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 task presents a challenge due to the lack of large-scale labeled datasets aligned with this specific distinction, which goes beyond traditional sentiment analysis and broad subjectivity detection. The methodology involves leveraging Large Language Models (LLMs) to generate synthetic training data. We employ parameter-efficient fine-tuning (PEFT), specifically LoRA, on a pre-trained language model (DistilRoBERTa-base) for the classification task. The performance of LoRA fine-tuning 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 serve as a call for more research on this topic. This work contributes by introducing a distinction for a new and important feature in semantic analysis.

1 Introduction

Hume's guillotine, or the **Is-Ought Problem**, is a philosophical thesis stating that "an ethical or judgmental conclusion cannot be inferred from purely descriptive factual statements". Whether "ought" can be derived from "is" remains debated. However, expanding on this thesis, we find that **evaluation** and **description** operate on highly parallel systems in formal reasoning. At a given time, evaluation does not affect the physical world. Conversely, while physical world conditions are variables in the overall evaluative outcome, they do not modify the system used for evaluation itself. This philosophical distinction highlights a need in natural language processing (NLP) to differentiate between text that makes judgments or assessments and text that describes facts or experiences. In the field of sentiment analysis, however, most research has focused on classification of sentiment polarity or broad subjectivity, leaving the specific distinction between evaluative judgments and non-evaluative text less explored. Before moving on, the distinction between the three axes: polarity, broad subjectivity, and the "evaluativeness" axis defined in this work, should be clarified.

2 Definitions and Comparison of Classification Axes

Based on the refined definitions relevant to this work, the three principal axes for text classification can be defined and compared as follows:

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

2.1 Objective vs. Subjective (Broad)

Objective: Text that pertains to facts, external reality, and things that are verifiable or measurable. It is independent of personal feelings, interpretations, or beliefs.

Subjective (Broad): Text that pertains to internal states, personal feelings, experiences, beliefs, opinions, judgments, and interpretations. It encompasses anything that is not purely objective. This broad definition includes feelings, experiences, and judgments.

2.2 Sentiment (Positive/Negative/Neutral)

Sentiment: The emotional tone or polarity expressed towards a subject. This is typically classified as Positive (favorable), Negative (unfavorable), or Neutral (neither clearly positive nor negative, or mixed).

2.3 Evaluative vs. Non-evaluative

Evaluative: Text that expresses an *evaluative judgment, stance, argument, or assessment* about a subject. According to the specific definition used in this context, this is considered a *subset* of broad subjective text.

Non-evaluative: Text that is either a verifiable *objective fact or report* OR a *subjective report of a personal experience, feeling, or non-evaluative description* that does *not* contain a generalized evaluative judgment or stance about the subject itself.

2.4 Comparison of Axes

While related, these axes capture distinct aspects of text, as illustrated in the following tables comparing the Evaluative/Non-evaluative axis with the other two axes.

Table 1: Comparison of Evaluative/Non-evaluative and Sentiment Axes

	Evaluative	**Non-evaluative**
Positive/Negative Sentiment	**Example:** "This restaurant has the best pizza in town." (Evaluative judgment, Positive Sentiment) **Example:** "The customer service was appalling." (Evaluative judgment, Negative Sentiment)	**Example:** "I was heartbroken to hear the news." (Pure personal feeling, implies Negative Sentiment, but not a judgment about the source of the news) **Example:** "Sales increased by 30% this quarter." (Objective fact, implies Positive Sentiment)
Neutral Sentiment	**Example:** "The article presents a comprehensive, though somewhat dry, analysis." (Evaluative judgment/stance, Mixed/Neutral Sentiment) **Example:** "The new policy is a necessary step, but it doesn't fully address the core issue." (Evaluative judgment/stance, Mixed/Neutral Sentiment)	**Example:** "The sky is blue." (Objective fact, Neutral Sentiment) **Example:** "I felt sleepy during the presentation." (Pure personal experience, Neutral Sentiment, not a judgment on the presentation's quality)

These comparisons highlight that the Evaluative/Non-evaluative axis represents a distinct categorization focused specifically on the presence or absence of **evaluation** or **judgment** about a subject, as defined in this work.

Table 2: Comparison of Evaluative/Non-evaluative and Objective/Subjective (Broad) Axes

	Evaluative	**Non-evaluative**
Objective	*This cell is inherently contradictory.* Evaluative text (evaluative judgment/stance) is a form of subjective text and cannot be objective. Objective text is factual and lacks personal evaluation.	**Example:** "Water boils at 100 degrees Celsius at standard atmospheric pressure." (Objective fact) **Example:** "The company is headquartered in Seattle." (Objective fact/report)
Subjective (Broad)	**Example:** "That book was incredibly insightful." (Judgment) **Example:** "I think his argument relies on flawed logic." (Evaluative stance/judgment)	**Example:** "I feel a bit under the weather today." (Pure personal feeling) **Example:** "I believe tomorrow will be sunny." (Personal belief/speculation)

2.5 Why is the "Evaluativeness" Axis Important?

Separating evaluative information, such as opinions and judgments, from descriptive or factual content is valuable for various downstream natural language processing applications. These include enhancing model interpretability, improving information retrieval by filtering biases, and enabling more accurate automated reasoning.

2.6 Research Challenge

A significant challenge for training machine learning models on this task is the lack of large-scale, publicly available datasets explicitly labeled according to this specific, nuanced definition. While datasets exist for sentiment or broad subjectivity, they do not precisely capture the distinction between evaluative judgments and non-evaluative subjective experiences as required by the task.

2.7 Adopted Methodology

Following a common strategy for tasks lacking large labeled datasets, this work leverages Large Language Models (LLMs) to generate synthetic training data. This process essentially distills knowledge from the LLMs' world modeling into a classifier (as discussed in Section 3.3). Careful prompt engineering is employed to guide the LLM (specifically, GPT-4.1 mini) to label text snippets according to the definition presented herein. For model training, we applied two approaches: standard direct fine-tuning and parameter-efficient fine-tuning (PEFT). Specifically, we used LoRA (detailed in Sections 3.4 and 4.4) on a pre-trained language model, DistilRoBERTa-base (Section 4.3), for classification. These models are trained and evaluated on data derived from the 20 Newsgroups dataset [Mitchell, 1997], using a partially manually-reviewed test set for robust evaluation.

The main contributions of this paper are:

1. defining and tackling the task of classifying text as evaluative vs. non-evaluative based on a specific nuance distinguishing evaluative judgment from objective facts and non-evaluative personal experiences.
2. demonstrating the effectiveness of using LLM-assisted labeling with careful prompt engineering for generating synthetic training data for this task.
3. evaluating the performance of parameter-efficient LoRA fine-tuning in comparison to standard direct fine-tuning of a DistilRoBERTa-base model on this LLM-labeled dataset.
4. analyzing the model's performance across different topics present in the dataset.

3 Related Work

This section provides an overview of existing research and techniques relevant to this study, establishing the background for the methodology employed.

3.1 Sentiment Analysis

Sentiment analysis [Liu, 2012] is a well-established field focusing on identifying and extracting insight from sentiment-rich data, with sentiment polarity classification and subjectivity detection being two main areas of focus.

3.2 Subjectivity Detection

Subjectivity detection aims to distinguish subjective content from objective factual content [Liu, 2012]. Datasets like the Multi-Perspective Question Answering (MPQA) corpus [Wiebe et al., 2005] and the Cornell Movie Review dataset [Pang and Lee, 2005] have been instrumental in this area. However, the definition of "subjectivity" in much of this work is broader than the "evaluative" definition used here (detailed in Section 2.3). It often includes personal beliefs, experiences, desires, and speculations alongside explicit evaluations. This task specifically seeks to isolate text that contains an active *judgment* or *stance* on a subject, treating reports of personal feelings or experiences that do not generalize to a judgment as non-evaluative, thus requiring a finer-grained distinction than typical subjectivity detection.

3.3 LLMs for Data Annotation

The advent of powerful Large Language Models has opened new avenues for data annotation, particularly for tasks where manual labeling is costly or time-consuming, or when prototyping a new type of label [Tan et al., 2024]. LLMs can generate synthetic data or provide initial labels together with some reasoning/analysis for why the label is chosen, which can be reviewed and corrected by humans later at a lower cost. Techniques like few-shot learning and prompt engineering are critical for guiding LLMs to adhere to specific labeling guidelines [Brown et al., 2020]. This approach is particularly relevant for tasks lacking existing large-scale labeled resources, although the potential for noise or bias introduced by the LLM, especially with complex or nuanced definitions, must be considered.

3.4 Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning large pre-trained language models for downstream tasks often involves updating millions or billions of parameters, requiring significant computational resources and storage. Parameter-Efficient Fine-Tuning (PEFT) methods address these challenges by training only a small fraction of additional parameters while keeping the vast majority of the pre-trained model frozen [Han et al., 2024]. Adapters insert small, task-specific feed-forward networks within the layers of the pre-trained model. LoRA (Low-Rank Adaptation) [Hu et al., 2021] is a popular PEFT adapter that injects low-rank matrices into the transformer layers. These methods offer advantages in terms of reduced training time, memory usage, model size, and modularity [Pfeiffer et al., 2020, Mangrulkar et al., 2022]. LoRA is explored in this work as a lightweight alternative to full fine-tuning for the specific classification task.

4 Methodology

This section details the specific methods and techniques applied in this study to tackle the evaluative vs. non-evaluative classification task, building upon the concepts introduced in the Related Work.

4.1 Data Source and Preparation

This study utilizes the 20 Newsgroups dataset [Mitchell, 1997], a collection of approximately 20,000 documents harvested from 20 different Usenet newsgroups. This dataset was chosen after an initial exploration of Wikipedia data proved challenging due to its rigorous Neutral Point of View (NPOV) policy, which limits the prevalence of overt opinions other than in citations. The 20 Newsgroups dataset offers a diverse range of topics and writing styles across various domains (e.g., politics, religion, science, computers). This makes it suitable for training and evaluating a text classifier (See Figure 1) and provides the additional benefit of analyzing model performance across different topics.

Preprocessing steps were applied to clean the raw text data. Pruning logic was implemented, inspired by MapReduce principles for efficient parallel processing, including the removal of Usenet headers, footers, and signature blocks, and handling excessive quoting from previous messages. Texts were segmented by paragraph or 1000 character segments, whichever is shorter.

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

4.2 Data Labeling Strategy (LLM-Assisted)

Given the absence of a large-scale dataset labeled according to the specific definition presented herein (Section 2.3), an LLM-assisted labeling strategy was employed as introduced in Section 3.3. GPT-4.1 mini was used to generate synthetic labels for the processed text snippets from the 20 Newsgroups dataset. The effectiveness of this approach relies heavily on careful prompt engineering. Prompts were designed that provided the LLM with the precise definition of evaluative and non-evaluative text, including examples illustrating the key distinctions, particularly between evaluative judgments and non-evaluative personal experiences. The LLM was instructed to classify each text snippet and provide a brief justification for its label. While cost-effective for generating a large volume of training data, it is acknowledged that LLM-generated labels may contain noise or systematic biases. A set of 500 samples was manually reviewed. Of these, 9 were identified as having incorrect labels. A similar number received the correct label but with incorrect reasoning provided by the LLM. A subset of this analysis can be found in the appendix (Section 9.1).

4.3 Model Architecture

The ‘DistilRoBERta-base‘ model from the Hugging Face Transformers library was chosen as the base architecture. DistilRoBERTa is a distilled version of RoBERTa, a robustly optimized BERT approach, known for its strong performance across various NLP tasks while being computationally more efficient than its larger counterparts. The model consists of a transformer encoder stack, and for classification, a linear classification head is added on top of the final hidden state representation of the input text.

4.4 Fine-tuning Strategies

Two fine-tuning strategies were explored: standard direct fine-tuning and Parameter-Efficient Fine-Tuning (PEFT) using LoRA (as introduced in Section 3.4):

- **Direct Fine-tuning:** This is the standard approach where the entire pre-trained model (or typically, the classification head and potentially the upper layers of the transformer) is trained on the downstream task data. This involves updating a large number of parameters.
- **LoRA Fine-tuning:** Parameter-efficient fine-tuning using LoRA was implemented. This method freezes the weights of the pre-trained DistilRoBERTa model and injects small, trainable low-rank decomposition matrices into the transformer layers.

5 Experimental Setup

The total dataset after pruning and preparation (and before labeling) is $\tilde{5}$0,000 rows. Since LLM labeling is costly, a subset of 10,000 text snippets were randomly sampled with the same proportion of text from each topic as the original dataset. This subset is then split into an 80% training set (8,000 samples) and a 20% test set (2,000 samples). The classification labels ("evaluative", "non-evaluative") were encoded into integers (e.g., 0 and 1).

The 'DistilRoBERta-base' model [Sanh et al., 2020] was fine-tuned using two distinct approaches: standard direct fine-tuning and LoRA adapter fine-tuning. Both models were trained on the same 8,000 LLM-labeled training samples. Training was performed using the Hugging Face Transformers library with PyTorch, and the adapter is trained using the 'peft' (parameter-efficient fine-tuning) library.

Key hyperparameters for training were set as follows: Learning rate lr $= 2 \times e^{-5}$, Batch size = 200, Number of epochs = 10. For LoRA, parameters were rank $r = 16$, scaling factor $\alpha = 64$, and dropout rate $d = 0.05$. In this configuration, only 1,034,498 parameters were trained, representing 1.24% of the base model's 83,154,436 parameters.

Training and evaluation were conducted on a single RTX2080Ti GPU with 11GB of Memory.

Model performance was evaluated exclusively on the 2,000-sample test set. The primary evaluation metric was the F1 Score, which balances Precision and Recall and is suitable for classification tasks. Accuracy, Precision, and Recall are also reported for a comprehensive view. Furthermore, F1 scores were computed broken down by the original 20 Newsgroup topics present in the test set to analyze performance variations across domains.

6 Results

This section presents the experimental results comparing the performance of the direct fine-tuned DistilRoBERTa model and the LoRA fine-tuned DistilRoBERTa model on the test set (2000 samples).

6.1 Overall Performance

Table 3 presents the overall accuracy and classification metrics for the positive class ('evaluative') for both models on the test set.

Table 3: Overall Classification Performance and Metrics for the 'Evaluative' Class on Test Set

Model	Overall Accuracy	'Evaluative' Precision	'Evaluative' Recall	'Evaluative' F1 Score
Direct Fine-tuning	0.8755	0.7845	0.7858	0.7852
LoRA Fine-tuning	0.8525	0.7536	0.7288	0.7410

The Direct Fine-tuning approach achieved higher overall accuracy (0.8755 vs 0.8525) and better performance on the 'evaluative' class (F1 score 0.7852 vs 0.7410) compared to LoRA Fine-tuning. This represents a difference of approximately 4.4 percentage points in the 'evaluative' F1 score.

While both models demonstrate reasonable performance, direct fine-tuning appears more effective for this task, suggesting the full model adaptation captures the nuances better than the low-rank approximation of LoRA.

6.2 Confusion Matrix

Figures 2 and 3 show the confusion matrices. The test set contains 579 evaluative samples and 1421 non-evaluative samples.

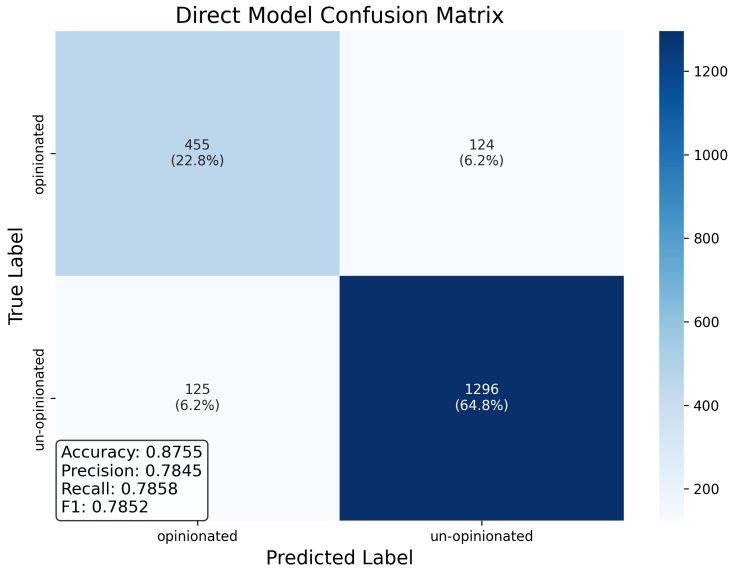

Figure 2: Direct Model Confusion Matrix on Test Set

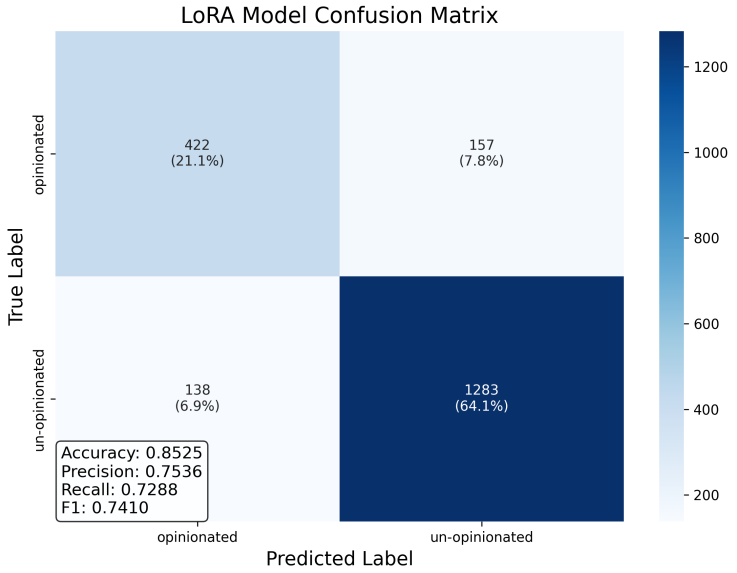

Figure 3: LoRA Model Confusion Matrix on Test Set

The Direct model exhibits fewer misclassifications in both directions (FN: 124 vs 157, FP: 125 vs 138), confirming its superior performance. Both models show a notable number of False Negatives, indicating difficulty in identifying some evaluative text.

6.3 Performance by Topic

Figures illustrating the F1 scores per topic (Figures 4 and 5) show considerable variation across the 20 Newsgroups topics.

The average F1 score across topics for the Direct model is approximately 0.779, while for the LoRA model it is approximately 0.651. For most topics, the LoRA model lags only slightly behind. However, for the least performing 4 topics, there is a significant increase in the performance gap, hinting at potential mechanisms related to the LoRA adapter's properties or similarities among these topics.

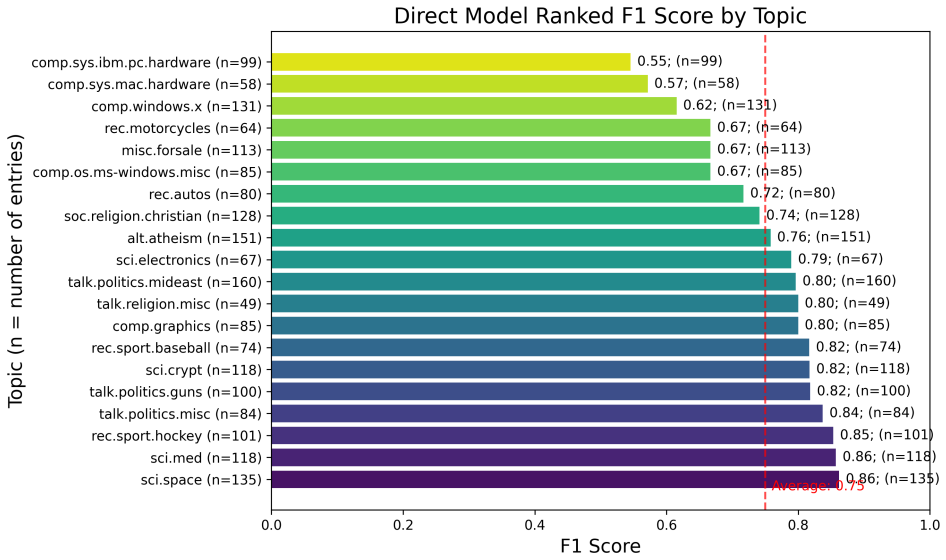

Figure 4: F1 Scores by Topic for Direct Model

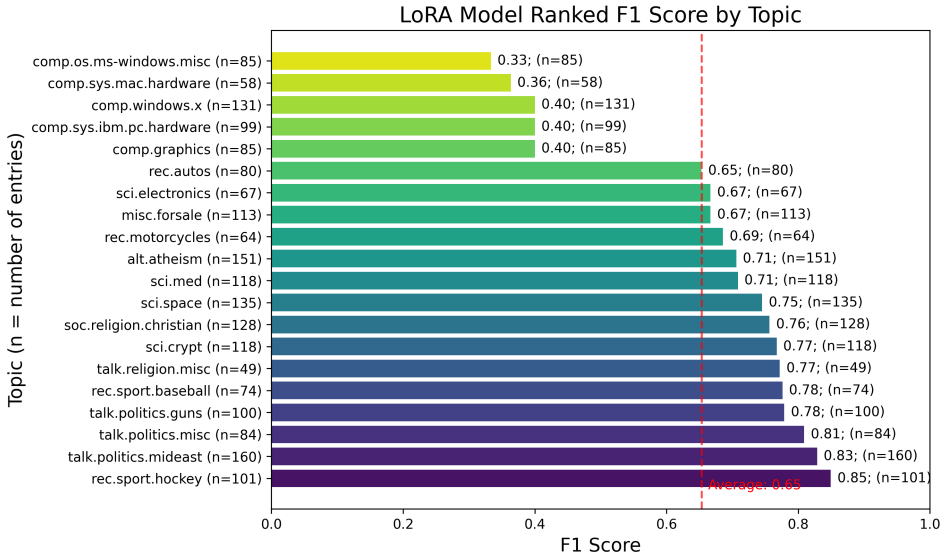

Figure 5: F1 Scores by Topic for LoRA Model

Performance is generally higher on topics like 'sci.space' (Direct F1: 0.8621, LoRA F1: 0.7451), 'sci.med' (Direct F1: 0.8571, LoRA F1: 0.7083), and 'rec.sport.hockey' (Direct F1: 0.8533, LoRA F1: 0.8493). Conversely, topics related to politics and religion such as 'comp.sys.ibm.pc.hardware' (Direct F1: 0.5455, LoRA F1: 0.4000), 'comp.sys.mac.hardware' (Direct F1: 0.5714, LoRA F1:

0.3636), ‘comp.windows.x‘ (Direct F1: 0.6154, LoRA F1: 0.4000), and ‘talk.religion.misc‘ (Direct F1: 0.8000, LoRA F1: 0.7719) show lower F1 scores.

6.4 Result Analysis

The manual analysis (Section 9.1) confirmed the LLM's capability in applying the definition, though occasional errors occurred. Out of 500 samples examined, 9 had incorrect labels, and a similar number had correct labels but incorrect reasoning. Errors were more common with complex text.

Evaluation on a test set showed promising results. The Direct fine-tuned model achieved an 'evaluative' F1 score of 0.7852 (overall accuracy 0.8755). The LoRA model underperformed it, achieving an 'evaluative' F1 of 0.7410 (overall accuracy 0.8525). While LoRA offers parameter efficiency and flexibility in terms of swapping adapters, direct fine-tuning provided better performance for this task's subtle distinctions. Performance varied across topics, with political and religious domains posing greater challenges.

Performance variation by topic highlights the domain dependency. The topics within close proximity in content also yielded relatively close classification results [1]. This underscores the need for potentially domain-specific models or adaptation strategies for optimal performance across diverse text types.

7 Limitations and Future Work

7.1 Limitations

- **LLM Training Data Noise:** The training data quality is limited by potential inaccuracies and biases in the LLM-generated labels, especially for a nuanced definition.
- **Definition Specificity:** The tailored definition of 'evaluative' might not be universally applicable without adaptation.
- **Dataset Domain:** Performance is likely tied to the characteristics of the 20 Newsgroups dataset; cross-domain generalization may vary.
- **Resource Constraints:** Limited compute restricted extensive hyperparameter tuning or the use of larger base models.
- **Binary Scope:** The model only provides a binary classification, lacking finer-grained distinctions within evaluative or non-evaluative text.
- **Human Review Scale:** The test set, while high-quality, is limited in size, making granular analysis (e.g., per-topic breakdown) less robust for some topics.
- **Choice of Base Model:** The use of a masked language model (DistilRoBERTa), designed for autoencoding tasks, might not be optimal compared to autoregressive LLMs that excel at understanding nuances and generating text, potentially impacting classification performance.

7.2 Future Work

Building on this work, future research could explore:

- **Improved LLM Labeling:** Experiment with larger LLMs, advanced prompting (e.g., chain-of-thought), or active learning to refine LLM training data quality.
- **Expanded Human Data:** Increase the size and diversity of the test set and potentially create a larger partially human-annotated training set for better evaluation and model training.
- **Cross-Domain Evaluation:** Test models on diverse datasets (social media, news) to assess generalization and apply domain adaptation techniques if needed.

[1] While no dependency between topic and classification ease can be definitively drawn from this single result, it's worth noting this outcome differs from previous observations using a different model (RoBERTa instead of DistilRoBERTa).

- **Multi-Class Modeling:** Develop models to distinguish more categories like 'Evaluative Judgment', 'Objective Fact', and 'Subjective Experience'.
- **PEFT Comparison:** Conduct a comprehensive study comparing various PEFT methods and configurations against direct fine-tuning on this task, including resource usage analysis.
- **Explainability:** Use techniques to understand which text features drive the classification decisions, particularly how the model differentiates judgment from experience.
- **Pipeline Integration:** Integrate the classifier into larger systems for opinion mining, bias detection, or automated reasoning by segmenting text and classifying units.
- **Leverage Subjectivity Detection:** Explore using an existing, robust subjectivity detector to filter or refine the LLM-generated labels, given the hierarchical relationship between broad subjectivity and the 'evaluative' distinction.

8 Conclusion

This study successfully defined and addressed the task of classifying text as evaluative vs. non-evaluative. We leveraged LLM-assisted labeling to create training data from the 20 Newsgroups dataset and fine-tuned a DistilRoBERTa-base model using both standard direct fine-tuning and LoRA.

This work contributes a methodology for tackling classification tasks with nuanced definitions lacking large gold-standard datasets by effectively using LLM-generated data. The ability to distinguish subjective judgments from other text forms is valuable for various downstream NLP applications, representing a step towards more sophisticated text understanding systems. Future work should focus on refining data quality, expanding evaluation, exploring multi-class distinctions, and further analyzing resource-performance trade-offs of PEFT methods.

9 Appendix

9.1 Selected Example of Manual Analysis of LLM-Generated Labels

Example 1:

- **Original Text:** "YHVH sends forth your mighty scepter from Zion. Rule in the midst of your foes Your people will offer themselves freely on the day you lead your host on the holy mountains."
- **LLM Classification:** non-evaluative
- **LLM Reasoning:** "The text is a passage that appears to be a religious or scriptural statement... It does not express a personal evaluative judgment or subjective stance. Instead, it presents descriptive or declarative content... Therefore, it is non-evaluative."
- **Human Classification:** Non-evaluative.
- **Synthetic Label Accuracy:** Good.
- **LLM Reasoning Correctness:** Correct. This is declarative or descriptive text within a specific domain (scripture), asserting beliefs or narratives rather than providing an evaluation or judgment of something.

Example 2:

- **Original Text:** "To what follows, our moderator has already answered the charge of arrogance more ably that I could have done so, so I will confine myself to answering the charge of illogic."
- **LLM Classification:** non-evaluative
- **LLM Reasoning:** "The text is responding to charges... It expresses a position regarding how the author intends to respond... it is more of a procedural statement about the discussion... Therefore, it is presenting a factual stance about discussion proceedings rather than an opinion on a topic."
- **Human Classification:** Non-evaluative.

- **Synthetic Label Accuracy:** Good.
- **LLM Reasoning Correctness:** Correct. The text is primarily procedural, outlining the author's planned response strategy within a discussion context, rather than offering an evaluation of a broader topic.

Example 3:

- **Original Text:** "Mission specialist astronauts, working with the commander and pilot, have overall responsibility for the coordination of Shuttle operations... Mission specialists are required to have a detailed knowledge of Shuttle systems... Mission specialists will perform extra-vehicular activities..." (Full text describing roles and duties)
- **LLM Classification:** non-evaluative
- **LLM Reasoning:** "The paragraph provides factual information about the roles and responsibilities... without expressing any evaluative judgment or opinion. It describes duties and knowledge required, which are objective facts... rather than a personal or subjective assessment."
- **Human Classification:** Non-evaluative.
- **Synthetic Label Accuracy:** Good.
- **LLM Reasoning Correctness:** Correct. This is a purely descriptive account of job responsibilities, factual and objective.

Example 4:

- **Original Text:** "It could be, perhaps, your head wasn't screwed on just right. In 1941, while the Jews were being assembled for their doom in the Nazi concentration camps, the Armenian volunteers in Germany formed the first Armenian battalion... under the command of the former guerilla leader Dro (the butcher)... and the architect of the cold-blooded genocide of 2.5 million Turks and Kurds... An Armenian National Council was formed by the notorious Dashnak Party leaders in Berlin... Encouraged by this, the Armenians summarily formed a provisional government that endorsed and espoused fully the principles of the Nazis..." (Text containing terms like "the butcher", "cold-blooded genocide", "notorious")
- **LLM Classification:** evaluative
- **LLM Reasoning:** "The paragraph presents historical information and claims... However, the language used includes subjective and emotionally charged terms such as "the butcher," "notorious," and "cold-blooded genocide," which indicate evaluative judgments and negative stances... These terms reflect evaluative viewpoints rather than neutral factual reporting."
- **Human Classification:** Evaluative.
- **Synthetic Label Accuracy:** Good.
- **LLM Reasoning Correctness:** Correct. The reasoning correctly identifies the strong, emotionally charged, and evaluative language as indicative of an evaluative stance, aligning with the definition of evaluative judgment used in this work.

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.

B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 371–378, 2005. URL http://www.cs.cornell.edu/people/pabo/movie-review-data/rotten_imdb.tar.gz.

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.