IMS_medicALY at #SMM4H 2024: Detecting Impacts of Outdoor Spaces on Social Anxiety with Data Augmented Ensembling

Amelie Wührl[1,2,✎], **Lynn Greschner**[2,✎],
Yarik Menchaca Resendiz[1,2,✎] and **Roman Klinger**[2]
[1]University of Stuttgart, Germany, [2]University of Bamberg, Germany
firstname.lastname@ims.uni-stuttgart.de, firstname.lastname@uni-bamberg.de

Abstract

Many individuals affected by Social Anxiety Disorder turn to social media platforms to share their experiences and seek advice. This includes discussing the potential benefits of engaging with outdoor environments. As part of #SMM4H 2024, Shared Task 3 focuses on classifying the effects of outdoor spaces on social anxiety symptoms in Reddit posts. In our contribution to the task, we explore the effectiveness of domain-specific models (trained on social media data – SocBERT) against general domain models (trained on diverse datasets – BERT, RoBERTa, GPT-3.5) in predicting the sentiment related to outdoor spaces. Further, we assess the benefits of augmenting sparse human-labeled data with synthetic training instances and evaluate the complementary strengths of domain-specific and general classifiers using an ensemble model. Our results show that (1) fine-tuning small, domain-specific models generally outperforms large general language models in most cases. Only one large language model (GPT-4) exhibits performance comparable to the fine-tuned models (52% F_1). Further, we find that (2) synthetic data does improve the performance of fine-tuned models in some cases, and (3) models do not appear to complement each other in our ensemble setup.

1 Introduction

Social Anxiety Disorder is a medical condition that can significantly impact an individual's life (Vilaplana-Pérez et al., 2021). Social media platforms have emerged as spaces where affected individuals can communicate their experiences and seek support. These platforms are rich with biomedical information, providing an opportunity for medical practitioners to gain novel insights about medical conditions. However, this data is highly diverse and annotation is expensive, especially for the medical domain. Access to a broad variety of classification and generative models has the potential to close this gap as it (1) allows to explore the capability of domain-specific and general models to solve these types of tasks, and (2) opens the possibility to generate synthetic training instances to complement sparse, human-labeled data. As part of #SMM4H 2024, Shared Task 3 focuses on classifying the effects of outdoor spaces on social anxiety symptoms in Reddit posts. Given the post, the goal is to predict the user's sentiment towards the effect of outdoor space in a multi-class classification setup with the target labels POSITIVE, NEGATIVE, NEUTRAL and UNRELATED.

With our contribution, we investigate three research questions (RQs):

RQ1 Given the sparsity of the data, do fine-tuned, domain-specific models outperform general models?

RQ2 Does incorporating synthetic data complement human-labeled data and enhance model robustness?

RQ3 Is Reddit's text diversity better captured by a set of different models in an ensemble setup?

2 System Description

We hypothesize that the diversity of texts shared on Reddit benefits from aggregating multiple approaches. Therefore, we design an ensemble model that takes as input the predictions of individual models varying in architecture and training procedure. The individual models are:

Fine-tuned language models. We fine-tune `BERT` (Devlin et al., 2019), `RoBERTa` (Liu et al., 2019), `DistilBERT` (Sanh et al., 2019) and `SocBERT` (Guo and Sarker, 2023) on the training split of the task data to obtain customized models. We set truncation and padding to True, batch size to 4, and train

✎The first three authors contributed equally.

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 83–87

for 3 epochs, with a learning rate of $5{\cdot}10^{-5}$ using the AdamW optimizer.

Few-shot prompting. To explore the capacity of general large language models (LLMs), we prompt `Mistral-7B-v0.1` (Jiang et al., 2023), `Llama-2-7b-chat-hf` (Touvron et al., 2023), `GPT-3.5` (OpenAI, 2022) and `GPT-4` (OpenAI, 2023) to generate labels. We provide them with task instructions in a one-shot setup, where the example instance is randomly chosen from the training data. Table 3 shows the prompt templates.

Ensemble. Our neural `Ensemble` consists of an input layer, a one dimensional convolutional layer and max-pooling layer, a dense layer with 128 neurons, and a classification layer (Dense) with 4 neurons. The model inputs include 8 predictions and probabilities from fine-tuned models (4 models trained with and 4 trained without synthetic data) and 4 predictions from LLMs, 12 in total. We train the ensemble using 400 instances from the validation data over 20 epochs and evaluate it using the remaining 200 instances from the same subset.

Synthetic data augmentation. We further hypothesize that additional training data for the minority classes POSITIVE, NEGATIVE, and NEUTRAL benefits model performance in the fine-tuning setup. To investigate this, we generate synthetic instances using `Mistral-7B` and `GPT-3.5`. We generate as many instances as needed to match the size of the largest class (UNRELATED, 1,131). Table 4 shows prompt templates and examples.

3 Results

Table 1 shows the performance ofe all classifiers on the shared task's validation set. Table 2 reports the results of three of our systems on the test set.

RQ1: How do domain-specific models compare to general models within this task? Overall, the models show a mixed performance. `GPT-4` achieves the best results (.52 F_1). The general-domain models are slightly more robust than `SocBERT` which is specialized for the social media domain (Δ .07 F_1). Compared to prompting, fine-tuning leads to more consistent performances across models.

RQ2: How do synthetic training instances affect classification performance? Table 1 reports the results for models fine-tuned with (+) and without additional synthetic training instances. We observe that for two out of four models (`DistilBERT`, `RoBERTa`), fine-tuning on the additional synthetic data leads to a more robust performance, compared to only training on gold data. This indicates that to a certain degree, the synthetic data is complementary to the human annotations.

	Model	P	R	F_1
Gold	BERT	0.58	0.47	**0.50**
	DistilBERT	0.70	0.37	0.39
	RoBERTa	0.58	0.43	0.43
	SocBERT	0.54	0.43	0.45
Gold+Syn	BERT+	0.58	0.47	**0.50**
	DistilBERT+	0.47	0.45	0.45
	RoBERTa+	0.47	0.48	0.47
	SocBERT+	0.54	0.43	0.45
Prompt	GPT-3.5	0.32	0.46	0.28
	GPT-4	0.56	0.55	**0.52**
	Llama-2	0.19	0.30	0.15
	Mistral	0.28	0.27	0.11
	Ensemble	0.56	0.49	0.51

Table 1: Macro F_1 of individual classifiers on the validation set. + indicates that the models are fine-tuned with additional synthetic data. We evaluate the ensemble on 200 unseen instances from the validation set.

Model	Acc	P	R	F_1
Ensemble	0.48	0.35	0.40	0.34
GPT-4	0.56	0.60	0.52	0.50
SocBert	0.62	0.63	0.53	0.56

Table 2: Performance of three classifiers – Domain-specific (SocBert), General-Domain (GPT-4), and the Ensemble model – on the Task 3 test set of #SMM4H 2024.

RQ3: Do domain-specific & general models complement each other? Table 2 reports the performance of our models on the test set. We submit the predictions from the best domain-specific (`SocBERT`), general-domain (`GPT-4`), and the `Ensemble` model. The best model is `SocBERT` (.56 F_1), followed by `GPT-4` (.50 F_1). The predictions from the `Ensemble` obtain an F_1-score of .34, indicating that (1) models do not complement each other or (2) the ensemble might benefit from additional features that go beyond prediction probabilities. The result may also be attributed to the limited amount of data available for training the `Ensemble`.

4 Conclusion

We present our contribution to the #SMM4H 2024 Shared Task 3 which focuses on classifying the effects of outdoor spaces on social anxiety symptoms in Reddit posts. We find that fine-tuning models

overall show a more robust performance compared to LLM prompting. Synthetic data may increase the robustness of the models. We can not show a superior performance of an ensemble. This leads to important future work: For the prompting approaches, we have to evaluate the impact of the prompt design for the task. For fine-tuned models, a thorough analysis of the synthetic data is key to gauging the impact of generated instances in more detail. For all models, an in-depth error analysis is crucial to understand model capabilities and the impact the individual predictions may have on the `Ensemble`. Further, testing alternative ensemble designs (e.g., Gradient-boosted Decision Trees) is key to understanding the interaction between the probability-based, class predictions we obtain from the fine-tuned models, and the class-only predictions from LLMs.

Acknowledgements

Yarik Menchaca Resendiz is funded by a `CONACYT` scholarship (2020-000009-01EXTF-00195). Lynn Greschner is funded by the `EMCONA` project (DFG, project number KL 2869/5-1). Amelie Wührl is funded by the `FIBISS` project (DFG, KL 2869/12-1) as well as the `CEAT` project (DFG, KL 2869/1-2.). We thank the reviewers for their valuable feedback.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuting Guo and Abeed Sarker. 2023. SocBERT: A pretrained model for social media text. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 45–52, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

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. *CoRR*, abs/1907.11692.

OpenAI. 2022. Gpt-3.5 turbo. `https://www.openai.com`. Accessed: [18.04.2024].

OpenAI. 2023. Gpt-4 turbo. `https://www.openai.com`. Accessed: [18.04.2024].

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Alba Vilaplana-Pérez, Ana Pérez-Vigil, Anna Sidorchuk, Gustaf Brander, Kayoko Isomura, Eva Hesselmark, Ralf Kuja-Halkola, Henrik Larsson, David Mataix-Cols, and Lorena Fernández de la Cruz. 2021. Much more than just shyness: the impact of social anxiety disorder on educational performance across the lifespan. *Psychological Medicine*, 51(5):861–869.

A Appendix

A.1 One-shot Prompting

We use the template provided in Table 3 to prompt the four LLMs.

A.2 Synthetic Data Generation

Prompt Design. Using `Mistral-7B`, we simulate a one turn user-assistant conversation. We instruct the model to produce a post by a Reddit user that describes how an outdoor space or activity affects their social anxiety symptoms POSITIVELY, NEGATIVELY or has a NEUTRAL effect. We randomly pick an instance from the training data as a one-shot example in the prompt. Similarly, we randomly choose an outdoor space/activity from the spaces mentioned in the training data as well as a persona for the Reddit user to ensure more variability. We provide the prompt template in Table 4. We follow a similar process for `GPT-3.5`, we generate new instances by randomly selecting a humman generated Reddit post and using it as a few-shot example for `GPT-3.5` to create another Reddit post, the newly generated post uses the same keywords (e.g., *beach*, *forest*), see Table 4 for examples.

Personas. 'teenager', 'young adult', 'middle-aged adult', 'senior citizen', 'child', 'adolescent', 'adult', 'elderly person', 'teacher', 'doctor', 'nurse', 'computer scientist', 'engineer', 'scientist', 'researcher', 'professor', 'academic', 'student', 'florist', 'farmer', 'chef', 'cook', 'baker', 'waiter', 'waitress', 'cashier', 'bank teller', 'receptionist', 'librarian', 'archivist', 'historian', 'writer', 'author', 'PhD student', 'graduate student', 'undergraduate student'.

Activities. 'ocean', 'swim', 'outdoors', 'running', 'go for a run', 'soccer', 'pond', 'golf', 'playground', 'rowing', 'coast', 'climb', 'bonfire', 'basketball', 'horses', 'snowboard', 'forest', 'hills', 'lawn', 'tennis', 'hill', 'bicycle', 'cabin', 'mountain', 'snowboards', 'surfing', 'backyard', 'fresh air', 'outside', 'bikes', 'waterfall', 'runners', 'backpacking', 'go for a walk', 'riding', 'runs', 'park', 'jog', 'river', 'nature', 'mountains', 'baseball', 'hiking', 'bench', 'outdoor', 'sea', 'skating', 'hike', 'skate', 'breeze', 'barbecue', 'bay', 'waves', 'runner', 'stream', 'swimming', 'shore', 'horse', 'biking', 'open space', 'sand', 'run', 'canoe', 'camping', 'benches', 'surf', 'jogging', 'boat', 'cruise', 'wildlife', 'parks', 'hikes', 'tree', 'camps', 'bleachers', 'walk', 'pool', 'bike', 'camp', 'sun', 'grass', 'garden', 'beach', 'lake', 'landscape', 'trees'.

Examples of generated instances. Using the prompt structures from Table 4.

Mistral I am a middle aged man who suffers from depression and anxiety. Running has always been a great way for me to clear my mind and relieve stress. However, lately I've noticed that running has also been triggering my anxiety. Every time I step outside and see other runners, I start to compare myself to them and feel like I'll never be able to catch up. This leads to feelings of shame and self-doubt, making it harder for me to enjoy the exercise. I've tried different types of runs, such as solo runs and group runs, but nothing seems to help. Has anyone else experienced this before? How do you cope with these feelings?

Being outside in nature has been a game-changer for my social anxiety. The vast openness and fresh air provide a sense of freedom and calm that I can't find anywhere else. Just the simple act of stepping outside can instantly ease my racing thoughts and tension.

GPT3.5 If you're feeling up to it, I highly recommend giving backpacking a try. Start with a short trip or even just a day hike to see how it resonates with you. You might be surprised at how much it can help alleviate some of your social anxiety symptoms. Remember, you're not alone in this journey. Feel free to reach out if you need any advice or support. Take care of yourself and keep pushing through! You got this.

Reddit post: "I struggle with social anxiety, and finding ways to cope with it has been a journey. One thing that has surprisingly helped me is spending time outside. There's something about nature that calms my anxieties and helps me feel more at ease. Whether it's taking a walk in the park, going for a hike, or simply sitting in my backyard, being outside helps me clear my mind and feel less overwhelmed by social situations.

Hey everyone, I just wanted to share my experience with being outside and how it has helped me with my social anxiety. For the longest time, I struggled with being around people and entering social situations made me extremely anxious. However, I found that spending time outside in nature has been incredibly beneficial for my mental health.

Prompt
We analyze effects of outdoor spaces on social anxiety symptoms in Reddit posts. You will be presented with a user-written post. The posts were filtered based on a list of nature-related keywords related to outdoor spaces and activities. Your task is to categorize posts into one of four categories: 0) unrelated: the nature-related keyword does not reference nature (e.g., it is used in a metaphor or idiomatic expression), the user is/has not personally experienced the nature-related keyword, or it. Note, that each post has only one classification. 1) positive effect: the nature-related space/activity helps the user's mental well-being. 2) neutral or no effect: the nature keyword is referencing nature, however, the user makes no mention of it having a positive or negative effect on the user's mental well-being. 3) negative: the nature-related space/activity has a negative effect on the user's mental well-being. Provide the output in a json format with the key being 'label' and the value being the category number as an integer. For example: if you believe the post should be categorized as 1) positive, your json output should be: **{'label': 1}** Now consider the following example: *Ï'm supposed to go on a hike with friends, but I'm feeling tense about it. they still haven't made any proper plans yet. I was kind of hoping they would forgot about it, so I wouldn't have to go through the hassle of getting ready and dealing with the crowds. Now I'll have to wake up early and be prepared, just in case.* Ẅhat is the correct category for this post? Here is the correct category formatted as json: **{'label': 3}**

Table 3: Template for few-shot classification with four LLMs: The task/instruction description is in monospace, class descriptions (0, 1, 2, and 3) are in normal font, few-shot examples are in *italics*, and the expected LLM output is in **bold**.

M.	Role	Prompt
Mistral 7-B	user	Imagine you are a person who is suffering from social anxiety. Write a Reddit post in which you describe the effects of an nature-related space or outdoor activity on your social anxiety symptoms. The outdoor space or activity could be something like 'surfing', 'backyard', 'fresh air' or 'basketball'. Your post should describe how the outdoor space or activity has a **<target sentiment>** effect on your symptoms, so the nature-related space/activity **<helps your mental well-being./does not help your mental well-being/has no effect on your symptoms.>**. Provide the output in json format with the key being 'post' and the value being the text of your post. Write a post for the outdoor space/activity * **<activity>***.
	assistant	Here is the post I came up with formatted as json: {'post': '**<random training instance for target sentiment>**'}
	user	Perfect! Let's try another one. Imagine you are a **<persona>**. Write a post for the outdoor space/activity ***<activity>***. Your post should describe how the outdoor space or activity has a **<target sentiment>** effect on your symptoms, so the nature-related space/activity **<helps your mental well-being./does not help your mental well-being/has no effect on your symptoms.>**. Only output the json, no additional text or explanation.
GPT 3.5	system	Imagine you are a person who is suffering from social anxiety. Write a Reddit post in which you describe the effects of nature-related space or outdoor activity on your social anxiety symptoms. Use the following example: **<keywords>** Reddit post: **<Reddit post example>**. Do not write more than 350 words and only write the post itself.
	user	Keywords: **<keywords>**

Table 4: Prompt template to generate additional training instances with Mistral-7B-v0.1 and GPT 3.5. We randomly sample an example instance from the training instances with the target sentiment and instruct the model to write from the perspective of a randomly sampled persona to increase variety in the synthetic data.