Sentiment Analysis in Customer Service Conversations: Comparing Fine-Tuning and From-Scratch Training Approaches with nanoGPT

Bilge Kağan ÖZKAN
Department of Information Systems
Middle East Technical University
Ankara, Turkey
ozkan.bilge@metu.edu.tr

***Abstract*—This study compares two approaches for sentiment analysis in customer service conversations: fine-tuning a pre-trained GPT-2 model versus training a similar architecture from scratch. I modified the nanoGPT architecture by adding a specialized sentiment classification head and implementing different learning rates for transformer and sentiment components. Our evaluation reveals that the fine-tuned model significantly outperforms the from-scratch model, achieving 90% versus 66.7% accuracy. This highlights the value of transfer learning for sentiment classification tasks, especially when working with limited domain-specific data.**

***Index Terms*—sentiment analysis, customer service, nanoGPT, fine-tuning, transfer learning**

I. Introduction

Sentiment analysis plays a critical role in understanding customer satisfaction and improving service quality. While transformer-based language models are widely used in natural language processing tasks, the question of fine-tuning pre-trained models versus training domain-specific models from scratch remains relevant, especially for specialized applications like sentiment analysis in customer service.

This study examines how nanoGPT (a smaller implementation of the GPT architecture) can be adapted for sentiment classification tasks by comparing two approaches: (1) fine-tuning a pre-trained GPT-2 model and (2) training the same architecture from scratch. This comparison provides insights into the effectiveness and efficiency of transfer learning for specialized classification tasks in customer service conversations.

II. Dataset

The dataset consists of customer service conversations labeled as negative, neutral, or positive. The test set was predefined, while training and validation sets were created using an 80/20 split with StratifiedKFold (n_splits=5). Key preprocessing steps included: preserving both customer and agent parts with special tags; replacing URLs, emails, and numerical values with tokens; extracting features like conversation length and question mark count; using GPT-2 tokenizer with 1024 token blocks; and mapping sentiment labels to numerical values (0, 1, 2).

For training, I used a feature-enhanced prompt format that included both conversation content and extracted numerical features. Our exploratory data analysis revealed class imbalance (neutral examples most common, positive least numerous), longer negative conversations, and distinct linguistic patterns for each sentiment class.

To address class imbalance, I implemented strategic sampling weights based on class frequency, conversation length, and customer text length, with additional weighting for underrepresented positive examples. I also used a weighted loss function during training to prioritize underrepresented classes.

III. Modeling

The base architecture implemented is nanoGPT, based on the GPT-2 transformer model with 12 attention heads, feed-forward networks, layer normalization, residual connections, 12 transformer layers, and 768-dimensional embeddings.

I modified this architecture with the following key changes:

- Replaced the language modeling head with a sentiment classification head
- Added a sentiment-specific attention mechanism focusing on emotionally relevant parts, particularly the last 100 tokens
- Applied different dropout rates for transformer backbone (0.1%) and sentiment head (0.2%)
- Adjusted the final layer to output three class probabilities

Two model variants were implemented: (1) **Fine-Tuned Model**: Initialized with pre-trained GPT-2 weights, using lower learning rate for transformer components (5e-5) and higher rate for the sentiment head (20x); and (2) **From-Scratch Model**: Initialized with random weights, using identical hyperparameters for fair comparison.

Both models were trained with batch size 12, gradient accumulation steps 40, maximum 2,000 iterations, cosine learning rate decay with warmup, and validation every 250 iterations. Training was tracked using WandB.

IV. Evaluation

I evaluated model performance using multiple metrics: (1) **Accuracy** for overall correctness and interpretability;

(2) **Weighted F1 Score** to account for class imbalance by weighting each class's contribution according to its frequency; (3) **Confusion Matrix** to identify specific error types; and (4) **Class-Specific Metrics** (precision, recall, F1) for detailed analysis of each sentiment class.

The selection of these metrics was guided by class imbalance considerations, the need for detailed error analysis, and the business context where correctly identifying negative sentiment may be more important than distinguishing between neutral and positive sentiments.

V. Results

The WandB experiment tracking tool revealed that the fine-tuned model significantly outperformed the from-scratch model across all metrics:

TABLE I
PERFORMANCE METRICS COMPARISON

Model	Acc.	W-F1	Neg-F1	Neu-F1	Pos-F1
Fine-tuned	0.900	0.898	1.000	0.870	0.820
From-scratch	0.667	0.599	0.950	0.667	0.180

The most striking difference was in the positive sentiment class, where the fine-tuned model correctly identified 7/10 positive examples, while the from-scratch model misclassified 9/10 positive examples as neutral.

The WandB tracking also revealed that the fine-tuned model showed faster convergence (500 iterations), more stable training curves, and required less computation time to achieve better results.

VI. Discussion

The comparison between fine-tuning and from-scratch training approaches revealed several important insights:

Advantages of Fine-Tuning: Superior performance across all metrics; faster convergence with pre-trained weights; and better recognition of positive sentiment by leveraging pre-acquired linguistic knowledge.

Limitations of From-Scratch Training: Poor distinction between positive and neutral sentiment (only 10% recall for positive examples); greater computational requirements despite identical architecture; and higher tendency toward overfitting.

The superior performance of the fine-tuned model can be attributed to linguistic knowledge transfer from pre-trained GPT-2, better contextual understanding from exposure to diverse texts, efficient parameter adaptation through different learning rates, and the sentiment-specific attention mechanism focusing on emotionally relevant content.

VII. Conclusion

This study demonstrates the clear advantages of fine-tuning pre-trained language models over training from scratch for sentiment analysis in customer service conversations. The fine-tuned model achieved 90% accuracy compared to only 66.7% for the from-scratch model, with particularly significant improvements in recognizing positive sentiment.

Our findings support that transfer learning is an efficient and effective approach for sentiment analysis tasks, even when adapting models to specialized domains like customer service. The modified nanoGPT architecture with a sentiment classification head provides a strong foundation, but its performance increases significantly when built upon pre-trained weights.

Future work could explore domain-specific pre-training, different model sizes, and more sophisticated attention mechanisms for sentiment analysis tasks, as well as techniques to improve positive sentiment recognition in from-scratch models.