Stance and Target Detection in Text using Large Language Models

April 15, 2025

Contents

Abstract

Stance detection (SD) aims to identify the expressed attitude towards a specific target within a text. Traditional methods often assume the target is known. This project addresses the more challenging Open-Target Stance Detection (OTSD) task, where the target must first be identified from the text before determining the stance, without prior knowledge of potential targets. We investigate the capabilities of Large Language Models (LLMs), specifically a Llama-based architecture, for this task. Following methodologies similar to recent research, we finetuned a base Llama model on a combined dataset derived from TSE and VAST. Evaluation was performed using both direct string matching and semantic evaluation assisted by other LLMs (Gemini, DeepSeek). Our results show that while the base model struggled with target identification (3% direct match accuracy), finetuning significantly improved this capability (17% direct match accuracy). Semantic evaluation revealed substantial gains in target matching (from 37.82% to 57.09%) after finetuning, with a modest improvement in stance detection (62.52% to 63.07%). This work highlights the potential of finetuning LLMs for the OTSD task, particularly for improving target identification.

1 Introduction

Stance Detection (SD) is a crucial task in natural language processing, focusing on determining the viewpoint (e.g., favor, against, neutral) expressed in a text towards a given entity or topic, known as the target [1]. Many real-world applications, such as social media analysis, opinion polling, and argument mining, benefit from accurate stance detection.

However, a significant limitation of many existing SD approaches is the requirement that the target of the stance be provided beforehand. This assumption does not hold in many practical scenarios where the target itself needs to be identified from the text. This leads to the more complex and realistic task of Open-Target Stance Detection (OTSD), where the model must first identify the relevant target(s) within the text and then determine the stance towards that identified target [1]. The OTSD task presents challenges in both target generation (TG) and subsequent stance classification, especially when targets are not explicitly mentioned.

Recent advancements in Large Language Models (LLMs) have shown promise in various zero-shot and few-shot learning scenarios. Their ability to understand context and generate text makes them potential candidates for tackling the complexities of OTSD. This project investigates the effectiveness of using a Llama-based LLM architecture for the OTSD task. We explore the performance of a base LLM compared to a version finetuned specifically for this task on relevant datasets. Our goal is to assess the impact of finetuning on both target identification and stance classification performance within the OTSD framework. We follow methodologies inspired by recent work in the field [1], adapting them to our specific model architecture and evaluation setup.

2 Methodology

Our approach to Open-Target Stance Detection (OTSD) involves identifying the target and determining the stance expressed towards it within a given text, leveraging the capabilities of Large Language Models (LLMs).

2.1 Model Architecture and Finetuning

We utilized a Llama-based LLM architecture as the foundation for our experiments. Specifically, the base model was a version of Llama 3.1 with 8 billion parameters.

Two primary models were compared:

1. **Base Llama Model:** The pre-trained Llama 3.1 8B model used without any task-specific finetuning. This serves as a baseline to understand the model's inherent capabilities for OTSD in a zero-shot setting.
2. **Finetuned Llama Model:** The same base Llama model was finetuned using an efficient strategy. We employed Parameter-Efficient Finetuning (PEFT) using Low-Rank Adaptation (LoRA). Key aspects of this finetuning approach include:
   - **LoRA Configuration:** Low-rank matrices were introduced with a rank (r) of 16 and an alpha (α) value of 16. This approach significantly reduces the number of trainable parameters compared to full finetuning.

- **Targeted Modules:** The LoRA updates were applied specifically to the query, key, value, and output projection layers within the attention mechanism, as well as the gate, up, and down projection layers of the feed-forward networks.
- **Quantization:** To reduce memory requirements during training and inference, the base model's weights were loaded using 4-bit quantization.
- **Memory Optimization:** An optimized gradient checkpointing technique was used to further conserve memory, allowing for training with longer sequence lengths on available hardware.

The finetuning process aimed to adapt the model's parameters to better handle the nuances of the OTSD task based on the training data.

The finetuning process involved training the model on input texts paired with their corresponding target and stance labels from the training dataset.

2.2 Dataset

To train and evaluate our models, we created a combined dataset leveraging two existing stance detection resources mentioned in related work [1]:

- **TSE Dataset:** A dataset designed for Target-Stance Extraction.
- **VAST Dataset:** A versatile stance detection dataset.

By combining these datasets, we aimed to create a more diverse and comprehensive dataset for the OTSD task. The combined dataset was preprocessed and formatted appropriately for model input.

2.3 Data Split

The combined dataset was split into training and testing sets:

- **Training Set (80%):** Used for finetuning the Llama-based model.
- **Testing Set (20%):** Held out and used exclusively for evaluating the performance of both the base and finetuned models. This ensures an unbiased assessment of the models' generalization capabilities.

2.4 Task Formulation

For both models, the task was formulated as follows: Given an input text, the model should generate both the most likely target discussed in the text and the stance (Favor, Against, or None) expressed towards that target.

3 Experimentation

We conducted experiments to evaluate the performance of the base and finetuned Llama models on the OTSD task using the held-out test set.

3.1 Experimental Setup

The experiments were run using the pre-defined test split (20

3.2 Evaluation Metrics

Evaluating OTSD requires assessing both the quality of the generated target and the accuracy of the predicted stance. We employed two distinct evaluation strategies:

1. **Direct Target Matching Accuracy:** This metric calculates the percentage of test samples where the exact string of the predicted target precisely matches the ground truth target string. While simple to compute, this metric is very strict and does not account for semantic similarity (e.g., "climate change action" vs. "acting on climate change"). It primarily measures the model's ability to replicate the exact target phrasing from the ground truth.

2. **LLM-based Semantic Evaluation:** Recognizing the limitations of direct matching, we employed other powerful LLMs (specifically, models from the Gemini and DeepSeek families) as evaluators. For each test sample, the evaluator LLM was provided with:

    - The input text
    - The ground truth target
    - The ground truth stance
    - The model's predicted target
    - The model's predicted stance

    The evaluator LLM was then prompted to assess the semantic match between the predicted and ground truth targets, and separately, the match between the predicted and ground truth stances. This resulted in two scores per sample: a Target Match Percentage and a Stance Match Percentage. The final reported scores are the average match percentages across the entire test set. This method provides a more nuanced evaluation of semantic correctness.

3.3 Models Compared

The core comparison was between:

- The performance of the **Base Llama Model** (zero-shot performance).
- The performance of the **Finetuned Llama Model** (performance after task-specific training).

4 Results and Analysis

The evaluation yielded distinct results for direct matching and semantic evaluation, highlighting the impact of finetuning.

4.1 Direct Target Matching

The direct target matching accuracy provides a baseline understanding of how often the models generated the exact target string.

Table 1: Direct Target Matching Accuracy (%)

Model	Direct Match Accuracy (%)
Base Llama Model	3%
Finetuned Llama Model	17%

As shown in Table 1, the direct matching accuracy was low for both models, which is expected given the strictness of the metric and the variability in how targets can be phrased. However, finetuning resulted in a notable increase, from 3% to 17%, indicating that the finetuned model learned to generate targets closer to the exact ground truth phrasing more often than the base model.

4.2 LLM-based Semantic Evaluation

The semantic evaluation provides a more meaningful assessment of the models' understanding and generation capabilities.

Table 2: LLM-based Semantic Evaluation Scores (%)

Model	Average Stance Match (%)	Average Target Match (%)
Base Llama Model	62.52%	37.82%
Finetuned Llama Model	63.07%	57.09%

Table 2 presents the average match percentages as determined by the evaluator LLMs. Key observations include:

- **Target Matching Improvement:** Finetuning led to a substantial improvement in the semantic accuracy of target generation, increasing the average target match score from 37.82% to 57.09%. This is a significant gain (+19.27 points) and suggests that the finetuning process effectively taught the model to identify the core semantic target of the text, even if not phrased identically to the ground truth.
- **Stance Matching Improvement:** The improvement in stance detection accuracy was much more modest, increasing slightly from 62.52% to 63.07% (+0.55 points). This indicates that while finetuning greatly helped with identifying *what* the stance was about (the target), it had a minimal impact on correctly classifying the stance itself (Favor, Against, None) compared to the base model's inherent capabilities.
- **Discrepancy between Metrics:** The large difference between direct match scores (3%/17%) and semantic target match scores (38%/57%) underscores the importance of using semantic evaluation for generative tasks like target identification. Models often capture the correct meaning without using the exact wording.

4.3 Analysis

The results strongly suggest that finetuning a Llama-based architecture on a combined TSE and VAST dataset significantly enhances its ability to perform the target generation aspect of the Open-Target Stance Detection task. The base model, while capable of reasonable stance classification given a target (implied by the 62.52% stance match), struggled significantly with identifying the correct target semantically (37.82%). Finetuning addressed this weakness considerably (57.09

The minimal improvement in stance detection post-finetuning could suggest several possibilities: a) the base model was already quite competent at stance classification given a reasonably identified target, b) the finetuning data or process might have been more effective for the TG sub-task than the SD sub-task, or c) stance detection itself might be more challenging or require different signals than target identification within this framework.

5 Conclusion

This project investigated the application of a Llama-based LLM to the Open-Target Stance Detection (OTSD) task, comparing a base model against a model finetuned on a combined TSE and VAST dataset. Our findings demonstrate that while the base LLM possesses some inherent capability for stance detection, it struggles with accurately identifying the target of the stance in an open setting without prior examples.

Finetuning proved highly effective in improving the model's target generation capabilities, as evidenced by the significant increase in semantic target matching scores (from 37.82% to 57.09%). The improvement in stance detection accuracy was marginal, suggesting that the primary benefit of finetuning in our setup was related to target identification. The discrepancy between low direct matching scores and higher semantic matching scores highlights the need for semantic evaluation methods in OTSD.

Future work could explore different LLM architectures, investigate more sophisticated finetuning strategies specifically targeting both TG and SD, incorporate larger or more diverse datasets, and refine the LLM-based evaluation protocol for potentially even more accurate assessment. Overall, our results confirm the potential of finetuned LLMs as a viable approach for addressing the challenges of Open-Target Stance Detection.

References

[1] Abu Ubaida Akash, Ahmed Fahmy, and Amine Trabelsi. Can Large Language Models Address Open-Target Stance Detection? *arXiv preprint arXiv:2409.00222*, 2024. `https://arxiv.org/abs/2409.00222`