CEUR-WS.org/Vol-3497/paper-055.pdf

Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms from Social Media

Notebook for the eRisk Lab at CLEF 2023

Ana-Maria Bucur[1,2,*]

[1] *Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania*
[2] *PRHLT Research Center, Universitat Politècnica de València, Spain*

Abstract

In this work, we present the contribution of the BLUE team in the eRisk Lab task on searching for symptoms of depression. The task consists of retrieving and ranking Reddit social media sentences that convey symptoms of depression from the BDI-II questionnaire. Given that synthetic data provided by LLMs have been proven to be a reliable method for augmenting data and fine-tuning downstream models, we chose to generate synthetic data using ChatGPT for each of the symptoms of the BDI-II questionnaire. We designed a prompt such that the generated data contains more richness and semantic diversity than the BDI-II responses for each question and, at the same time, contains emotional and anecdotal experiences that are specific to the more intimate way of sharing experiences on Reddit. We perform semantic search and rank the sentences' relevance to the BDI-II symptoms by cosine similarity. We used two state-of-the-art transformer-based models (MentalRoBERTa and a variant of MPNet) for embedding the social media posts, the original and generated responses of the BDI-II. Our results show that using sentence embeddings from a model designed for semantic search outperforms the approach using embeddings from a model pre-trained on mental health data. Furthermore, the generated synthetic data were proved too specific for this task, the approach simply relying on the BDI-II responses had the best performance.

Keywords

depression symptoms, Beck's Depression Inventory, ChatGPT, Large Language Models

1. Introduction

Depression is one of the most prevalent mental disorders, with 5% of adults[1] suffering from it. Even if there is effective treatment, depression remains undiagnosed in some individuals due to the lack of access to medical services or stigma around mental illnesses [1]. For depression screening, mental health professionals use different scales, such as Center of Epidemiological Scales-Depression (CES-D) [2], Patient Health Questionnaire-9 (PHQ-9) [3], Beck's Depression Inventory-II (BDI-II) [4] and Hamilton Rating Scale for Depression (HRSD) [5]. With the rise in social media use and the anonymity and support provided on these platforms [6], researchers from both natural language processing and psychology began using social media data to search

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[*] Corresponding author.
ana-maria.bucur@drd.unibuc.ro (A. Bucur)
0000-0003-2433-8877 (A. Bucur)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://www.who.int/news-room/fact-sheets/detail/depression

for symptoms or signs of mental disorders in online users. In recent years, the field of mental illnesses detection shifted from black-box approaches providing only binary labels [7, 8] to explainable, interpretable approaches [9, 10] incorporating information from the depression screening scales. With the recent advancement in Large Language Models (LLMs) [11, 12], there have been efforts in testing their capabilities on mental health assessment.

The eRisk lab on Early Risk Prediction on the Internet of mental disorders started in 2017, with the pilot task of early risk detection of depression from social media data. From then on, the lab organized several tasks yearly and expanded to other mental illnesses such as eating disorders, pathological gambling and self-harm. The tasks consisted of detecting these mental health problems from social media data as early as possible, or automatically filling in questionnaires used by mental health professionals to diagnose depression or eating disorders. In the current edition, the task consists in retrieving and ranking social media posts with depression symptoms from the BDI-II questionnaire.

In this work, we present our proposed method for searching symptoms of depression, as part of the eRisk Lab. Inspired by recent works on generating and augmenting data using LLMs [13, 14], we follow a similar approach, and generate synthetic Reddit posts for each of the BDI-II symptoms such that the generated data has more diversity than the responses from BDI-II. We hypothesize that by generating synthetic data similar to the BDI-II responses with ChatGPT, we will add more diversity to the data and will be able to retrieve more relevant sentences. We aim for the generated data to resemble Reddit posts, in which users share their experiences more intimately. We explore different approaches based on pre-trained transformer-based models for encoding the social media data, the BDI-II responses and the synthetic data generated by LLMs. We perform semantic search and use cosine similarity to get the most relevant social media posts to the original and generated queries. Our results infirm our hypothesis that generated data improve the results. The semantic search model utilizing the original BDI-II responses as queries performs better than the model using generated data. The data generated by ChatGPT is too specific, and future work needs to be done to manipulate the prompt such that data is semantically similar and more diverse than the BDI-II responses, but, at the same time, has fewer specific details. However, the generated text is informative and generating mental health data with LLMs is a promising research direction.

2. Related work

Approaches in NLP for mental disorders detection from social media data achieved state-of-the-art results by using Convolutional Neural Networks (CNN) [8, 15], Recurrent Neural Networks (RNN) [16, 17], Hierarchical Attention Networks (HAN) [18, 19], and transformer-based architectures [7, 20, 21, 22]. However, most methods output binary labels for classification, operate as black boxes and are not interpretable. They cannot be used in real-life scenarios due to the lack of trust from mental health professionals [23].

Recently, there have been efforts in augmenting mental disorder detection methods with information from clinical questionnaires such as CES-D, PHQ-9, BDI-II, and HRSD. Nguyen et al. [23] proposed several approaches for depression detection that were constrained by the presence of the symptoms from PHQ-9. The proposed models consisted of two components, a questionnaire model that predicted the PHQ-9 symptoms and a depression model which used the symptom features for prediction. The authors showed that the models constrained on PHQ-9 had comparable performance to unconstrained methods, could better generalize to other datasets and are interpretable. Similarly, Zhang et al. [9] performed symptom-assisted mental disorders identification, achieving better results than baselines that use only text. Furthermore, their method was interpretable and provided symptom-based explanations for several mental health disorders, such as depression, anxiety, bipolar disorder, obsessive-compulsive disorder, eating disorders, ADHD and post-traumatic stress disorder. Psychiatric scales were also used for screening risky posts with HANs for early risk detection of depression [10]. Liu et al. [24] crawled data from different subreddits corresponding to 13 depression symptoms (e.g., *r/insomnia*, *sleep* for sleep problems, *r/chronicfatigue*, *r/Fatigue* for fatigue, etc.). Different models were trained on the data to detect each symptom. The predictions of these symptom detection models on Facebook data were validated against PHQ-9, General Anxiety Disorder-7 (GAD-7) and UCLA Loneliness Scale (UCLA-3) filled in by individuals. The authors showed that the automatically predicted symptoms were significantly associated with the symptoms checked by the self-report surveys, except for fatigue.

With recent advancements in LLMs [11, 12], there have been efforts in evaluating them for mental health assessment [25, 26]. Yang et al. [25] compared ChatGPT[2] with three supervised baselines and showed that, even if ChatGPT can achieve good results in a zero-shot classification setting, it lacks behind transformer-based specialized models for downstream tasks such as suicide and depression identification from social media data. Amin et al. [26] performed an interpretable mental health analysis through emotional reasoning using ChatGPT on 11 datasets across 5 tasks related to depression, stress and suicide ideation. Their results showed that zero-shot ChatGPT performed better than traditional neural network architectures but could not surpass the performance of specialized transformer-based models. The authors performed human evaluations and tested the impact of emotional reasoning in mental health assessment. Using emotional reasoning improved ChatGPT's performance, and the model could generate explanations for its predictions.

Besides mental health assessment, other applications of LLMs are generating and augmenting data [14, 27, 28]. Meyer et al. [13] evaluated the synthetic data generated by GPT-3 [11] for conversational tasks. The authors showed that the performance of classifiers trained on synthetic data performed worse than classifiers trained on fewer samples of real user-generated data. The data generated by GPT-3 has less variability than the real data. However, generating synthetic data might be a suitable approach in a scenario with a very small amount of data or resources available.

In line with these approaches of using LLMs to generate synthetic data, we use ChatGPT to generate data similar to the BDI-II questionnaire responses, simulating how social media users disclose their feelings and experiences on Reddit. We use the original BDI-II responses and the generated data as queries for semantic search and retrieve the most relevant sentences by their cosine similarity to the queries.

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

3. eRisk 2023 - Task 1: Search for symptoms of depression

The first task from the eRisk 2023 Lab [29] consists of ranking sentences from social media posts according to their relevance to the symptoms from Beck Depression Inventory–II (BDI-II) [4]. The BDI-II is a questionnaire used by mental health professionals to screen for depression and consists of 21 questions related to symptoms of depression such as sadness, pessimism, loss of pleasure, loss of interest, tiredness and others. Each question corresponds to one of the symptoms. BDI-II is a Likert scale survey, for each question there are 4 possible responses measuring the intensity of the symptom from the absence of it, to its maximum intensity (with the exception of item 16 and 18, which have 7 possible responses). The challenge consists of ranking the sentences from Reddit by their relevance to each of the symptoms of the BDI-II. A given sentence is considered relevant to a symptom if it contains information about the user's mental state regarding the symptom, even if the user mentions that they do not suffer from the given symptom. The data for this task was compiled from the eRisk past data and was organized as TREC formatted sentences for each user. A total of approx 4 million of sentences from 3,107 users were provided for this task.

For evaluating the systems' performance and assess the sentences' relevance to the BDI-II symptoms, top-k pooling was used, with k equal to 50. The top 50 relevant sentences for each symptom from each system were combined in a pool of relevant sentences. These sentences were further assessed as being relevant or not to the symptoms by three annotators. A sentence was considered relevant to a symptom if it contained information about the state of the individual and is topically-related to the BDI-II symptoms.

4. Method

To search for symptoms of depression in Reddit data, we proposed an approach based on semantic search using as queries the corresponding responses for each item from BDI-II. Inspired by previous works that use LLMs to generate synthetic data [13, 14], we also experimented with generating synthetic Reddit posts with ChatGPT to be used as queries. We aimed for the generated data to have more diversity than the BDI-II responses, while preserving the meaning, and to be expressed more intimately, specific to Reddit.

Synthetic data provided by LLMs have been successfully used in other works [14, 28, 30, 31] and have proved to be a reliable method for augmenting and fine-tuning downstream models. We generated synthetic data using `text-davinci-3` [11] for each item of the BDI-II questionnaire. We used the OpenAI text completion API[3] and designed a prompt such that the answers had more diversity than the BDI-II responses and conveyed the intimate way of sharing experiences and feelings specific to Reddit [6]. In Table 1, we showcase the prompt we used to instruct the model, similar to the approach of Wang et al. [30]. However, our prompt was simpler, and geared towards simulating user responses, not tasks with their outcomes. We included instructions that limited the size of the text, ensured semantic diversity in the generated texts and ensured that the generated data contained emotional and anecdotal experiences that aligned with each BDI-II item. In Algorithm 1, we showcase our algorithm for generating data using

[3] https://platform.openai.com/docs/guides/gpt/completions-api

the OpenAI API. Each completion was post-processed by removing trailing quotation marks, enumeration numbers, and splitting by *newline* to obtain individual texts. BDI-II contains 21 items related to depression symptoms, with a total of 90 possible responses measuring the intensity of symptoms. For each of these 90 responses, we generated 30 synthetic Reddit posts, totaling 2,700 generated texts. We show in Table 2 some generated examples for the first symptom of the BDI-II questionnaire. The generated texts were longer than the BDI-II responses and had greater diversity. Some examples even contained self-disclosure, which is specific for Reddit data [6], such as "My cat passed away", "I just broke up with my partner". We hypothesized that, by augmenting the queries with the synthetically generated data, we would be able to retrieve more relevant sentences.

```
You are asked to come up with a set of "{N}" diverse reddit posts
that are examples to the BDI depression questionnaire for the
"{symptom}" symptom. For this symptom, the BDI answer of
interest is "{item}".
These examples will be given to a ranking model that will compute
the similarity between the answer item text and the reddit post.

Here are the requirements:
1. The language used for the reddit posts should be diverse. For
example, you should combine descriptions of past experiences with
feelings or events.
2. The reddit posts should be in English.
3. The reddit posts should be 2 to 3 sentences long.
4. The reddit posts should provide substantial content to make
ranking feasible.
5. The reddit posts should be specific and not just describe
general situations, but rather specific personal experiences and
self-disclosure.
6. The reddit posts should, as much as possible, not contain the
exact words of the BDI item.

List of "{N}" reddit posts:
```

Table 1
Instructions for the data generation that ensure diversity and adherence to the BDI-II responses

We pre-processed all the posts from each Reddit user by removing URLs and texts not in English, detected by the *polyglot* package. We computed the sentence embeddings for all texts (original posts, BDI-II responses, generated data) using two state-of-the-art methods, a variant of MPNet [32] and MentalRoBERTa [33]. The MPNet variant we used, `multi-qa-mpnet-base-dot-v1`[4] was explicitly designed for semantic search. MentalRoBERTa[5] was chosen because it was trained on a large corpus of mental health data, mainly from Reddit. The MentalRoBERTa model has shown good downstream performance for mental disorders detection from social media data [34, 35].

[4] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1
[5] https://huggingface.co/mental/mental-roberta-base

Algorithm 1 Generate synthetic data based on BDI-II items and responses

```
Require: instruction prompt $P$, questionnaire $Q$
  $D \leftarrow \{\}$
  for symptom $s$ and item $q \in Q$ do
    $p$ = $P$.format(N = $N$, item = $q$, symptom = $s$)
    $d$ = openai.Completion.create(model = "text-davinci-003", prompt = $p$)
    $d$ = postprocess($d$)
    $D = D \cup \{$d$\}$
  end for
      return $D$
```

Table 2
Examples generated by GPT 3 for the first symptom of BDI-II: Sadness

BDI-II Response	Generated text
I do not feel sad.	My cat passed away last week and although I miss her terribly, I'm not feeling any sadness about it. I just broke up with my partner, but I'm focusing on all the good memories we shared and not letting myself be consumed by sadness.
I feel sad much of the time.	I feel like I'm stuck in an endless cycle of sadness and emptiness. It feels like no matter how hard I try, I can't shake this sadness that's been following me around.
I am sad all the time.	I'm feeling so empty and down lately, like nothing can make me happy anymore. I feel so overwhelmed with sadness that it's hard to get out of bed in the morning.
I am so sad or unhappy that I can't stand it.	I feel like I'm stuck in a dark hole with no way out and it's suffocating me. I'm so overwhelmed by sadness that I can barely function anymore.

We performed semantic search and used cosine similarity between embeddings to get the most relevant social media posts to the original BDI-II responses and generated queries. We retrieved the top 50 sentences with the highest cosine similarity to the queries. Given that we could submit a maximum of 1,000 results for each submission, we sorted the retrieved sentences by cosine similarity scores and kept only the most relevant 1,000 sentences.

We experimented with different queries and embedding methods, as explained below:

SemSearchOnBDI2Queries We performed semantic search using the original 90 BDI-II responses as queries. All texts were encoded using `multi-qa-mpnet-base-dot-v1`.

SemSearchOnGeneratedQueries We performed semantic search using the 2,700 generated synthetic Reddit texts as queries, with `multi-qa-mpnet-base-dot-v1` embeddings.

SemSearchOnAllQueries We use all the original and generated queries and perform semantic search on texts encoded with `multi-qa-mpnet-base-dot-v1`.

SemSearchOnBDI2QueriesMentalRoberta We use the original BDI-II responses as queries, but MentalRoBERTa is used for embedding the data.

SemSearchOnGeneratedQueriesMentalRoberta We perform semantic search using the generated data as queries; texts were encoded using MentalRoBERTa.

5. Results

Table 3
Ranking-based evaluation for Task 1 (majority voting)

Team	Run	AP	R-PREC	P@10	NDCG@1000
Formula-ML	SentenceTransformers_0.25	**0.319**	**0.375**	**0.861**	**0.596**
OBSER-MENH	salida-distilroberta-90-cos	0.294	0.359	0.814	0.578
uOttawa	USESim	0.160	0.248	0.600	0.382
NailP	T1_M2	0.095	0.146	0.519	0.226
RELAI	bm25|mpnetbase	0.048	0.081	0.538	0.140
UNSL	Prompting-Classifier	0.036	0.090	0.229	0.180
UMU	LexiconMultilingualSentenceTransformer	0.073	0.140	0.495	0.222
GMU	FAST-DCMN-COS-INJECT_FULL	0.001	0.003	0.014	0.005
Mason-NLP	MentalBert	0.035	0.072	0.286	0.117
BLUE	SemSearchOnBDI2Queries	0.104	0.126	0.781	0.211
BLUE	SemSearchOnAllQueries	0.065	0.086	0.629	0.160
BLUE	SemSearchOnGeneratedQueries	0.052	0.074	0.586	0.139
BLUE	SemSearchOnBDI2QueriesMentalRoberta	0.027	0.044	0.386	0.089
BLUE	SemSearchOnGeneratedQueriesMentalRoberta	0.029	0.063	0.367	0.105

Table 4
Ranking-based evaluation for Task 1 (unanimity)

Team	Run	AP	R-PREC	P@10	NDCG@1000
Formula-ML	SentenceTransformers_0.25	0.268	**0.360**	**0.709**	**0.615**
Formula-ML	SentenceTransformers_0.1	**0.293**	0.350	0.685	0.611
OBSER-MENH	salida-distilroberta-90-cos	0.281	0.344	0.652	0.604
uOttawa	USESim	0.139	0.232	0.438	0.380
NailP	T1_M2	0.090	0.143	0.410	0.229
UMU	LexiconMultilingualSentenceTransformer	0.059	0.125	0.333	0.209
RELAI	bm25|mpnetbase	0.039	0.069	0.343	0.124
UNSL	Prompting-Classifier	0.020	0.063	0.090	0.157
GMU	FAST-DCMN-COS-INJECT_FULL	0.001	0.003	0.014	0.006
Mason-NLP	MentalBert	0.024	0.054	0.190	0.099
BLUE	SemSearchOnBDI2Queries	0.129	0.167	0.643	0.260
BLUE	SemSearchOnAllQueries	0.067	0.105	0.452	0.177
BLUE	SemSearchOnGeneratedQueries	0.052	0.088	0.381	0.147
BLUE	SemSearchOnBDI2QueriesMentalRoberta	0.032	0.058	0.300	0.104
BLUE	SemSearchOnGeneratedQueriesMentalRoberta	0.018	0.059	0.186	0.085

The results of the eRisk Lab task on searching for depression symptoms are presented in Tables 3 and 4. We show the results of all 5 runs submitted by us and the best-performing run from each other team. The metrics used for evaluating the relevance of the sentences were Average Precision (AP), R-Precision, Precision at 10 (P@10), and Normalized Discounted Cumulative Gain at 1000 (NDCG@1000). Table 3 presents the systems' performance compared to the gold standard obtained from majority voting of the relevant sentences assessed by the annotators. Table 4 presents the systems' performance compared to the gold standard obtained from the sentences considered relevant by all three annotators. Comparing our proposed methods, the model using only the BDI-II responses as queries, SemSearchOnBDI2Queries, performed best in both ranking-based evaluation settings, majority voting and unanimity, achieving 0.104 AP in the first scenario, and 0.129 AP in the second one. The second-best model was the one that used as queries all the texts (original and generated), SemSearchOnAllQueries, with an AP of 0.065 in majority voting evaluation, and 0.067 in unanimity evaluation. The model using only generated data as queries, SemSearchOnGeneratedQueries, had the lowest performance from the models using embeddings from MPNet. The SemSearchOnBDI2Queries model had a good P@10 of 0.781 for majority voting ranking-based evaluation and 0.643 for unanimity evaluation, showing that our semantic search method using MPNet embeddings on the original BDI-II queries was best at retrieving relevant sentences in top 10 documents. Even if the embeddings provided by the pre-trained model on mental health data, MentalRoBERTa, had a good performance for detection tasks [34, 35], it had the lowest performance for symptoms retrieval.

However, our proposed methods ranked fourth compared to all the systems developed by other participants in the eRisk task. Our hypothesis that the synthetically generated queries will improve performance was proved false. We aimed for variability, as the BDI-II responses were short and standard, but the texts generated by ChatGPT might be too specific. Some of the generated texts provided too many details, which were not helpful for semantic search: "I just got back from a great vacation and it's been really hard to get back into the swing of things - not feeling particularly sad, but definitely a bit down.", "I don't know what to do with myself anymore - no matter how hard I try, I can't shake this overwhelming sense of gloom.". For future work, we would like to experiment with different prompts to generate data semantically similar and more diverse than the BDI-II responses, with fewer specific details.

6. Conclusions

In this work, we presented the contributions of the BLUE team in the eRisk Lab task on retrieving relevant social media text relevant to the symptoms of depression from the BDI-II questionnaire. We performed semantic search using the original BDI-II responses and synthetically generated texts as queries. We hypothesized that, by using ChatGPT to generate synthetic data similar to Reddit posts in which users disclose their feelings and experiences, we could retrieve more relevant sentences for each BDI-II item. We experimented with two pretrained transformer-based methods to encode the queries and social media posts, MentalRoBERTa and a variant on MPNet designed specifically for semantic search. Our hypothesis was proved false; the model performing semantic search using as queries the original BDI-II responses outputted more relevant sentences than the one using generated data. The synthetic data generated by ChatGPT was too specific for retrieving depression symptoms, and future work needs to be done for prompt manipulation such that the model can generate suitable data for this task.

References

[1] A. Handy, R. Mangal, T. S. Stead, R. L. Coffee Jr, L. Ganti, Prevalence and impact of diagnosed and undiagnosed depression in the united states, Cureus 14 (2022).

[2] W. W. Eaton, C. Muntaner, C. Smith, A. Tien, M. Ybarra, Center for epidemiologic studies depression scale: Review and revision, The use of psychological testing for treatment planning and outcomes assessment (2004).

[3] K. Kroenke, R. L. Spitzer, J. B. Williams, The phq-9: validity of a brief depression severity measure, Journal of general internal medicine 16 (2001) 606–613.

[4] A. T. Beck, R. A. Steer, G. Brown, Beck depression inventory–ii, Psychological assessment (1996).

[5] M. Hamilton, A rating scale for depression, Journal of neurology, neurosurgery, and psychiatry 23 (1960) 56.

[6] M. De Choudhury, S. De, Mental health discourse on reddit: Self-disclosure, social support, and anonymity, in: Proceedings of ICWSM, volume 8, 2014, pp. 71–80.

[7] A.-M. Bucur, A. Cosma, L. P. Dinu, Early risk detection of pathological gambling, self-harm and depression using bert, in: CLEF (Working Notes), 2021.

[8] A. Yates, A. Cohan, N. Goharian, Depression and self-harm risk assessment in online forums, in: Proceedings of EMNLP, 2017, pp. 2968–2978.

[9] Z. Zhang, S. Chen, M. Wu, K. Zhu, Symptom identification for interpretable detection of multiple mental disorders on social media, in: Proceedings of EMNLP, 2022, pp. 9970–9985.

[10] Z. Zhang, S. Chen, M. Wu, K. Zhu, Psychiatric scale guided risky post screening for early detection of depression, in: Proceedings of IJCAI, 2022.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Proceedings of NeurIPS 33 (2020) 1877–1901.

[12] OpenAI, Gpt-4 technical report, arXiv (2023).

[13] S. Meyer, D. Elsweiler, B. Ludwig, M. Fernandez-Pichel, D. E. Losada, Do we still need human assessors? prompt-based gpt-3 user simulation in conversational ai, in: Proceedings of CUI, 2022, pp. 1–6.

[14] H. Dai, Z. Liu, W. Liao, X. Huang, Z. Wu, L. Zhao, W. Liu, N. Liu, S. Li, D. Zhu, et al., Chataug: Leveraging chatgpt for text data augmentation, arXiv preprint arXiv:2302.13007 (2023).

[15] G. Rao, Y. Zhang, L. Zhang, Q. Cong, Z. Feng, Mgl-cnn: a hierarchical posts representations model for identifying depressed individuals in online forums, IEEE Access 8 (2020) 32395–32403.

[16] M. Trotzek, S. Koitka, C. M. Friedrich, Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, IEEE Transactions on Knowledge and Data Engineering 32 (2018) 588–601.

[17] R. Skaik, D. Inkpen, Using twitter social media for depression detection in the canadian population, in: Proceedings of AICCC, 2020, pp. 109–114.

[18] A.-S. Uban, P. Rosso, Deep learning architectures and strategies for early detection of self-harm and depression level prediction, in: CLEF (Working Notes), volume 2696, 2020, pp. 1–12.

[19] A. S. Uban, B. Chulvi, P. Rosso, Multi-aspect transfer learning for detecting low resource mental disorders on social media, in: Proceedings of LREC, 2022, pp. 3202–3219.

[20] D. Owen, J. Camacho-Collados, L. E. Anke, Towards preemptive detection of depression and anxiety in twitter, in: Proceedings of SMM4H Workshop, 2020, pp. 82–89.

[21] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020 (2020).

[22] A.-M. Bucur, A. Cosma, P. Rosso, L. P. Dinu, It's just a matter of time: Detecting depression with time-enriched multimodal transformers, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 200–215.

[23] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, A. Cohan, Improving the generalizability of depression detection by leveraging clinical questionnaires, in: Proceedings of ACL, 2022, pp. 8446–8459.

[24] T. Liu, D. Jain, S. R. Rapole, B. Curtis, J. C. Eichstaedt, L. H. Ungar, S. C. Guntuku, Detecting symptoms of depression on reddit, in: Proceedings of WebSci, 2023, pp. 174–183.

[25] K. Yang, S. Ji, T. Zhang, Q. Xie, S. Ananiadou, On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis, arXiv preprint arXiv:2304.03347 (2023).

[26] M. M. Amin, E. Cambria, B. W. Schuller, Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt, IEEE Intelligent Systems 38 (2023) 2.

[27] Y.-J. Lee, C.-G. Lim, Y. Choi, J.-H. Lm, H.-J. Choi, Personachatgen: Generating personalized dialogues using gpt-3, in: Proceedings of CCGPK Workshop, 2022, pp. 29–48.

[28] S. Ubani, S. O. Polat, R. Nielsen, Zeroshotdataaug: Generating and augmenting training data with chatgpt, arXiv preprint arXiv:2304.14334 (2023).

[29] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association, CLEF 2023, Springer International Publishing, Thessaloniki, Greece, 2023.

[30] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language model with self generated instructions, arXiv preprint arXiv:2212.10560 (2022).

[31] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[32] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, Proceedings of NeurIPS 33 (2020) 16857–16867.

[33] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, Mentalbert: Publicly available pretrained language models for mental healthcare, in: Proceedings of LREC, 2022, pp. 7184–7190.

[34] A. Aich, A. Quynh, V. Badal, A. Pinkham, P. Harvey, C. Depp, N. Parde, Towards intelligent clinically-informed language analyses of people with bipolar disorder and schizophrenia, in: Findings of EMNLP, 2022, pp. 2871–2887.

[35] D. Owen, D. Antypas, A. Hassoulas, A. F. Pardiñas, L. Espinosa-Anke, J. C. Collados, et al., Enabling early health care intervention by detecting depression in users of web-based forums using language models: Longitudinal analysis and evaluation, JMIR AI 2 (2023).