Saivineetha at CheckThat! 2025: Exploring Fine-Tuning and Zero-Shot Approaches for Claim Normalization*

Notebook for the CheckThat! Lab at CLEF 2025

Baddepudi Venkata Naga Sri Sai Vineetha[1,†]

[1]*Independent Contributor (TCS Research)*

Abstract

This paper presents our participation in the CLEF 2025 CheckThat! Lab's Task 2 which focuses on claims extraction and normalization. This task aims to decompose social media posts into simpler, comprehensible forms. The task spans across 20 languages - English, Arabic, Bengali, Czech, German, Greek, French, Hindi, Korean, Marathi, Indonesian, Dutch, Punjabi, Polish, Portuguese, Romanian, Spanish, Tamil, Telugu, Thai. Our study focuses on two different languages, Hindi and Telugu. Our approach involves Parameter-Efficient Fine-Tuning (PEFT) on multi-lingual Large Language Model (LLM) for Hindi dataset and zero-shot inferencing for Telugu dataset. Our proposed method is ranked third in Hindi with METEOR score of 0.2996 and fourth in Telugu with METEOR score of 0.3774 in the organizer's leaderboard.

Keywords

Large Language Model (LLM), Claim Normalization, Parameter-Efficient Fine-tuning (PEFT), Zero-shot inference

1. Introduction

Social media platforms are rife with misinformation, often embedded in the posts that are verbose and ambiguous. It is challenging for human fact checkers to verify claims from social media posts. Moreover, false news spreads faster than the capacity of fact-checkers to verify the claims in them. Therefore, automated systems are needed to increase the efficiency of fact-checkers. Several studies have explored the downstream tasks present in fact-checking pipeline such as detecting claims, evaluating the worthiness for fact-checking etc. Claim normalization involves simplifying the social media posts into more understandable forms to identify the core claims present in them [1].

In this paper, we detail our approach to training large language models for generating normalized claims from the given noisy, unstructured social media posts, specifically for Task 2 of CLEF 2025 CheckThat! Lab [2] [3] [4]. This task involves generating the simple and concise claims for the given social media posts in 20 different languages.

In this paper, we detail our approach for generating normalized claims for two languages Hindi and Telugu. For Hindi, our approach is to perform PEFT fine-tuning [5] of multi-lingual LLMs on the dataset provided. PEFT allows to fine-tune LLMs LLaMa, Gemma, Mistral etc for multilingual tasks and for domain specific tasks. We employed Gemma 2 9B [6] model. For Telugu, we performed zero-shot prompting of LLMs using Gemma 3 12B [7] model. Gemma model's ability to handle diverse languages makes them a suitable choice for this multilingual claim normalization task, enabling us to generate concise normalized claims for the given social media posts. In the task's official leaderboard, we are ranked third (out of 11) in Hindi and fourth (out of 9) in Telugu.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[†] This work was conducted independently and does not represent the views or positions of TCS Research.

saivineetha1998@gmail.com (B. V. N. S. S. Vineetha)

https://www.linkedin.com/in/bvnssaivineetha/ (B. V. N. S. S. Vineetha)

0009-0005-8421-9310 (B. V. N. S. S. Vineetha)

2. Related Work

Sundriyal et al. [1] introduced the novel task of Claim Normalization which aims to simplify the noisy social media posts into a simple, concise form. The authors proposed CACN, an approach that leverages chain-of-thought prompting and claim worthiness estimation to interpret complex claims. The authors have also introduced CLAN dataset which has more than 6k social media posts and their normalized claims. Papageorgiou et al. [8] employed pre-trained LLM to extract factual sentences from news text. Fact and claim extractions helps in fact-checking in news articles. They also proposed an approach using graph convolutional networks to capture more complex relations from the text. Wang et al. [9] introduced a task of claim clarification. Claim clarification involves rewriting the ambiguous parts of claims thereby enhancing the content and removing redundant information. They evaluated the performance of claim clarification task across various LLMs. The authors proposed a semantic evaluation approach based on sliding window.

3. Methodology

3.1. Datasets

The dataset for Claim Normalization task comprises of noisy, unstructured social media post and their normalized claims. The task comprises of two settings - monolingual and zero-shot. Monolingual consists of training, development and test datasets provided for specific languages. This approach ensures that the model learns the language specific patterns. The datasets of English, German, French, Spanish, Portugese, Hindi, Marathi, Punjabi, Tamil, Arabic, Thai, Indonesian, and Polish follow this setting. Zero-shot has only test data and no training or development datasets are provided. This approach evaluates the performance of the model to unseen languages. This includes datasets of languages Dutch, Romanian, Bengali, Telugu, Korean, Greek, and Czech.

For this task, we conducted experiments on languages Hindi and Telugu. For Hindi, the dataset is monolingual. Train data has 1081 posts and their normalized claims, development dataset has 50 posts, while test dataset has about 100 posts. For Telugu, the dataset is zero-shot. It has 116 posts.

3.2. Proposed Approach

In this research, inspired by the performance of multi-lingual LLMs we choose Gemma 2 -9B model to perform fine-tuning for Hindi and Gemma 3 -12B model for zero-shot prompting for Telugu. In this study, we utilized the multilingual Gemma 2 9B model for fine-tuning on the Hindi dataset. This model was selected due to its strong support for multiple languages and its compatibility with Kaggle's GPU infrastructure, which enabled efficient fine-tuning within the available computational constraints. For the Telugu dataset, we employed the Gemma 3 12B model in inference-only mode, as the objective was limited to evaluating performance without additional fine-tuning.

3.2.1. Hindi

The Gemma 2 9B instruct model [6] is a multi-lingual model by Google which has been trained on diverse languages. We further fine-tuned this instruct model on Hindi dataset having posts and normalized claims provided for the task. We performed Parameter Efficient Fine-Tuning (PEFT) using quantization with Low Rank Adaptors (QLoRA) with 4-bit quantization. The dataset provided has two columns post and normalized claim. The dataset is formatted with the prompt defined in Fig 1. Table 1 represents the hyperparameters used for fine-tuning.

We performed PEFT fine-tuning on train dataset and evaluation on development dataset.

You are an AI assistant fluent in Hindi, trained to interpret and refine informal content. Your task is to read a noisy or unstructured social media post written in Hindi and convert it into a clear, concise, and structured statement in proper Hindi. Your goal is to provide a highly accurate and reliable answer. If you are not fully confident about the correct answer, clearly state that you are unsure or provide the most likely possibilities. Do not guess or provide potentially incorrect information.

Instruction:

Given a noisy, unstructured social media post in Hindi, simplify it into a concise form in Hindi. In other words, generate the normalized claims for the given social media posts.

Post:

<input post>

Normalized Claim:

Figure 1: Prompt template for instruction fine-tuning for Hindi.

Table 1
Hyperparameter settings used for model fine-tuning

Hyperparameter	Value
LoRA Rank	16
LoRA Dropout	0
LoRA Alpha	16
Batch Size	8
Number of Epochs	7
Learning Rate	2e-4
Optimizer	AdamW
Scheduler	Linear

3.2.2. Telugu

The Gemma 3 12B instruct model [7] is a multi-lingual LLM which excels in various languages and text generation tasks. We performed zero-shot prompting using Telugu dataset having the social media posts. Fig 2 represents the prompt used for inference.

Table 2
Hyperparameter settings used for zero-shot inferencing

Hyperparameter	Value
Max new tokens	256
Temperature	0.1
Top p	0.9

You are an AI assistant fluent in Telugu, trained to interpret and refine informal content. Your task is to read a noisy or unstructured social media post written in Telugu and convert it into a clear, concise, and structured statement in proper Telugu. This final version should express the core message or claim made in the post, without adding or removing any essential meaning.

Do not guess or make assumptions. If the post is unclear or ambiguous, explicitly mention that in Telugu. Provide the most accurate normalized claim possible based only on the information available.

Instruction:

Simplify the following unstructured Telugu social media post into a formal and concise claim in Telugu. Only return the normalized claim, only in Telugu.

Post:

<input post>

Normalized Claim:

Figure 2: Prompt template for zero-shot inference for Telugu.

4. Experimental Results and Analysis

4.1. Baseline Model

We used Google's UMT5 base as baseline model as mentioned in CheckThat! Lab task 2. We fine-tune UMT5 model with Hindi dataset for 5 epochs with learning rate 5e-4. For Telugu dataset, we performed inferencing on UMT5 model using the same prompt as in Fig 2.

4.2. Results and Analysis

The official metric for Claim Normalization competition for Task 2 of CheckThat! Lab is METEOR score. The METEOR score is calculated between the predicted normalized claims available from the model and the test gold outputs which are manually curated. The METEOR score lies between 0 and 1 where 1 indicates the predicted claims more closely match the test gold outputs.

We have used METEOR [10], ROUGE-2 [11], BLEU [12] and cosine similarity metrics to compare the predicted and actual test data.

4.2.1. Hindi

We evaluated the normalized claims obtained from the fine-tuned model and compared the metrics against the test gold outputs released for this task. We perform comparison of baseline model Google's UMT5 base which is finetuned on Hindi dataset and our proposed model.

Table 3 shows the comparison of the baseline model with the proposed model for Hindi language. The METEOR score of the proposed model increased by 9 times, ROUGE-2 score approximately increased by 3 times. BLEU score increased by 10% and cosine similarity by 2 times.

Table 3
Comparison of baseline and proposed model across evaluation metrics for Hindi language

Model	METEOR	ROUGE-2	BLEU	Cosine Similarity
Baseline Model (Fine-tuned UMT5 base model)	0.101	0.097	3.431e-156	0.231
Proposed Model	0.938	0.276	3.436e-157	0.436

4.2.2. Telugu

We evaluated the normalized claims obtained from zero-shot inference and compared the metrics against the test gold outputs released for this task. We performed inference on Google's UMT5 base model as baseline results for Telugu language. We perform the comparison between these and the zero-shot inference with the proposed prompt.

Table 4 shows the comparison of the baseline model with the inference performed for Telugu language. The METEOR score of the proposed model increased approximately by about 3 times, ROUGE-2 score increased by 2 times, BLEU score increased by 10% and cosine similarity by approximately 1.5 times.

Table 4
Comparison of baseline and proposed model across evaluation metrics for Telugu language

Model	METEOR	ROUGE-2	BLEU	Cosine Similarity
Baseline Model (UMT5 base model)	0.064	0.084	2.224e-156	0.282
Proposed Model	0.356	0.210	2.981e-157	0.398

5. Conclusion

This study aims to build a model that helps to normalize claims from social media posts as part of CheckThat! Lab 2025. We explored fine-tuning LLM on Hindi language for new dataset. This highlights the capabilities of LLM models when trained on new domain or language such as Hindi language in this task. We have performed zero-shot inferencing on Telugu dataset. This shows the capabilties of zero-shot inference on LLMs using clear and structured prompts. Our submissions attained third place in Hindi with METEOR score of 0.2996 and fourth in Telugu with METEOR score of 0.3774.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] M. Sundriyal, T. Chakraborty, P. Nakov, From chaos to clarity: Claim normalization to empower fact-checking, arXiv preprint arXiv:2310.14338 (2023).

[2] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The CLEF-2025 CheckThat! Lab: Subjectivity, Fact-Checking, Claim Normalization, and Retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.

[3] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. Venktesh, Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, Fact-Checking, Claim Normalization, and Retrieval, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[4] M. Sundriyal, T. Chakraborty, P. Nakov, Overview of the CLEF-2025 CheckThat! Lab Task 2 on Claim Normalization, 2025.
[5] D. K. Gajulamandyam, S. Veerla, Y. Emami, K. Lee, Y. Li, J. S. Mamillapalli, S. Shim, Domain Specific Finetuning of LLMs Using PEFT Techniques, in: 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2025, pp. 00484–00490.
[6] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al., Gemma 2: Improving open language models at a practical size, arXiv preprint arXiv:2408.00118 (2024).
[7] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al., Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025).
[8] E. Papageorgiou, I. Varlamis, C. Chronis, Harnessing Large Language Models and Deep Neural Networks for Fake News Detection, Information 16 (2025) 297.
[9] Y. Wang, B. He, X. Chen, L. Sun, Can LLMs Clarify? Investigation and Enhancement of Large Language Models on Argument Claim Optimization, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 4066–4077.
[10] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
[11] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013.
[12] K. Papineni, S. Roukos, T. Ward, W. jing Zhu, Bleu: a method for automatic evaluation of machine translation, 2002, pp. 311–318.