CUFE at StanceEval2024: Arabic Stance Detection with Fine-Tuned Llama-3 Model

Michael Ibrahim
Computer Engineering Department, Cairo University
1 Gamaa Street, 12613
Giza, Egypt
michael.nawar@eng.cu.edu.eg

Abstract

In NLP, stance detection identifies a writer's position or viewpoint on a particular topic or entity from their text and social media activity, which includes preferences and relationships. Researchers have been exploring techniques and approaches to develop effective stance detection systems. Large language models' latest advancements offer a more effective solution to the stance detection problem. This paper proposes fine-tuning the newly released 8B-parameter Llama 3 model from Meta GenAI for Arabic text stance detection. The proposed method achieved a Macro average F_1 score of 0.7647 in the StanceEval 2024 Task on stance detection in Arabic language (Alturayeif et al., 2024).

1 Introduction

The transformer architecture has led to the creation of vast self-supervised neural networks with tens to hundreds of billions of parameters, trained on trillions of tokens, revolutionizing natural language processing. Recent models based on the transformer architecture include GPT series (Brown et al., 2020), LLaMA (Touvron et al., 2023), and others. As of late 2024, the release of LLaMA-3 sparked the growth of a thriving open-source community, resulting in the development of near-state-of-the-art models like Mistral (Jiang et al., 2023) and Mixtral (Jiang et al., 2024).

With minimal finetuning, these LLMs can adjust to new tasks and instructions. The NLP community utilizes PEFT techniques for cost-effective fine-tuning instead of full-parameter tuning. These techniques in PEFT can significantly reduce a model's computational footprint, as per the study (Houlsby et al., 2019). LoRA (Low-Rank Adaptation) (Hu et al., 2021) efficiently balances fine-tuning performance and efficiency using low-rank matrix approximations. Through quantization techniques, LLM model weights compression reduces storage and memory usage during training and inference. Using NormalFloat Quantization (Chen et al., 2023), these methods can maintain model performance while significantly decreasing LLMs' resource needs, enabling resource-effective fine-tuning via integration with PEFT methods such as LoRA.

LLMs are capable of stance detection and various NLP tasks through their adaptable prompting methods. Their advanced language skills allow them to accurately align textual context with target subjects, mirroring the author's intended stance (Chowdhery et al., 2023). Traditional NLP model training methods differ significantly from prompting strategies. Using these strategies, LLMs can make predictions with less fine-tuning, demonstrating their adaptability for various tasks. This change in approach simplifies the application of LLMs and broadens their usefulness, excelling at tasks involving intricate language subtleties like stance detection (Bang et al., 2023).

In this paper, a 8-Billion parameter LLAMA-3 model was fine-tuned for Arabic stance detection, the rest of the paper is organized as follows. The related work is summarized in Section 2. The dataset used for training and validation was detailed in Section 3. In Section 4, the system is presented. Section 5 summarizes the study's key findings.

2 Related Work

2.1 Arabic Stance Detection

The paper (Alzanin et al., 2023) suggests utilizing a Genetic Algorithm (GA) and ensemble learning for an efficient Arabic multilabel classification method. The study examines the impact of Arabic text representation on classification by utilizing Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques. It compares the Logistic Regression Classifier with both single

Proceedings of The Second Arabic Natural Language Processing Conference, pages 807–810

and ensemble methods, specifically the Extra Trees Classifier and Random Forest Classifier. According to the experimental findings on the MAWQIF dataset, the suggested strategy surpasses related methods with F_1 score of 80.88% for sentiment analysis and 68.76% for stance detection. The F_1 score result of the classifier increased from 65.62% to 68.80% due to data augmentation and feature selection. The study demonstrates that GA-based feature selection, in conjunction with ensemble learning, enhances Arabic stance detect.

The paper (Haouari and Elsayed, 2024) worked on identifying the stance towards authorities on Twitter rumors, classifying these as either supporting the rumor, denying it, or taking no stance. They developed the AuSTR (Authority Stance towards Rumors) dataset, obtained from Arabic Twitter's authority timelines for this task. Then, they evaluated the effectiveness of using existing Arabic stance datasets to expand AuSTR and train BERT models for this task. A model solely trained on AuSTR with a class-balanced focus loss performs as well as existing datasets supplemented with AuSTR, with an average Macro F_1 score of 0.84.

The paper (Alturayeif et al., 2023) introduces novel multi-task learning models, Parallel Multi-Task Learning (PMTL) and SequentialMulti-Task Learning (SMTL), for stance detection that include sentiment analysis and sarcasm detection tasks. They proposed and showed the effectiveness of various task weighting methods within theses models. This work is the initial effort employing sarcasm detection for enhancing stance detection and introducing task weighting in a multi-task stance detection model. The performance difference between PMTL and SMTL architectures is insignificant. The advantages of auxiliary tasks is emphasized as a solution to data scarcity, but their impact can fluctuate depending on the dataset and task. The SMTL with hierarchal weighting (SMTL-HW) model, which was extensively tested, delivers top-tier performance. The SMTL-HW model, with its sequential architecture and hierarchical task weighting, has been empirically shown to be effective. These features enhance task representations, boosting stance detection performance. The proposed SMTL-HW achieved a Macro average F_1 score of 0.6865 on the MAWQIF dataset, while other methods like SMTL and PMTL achieves an average Macro F_1 score of 0.6266 and 0.6421 respectively.

2.2 LLM Applications

Currently, LLMs are being extensively used for various NLP tasks. The optimal use of these models remains undetermined. Three methods are used for building applications with LLMs.

- **Zero-shot prompting** It is querying prompts that aren't in LLM's training data to elicit responses. These prompts typically consist of elaborate directions and a major query. Creating accurate prompts is crucial for optimizing the performance of large language models.
- **Few-shot learning** Few-shot learning involves giving LLMs a limited number of examples to produce appropriate responses. Zero-shot prompting refers to a method without providing any examples. In few-shot learning, examples are incorporated into the prompt template to guide the model's response.
- **Fine-tuning**. Task adaptation can be achieved using the methods mentioned above without requiring additional training on LLMs, while fine-tuning necessitates further training using task-specific data. This method is most effective when using tailored datasets.

According to (Yang et al., 2024), LLMs perform effectively in most NLP tasks based on their examination of 'use cases' and 'no use cases' for particular downstream tasks using the mentioned methods.

3 Mawqif Dataset

The MAWQIF dataset (Alturayeif et al., 2022) was for fine-tuning the LLama 3 model, this dataset differs from previously published datasets as it covers Middle Eastern hot topics beyond, unlike other previously published datasets that prioritize political topics like elections and referendums over a broader range. the initial considerations. Mawqif explores three hot topics: COVID-19 vaccine, digital transformation, and women empowerment. The Twitter platform was the source of the collected dataset. Three to seven annotators were responsible for annotating each tweet–target pair. Annotation of a row stops when its confidence score exceeds 0.7 or when it already has seven annotations.

The MAWQIF dataset consists of 4,121 tweets classified under three categories: COVID-19 vaccine (1,373), digital transformation (1,348), and

women empowerment (1,400). Each tweet in this dataset is tagged for its stance, sentiment, and sarcasm. Table 1 presents select examples from the MAWQIF dataset.

The distribution of labels differs among the three targets. Digital transformation discussions on Twitter usually express a positive outlook. Women empowerment and digital transformation topics exhibit a higher frequency of positive sentiment compared to the COVID-19 vaccine topic, which only has 25% positive tweets. In COVID19 vaccine tweets, sarcasm is used more often. 34% of seemingly positive tweets are actually sarcastic, while 31% of negative tweets exhibit a non-negative sentiment.

4 Methodology

The Llama-3 model was finetuned using formatted prompt-response triplets: Target topic [target], Arabic Tweet [text] and Stance [stance]. The dev set, like the training set, is formatted similarly. This prompt customizes the model for the DA-MSA machine translation assignment. The prompt is formatted as follows:

<s> [INST] «SYS» You are an expert Arabic stance analyzer! «/SYS» Consider the subsequent Arabic passage which discusses [target] and determine if it is Favor, None, or Against, and return the answer as the corresponding stance label "Favor" or "None" or "Against". [text] [/INST] [stance] </s>

The learning rate for finetuning the Llama-3 model was set to 1e-4 with the Adam optimizer used during training. The evaluation is conducted every 50 steps with a batch size of 1. The low-rank approximation rank is set to 64, and its scaling factor for adaptation is set to 16 in the LoRA configurations. The model trainable parameters are all linear layers: "q_proj", "up_proj", "o_proj", "k_proj", "down_proj", "gate_proj", and "v_proj". A 0.05 dropout is applied in the LoRA layer. The model's weights are quantized to 4-bit precision and mixed-precision training with float16 and float32 is enabled to reduce memory requirements and accelerate the training process. This model was trained on Google Colab with a single NVIDIA A100 GPU with 40GB of memory. The finetuning of the model and the generation of the test results took almost one hour. The code used for training and generating the results can be found on the following GitHub repository. [1]

The proposed system achieved an average Marcro F_1 score of 0.7647 on the testing dataset, and it attained an F_1 score of 0.8363 for detecting stance in women empowerment Arabic tweets, an F_1 score of 0.8038 for detecting stance in Covid-19 vaccine Arabic tweets, and an F_1 score of 0.6539 for detecting stance in digital transformation Arabic tweets.

5 Conclusion

Stance detection reveals valuable insights into public opinion, trends, and behaviors across various topics. LLMs' text processing and analysis skills make them essential in social media navigation. Through fine-tuning, these models become more precise and adaptable, essential for understanding and engaging with societal discourse in meaningful ways. In this paper, PEFT methods like quantization and LoRA were used to finetune a LLama-3 model with 8 billion parameters model, this model was adapted to detect stances in Arabic tweets. The proposed method is efficient and it achieved an average Macro F_1 score in the StanceEval 2024 Task on stance detection in Arabic language.

References

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. Enhancing stance detection through sequential weighted multi-task learning. *Social Network Analysis and Mining*, 14(1):7.

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. Stanceeval 2024: The first arabic stance detection shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label arabic dataset for target-specific stance detection. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Samah M Alzanin, Abdu Gumaei, Md Azimul Haque, and Abdullah Y Muaad. 2023. An optimized arabic multilabel text classification approach using genetic algorithm and ensemble learning. *Applied Sciences*, 13(18):10264.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei

[1] https://github.com/MichaelIbrahim-GaTech/StanceEval2024.git

Target	Tweet	Stance	Sentiment	Sarcasm
Covid-19 Vaccine	حاشتنا كورونا وطبنا منها ولله الحمد ومانحتاج تطعيم ولاتحسفنا أبدا we were diagnosed with Corona and recovered from it, thank God, we do not need a vaccination and we will never regret it	Against	Positive	No
Digital Trans.	مليون كتاب!! اين التحول الالكتروني للمناهج؟ كمية هدر سنوي للكتب مؤسفة نتمنى احلال الاجهزة اللوحية بدلاً من الكتب Millon books!! Where is the digital transformation of curricula? The amount of annual waste of books is unfortunate. We wish to replace books with tablets	Favor	Negative	No
Women Empow.	#القبض _علي _مدعيه _النبوه فاهمة تمكين المرأة غلط #Arrest_of_the_imposter_of_prophecy she misunderstood women's empowerment	None	Neutral	Yes

Table 1: Examples from the MAWQIF dataset

Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. 2023. Parameter-efficient fine-tuning design spaces. *arXiv preprint arXiv:2301.01821*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Fatima Haouari and Tamer Elsayed. 2024. Are authorities denying or supporting? detecting stance of authorities towards rumors in twitter. *Social Network Analysis and Mining*, 14(1):34.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.