading comprehension tasks [2].

Similarly, Liu et al. researched using A Robustly Optimized BERT Pretraining Approach (RoBERTa), to optimize the pre-trained process of BERT to reveal its capabilities in multi-tasks, including multilingual sentiment analysis. They improved training by removing the next sentence prediction objective, utilizing dynamic masking, increasing training duration, and incorporating larger and more diverse datasets [3].

Similarly, Sun et al. explored fine-tuning models on specific datasets and demonstrated that this method conspicuously improves model performance. The researchers found that fine-tuned models achieved better accuracy and understanding of context than general pre-trained models. However, despite the improvements as mentioned earlier, the model still faces challenges in handling more complex or domain-specific tasks [4].

This study adopts a fine-tuned transformer-based model to implement sentiment analysis on specific domain datasets. By taking advantage of BERT, this research aims to solve existing challenges, including class imbalance, domain-specific language nuances, and sentiment's context-sensitive nature.

2. Method

2.1. Dataset

IMDb and Amazon Polarity are standard datasets that are widely used in sentiment analysis [5-7]. This research started with fine-tuning the DeBERTa (Decoding-Enhanced BERT with Disentangled Attention) model, then the model was tested on the Amazon Polarity dataset to evaluate its generalization ability [8].

The IMDb dataset consists of 50000 labeled movie reviews, divided into a training set (25000 reviews) and a test set (25000 reviews). Each review is classified as positive or negative, making it a typical binary classification task. The Amazon Polarity dataset is composed of more large-scale reviews than the IMDb dataset, with 400000 reviews in the test set, categorized into 2 classes: positive reviews (label 1) and negative reviews (label 0).

These two datasets have distinct characteristics. Each of the comments in the IMDb dataset is longer and more structured, and texts in the Amazon Polarity dataset are shorter but with a more diverse set of features, including many colloquial expressions.

2.2. Model

Decoding-enhanced BERT with Disentangled Attention (DeBERTa), which is an advanced language model and performs well in natural language processing tasks. In the transaction of fine-tuning, the IMDb dataset is utilized for training DeBERTa model, and the Amazon Polarity dataset is used for assessing the generalization capability of the fine-tuned model on different domains of data. In particular, DeBERTa performs exceptionally well in text classification, sentiment analysis, and similar tasks, especially when handling long texts and complex semantic relationships.

The research utilizes the DeBERTa-base provided by Hugging Face as the fundamental model and AutoTokenizer and AutoModelForSequenceClassification to load it in. The model is initialized as a binary classification task (num_labels=2) to adapt to the sentiment classification task for positive and negative reviews on the IMDb dataset. To accelerate the training and inference processes, the model is deployed onto a GPU device.

DOI: 10.54254/2755-2721/2025.KL22625

In the procedure of processing, this research utilizes the DeBERTa tokenizer to tokenize, pad, and truncate the comments, assuring that each comment is all set to the length of 256 tokens. After tokenization, the data will be converted into PyTorch tensor format by removing the original text column and retaining only the formats required by the model.

The training process of the model is managed through the Hugging Face Trainer class. Table 1 shows detailed training data.

Table 1: Detailed training data

Evaluation Strategy	Epoch
Learning Rate	2e-05
Weight Deacy	0.01
Logging Steps	50
Train Batch Size	4

3. Experimental Result and Discussion

3.1. Result

DeBERTa demonstrates powerful performance in sentiment analysis tasks. Table 2 shows detailed test data.

Table 2: Test data

Accuracy	0.93206
Precision	0.91280
Recall	0.94977
F1 Score	0.93324

3.2. Discussion

The experimental result on Amazon Polarity demonstrates that DeBERTa after fine-tuning performs well in sentiment analysis tasks. The model achieved an accuracy of 93.21%, indicating its exceptional performance in telling positive and negative comments. A precision of 91.28% suggests that most samples predicted as positive are actually true, which is vital in situations requiring high reliability on avoiding wrongly detecting such as e-commerce review analysis. The recall of 94.98% shows that the model captures the majority of the true positive samples, reducing missing detections. Lastly, the F1 score of 93.32%, which gets a great balance between precision and recall.

Although the experimental results demonstrate an advance, several limitations still exist. The IMDb dataset reviews focus on movies and more reasoned, while Amazon Polarity contains colloquial terms (e.g., “battery life”, “texture”) on diverse products (e.g., phones, clothes). The model may get confused with such not obvious comments. Furthermore, this phenomenon may get worse when facing extremely imbalanced situations.

In addition, the cost of training a small, experienced model is acceptable, but when it comes to training a more powerful and generative model on all sides of reviews or comments, it is costly and expensive. To fix this problem, researchers can adopt several ways to cut down costs, they can use open-source models and fine-tune them instead of training a model from scratch. Additionally, they can develop more advanced ways to optimize the training efficiency.

DOI: 10.54254/2755-2721/2025.KL22625

4. Conclusion

This study conducted cross-domain sentiment analysis experiments on IMDb and Amazon Polarity dataset by fine-tuning DeBERTa model. The experimental result demonstrated that DeBERTa had a great performance in handling both long and short text, especially in capturing intricate semantic relationships and context-dependent sentiments. Specifically, DeBERTa achieved an accuracy of 93.21% and an F1 score of 93.32% on the IMDb dataset, showcasing high performance in structured long-text sentiment classification tasks. Additionally, the model also showed exceptional generalization capabilities on the Amazon Polarity dataset, which could deal with diverse short comments.

Compared to traditional BERT and RoBERTa models, DeBERTa reinforces its ability to capture contextual information through decoupled attention mechanisms. Furthermore, this study optimized the model's performance on specific datasets through fine-tuning strategies, verifying the importance of fine-tuning in enhancing the model's performance.

However, there are still some limitations in DeBERTa. Firstly, the model may be interrupted when facing extremely imbalanced datasets, researchers in the future may explore more efficient models. Then, although DeBERTa did well in capturing contextual information, its high computational complexity, long time of training and inference can be a block in practical applications.

Overall, this study highlights DeBERTa's strong performance in sentiment analysis, especially its ability to adapt to different types of data—from lengthy movie reviews to short, casual product comments. The result not only confirms its effectiveness in capturing nuanced emotions but also offers practical insights for improving real-world applications. These findings pave the way for future work on building lighter, faster models without sacrificing accuracy.

References

[1] *Muzumdar, P. (2021). Impact of review valence and perceived uncertainty on purchase of time-constrained and discounted search goods. ArXiv, abs/2110.09313.*

[2] *Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). 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 (Vol. 1, pp. 4171–4186). Association for Computational Linguistics.*

[3] *Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. ArXiv, abs/1907.11692.*

[4] *Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), Chinese Computational Linguistics (CCL 2019). Lecture Notes in Computer Science, vol 11856. Springer, Cham.*

[5] *He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. In the International Conference on Learning Representations.*

[6] *Maas, A. L., Daly, R. E., Pham, P. T., Huang, D. T., Ng, V., Potts, C., & Dredze, M. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), 142-150.*

[7] *McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. In Proceedings of the 7th ACM Conference on Recommender Systems (pp. 165-172).*

[8] *Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems 28 (NIPS 2015).*