AAST-NLP@#SMM4H'24: Finetuning Language Models for Exact Age Classification and Effect of Outdoor Spaces on Social Anxiety

Ahmed El-Sayed and **Omar Nasr** and **Noha S. Tawfik**
Arab Academy for Science, Technology & Maritime Transport
{ahmedelsayedhabashy,omarnasr52}@gmail.com, noha.abdelsalam@aast.edu

Abstract

This paper evaluates the performance of "AAST-NLP" in the Social Media Mining for Health (SMM4H) Shared Tasks 3 and 6, where more than 20 teams participated in each. We leveraged state-of-the-art transformer-based models, including Mistral, to achieve our results. Our models consistently outperformed both the mean and median scores across the tasks. Specifically, an F1-score of **0.636** was achieved in classifying the impact of outdoor spaces on social anxiety symptoms, while an F1-score of **0.946** was recorded for the classification of self-reported exact ages

1 Introduction

The widespread use of social media platforms has encouraged the employment of Natural Language Processing (NLP) in all domains, specifically in health-related applications (Correia et al., 2020). These platforms are considered unfiltered, real-time, valuable sources for analysis As millions of users openly share their personal health narratives and experiences on a daily basis. The Social Media Mining for Health Applications (SMM4H-2024) workshop provides a chance to develop natural language processing (NLP) models that automatically extract meaningful information from social media data through seven shared tasks in various contexts, including pharmaceutical, social, health, and clinical. In this paper, we describe our team's participation in both Task 3, Self-reported Exact Age Classification, and Task 6, Multi-class Classification of the Effects of Outdoor Spaces on Social Anxiety Symptoms in Reddit. The former task explores the extraction of patient demographics, enabling large-scale observational studies. Similarly, the latter investigates outdoor spaces mentioned in social media and aims to qualitatively assess their effects and relation to Social Anxiety Disorder (SAD).

2 Tasks & Datasets Description

2.1 Task 3

Social Media Mining for Health (SMM4H-24) (Xu et al., 2024) organized a total of 7 tasks, each with its own theme targeted at social media mining tasks. Task 3 involves categorizing posts mentioning specific outdoor keywords into four groups based on their effects on social anxiety symptoms: positive effect, neutral or no effect, negative effect, or unrelated mentions. The dataset includes posts from r/socialanxiety subreddit, filtered for users aged 12-25 and mentions 80 green/blue space keywords. The dataset consists of 1800 training examples, 600 validation examples and 600 testing examples. The training dataset was severely imbalanced with 1131 being labeled as unrelated, 395 as neutral, 160 as positive and 114 negative.

2.2 Task 6

Task 6 involves classifying social media posts according to self exact-age reporting.This task expands the scope of social media data usage by accurately extracting the precise ages reported by users rather than categorizing them into ranges of age groups as is typically done. This task builds upon the tasks established in SMM4H 2022 (Weissenbacher et al., 2022), which addressed a similar thematic task. The key distinction lies in adapting this year's task towards cross-platform evaluation, thereby extending the scope and applicability of the original concept. The dataset includes posts from X (previously Twitter) and Reddit. In addition to the annotated datasets used in training, validation and testing phases, an additional 100000 unlabeled Reddit posts from r/AskDocs where provided. The provided dataset is imbalanced, with 5966 posts provided having no exact age mention and 2834 posts with exact age mentions. The validation set follows a similar distribution with 2435 and 1765 posts for the respective classes.

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 110–113

3 Proposed System

3.1 Data Preprocessing

3.1.1 Task 3

The social media data in task 3 consisted of multi-sentence posts, many of which were not relevant to our objective. This required preprocessing the data to filter meaningful sentences. Consequently, we investigated three distinct methodologies for sentence filtering. These methodologies encompassed isolating solely the sentence containing the specified keyword, retrieving the sentence featuring the specified keyword along with two additional sentences (one preceding and one succeeding it), and extracting two preceding and two succeeding sentences in conjunction with the sentence containing the keyword. Our objective was to extract the sentence containing the necessary context to classify it into the correct category. Adding extra sentences, besides those containing the specified keyword, aimed to provide additional context that could enhance our model's capability. Our findings show that utilizing solely the sentence containing the keyword yielded superior performance in both F1-Score and training time. Moreover, we eliminated emojis, hyperlinks, punctuation, extra spaces, and line breaks.

3.1.2 Task 6

Similar to task 3, the posts included numerous sentences lacking relevance to our designated task. Opposed to experimenting with different sentence combinations, we opted to extract only the sentences containing digits as these were most likely to include age information. The common abbreviations used on social media, such as "yo" for years old and "bday" for birthday," were addressed passively. Additionally, certain social media patterns were challenging for the model to understand and required further resolution for training. These included:

- Samples starting with words matching the pattern `(\d+)([FMfm])` indicated a male or female mentioning their age.
- Some samples contained a reversed version of this pattern, so `([FMfm])(\d+)` was used instead.
- Samples with the pattern `(\d+)y` implied a person stating their age in years.
- Samples contained `(\d{2})s` implied an age group rather than an exact age mention yet it was often misclassified so they were replaced by their written counterparts foe examples "20s" would be replaced by twenties.

Finally, All of the posts had their emojis, hyperlinks and punctuation marks removed. Task 6 organizers provided an extra substantial volume of unlabeled Reddit posts with sentences that included 2-digit numbers. We employed a rule-based approach to label the provided in order to increase the size of our training dataset. The rules used were directly extracted from the task annotation guidelines and domain knowledge. For instance, any sentence that had a digit that matched the regex pattern `(\d+)([FMfm])` would be labeled as containing an exact age mention. The Reddit platform, in particular, has a number of abbreviations that correspond to exact-age mentions such as "21F" (which would man a 21 years old female). Those abbreviations were resolved through the regex patterns. The abbreviations were resolved using regex through the following patterns `(\d+)([FMfm])`, `([FMfm])(\d+)`, `(\d+)([FMfm])` and `(\d+)(?:\.,|;)`. Additionally, emojis, punctuation, and hyperlinks were removed.

3.2 Language Models

For both tasks, multiple Language Models renowned for their State of the Art (SOTA) performance on Natural Language Processing (NLP) tasks were experimented with, including RoBERTa (Liu et al., 2019), BERTweet (Nguyen et al., 2020), and XLM-RoBERTa (Conneau et al., 2020). Dice Loss (Li et al., 2020) was utilized in fine-tuning our language models due to their proven performance in tackling NLP tasks in recent studies. It is particularly effective when handling imbalanced datasets because it emphasizes the correct classification of underrepresented classes and maximizes the overlap between predicted probabilities and actual outputs. This leads to improved model performance, where minority classes are crucial compared to other loss functions. For task 6, We also experimented with Mistral-7b (Jiang et al., 2023) through HuggingFace's quantized version.[1]. To achieve better performance with Mistral; we employed prompt engineering, which involves, in other words, crafting strategic prompts to guide AI systems, ensuring effective outputs (Chen et al., 2023). It includes

[1] https://huggingface.co/unsloth/mistral-7b-bnb-4bit

selecting language, context, and constraints to generate relevant responses. Several schemes were tested, but the one used for the testing set was based on Chain-of-Thought (Wei et al., 2023) by feeding the language model samples from the training and validation sets following an iterative approach to avoid misclassified examples.

3.3 Hyperparamters and Training

The hyperparameters reported in table 1 represent those employed in training the model used to submit the final results for both tasks

Hyperparameter	Task 3	Task 6
Epochs	30	30
Learning Rate	2e-5	1e-5
Batch Size	8	16
Max length	200	80
Optimizer	Adam	Adam
Early Stopping Patience	5	5
Reduce On Plateau	2	2
Loss Function	Dice Loss	Dice Loss

Table 1: Training Hyperparameters. Parameters shown for RoBERTa and BERTweet-Large/RoBERTa-Large for tasks 3 and 6, respectively.

The training procedure was conducted using Kaggle's [2] free-to-use platform, which provides 29 GB of RAM, a 16 GB NVIDIA P100 GPU, and Python. The autofit functionality from ktrain (Maiya, 2022) was utilized, incorporating a triangular learning rate policy (Smith, 2017).

4 Results

4.1 Task 3

Our experiments concluded that Roberta-Large[3] had the best performance of the three models when evaluated during the validation stage and was then used in the testing phase.

Table 2 illustrates our model's performance on the test set. Our F1 and recall results exceed the mean and median of the other submissions by a large margin, showcasing our model's robustness and efficiency. Our model's precision also surpasses the mean and median, though by a narrower margin.

[2] https://www.kaggle.com/

[3] https://huggingface.co/FacebookAI/roberta-large

	Precision	Recall	F1-Score
Validation	0.68	0.65	0.66
Mean	0.5649	0.5379	0.5186
Median	0.63	0.5885	0.5795
RoBERTa	**0.631**	**0.644**	**0.635**

Table 2: Task 3 Results.

4.2 Task 6

For task 6, the results revealed that BERTweet exhibited superior performance compared to the other two models, An Ensemble of BERTweet and RoBERTa were used in the testing phase. Moreover, high results were achieved through prompt engineering in our zero-shot experiments with Mistral despite not utilizing any pre-training.

Ensembling via majority voting was used for the final submission on the test set, the three models used were all given equal weights. Table 3 illustrates the results obtained on the test set.

	Precision	Recall	F1-Score
Validation	0.93	0.96	0.94
Mean	0.924	0.926	0.924
Median	**0.934**	0.949	0.936
Ensemble	0.932	**0.959**	**0.946**

Table 3: Task 6 Results.

5 Error Analysis

In our error analysis of the validation set for Task 3, we observed that one of the main reasons for performance degradation was the misclassification of unrelated and neutral instances. Upon checking samples of these instances manually, we identified that there exist some examples to be mislabelled or at least confusing even on the human level. Eliminating such ambiguous data in following versions of the task will definitely increase the training data quality and reflect such quality on the models' performance. Additionally, in Task 6, mislabeled examples in both the training and validation sets pose a risk to the effectiveness of the training and evaluation processes. One clear example of that is the post with ID: 8812 in the validation set which was wrongfully classified as having no exact age mention.

6 Conclusion

In this work, approaches were presented to address the challenges of self-reported exact age classification and classification of effects of outdoor spaces on social anxiety symptoms. For the first task, an ensemble of RoBERTa, BERTweet, and Mistral was employed to handle the problem. Despite the achieved performance, it is believed that there is room for improvement in the approach in the future. Areas of potential improvement include refining the preprocessing procedure and fine-tuning Mistral language model.

References

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *Preprint*, arXiv:2310.14735.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Rion Brattig Correia, Ian B. Wood, Johan Bollen, and Luís M. Rocha. 2020. Mining social media data for biomedical signals and Health-Related behavior. *Annual review of biomedical data science*, 3(1):433–458.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced nlp tasks. *Preprint*, arXiv:1911.02855.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Arun S. Maiya. 2022. ktrain: A low-code library for augmented machine learning. *Preprint*, arXiv:2004.10703.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *Preprint*, arXiv:2005.10200.

Leslie N. Smith. 2017. Cyclical learning rates for training neural networks. *Preprint*, arXiv:1506.01186.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications (#SMM4H) shared tasks at COLING 2022. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 221–241, Gyeongju, Republic of Korea. 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. 2024. Overview of the 9th social media mining for health (#SMM4H) research and applications workshop and shared tasks at ACL 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.