Domain-specific Continued Pretraining of Language Models for Capturing Long Context in Mental Health

Shaoxiong Ji [1] **Tianlin Zhang** [2] **Kailai Yang** [2] **Sophia Ananiadou** [2]
Erik Cambria [3] **Jörg Tiedemann** [1]

[1] University of Helsinki [2] The University of Manchester [3] Nanyang Technological University
{shaoxiong.ji; jorg.tiedemann}@helsinki.fi; cambria@ntu.edu.sg
{kailai.yang,tianlin.zhang}@postgrad.manchester.ac.uk
{sophia.ananiadou}@manchester.ac.uk

Abstract

Pretrained language models have been used in various natural language processing applications. In the mental health domain, domain-specific language models are pretrained and released, which facilitates the early detection of mental health conditions. Social posts, e.g., on Reddit, are usually long documents. However, there are no domain-specific pretrained models for long-sequence modeling in the mental health domain. This paper conducts domain-specific continued pretraining to capture the long context for mental health. Specifically, we train and release MentalXLNet and MentalLongformer based on XLNet and Longformer. We evaluate the mental health classification performance and the long-range ability of these two domain-specific pretrained models. Our models are released in HuggingFace[1].

1 Introduction

Natural Language Processing (NLP) applied to mental healthcare (Le Glaz et al., 2021; Zhang et al., 2022) has received much attention with specific applications to bipolar disorder detection (Harvey et al., 2022), depression detection (Ansari et al., 2023), and suicidal ideation detection (Ji et al., 2021). Recent work applied pretrained language models and domain-specific continued pretraining in the mental health domain with public models released, such as MentalBERT and MentalRoBERTa (Ji et al., 2022b). However, due to the quadratic complexity of self-attention in the transformer network (Vaswani et al., 2017), the pretrained Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and its domain-specific variants have limited ability to capture long-range context, and the pretrianed models can only process sequence within 512 tokens in the downstream applications.

To address this issue, efficient transformers are proposed, such as Longformer (Beltagy et al., 2020) and Transformer-XL (Dai et al., 2019) to capture long context. Qin et al. (2023) conducted a systematic analysis on the long-range ability of efficient transformers. In mental healthcare, texts such as self-reported mental conditions are usually long documents. For example, in social network analysis on Reddit, users' posts are long, and each user might have multiple posts.

This paper focuses on mental health analysis with long documents. We conduct domain-specific continued pretraining with Longformer and XL-Net architectures in the mental health domain. Our contributions are as follows. We train and release two domain-specific language models, i.e., MentalXLNet and MentalLongformer. We evaluate the performance of these two models on various mental healthcare classification datasets. Finally, we discuss the long-range ability of these two models and summarize how to choose pretrained language representations for specific applications.

2 Methods and Materials

This section presents the methods and materials. The self-attention in the standard transformer architecture suffers from quadratic complexity with sequence length. As a result, the BERT model pretrained with a masked language modeling (MLM) objective limits the maximum sequence length to 512 tokens.

We introduce two transformer networks for long documents and domain-specific pretraining to continue the pretraining in the mental healthcare domain. Table 1 summarizes the learning objectives and sequence lengths of existing pretrained models for mental healthcare and models trained in this paper. For downstream classification tasks, the max sequence length of BERT and RoBERTa is 512.

---

[1] https://huggingface.co/AIMH

Model	Objective	Seq. Length
MentalBERT	MLM	128
MentalRoBERTa	MLM	128
MentalXLNet	PLM	512
MentalLongformer	MLM	4096

Table 1: A summary of pretrained models for mental healthcare

2.1 Transformers for Long Sequence Modeling

Longformer Longformer (Beltagy et al., 2020) proposes an efficient attention mechanism with a linear complexity that leverages local windowed attention and task-motivated global attention. It is better at autoregressive language modeling on long sequences than prior works. When pretrained with MLM objective, Longformer achieves better long sequence modeling capacity on various downstream tasks.

XLNet Transformer-XL (Dai et al., 2019) solves the context fragmentation issue of fixed-length contexts by devising the recurrent operations for segments in the self-attention network. XLNet (Yang et al., 2019) combines the best of autoregressive and autoencoding language modeling and adopts the Permutation Language Modeling (PLM) objective. It captures bidirectional context and avoids the discrepancy of masked positions between pretraining and fine-tuning in masked language models. XLNet has been utilized to train domain-specific models, e.g., Clinical XLNet (Huang et al., 2020) in the clinical domain.

2.2 Domain-specific Continued Pretraining

We use the same corpus for pretraining as used in MentalBERT (Ji et al., 2022b). The pretraining corpus is sourced from Reddit, an online forum of anonymous communities where people with similar interests can engage in discussions. For the purpose of our study, we focus on subreddits related to the mental health domain. The selected subreddits for mental health-related content include "r/depression", "r/SuicideWatch", "r/Anxiety", "r/offmychest", "r/bipolar", "r/mentalillness", and "r/mentalhealth". We obtain the posts of users from these subreddits through web crawling. Even though user profiles are publicly available, we do not collect them while gathering the pretraining corpus. We conduct continued pretraining on the cluster with Nvidia A100

and V100 GPUs and utilize four GPUs in a single computing node.

3 Evaluation

Our study aims to investigate the effectiveness of using pretrained language models in detecting binary and multi-class mental disorders such as stress, anxiety, and depression and capturing long context in self-reported mental health texts. To achieve this, we fine-tune the language models in downstream tasks. For the classification model, we adopt the pooled representations of the last hidden layer and the multi-layer perceptron (MLP) with the hyperbolic tangent activation function. The learning rate of the transformer text encoder is set to 1e-05, and the learning rate of the classification layers is 3e-05. We use the Adam optimizer (Kingma and Ba, 2014) for training.
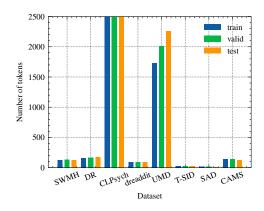

Figure 1: Median number of tokens in the datasets

3.1 Datasets

We consider various mental health classification tasks, i.e., depression, stress, and suicidal ideation, and the cause classification task on stress and mental disorder. For depression detection, we use two datasets, CLPsych15 (Coppersmith et al., 2015) and Depression_Reddit (DR) (Pirina and Çöltekin, 2018), collected from Reddit. These two datasets contain binary labels. For stress, we adopt two datasets, i.e., Dreaddit (Turcan and McKeown, 2019) and SAD (Mauriello et al., 2021), collected from Reddit and short text messages, respectively. Dreaddit is a dataset for binary stress classification. The SAD dataset is annotated with stress causes, including school, financial problems, family issues, social relationships, work, health or physical pain, emotional turmoil, everyday decision-making, and other uncategorized causes. We use the T-SID

Category	Platform	Dataset	Train	Validation	Test
Assorted	Reddit	SWMH (Ji et al., 2022a)	34,823	8,706	10,883
Depression	Reddit	Depression_Reddit (Pirina and Çöltekin, 2018)	1,004	431	406
Depression	Reddit	CLPsych15 (Coppersmith et al., 2015)	457	197	300
Stress	Reddit	Dreaddit (Turcan and McKeown, 2019)	2,270	568	715
Suicide	Reddit	UMD (Shing et al., 2018)	993	249	490
Suicide	Twitter	T-SID (Ji et al., 2022a)	3,072	768	960
Stress	SMS-like	SAD (Mauriello et al., 2021)	5,548	617	685
Assorted	Reddit	CAMS (Garg et al., 2022)	2,208	321	626

Table 2: A summary of datasets. Note we hold out a portion of the original training set as the validation set if the original dataset does not contain a validation set.

dataset (Ji et al., 2022a) for suicidal ideation detection, which also contains tweets with depression and post-traumatic stress disorders.

We also use Reddit posts collected by the UMD (Shing et al., 2018). SWMH (Ji et al., 2022a) contains posts with several mental health conditions and suicidal ideation. The last dataset used in our experiments is CAMS (Garg et al., 2022) for causal analysis of mental health issues. We utilize the causal categorization, which consists of six categories, i.e., bias or abuse, jobs and career, medication, relationships, alienation, and no reason. Datasets are summarized in Table 2 with the median number of tokens of each split shown in Figure 1. The CLPsych and UMD datasets containing multiple posts are extremely long, with over 2,500 and 1,500 tokens in the median, respectively. We analyze the long-range ability of MentalXLNet and MentalLongformer on these two datasets.

3.2 Baselines

We compare our models with three categories of models, i.e., models from the original checkpoint, models with domain-specific continued pretraining, and ChatGPT[2] in the zero-shot setting.

Models from the Huggingface Checkpoint We consider BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and Longformer (Beltagy et al., 2020), using the model checkpoint released in HuggingFace transformers (Wolf et al., 2020).

Models with Domain-specific Continued Pretraining We then compare our models with two domain-specific models for mental health, i.e., MentalBERT and MentalRoBERTa (Ji et al., 2022b).

---

[2] https://openai.com/blog/chatgpt

Zero-shot ChatGPT We compare our models with zero-shot ChatGPT referring to Yang et al. (2023). ChatGPT_{ZS} uses ChatGPT as the inference engine with simple manual prompts, ChatGPT_V, ChatGPT_{N_sen}, and ChatGPT_{N_emo} inject distant supervision signals from lexicons, i.e., NRC EmoLex sentiments (Mohammad and Turney, 2013), VADER sentiments (Hutto and Gilbert, 2014), and SenticNet (Cambria et al., 2022), to the prompts. ChatGPT_{CoT} and ChatGPT_{CoT_emo} utilize the Chain-of-Thought method (Wei et al., 2022) and its enhancement with emotion (Amin et al., 2023).

3.3 Main Results

Table 3 presents the results on eight test sets, comparing the models trained in this paper with baselines in a supervised setting and with ChatGPT in the zero-shot setting. Table 4 only compares the performance with supervised baselines because ChatGPT does not achieve strong performance compared to strong supervised methods, and we have no budget to run more experiments with ChatGPT. MentalXLNet and MentalLongformer outperform other baselines in most cases, especially on datasets with longer sequences, such as CLPsych15. MentalRoBERTa performs best on Dreaddit, showing its robust performance on short sequences, while MentalLongformer also gets a comparable performance. On the SMS-like SAD dataset, Longformer achieves the best performance, while the MentalLongformer trained on Reddit gets a slightly worse performance.

3.4 Results of Long-range Ability

We analyze the long-range ability by inputting various lengths of text into the model. Table 5 shows the results on UMD and CLPsych15 datasets with longer documents than others. The results indicate an increasing trend in recall and F1 scores with

Model	DR		CLPsych15		Dreaddit		T-SID		SAD		CAMS	
	Rec.	F1	Rec.	F1	Rec.	F1	Rec.	F1	Rec.	F1	Rec.	F1
BERT	91.13	90.90	64.67	62.75	78.46	78.26	88.44	88.51	62.77	62.72	40.26	34.92
RoBERTa	95.07	95.11	67.67	66.07	80.56	80.56	88.75	88.76	66.86	67.53	41.18	36.54
XLNet	90.89	90.44	69.83	69.12	78.88	78.84	86.04	86.18	67.30	67.30	50.64	49.16
Longformer	95.81	95.74	75.67	75.47	81.54	81.45	89.58	89.63	**69.20**	**69.01**	49.52	49.42
MentalBERT	94.58	94.62	64.67	62.63	80.28	80.04	88.65	88.61	67.45	67.34	45.69	39.73
MentalRoBERTa	94.33	94.23	70.33	69.71	**81.82**	**81.76**	88.96	89.01	68.61	68.44	50.48	47.62
ChatGPT$_{ZS}$	82.76	82.41	60.33	56.31	72.72	71.79	39.79	33.30	55.91	54.05	32.43	33.85
ChatGPT$_V$	79.51	78.01	59.20	56.34	74.23	73.99	40.04	33.38	52.49	50.29	28.48	29.00
ChatGPT$_{N\_sen}$	80.00	78.86	58.19	55.50	70.87	70.21	39.00	32.02	52.92	51.38	26.88	27.22
ChatGPT$_{N\_emo}$	79.51	78.41	58.19	53.87	73.25	73.08	39.00	32.25	54.82	52.57	35.20	35.11
ChatGPT$_{CoT}$	82.72	82.90	56.19	50.47	70.97	70.87	37.66	32.89	55.18	52.92	39.19	38.76
ChatGPT$_{CoT\_emo}$	83.17	83.10	61.41	58.24	75.07	74.83	34.76	27.71	58.31	56.68	43.11	42.29
MentalXLNet	95.32	95.24	71.67	71.49	80.42	80.41	89.17	89.12	**69.20**	68.76	**50.80**	**50.08**
MentalLongformer	**96.55**	**96.53**	**77.00**	**76.32**	81.12	81.05	**89.90**	**89.89**	68.76	68.44	49.20	48.74

Table 3: Results of mental health classification. The bold text represents the best performance. Note that: for Longformer and MentalLongformer, the best results are reported with longer texts as inputs.

Model	UMD		SWMH	
	Rec.	F1	Rec.	F1
BERT	61.63	58.01	69.78	70.46
RoBERTa	59.39	60.26	70.89	72.03
XLNet	63.06	60.09	70.60	70.57
Longformer	**74.29**	**72.85**	71.86	71.79
MentalBERT	64.08	58.26	69.87	71.11
MentalRoBERTa	57.96	58.58	70.65	**72.16**
MentalXLNet	63.06	60.29	71.07	71.18
MentalLongformer	73.06	72.47	**72.05**	72.08

Table 4: Results on mental health classification on UMD and SWMH

Dataset	Seq. Len.	Longformer		MentalLongformer	
		Rec.	F1	Rec.	F1
UMD	512	65.10	58.36	62.24	**59.74**
	1024	63.27	62.34	**64.29**	**66.22**
	1536	67.55	66.90	67.55	66.90
	2048	65.92	67.90	**70.82**	**69.19**
	2560	68.98	68.10	**72.04**	**72.53**
	3072	71.43	72.15	**72.65**	69.76
	3584	62.65	66.08	**72.45**	**72.13**
	4096	74.29	72.85	73.06	72.47
CLP	512	64.33	63.44	59.00	54.85
	1024	70.67	69.68	**71.33**	**70.76**
	1536	69.00	67.27	**71.00**	**69.57**
	2048	75.33	75.26	**72.32**	72.00
	2560	75.00	74.57	**76.00**	**75.69**
	3072	65.33	62.53	**72.33**	**70.97**
	3584	72.00	70.91	**75.00**	**74.31**
	4096	75.67	75.47	**77.00**	**76.32**

Table 5: Long-range ability analysis on UMD and CLPsych15. The bold text indicates that MentalLongformer achieves better scores.

a certain degree of fluctuation when the sequence length increases. Domain-specific continued pretraining continues to improve performance in most cases.

These results show the long-range ability of Longformer and XLNet and their domain-specific variants and also verify the effectiveness of domain-specific continued pretraining. However, there is no clear explanation for the fluctuation due to the black-box nature of transformers and the lack of human-grounded evaluation. One possible guess is that longer texts provide more information but can also introduce redundancy that impairs the model performance.

4 Related Work

Mental health surveillance in social media has gained increasing research attention from the NLP community. Le Glaz et al. (2021) conducted a systematic review of machine learning and NLP in mental health, summarized the state of the art in this field, discussed the challenges and opportunities, and provided recommendations for future research.

Emotion information in social posts is an important cue for mental health detection. Zhang et al. (2023) surveyed emotion fusion methods for mental illness detection from social media. Various approaches used NLP and machine learning for mental health classification. Krishnamurthy et al. (2016) presented a hybrid statistical and semantic model for identifying mental health and behavioral disorders using social network analysis. Ji et al. (2018) evaluated supervised learning

methods for detecting suicidal ideation in online user content. Nijhawan et al. (2022) studied stress detection. Spruit et al. (2022) explored language markers of mental health in psychiatric stories.

These papers highlight the potential of natural language processing and machine learning for improving the early detection of mental health conditions. Ive et al. (2020) focused on the generation and evaluation of artificial mental health records for natural language processing. Ji (2022) emphasized the importance of intention understanding in suicide risk assessment with pretrained language models.

5 Conclusion

We train and release two domain-specific language models in mental health, i.e., MentalXLNet and MentalLongformer. We empirically analyze the performance of these two models on various mental health classification datasets. We validate that the domain-specificity of pretrained language models can improve the performance of downstream tasks. For short texts within 512 tokens, we recommend MentalRoBERTa and MentalXLNet. For longer texts, MentalLongformer is a better choice.

Ethical Statement

Privacy is important in the mental health domain. We use social media posts that are manifestly public and do not collect user profiles when pretraining language models and fine-tuning classification models. We use the corpus collected from Reddit and do not interact with users who post on Reddit. Language models can be biased if the collected social posts contain inherent biases. Models pretrained and fine-tuned in this paper can not replace psychiatric diagnoses. We recommend individuals experiencing a mental health condition seek help from mental health professionals.

Acknowledgments

Limitations

We conduct domain-specific continued pretraining on a Reddit corpus, which means the domain-specificity primarily applies to Reddit social media. Social media data can be biased. However, we did not investigate this issue due to the lack of resources. This could be a future research task.

References

Mostafa Amin, Erik Cambria, and Björn Schuller. 2023. Will affective computing emerge from foundation models and General AI? A first evaluation on Chat-GPT. *IEEE Intelligent Systems*, 38(2).

Luna Ansari, Shaoxiong Ji, Qian Chen, and Erik Cambria. 2023. Ensemble hybrid learning methods for automated depression detection. *IEEE Transactions on Computational Social Systems*, 10:211–219.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *LREC*, pages 3829–3839.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on CLPsych*, pages 31–39.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

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*.

Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, and Vijay Mago. 2022. CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts. In *LREC*, pages 6387–6396.

Daisy Harvey, Fiona Lobban, Paul Rayson, Aaron Warner, Steven Jones, et al. 2022. Natural language processing methods and bipolar disorder: scoping review. *JMIR Mental Health*, 9(4):e35928.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward Moseley, Chih-Ying Deng, Naomi George, and Charolotta Lindvall. 2020. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 94–100.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *NPJ digital medicine*, 3(1):69.

Shaoxiong Ji. 2022. Towards intention understanding in suicidal risk assessment with natural language processing. In *Findings of EMNLP*, pages 4028–4038. Association for Computational Linguistics.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2022a. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, 34:10309–10319.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2021. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8:214–226.

Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised learning for suicidal ideation detection in online user content. *Complexity*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022b. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*, pages 7184–7190, Marseille, France. European Language Resources Association.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Madan Krishnamurthy, Khalid Mahmood, and Pawel Marcinek. 2016. A hybrid statistical and semantic model for identification of mental health and behavioral disorders using social network analysis. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1019–1026. IEEE.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708.

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. *arXiv preprint arXiv:1907.11692*.

Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. 2021. Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Tanya Nijhawan, Girija Attigeri, and T Ananthakrishna. 2022. Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, 9(1):1–24.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12.

Guanghui Qin, Yukun Feng, and Benjamin Van Durme. 2023. The NLP task effectiveness of long-range transformers. In *Proceedings of EACL*.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on CLPsych*, pages 25–36.

Marco Spruit, Stephanie Verkleij, Kees de Schepper, and Floortje Scheepers. 2022. Exploring language markers of mental health in psychiatric stories. *Applied Sciences*, 12(4):2179.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. On the evaluations of ChatGPT and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Tianlin Zhang, Annika Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine*, 5.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.