# Beyond Sentiment: Classifying Evaluative vs. Non-evaluative Text using LLM-Labeled Data and Adapter Fine-tuning

**Bomin Zhang**
University of Maryland
College Park, MD 20742
bominz2@umd.edu

## Abstract

This paper addresses the task of classifying text based on a nuanced definition: distinguishing evaluative judgments (evaluative) from non-evaluative statements and descriptions of direct experience (non-evaluative). This challenge stems from the lack of large-scale labeled datasets aligned with this specific distinction, which goes beyond traditional sentiment analysis and broad subjectivity detection. We leverage Large Language Models (LLMs) to generate synthetic training data. Parameter-efficient fine-tuning (PEFT), specifically LoRA, is employed on a pre-trained language model (DistilRoBERTa-base) for the classification task. LoRA fine-tuning performance is compared against standard direct fine-tuning on the 20 Newsgroups dataset, using a partially human-reviewed test set for robust evaluation. Findings indicate some success and suggest avenues for further research on this novel semantic analysis feature.

## 1 Introduction

**Hume's guillotine**, or the **Is-Ought Problem**, highlights a philosophical distinction where evaluative conclusions cannot strictly be inferred from purely factual statements. Extending this, **evaluation** and **description** can be seen as operating on parallel systems in formal reasoning. Evaluation doesn't alter the physical world, while physical conditions inform, but don't modify, the evaluative system itself. This philosophical divide underscores a need in NLP to differentiate text making judgments from text describing facts or experiences. Traditional sentiment analysis focuses on polarity, and broad subjectivity detection includes feelings and experiences alongside judgments, leaving the specific distinction between evaluative judgments and non-evaluative text less explored. We clarify these distinctions next.

## 2 Definitions and Comparison of Classification Axes

We define and compare three key axes for text classification relevant to this work:

### 2.1 Objective vs. Subjective (Broad)

**Objective:** Verifiable facts, external reality, independent of personal feelings. **Subjective (Broad):** Internal states, personal feelings, experiences, beliefs, opinions, judgments. Includes anything not purely objective.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

### 2.2 Sentiment (Positive/Negative/Neutral)

**Sentiment:** Emotional tone or polarity (Positive, Negative, Neutral) towards a subject.

### 2.3 Evaluative vs. Non-evaluative (This Work)

**Evaluative:** Text expressing an *evaluative judgment, stance, argument, or assessment* about a subject. Considered a *subset* of broad subjective text in this context. **Non-evaluative:** Text that is either a verifiable *objective fact/report* OR a *subjective report of personal experience, feeling, or non-evaluative description* without a generalized evaluative judgment.

### 2.4 Comparison of Axes

These axes capture distinct aspects. Table 1 briefly illustrates the difference between the Evaluative/Non-evaluative axis and the other two.

Table 1: Comparison of Evaluative/Non-evaluative with Sentiment and Objective/Subjective

| Axis / Class | **Evaluative** | **Non-evaluative** |
|---|---|---|
| **Sentiment (Pos/Neg)** | Ex: "Best pizza in town." (Eval, Pos)<br>Ex: "Service appalling." (Eval, Neg) | Ex: "Heartbroken by news." (Pure feeling, implies Neg)<br>Ex: "Sales up 30%." (Objective, implies Pos) |
| **Sentiment (Neutral)** | Ex: "Comprehensive, dry analysis." (Eval, Mixed/Neutral)<br>Ex: "Necessary step, but fails..." (Eval, Mixed/Neutral) | Ex: "Sky is blue." (Objective, Neutral)<br>Ex: "Feel sleepy." (Pure experience, Neutral) |
| **Objective** | *Contradictory.* Evaluative is subjective. | Ex: "Water boils at 100C." (Objective) |
| **Subjective (Broad)** | Ex: "Book incredibly insightful." (Judgment)<br>Ex: "Argument relies on flawed logic." (Evaluative stance) | Ex: "Feel under weather." (Pure feeling)<br>Ex: "Believe tomorrow sunny." (Belief) |

This comparison shows the Evaluative/Non-evaluative axis focuses uniquely on the presence of **evaluation** or **judgment**.

### 2.5 Importance of the "Evaluativeness" Axis

Separating evaluative from descriptive/factual content aids NLP applications by enhancing interpretability, filtering bias in information retrieval, and improving automated reasoning.

### 2.6 Research Challenge

A key challenge is the lack of large datasets explicitly labeled per this nuanced definition, unlike sentiment or broad subjectivity data.

### 2.7 Adopted Methodology

We address this by leveraging LLMs (GPT-4.1 mini) for synthetic training data generation via careful prompt engineering [Tan et al., 2024]. We apply parameter-efficient fine-tuning (PEFT), specifically LoRA [Hu et al., 2021], on DistilRoBERTa-base [Sanh et al., 2020]. This is compared against standard direct fine-tuning. Training and evaluation use data from the 20 Newsgroups dataset [Mitchell, 1997] with a partially human-reviewed test set.

Contributions:

1. Defining and tackling the evaluative vs. non-evaluative classification based on a novel nuance.
2. Demonstrating LLM-assisted labeling effectiveness via prompt engineering for synthetic data.
3. Evaluating LoRA vs. direct fine-tuning of DistilRoBERTa-base on LLM-labeled data.
4. Analyzing model performance across dataset topics.

# 3 Related Work

Relevant areas include:

## 3.1 Sentiment Analysis

Focuses on polarity and broad subjectivity [Liu, 2012].

## 3.2 Subjectivity Detection

Distinguishes subjective from objective [Liu, 2012], using corpora like MPQA [Wiebe et al., 2005]. However, its "subjective" is broader, including experiences/beliefs without explicit judgment, differing from this work's 'evaluative'.

## 3.3 LLMs for Data Annotation

LLMs aid annotation for tasks lacking data [Tan et al., 2024]. Prompt engineering is crucial [Brown et al., 2020]. LLM bias/noise must be considered.

## 3.4 Parameter-Efficient Fine-Tuning (PEFT)

Trains a small fraction of parameters to adapt large models efficiently [Han et al., 2024]. LoRA [Hu et al., 2021] injects low-rank matrices. PEFT saves time, memory, and storage [Pfeiffer et al., 2020, Mangrulkar et al., 2022].

# 4 Methodology

## 4.1 Data Source and Preparation

Using the 20 Newsgroups dataset [Mitchell, 1997] due to its diverse topics (Figure 1). Preprocessing included removing headers, footers, and segmenting text.


Figure 1: Distribution of documents across the 20 newsgroup topics.

## 4.2 Data Labeling Strategy (LLM-Assisted)

GPT-4.1 mini generated labels for 10,000 sampled snippets. Prompts included precise definitions and examples. LLM noise exists; manual review of 500 samples found 9 incorrect labels and 9 correct labels with incorrect reasoning. More details in Appendix 10.

### 4.3 Model Architecture

'DistilRoBERTa-base' [Sanh et al., 2020], a distilled RoBERTa model, was used with a linear classification head.

### 4.4 Fine-tuning Strategies

- **Direct Fine-tuning:** Standard approach, trains substantial model parameters.
- **LoRA Fine-tuning:** PEFT method, freezes base model and trains low-rank matrices injected into transformer layers.

## 5 Experimental Setup

From 50,000 pruned texts, 10,000 were sampled preserving topic distribution. This split into 80% (8,000) training and 20% (2,000) test sets, LLM-labeled. Labels encoded: "evaluative", "non-evaluative".

Models fine-tuned on 'DistilRoBERTa-base' using Hugging Face Transformers/PyTorch and 'peft'. Hyperparameters: lr = $2 \times 10^{-5}$, Batch size = 200, Epochs = 10. LoRA parameters: rank $r = 16$, scaling $\alpha = 64$, dropout $d = 0.05$. Trains 1,034,498 parameters (1.24% of base model). Hardware: Single RTX2080Ti (11GB). Evaluation metrics: F1 Score (primary), Accuracy, Precision, Recall. Also F1 per topic.

## 6 Results

Evaluated on 2000-sample test set.

### 6.1 Overall Performance

Table 2 shows performance for the 'evaluative' class.

Table 2: Overall Performance ('Evaluative' Class)

| Model | Acc. | 'Eval' P | 'Eval' R | 'Eval' F1 |
|---|---|---|---|---|
| Direct FT | 0.8755 | 0.7845 | 0.7858 | 0.7852 |
| LoRA FT | 0.8525 | 0.7536 | 0.7288 | 0.7410 |

Direct fine-tuning outperformed LoRA (0.7852 vs 0.7410 'evaluative' F1), suggesting fuller model adaptation was more effective.

### 6.2 Confusion Matrix

Test set has 579 evaluative, 1421 non-evaluative samples.

Figure 2 shows direct FT had fewer misclassifications (FN: 124 vs 157, FP: 125 vs 138). Both show notable False Negatives.

### 6.3 Performance by Topic

Figures 3 and 4 show F1 scores per topic.

Average F1 is 0.779 for Direct, 0.651 for LoRA. LoRA lags more significantly in the lowest-performing topics ('comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware', 'comp.windows.x', 'talk.religion.misc'), often related to politics/religion which were challenging domains. Topics like 'sci.space' and 'rec.sport.hockey' had higher scores.

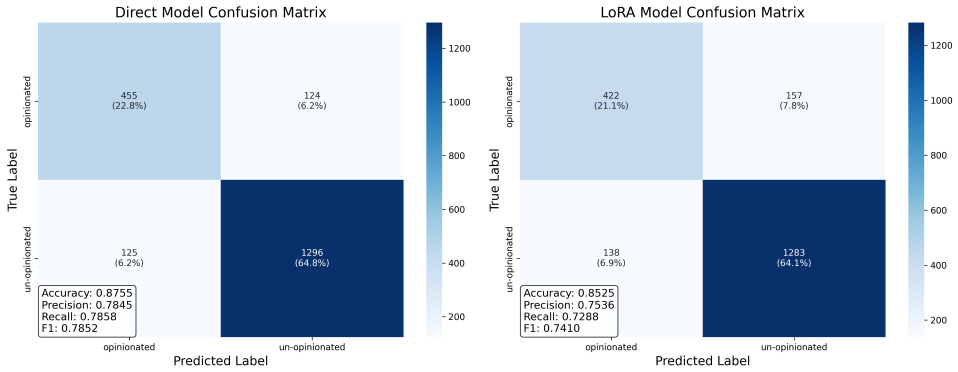


Figure 2: Confusion Matrices: Direct Model (Left) and LoRA Model (Right) on Test Set

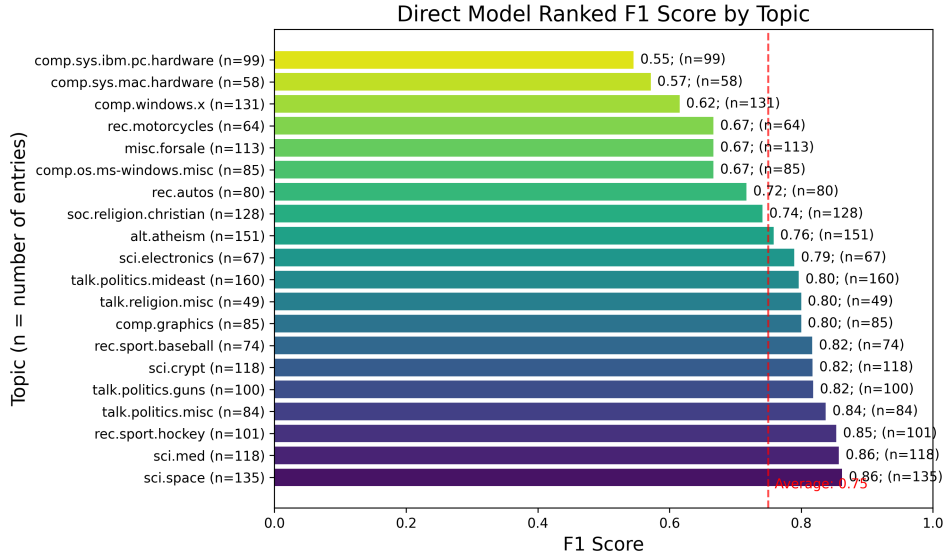


Figure 3: F1 Scores by Topic for Direct Model

# 7 Analysis

Manual analysis confirmed LLM's ability to apply the definition, though errors occurred ( 9/500 incorrect labels). Evaluation showed direct FT superior to LoRA, with an 'evaluative' F1 of 0.7852 vs 0.7410. LoRA's parameter efficiency came at a performance cost. Performance varied by topic, highlighting domain dependence; political/religious domains were harder. Close topics showed relatively close results, aligning with previous findings using similar models[1].

# 8 Limitations and Future Work

## 8.1 Limitations

- LLM training data noise/bias.
- Definition specificity; not universally applicable.
- Performance tied to 20 Newsgroups domain.
- Resource limits restricted hyperparameter tuning, larger models.
- Binary classification only.
- Limited human test set size impacts granular analysis robustness.

[1]Note: This observation is specific to this experiment and model choice.

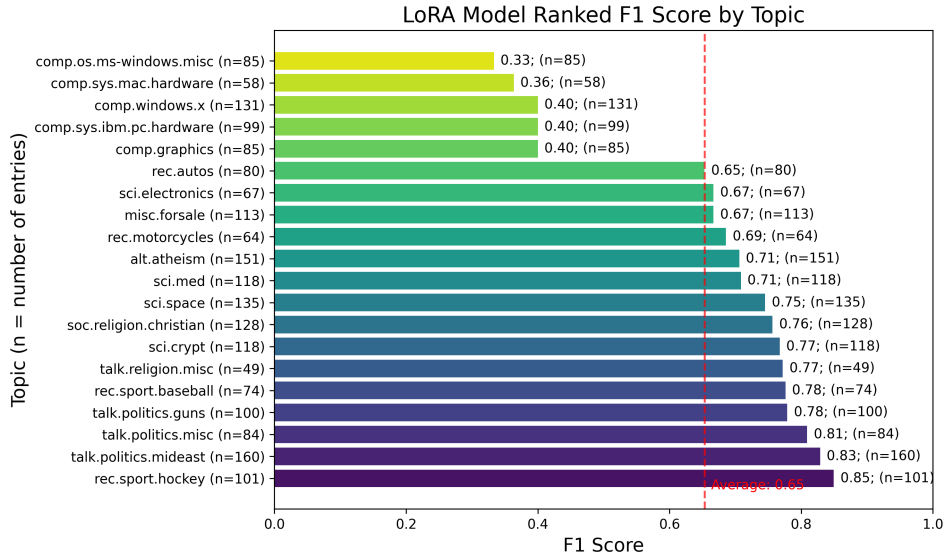


Figure 4: F1 Scores by Topic for LoRA Model

- Masked language model (DistilRoBERTa) might not be optimal vs. autoregressive LLMs for nuances.

### 8.2 Future Work

- Improve LLM labeling (larger LLMs, advanced prompting).
- Expand human data (larger test set, partial training set).
- Cross-domain evaluation and adaptation.
- Multi-class modeling (Judgment, Fact, Experience).
- Comprehensive PEFT comparison.
- Explainability: understand features driving classification.
- Integrate classifier into larger NLP pipelines.
- Leverage existing subjectivity detectors to refine LLM labels.

## 9 Conclusion

This study successfully defined and addressed the task of classifying evaluative vs. non-evaluative text using LLM-assisted labeling on the 20 Newsgroups dataset. We compared standard direct fine-tuning with LoRA on DistilRoBERTa-base. Direct fine-tuning achieved an 'evaluative' F1 of 0.7852, outperforming LoRA (0.7410), demonstrating that full model adaptation was more effective for this task's nuanced distinction, despite LoRA's parameter efficiency. Performance varied across topics, highlighting domain challenges. This work contributes a methodology for tackling tasks lacking large datasets via LLMs and introduces a valuable semantic distinction for NLP applications. Future work should refine data quality, expand evaluation, explore multi-class distinctions, and analyze PEFT trade-offs further.

## 10 Appendix

### 10.1 Manual Analysis of LLM-Generated Labels

A manual review of 500 LLM-labeled samples was conducted to assess label quality. The analysis confirmed that GPT-4.1 mini was largely capable of applying the specific definition of 'evaluative' vs. 'non-evaluative' provided through prompt engineering. Out of the 500 reviewed samples, approximately 9 samples were identified as having an incorrect classification label assigned by the LLM. A similar number of samples received the correct label but were accompanied by incorrect or

questionable reasoning provided by the LLM. Errors were more prevalent in text snippets with complex sentence structures, subtle language, or domain-specific jargon that was not explicitly covered in the prompt examples. Overall, the LLM-assisted labeling proved a viable, cost-effective method for generating a training dataset for this novel classification task, despite the presence of a small degree of noise. Due to space constraints, detailed examples from this analysis are omitted.

## References


T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language Models are Few-Shot Learners. *Abstract Report*, 33, 2020.

Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. URL `https://arxiv.org/abs/2403.14608`.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

B. Liu. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies*, volume 5, pages 16–29. Morgan & Claypool Publishers, 2012.

S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

T. Mitchell. Twenty Newsgroups. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5C323.

J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, and I. Gurevych. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online, 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.7`.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL `https://arxiv.org/abs/1910.01108`.

Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu. Large language models for data annotation and synthesis: A survey, 2024. URL `https://arxiv.org/abs/2402.13446`.

J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinion and emotion in language. In *LREC 2005*, 2005.