# Dolomites@#SMM4H 2024
# Helping LLMs "*Know The Drill*" in Low-Resource Settings A Study on Social Media Posts

**Giuliano Tortoreto** [†], **Seyed Mahed Mousavi** [‡]
[†] WISECode LLC
[‡] Signals and Interactive Systems Lab, University of Trento, Italy
gtortoreto@wisecode.ai, mahed.mousavi@unitn.it

## Abstract

The amount of data to fine-tune LLMs plays a crucial role in the performance of these models in downstream tasks. Consequently, it is not straightforward to deploy these models in low-resource settings. In this work, we investigate two new multi-task learning data augmentation approaches for fine-tuning LLMs when little data is available: "In-domain Augmentation" of the training data and extracting "*Drills*" as smaller tasks from the target dataset. We evaluate the proposed approaches in three natural language processing settings in the context of SMM4H 2024 competition tasks: multi-class classification, entity recognition, and information extraction. The results show that both techniques improve the performance of the models in all three settings, suggesting a positive impact from the knowledge learned in multi-task training to perform the target task.

## 1 Introduction

Collecting an adequate amount of data to fine-tune Large Language Models (LLM) in low-resource settings is an expensive practice. To address the lack of enough data in such settings, data augmentation techniques have been commonly used in different natural language processing tasks (Wei and Zou, 2019; Feng et al., 2021) as a low-cost solution. Such techniques focus on generating new examples close to the original samples and data distribution, thus increasing the number of training samples.

More recently, data augmentation has been studied from the perspective of multi-task learning. In Multi-Task Learning Data Augmentation (MTL-DA), the dataset of the target task is augmented with the data of other auxiliary tasks that can improve the model performance, despite having different characteristics (Sánchez-Cartagena et al., 2021). The model is then jointly optimized in a multi-task manner on the augmented data, resulting in more robust performance compared to traditional augmentation techniques (Wei et al., 2021).

In this work, we investigate two new MTL-DA techniques to fine-tune LLMs for low-resource tasks: **a) In-domain Augmentation** where we augment the original target dataset with the data (and tasks) collected from the same source (e.g. Reddit Social Forum); and **b) Drills** where we extract from the target task a set of smaller tasks and replicate the dataset to cover the corresponding samples for the extracted drills. We study the proposed techniques on three language processing tasks in the context of SMM4H 2024 challenge[1]: *I) Task 3*: four-class classification of Reddit social forum posts; *II) Task 4*: an entity recognition task on Reddit posts to identify two entity types (Ge et al., 2024); *III) Task 6*: information extraction from tweets and Reddit posts. A complete description of the tasks is presented in §A.

The results indicate the effectiveness of both augmentation techniques by achieving higher $\boldsymbol{F}_1$ scores on the validation sets, compared to target task fine-tuning (note that the competition test sets are not publicly available), as well as scoring two to five points higher than overall scores in SMM4H competition.

## 2 Approach

### 2.1 Proposed Techniques

We experiment with two new MTL-DA techniques to fine-tune LLMs for low-resource settings:

**a) In-domain Augmentation** We jointly fine-tune the model on an augmented dataset, consisting of the data of the target task and other tasks collected from the same source. Here, we leverage the fact that Reddit is the common data source across all tasks. We then optimize/evaluate the performance of the model on each target task.

**b) Drills** We extract smaller tasks for each tar-

[1] SMM4H 2024: "The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks" (Xu et al.)



*Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks*, pages 17–22

| Approach | Task 3 | | | Task 4 | | | Task 6 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R |
| **GPT-4** *(In-Context Learning)* | | | | | | | | | |
| *0-shot* | 46.4 | 47.4 | 55.5 | 57.2 | 63.3 | 57.6 | 88.6 | 86.2 | 91.2 |
| *5-shot* | 50.1 | 51.1 | 62.4 | 59.4 | 56.5 | 67.6 | 85.6 | 80.7 | 91.2 |
| **Mistral** *(fine-tuning)* | | | | | | | | | |
| *Target Task* | 52.7 | 60.3 | 50.3 | 50.5 | 56.3 | 52.6 | **89.5** | 95.0 | 84.6 |
| *Target Task + Drills* | 46.6 | 55.5 | 44.6 | 55.4 | 65.9 | 51.2 | 86.2 | 92.0 | 81.1 |
| *In-domain* | 52.9 | 61.7 | 50.2 | **58.5** | 62.9 | 55.7 | 87.5 | 94.2 | 81.8 |
| *In-domain + Drills* | **53.8** | 57.1 | 52.5 | 56.0 | 60.1 | 53.5 | 85.8 | 94.1 | 78.8 |

Table 1: Evaluation of the proposed MTL-DA techniques (In-domain Augmentation and Drills) on the validation set of each task, compared to In-Context Learning (ICL) in {*zero*,*few*}-shot settings (the test set is not publicly available). While ICL with GPT-4 achieves the highest ***R***ecall in all tasks, MTL-DA approaches generally achieve higher ***P***recision and thus competitive $F_1$ scores. In-domain augmentation of the training set achieves competitive results in all tasks and manages to obtain the highest $F_1$ score in Task 4, the highest score in Task 3 by additional Drills, and the second best $F_1$ score in Task 6.

get task to augment the dataset. Regarding Task 3, we extracted two drills of "*sentiment classification: Pos|Neg|Neut.*" and "*post relevance classification: Y|N*". As a result, the training set was replicated twice to include the additional training examples. Regarding Task 4, we extracted two binary classification drills as "*includes social impact: Y|N*" and "*includes clinical: Y|N*", as well as two sequence extraction drills as "*social impact extraction*" and "*clinical impact extraction*". As a result, the training set was replicated four times to include the additional training examples. Concerning Task 6, we extracted one binary classification drill, in addition to the main task, as "*Source Platform Classification:Reddit|Twitter*", and duplicated the training set.

### 2.2 Model

We experimented with Mistral (7B) (Jiang et al., 2023), a decoder-only LLM. We applied QLoRA (Dettmers et al., 2024) to fine-tune the model, due to the limited GPU memory. Details of our implementation are provided in §B.

## 3 Evaluation

Since the test sets of the competition tasks are not publicly available, we analyze the performance of the model and the impact of the MTL-DA techniques on the validation set. We consider the *vanilla fine-tuning* of the model for the target downstream task as the baseline in our analysis. Furthermore, we compare the performance of the model with GPT-4[2] via In-Context Learning (ICL) in {0,5}-shot settings in each task as a low-cost alternative. We then present the scores obtained by the best-performing models and MTL-DA techniques in SMM4H competition, compared to the overall mean and median scores obtained by all participants.

### 3.1 Validation Set

The performance of Mistral with different combinations of the proposed augmentation techniques on the validation set is presented in Table 1. MTL-DA techniques result in considerable improvements in the model performance for Tasks 3 and 4. However, vanilla fine-tuning of the model achieves the highest score in Task 6, suggesting that the relatively bigger training data for Task 6 is adequate to train the model without additional augmentation techniques. In-domain Augmentation shows promising results in all tasks, suggesting the positive impact of joint multi-task learning. Introducing the Drills improved the model in Tasks 3 and 4 when combined with In-domain Augmentation. Regarding ICL, 5-shot learning achieves the highest recall score in all tasks. However, fine-tuned models manage to obtain higher precision and, accordingly, outperforming $F_1$ scores.

### 3.2 Error Analysis

To gain better insight into the impact of MTL-DA techniques, we manually controlled the fraction of the validation set in which models with different

[2]GPT-4-Turbo (gpt-4-0125-preview)

| Approach | Task 3 | | | Task 4 | | | Task 6 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $F_1$ | *P* | *R* | *Rel* $F_1$ | *Str* $F_1$ | *Tok* $F_1$ | $F_1$ | *P* | *R* |
| **SMM4H Competition** | | | | | | | | | |
| *Overall Mean* | 52.7 | 57.2 | 54.4 | 40.7 | 14.6 | 41.7 | 92.4 | 92.4 | 92.6 |
| *Overall Median* | 59.6 | 63.1 | 60.1 | 40.4 | 16.4 | 48.9 | 93.6 | 93.4 | 94.9 |
| **Our Scores** | | | | | | | | | |
| *Target Task* | / | / | / | / | / | / | 91.4 | 97.7 | 86.0 |
| *In-domain* | 55.8 | 66.3 | 53.0 | **44.9** | **20.1** | 49.6 | / | / | / |
| *In-domain + Drills* | **64.2** | 67.0 | 62.3 | 36.8 | 16.4 | 37.5 | **95.7** | 96.5 | 94.9 |

Table 2: The performance of the proposed approaches on the test sets of each task, in addition to the overall mean and median of all participants in the SMM4H 2024 competition. The results indicate the positive impact of MTL-DA techniques by achieving two to five points higher than overall scores. Note that the metrics used for Task 4 are Relaxed $F_1$ (For Ranking), Strict $F_1$, Token-level $F_1$.

MTL-DA techniques provided different predictions (disagreements). Regarding **Task 3** we observed that the model with vanilla fine-tuning on the target task excels at correctly predicting the label for emotionally nuanced samples. Meanwhile, the augmentation techniques further improve the model's ability in correctly predicting the label for neutral/unrelated samples. In **Task 4**, we noticed that the model with MTL-DA techniques performs better on samples that include keywords for medical side-effects/symptoms and their associated impacts such as "*brain zaps*", "*difficult urinating*" and "*hair loss*". Regarding **Task 6**, we observed that the model with augmentation techniques is more effective in identifying explicit age mentions. However, the model with vanilla fine-tuning on the target task shows more robustness in predicting cases lacking exact age mentions, which is frequent in discussions focused on age-related health concerns. More details are presented in §D.

### 3.3 SMM4H Test Sets

Given that the submissions for the competition are limited to 3, we applied Borda Count to determine the three top models in the validation set. The results of the best-performing models on each task, presented in Table 2, indicate the positive impact of augmentation techniques. Similar to the performance observed on the validation set, the model with *In-domain augmentation* and with *In-domain augmentation + Drills* achieved the highest performance in Tasks 3 and 4, respectively, each scoring approximately four points higher than the overall median of the competition. Interestingly, while vanilla fine-tuning achieved the highest performance on the validation set in Task 6, it is outperformed on the test set by the model with *In-domain augmentation + Drills* by approximately four points. This result suggests that augmentation techniques can increase the model's robustness to handle unseen samples.

## 4 Conclusion

We studied two new Multi-Task Learning Data Augmentation techniques to address the lack of adequate data to fine-tune LLMs in low-resource settings. We evaluated the proposed techniques in three tasks in the context of the SMM4H 2024 competition, and we observed a) the positive impact of augmentation techniques in improving model performance; and b) the potential of in-context learning as a low-cost surrogate to fine-tuning. Nevertheless, besides further ablation studies, there are a few unanswered questions: a) "*What is the best set of drills given a task?*"; b) "*When should a dataset be considered in-domain?*"; c) "*Would the same task on different domains also contribute positively?*"; d) "*How does the introduced approach compare to existing data augmentation techniques?*"; and e) "*Does the replication of the dataset potentially lead to overfitting?*". We aim to explore these questions in future work.

## Acknowledgement

We acknowledge the support of the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU.

## References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning

of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Yao Ge, Sudeshna Das, Karen O'Connor, Mohammed Ali Al-Garadi, Graciela Gonzalez-Hernandez, and Abeed Sarker. 2024. Reddit-impacts: A named entity recognition dataset for analyzing clinical and social effects of substance use derived from social media. *Preprint*, arXiv:2405.06145.

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*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2021. Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8502–8516, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Chengyu Huang, Shiqi Xu, and Soroush Vosoughi. 2021. Text augmentation in a multi-task view. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2888–2894, Online. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks", month = Aug, year =*.

# Appendix

## A SMM4H 2024 Tasks

We evaluate the proposed techniques on three tasks: **Task 3: Multi-class classification of effects of outdoor spaces on social anxiety symptoms** consisting of categorizing posts on the Reddit platform into four classes `positive effect`, `negative effect`, `neutral`, and `unrelated`. A total of 3000 annotated posts are provided into sets of 1800, 600, and 600 posts as training, validation, and test sets, respectively. **Evaluation:** To evaluate the model, we compute the macro-average F1-score over all 4 classes as done in the competition.
**Task 4: Extraction of the clinical and social impacts of nonmedical substance use from Reddit** consisting of an entity recognition tasks on Reddit posts for two entity types `clinical impact` and `social impact`. A total of 1380 posts are provided into sets of 843, 259, and 278 posts as training, validation, and test sets respectively. **Evaluations:** The model is tasked to predict the relevant spans. We automatically post-process the predicted tag for each token in the predicted span for both social and clinical impact. The prediction is then evaluated as a traditional entity tagging task (Note that it is different than the evaluation set-up deployed by SMM4H competition organizers).
**Task 6: Self-reported exact age classification with cross-platform evaluation in English** consists of extracting self-reported exact age from posts on X (Twitter) and Reddit. The training data consists of 8.8k tweets as well as 100k unannotated Reddit posts, while the validation set consists of 2.2k tweets and 1000 Reddit posts. The test data for this task includes 2.2k tweets and 14.4k Reddit posts. **Evaluations:** To evaluate the model, we compute the macro-average F1-score for the self-report class (post contains exact age) as done in the Task 6 competition.

## B Implementation Details

QLoRA (Dettmers et al., 2024) introduces additional adapter layers that are fine-tuned for the downstream task. We trained the model with two different GPUs on a cloud hosting platform: an A100 (80GB) and an A6000 (48GB). Considering our limitation of GPU power, we applied quantization to the LORA layers. We used 100 warm-up steps and, as suggested in QLora (Dettmers et al., 2024), we kept a ratio of $\alpha = 2 \times rank$ with rank=4,8,32 for 30 epochs. We applied LoRA adapters to every linear layer *"q_proj", "k_proj", "v_proj", "o_proj","gate_proj", "up_proj", "down_proj"*, and *"lm_head"*. We used a decaying learning rate that starts from 2e-4. We experimented with 2 parameters for the max sequence length, 796 and 1496. The best-performing max sequence length was 1496. This is probably due to the fact that Reddit posts' length required a longer number of tokens. To reduce GPU memory occupation, we halved batch size from two to one and doubled gradient accumulation steps from 16 to 32. To keep the memory footprint small in QLoRA, we enabled model loading in 4 bits and the computation type in `bfloat16`. To keep the memory footprint of the optimizer small, we used Paged AdamW 8bit (a version of Adam Optimizer that leverages CUDA paging). Each entry is packed with multiple sequences, as in T5 (Raffel et al., 2020). The hyperparameter configuration selected is rank=32, alpha=64 LoRA dropout = 0.1.

## C Fine-Tuning Costs

Using the A100 GPU, the longest training run took 24 hours on the model that included task drills and tasks for all three SMM4H tasks, summing up to a cost of $76. At the end of the study, the total cost, including all experiments, was $1600. The cost per hour of A100 GPU (80GB) and the A6000 (48GB) was respectively $3.18 and $1.89.

## D Error analysis

### D.1 Task 6

In Task 6, the Target model proved to be more effective in predicting the absence of the exact age in posts. We can observe this especially in discussions focused on health conditions broadly associated with age ranges ("cataracts with age...", "diagnosed in my late 20s...") rather than specific numerical ages. Conversely, the In-domain + drills model better recognized posts where age is explicitly stated ("I'm 28 going on 29...", "I'm 19..."). This indicates that the Target model tends to predict more effectively the lack of exact self-reported age, while the In-domain + drills model is more reliable when exact ages are directly provided within the text.

### D.2 Task 3

The target task model performs better in recognizing the emotional nuances of outdoor experiences.

It accurately predicts posts expressing anxiety, reporting social difficulties ("Crushed by SA...") or highlighting positive achievements ("I GOT TO RIDE IN A TESLA...").

In contrast, the In-domain + Drills model demonstrates strength in identifying neutral activities involving outdoor spaces ("Went outside for the first time in awhile...") and posts where the outdoor setting is peripheral to the core discussion ("Social anxiety has ruined my diet..."). This suggests that the In-domain + Drills model is better at recognizing when the impact of outdoor spaces is primarily neutral or indirect.

### D.3 Task 4

In Task 4, the in-domain model outperforms the target model in recognizing specialized medical terminology and its associated impacts. It accurately identifies clinically significant terms such as "brain zaps" and "hair loss" (potential side effects), as well as the severity of "cravings for heroin" (addiction). The in-domain model also demonstrates a better grasp of the nuanced social impacts related to health. Although the in-domain model's higher recall leads to some incorrectly predicted social impacts, in general it achieves higher performance in the identification and classification of clinically relevant keywords.