IMDB Sentiment Classification using GPT-2 and DistilGPT-2

Ameer Nessaee

February 23, 2025

1 Introduction

Sentiment analysis is a crucial task in natural language processing, with applications ranging from product reviews to social media analysis. Large language models like GPT-2 [2] have shown remarkable capabilities in understanding and generating human-like text. This study explores the use of GPT-2 and its distilled version, which follows similar principles to DistilBERT [3], for sentiment classification on the IMDB movie review dataset [1]. `https://github.com/nessaee/ECE696-310/tree/main/projects/1`.

2 Experimental Setup

2.1 Dataset

We utilize the IMDB movie review dataset [1], a balanced binary sentiment classification benchmark containing 50,000 reviews. The dataset is evenly split into 25,000 training and 25,000 test samples, with reviews labeled as either positive or negative sentiment. Each review consists of multiple sentences of variable length, with a median length of 230 words. The dataset is publicly available and widely used for evaluating sentiment analysis models.

2.2 Model Architectures

Our experiments compare two transformer-based language models: GPT-2 Small (124M parameters) and DistilGPT-2 (82M parameters). Both models employ byte-pair encoding tokenization with a 50,257-token vocabulary. GPT-2 consists of 12 transformer blocks with 768-dimensional hidden states and 12 attention heads. DistilGPT-2, derived through knowledge distillation, maintains the same hidden state and attention head dimensions but reduces the number of transformer blocks to 6, achieving a 34% reduction in parameter count while preserving the core architectural elements.

Both models share key architectural components: GELU activation functions, pre-norm layer normalization, learned absolute positional embeddings, and a dropout rate of 0.1. The feed-forward network dimension is set to 3,072 (4x hidden size) in both architectures. The key distinction lies in DistilGPT-2's reduced depth, which offers potential computational advantages while testing the efficacy of model compression for sentiment analysis.

3 Methodology

Our methodology comprises three main phases: baseline evaluation, model adaptation, and fine-tuning. For baseline evaluation, we assess both models' zero-shot sentiment classification capabilities using their pre-trained weights. This provides insights into their inherent ability to transfer general language understanding to sentiment analysis.

Model adaptation is achieved through HuggingFace's transformers library, which allows for augmenting each model with a binary classification head. This head maps the 768-dimensional hidden states to sentiment predictions while preserving the pre-trained weights. The tokenizer handles padding and special tokens automatically, maintaining sequence lengths at 512 tokens.

The fine-tuning process applies the AdamW optimizer with a learning rate of $5e^{-5}$ and batch size of 8 for both models. We evaluate performance using classification accuracy and F1 score on the validation and test sets, considering the performance delta between baseline and fine-tuned states at each epoch. All experiments are conducted with 5-7 random seeds to ensure statistical reliability, with results reported as means across runs.

4 Results and Discussion

4.1 Model Performance and Training Dynamics

As shown in Table 1, both models start with baseline performance near random chance (DistilGPT2: 51.9%, GPT2-Small: 50.1%). However, they demonstrate remarkable improvement through fine-tuning. DistilGPT2 achieves 87.3% accuracy after the initial training epoch, improving to 94.3% and ultimately reaching 97.0% accuracy by the third epoch. The precision, recall, and F1 scores consistently match the accuracy, indicating balanced performance across classes. GPT2-Small shows slightly superior performance throughout training, starting at 88.8% accuracy and improving to 95.4% and 98.0% in subsequent epochs.

Table 2 reveals that GPT2-Small achieves more substantial relative gains, particularly in F1 score. By the final epoch, GPT2-Small shows a 95.71% improvement in accuracy and an impressive 188.26% improvement in F1 score over its baseline. In comparison, DistilGPT2 achieves an 86.93% improvement in accuracy and a 93.96% improvement in F1 score. This difference in relative improvement suggests that while both models reach high absolute performance, GPT2-Small may be more effectively leveraging its larger parameter count during fine-tuning. Furthermore, it appears that the GPT2-small baseline model demonstrates less balanced performance (as shown by its F1 score), but reduces this imbalance during fine-tuning.

4.2 Architectural Impact

The results demonstrate an interesting trade-off between model size and performance. While GPT2-Small achieves marginally better final metrics (98.0% vs 97.0% accuracy), DistilGPT2's performance is remarkably strong considering its reduced architecture. The consistent precision-recall balance in both models (evidenced by matching scores across metrics) suggests that the knowledge distillation process effectively preserves the essential features for sentiment classification.

The rapid improvement in early epochs, followed by continued but diminishing gains, indicates efficient transfer learning from the pre-trained weights. This is particularly noteworthy for DistilGPT2, which achieves 97.0% accuracy with fewer parameters, suggesting that sentiment analysis may not require the full capacity of the larger model.

References

[1] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

[3] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Table 1: Performance Comparison of DistilGPT2 and GPT2-Small Models

Model	**Epoch**	**Accuracy**	**Precision**	**Recall**	**F1**
DistilGPT2	Baseline	0.519	0.520	0.520	0.500
	0	0.873	0.874	0.873	0.873
	1	0.943	0.943	0.943	0.943
	2	**0.970**	**0.970**	**0.970**	**0.970**
GPT2-Small	Baseline	0.501	0.520	0.500	0.340
	0	0.888	0.888	0.888	0.888
	1	0.954	0.954	0.954	0.954
	2	**0.980**	**0.980**	**0.980**	**0.980**

Table 2: Improvement Percentages Over Baseline

Model	**Epoch**	**Accuracy**	**Precision**	**Recall**	**F1**
DistilGPT2	0	68.35%	68.06%	67.96%	74.66%
	1	81.84%	81.42%	81.42%	88.68%
	2	86.93%	86.50%	86.50%	93.96%
GPT2-Small	0	77.26%	70.79%	77.54%	161.09%
	1	90.50%	83.46%	90.80%	180.59%
	2	**95.71%**	**88.48%**	**96.02%**	**188.26%**