1024m at SMM4H 2024: Tasks 3, 5 & 6 - Ensembles of Transformers and Large Language Models for Medical Text Classification

Ram Mohan Rao Kadiyala
University of Maryland, College Park
rkadiyal@terpmail.umd.edu

M.V.P. Chandra Sekhara Rao
RVR&JC College of Engineering
mvpcs@rvrjc.ac.in

Abstract

Social media is a great source of data for users reporting information and regarding their health and how various things have had an effect on them. This paper presents various approaches using Transformers and Large Language Models and their ensembles, their performance along with advantages and drawbacks for various tasks of SMM4H'24 - Classifying texts on impact of nature and outdoor spaces on the author's mental health (Task 3), Binary classification of tweets reporting their children's health disorders like Asthma, Autism, ADHD and Speech disorder (task 5), Binary classification of users self-reporting their age (task 6).

1 Introduction

Social media has become a key way for people to share their experiences and feelings. This has opened up new opportunities for researchers to understand how different aspects of life affect our well-being. The paper explores three tasks of SMM4H 2024(Xu et al., 2024) - 4-way classification of texts based on effect of nature, outdoor spaces and activities on author's mental health (Task 3), Binary classification of texts reporting health disorders in author's child including ADHD, Autism, Asthma and Speech disorder (Task 5)(Klein et al., 2024), Binary classification of texts self-reporting author's exact age directly / indirectly (Task 6).The paper explores usage of transformer models like RoBERTa(Liu et al., 2019), DeBERTa(He et al., 2021), Longformer(Beltagy et al., 2020) and LLs including both proprietary and open-source like GPT-4(OpenAI, 2024), Claude-Opus(Anthropic, 2024), Llama-3 8B(Touvron et al., 2023), Mistral 7B(Jiang et al., 2023), Gemma 7B(GemmaTeam, 2024), and ensembles along with advantages and drawbacks of each approach using the models. Similar previous works can be found in (Weissenbacher et al., 2022), (Magge et al., 2021) and (Klein et al., 2020).

2 Datasets

The dataset for Task 3 consists of 3000 reddit posts from r/socialanxiety belonging to four classes based on self reported impact of outdoor spaces and activities on the author's mental health - 0: unrelated to the task, 1: had a positive impact, 2: is neutral or had no effect, 3: had a negative effect. The dataset for Task 5 consists of 9734 tweets belonging to two classes - 1: users reporting having a child having ADHD, Asthma, Autism or Speech disorder and the rest as class 0. Similarly for Task 6, the dataset of 21200 texts consists of both tweets and reddit posts from r/AskDocs for two classes - Class 1 being texts through which the author's current age in years may be determined and rest as Class 0. The distribution of labels for the three tasks can be seen in Table 1, Table 2 and Table 3.

	Training	**Development**	**Testing**
Class 0	1131	377	?
Class 1	160	54	?
Class 2	395	131	?
Class 3	114	38	?
Total	1800	600	600

Table 1: Dataset split and class distribution : Task 3

	Training	**Development**	**Testing**
Class 0	5118	254	?
Class 1	2280	135	?
Total	7398	389	1947

Table 2: Dataset split and class distribution : Task 5

	Training	**Development**	**Testing**
Class 0	5966	2435	?
Class 1	2834	1765	?
Total	8800	4200	8200

Table 3: Dataset split and class distribution : Task 6

	F1	P	R
Bart-Large* (2-stage)	0.673	0.666	**0.687**
Bart-Large (direct)	0.654	0.676	0.643
Bart-Large (2-stage)	**0.679**	**0.677**	0.682
Mean	0.519	0.565	0.538
Median	0.580	0.630	0.589

Table 4: Precision, Recall and F1 on Test set compared to other participants : Task 3

* indicates model is trained without using validation set

	F1	P	R
Bart-Large* (direct)	0.912	0.896	**0.929**
Bart-Large (direct)	**0.918**	**0.923**	0.912
Mean	0.822	0.818	0.838
Median	0.901	0.885	0.917

Table 5: Precision, Recall and F1 on Test set compared to other participants : Task 5

* indicates model is trained without using validation set

3 Systems Description

While using transformers, For Task 3 two approaches were tested - one where classification was done directly in a 4 way and the other where classification was done is two stages, this involved first classifying the text whether it is related to the task or not i.e class 0 or not and then classifying the effect on the user is the second stage. For Task 5 and 6 it was done directly as a binary classification task[12]. In LLM approaches, The proprietary versions were used as zero shot and with the rest of the LLMs, they were tested in a zero-shot and fine-tuned manner. Additionally they were tested in a two stage classification for Task 3. In the case of ensembles, It was through majority voting in a set of models, through and-rule for high precision requirement and through or-rule for high recall requirements. For Task 5 and 6, while using LLMs, classification was done by dividing the criteria into parts and aggregating the individual results. i.e In the case of Task 5, individual prompts test for ADHD, Asthma, etc.. separately and or-rule is used for generating final label. Similarly and-rule was used for Task 6. The performance of different approaches can be seen in Table 7, Table 8 and Table 9.The data during training was shuffled after every epoch and also internally in each mini-batch.

[1]Code available at: https://github.com/1024-m/ACl-2024-SMM4H-Task-3-5-6

[2]Models available at: https://huggingface.co/1024m

	F1	P	R
Bart-Large (direct)	**0.959**	**0.953**	**0.965**
GPT-4 (and-rule)	0.922	0.895	0.951
Mean	0.924	0.924	0.926
Median	0.936	0.934	0.949

Table 6: Precision, Recall and F1 on Test set compared to other participants : Task 6

4 Error Analysis

The LLMs performed equally good on all kinds of data while transformers models performed less effectively when the kind of language used is off from rest of the data or when criteria for classification was mentioned in one sentence and referred to the conditions indirectly later on. It was observed that positively labelled samples were predicted correctly by either the LLM approach or transformers, hence ensembles of both had recall over 0.99 with just 1 percent drop in F1 scores in Task 5 and 6. Many of the positively misclassified samples were in the format of advertisements where the title appears to match the criteria for positive classification. This is one area where LLMs were still able to distinguish effectively while other models did not.

5 Conclusion

the performance of some of the models compared to others on the test set can be seen in Table 4, Table 5 and Table 6. The LLM approach did yield comparatively good results despite using in a 4bit precision due to lack of computational resources. It is likely the performance would be better that the current models in full precision. Many of the positive label texts have been filtered out during the data collection process. For example, texts self-reporting age in text format instead of numerical. Due to this, a higher focus on recall is necessary. A custom metric with higher importance to recall is better suited for Task 5 and 6 compared to F1 scores. Ensemble approaches like majority voting and filtering guaranteed positive label texts using LLM predictions could improve performance without a significant drop in the F1 scores. Finally, the performance improved on all the tasks while using dev set as additional training data compared to just the training data, hinting at the possibility of improving the performance by adding more training data. Augmentation through paraphrasing existing data however did not improve the results.

	Direct Classification			2-Stage Classification		
Model	**Macro-F1**	**Precision**	**Recall**	**Macro-F1**	**Precision**	**Recall**
Transformers (fine-tuned)						
longformer-large	0.603	0.610	0.596	0.667	**0.671**	0.660
RoBERTa-large	0.595	0.601	0.585	0.664	0.669	0.652
BART-large	0.603	0.597	0.611	**0.670**	0.652	**0.687**
DeBERTa-large	0.601	0.598	0.606	0.661	0.657	0.669
Proprietary LLMs (zero-shot)						
GPT-4	0.536	0.545	0.546	0.584	0.592	0.571
Claude-Opus	0.504	0.492	0.605	0.579	0.565	0.594
Open-source LLMs (fine-tuned)						
LLaMa-3-8B	0.643	0.622	0.653	-	-	-
Mistral-7B	0.637	0.621	0.646	-	-	-
Gemma-7B	0.639	0.624	0.644	-	-	-

Table 7: performance of different approaches on Dev set : Task 3

	Direct Classification			Or-rule Classification		
Model	**Class1-F1**	**Precision**	**Recall**	**Class1-F1**	**Precision**	**Recall**
Transformers (fine-tuned)						
longformer-large	0.937	**0.940**	0.933	-	-	-
RoBERTa-large	0.926	0.926	0.926	-	-	-
BART-large	**0.940**	0.933	0.947	-	-	-
DeBERTa-large	0.927	0.914	0.941	-	-	-
Proprietary LLMs (zero-shot)						
GPT-4	0.786	0.862	0.956	0.859	0.785	0.948
Claude-Opus	0.689	0.809	**0.985**	0.851	0.782	0.943
Open-source LLMs (fine-tuned)						
LLaMa-3-8B	0.925	0.939	0.911	-	-	-
Mistral-7B	0.921	0.921	0.921	-	-	-
Gemma-7B	0.920	0.934	0.907	-	-	-

Table 8: performance of different approaches on Dev set : Task 5

	Direct Classification			And-rule Classification		
Model	**Class1-F1**	**Precision**	**Recall**	**Class1-F1**	**Precision**	**Recall**
Transformers (fine-tuned)						
longformer-large	0.898	0.884	0.914	-	-	-
RoBERTa-large	0.891	0.862	0.920	-	-	-
BART-large	**0.901**	0.878	0.926	-	-	-
DeBERTa-large	0.894	0.869	0.923	-	-	-
Proprietary LLMs (zero-shot)						
GPT-4	0.861	0.791	**0.960**	0.897	0.870	0.925
Claude-Opus	0.858	0.794	0.952	0.893	0.873	0.937
Open-source LLMs (fine-tuned)						
LLaMa-3-8B	0.898	**0.912**	0.886	-	-	-
Mistral-7B	0.894	0.908	0.883	-	-	-
Gemma-7B	0.894	0.901	0.889	-	-	-

Table 9: performance of different approaches on Dev set : Task 6

References

Anthropic. 2024. Proprietary documentation of Company Name. [link].

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

GemmaTeam. 2024. Gemma: Open models based on gemini research and technology.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

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.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (#SMM4H) shared tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–36, Barcelona, Spain (Online). Association for Computational Linguistics.

Ari Z Klein, José Agustín Gutiérrez Gómez, Lisa D Levine, and Graciela Gonzalez-Hernandez. 2024. Using longitudinal twitter data for digital epidemiology of childhood health outcomes: An annotated data set and deep neural network classifiers. *J Med Internet Res*, 26:e50652.

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.

Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima López, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (#SMM4H) shared tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 21–32, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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 applications (#SMM4H) 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.