# MentalBERT:
# Publicly Available Pretrained Language Models for Mental Healthcare

**Shaoxiong Ji[†], Tianlin Zhang[‡], Luna Ansari[†], Jie Fu[§], Prayag Tiwari[†], Erik Cambria[¶]**

[†] Aalto University, Finland [‡] The University of Manchester, UK

[§] Mila, Québec AI Institute, Canada [¶] Nanyang Technological University, Singapore

{shaoxiong.ji; luna.ansari; prayag.tiwari}@aalto.fi

tianlin.zhang@postgrad.manchester.ac.uk

fujie@mila.quebec  cambria@ntu.edu.sg

## Abstract

Mental health is a critical issue in modern society, and mental disorders could sometimes turn to suicidal ideation without adequate treatment. Early detection of mental disorders and suicidal ideation from social content provides a potential way for effective social intervention. Recent advances in pretrained contextualized language representations have promoted the development of several domain-specific pretrained models and facilitated several downstream applications. However, there are no existing pretrained language models for mental healthcare. This paper trains and releases two pretrained masked language models, i.e., MentalBERT and MentalRoBERTa, to benefit machine learning for the mental healthcare research community. Besides, we evaluate our trained domain-specific models and several variants of pretrained language models on several mental disorder detection benchmarks and demonstrate that language representations pretrained in the target domain improve the performance of mental health detection tasks.

**Keywords:** Pretrained Language Models, Mental Healthcare, MentalBERT

## 1. Introduction

Mental health is a global issue, especially severe in most developed countries and many emerging markets. According to the mental health action plan (2013 - 2020) from the World Health Organization, one in four people worldwide suffer from mental disorders to some extent. Moreover, three out of four people with severe mental disorders do not receive treatment, worsening the problem. During some periods like the pandemic, people struggle with mental health issues, and many may not get mental health practitioners' help. Previous studies reveal that suicide risk usually has a connection to mental disorders (Windfuhr and Kapur, 2011). Partly due to severe mental disorders, 900,000 people commit suicide each year worldwide, making suicide the second most common cause of death among the young. Suicide attempters have been reported as suffering from mental disorders, with an investigation on a shift from mental health to suicidal ideation conducted by language and interactional measures (De Choudhury et al., 2016).

Early identification is a practical approach to mental illness and suicidal ideation prevention. Except for traditional proactive screening, social media is a good channel for mental health care. Social media platforms such as Reddit and Twitter provide anonymous space for users to discuss stigmatic topics and self-report personal issues. Social content from users who wrote about mental health issues and posted suicidal ideation has been widely used to study mental health issues (Ji et al., 2018; Tadesse et al., 2019). Machine learning-based detection techniques can empower healthcare

workers in early detection and assessment to take an action of proactive prevention. Recent advances in deep learning facilitate the development of effective early detection methods (Cambria et al., 2010; Cambria et al., 2012; Ji et al., 2022). A new trend in natural language processing (NLP), contextualized pretrained language models, has attracted much attention for various text processing tasks. The seminal work on a pretrained language model called BERT (Devlin et al., 2019) utilizes bidirectional transformer-based text encoders and trains the model on a large-scale corpus. With the success of BERT, several domain-specific pretrained language models for learning text representations have also been developed and released, such as biomedical BERT (Lee et al., 2020) and clinical BERT (Alsentzer et al., 2019; Huang et al., 2019) for the biomedical and clinical domain, respectively. However, there are no pretrained language models customized for the domain of mental healthcare.

Our paper trains and releases two representative bidirectional masked language models, i.e., BERT and RoBERTa (Liu et al., 2019), with corpus collected from social forums for mental health discussion. The pretrained models in the mental health domain are dubbed MentalBERT and MentalRoBERTa. To the best of our knowledge, this work is among the first to pre-train language models for mental healthcare. Besides, we conduct a comprehensive evaluation on several mental health detection datasets with pretrained language models in different domains. We release the pretrained MentalBERTs with Huggingface's model repository[1].

---

[1] http://huggingface.co/mental

## 2. Methods and Setup

This section introduces the language model pretraining technique and the pretraining corpus we collected. We then present the downstream tasks that we aim to solve by fine-tuning the pretrained models, and describing the setup of language model fine-tuning. Note that we aim to provide publicly available pretrained text embeddings as language resources and evaluate the usability in downstream tasks rather than propose novel pretraining techniques.

### 2.1. Language Model Pretraining

We follow the standard pretraining protocols of BERT and RoBERTa with Huggingface's Transformers framework (Wolf et al., 2020). These two models work in a bidirectional manner, and we follow their mechanism and adopt the same loss of masked language modeling during pretraining. We use the base network architecture for both models. The BERT model we use is base uncased, which is 12-layer, 768-hidden, and 12-heads and has 110M parameters. For the pretraining of RoBERTa-based MentalBERT, we apply the dynamic masking mechanism that converges slightly slower than the static masking. Instead of the domain-specific pretraining (Gu et al., 2020) that trains language models from scratch, we adopt the training scheme similar to the domain-adaptive pretraining (Gururangan et al., 2020) that continues the pretraining in specific downstream domains. Specifically, we start the training of language models from the checkpoint of original BERT and RoBERTa. In this way, we can utilize the learned knowledge from the general domain and save computing resources, and continued pretraining makes the model adaptive to the target domain of mental health.

We use four Nvidia Tesla v100 GPUs to train the two language models. The computing resources are also one of the main assets of this paper. We set the batch size to 16 per GPU, evaluate every 1,000 steps, and train for 624,000 iterations. Training with four GPUs takes around eight days, i.e., around 32 days with only one GPU.

### 2.2. Pretraining Corpus

We collect our pretraining corpus from Reddit, an anonymous network of communities for discussion among people of similar interests. Focusing on the mental health domain, we select several relevant subreddits (i.e., Reddit communities that have a specific topic of interest) and crawl the users' posts. We do not collect user profiles when collecting the pretraining corpus, even though those profiles are publicly available. The selected mental health-related subreddits include "r/depression", "r/SuicideWatch", "r/Anxiety", "r/offmychest", "r/bipolar", "r/mentalillness/", and "r/mentalhealth". Eventually, we make the training corpus with a total of 13,671,785 sentences.

### 2.3. Downstream Task Fine-tuning

We apply the pretrained MentalBERT and Mental-RoBERTa in binary mental disorder detection and multi-class mental disorder classification of various mental disorders such as stress, anxiety, and depression. We fine-tune the language models in downstream tasks. Specifically, we use the embedding of the special token `[CLS]` of the last hidden layer as the final feature of the input text. We adopt the multilayer perceptron (MLP) with the hyperbolic tangent activation function for the classification model. We set the learning rate of the transformer text encoder to be 1e-05 and the learning rate of classification layers to be 3e-05. The optimizer is Adam (Kingma and Ba, 2014).

## 3. Results

### 3.1. Datasets

We evaluate and compare mental disorder detection methods on different datasets with various mental disorders (e.g., depression, anxiety, and suicidal ideation) collected from popular social platforms (e.g., Reddit and Twitter). Table 1 summarizes the datasets used in this paper. We carefully choose those benchmarks to cover a relatively wide range of mental health categories and social platforms. Some datasets do not provide a validation set. Thus, we partition a small set from the original training set to make the validation set.

**Depression**  Depression is one of the most common mental disorders discussed on many social platforms. We take it as a representative to evaluate the performance of different pretrained models. The first dataset for depression comes from the CLPsych 2015 Shared Task (Coppersmith et al., 2015)[2]. The first task of CLPsych 2015 contains user-generated posts from users with depression on Twitter. The train partition consists of 327 depression users, and the test data contains 150 depression users. Note that there is an unknown data missing issue in the dataset of the CLPsych 2015 shared task. The second dataset used is from eRisk shared task 1 (Losada and Crestani, 2016), which is a public competition for early risk detection in health-related areas. The eRisk dataset contains posts from 2,810 users, where 1,370 users express depression in their posts and 1,440 act as the control group without depression. (Pirina and Çöltekin, 2018) collected additional social data form Reddit and combined them with previously collected data to identify depression[3]. We term this dataset as *Depression_Reddit* in this paper.

**Suicidal Ideation**  We use data collected from Reddit and Twitter to test the performance. Firstly, we use the UMD Reddit Suicidality Dataset (Shing et al., 2018) that has a total of 865 users in the subreddit of

---

Table 1: A summary of datasets. Note we hold out a portion of original training set as the validation set if the original dataset does not contain a validation set.

| Category | Platform | Dataset | train | validation | test |
|---|---|---|---|---|---|
| Assorted | Reddit | SWMH (Ji et al., 2022) | 34,823 | 8,706 | 10,883 |
| Depression | Reddit | eRisk18 T1 (Losada and Crestani, 2016) | 1,533 | 658 | 619 |
| 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., 2022) | 3,072 | 768 | 960 |
| Stress | SMS-like | SAD (Mauriello et al., 2021) | 5,548 | 617 | 685 |

"SuicideWatch" in Reddit[4]. The raw data annotation labels the user posts with four levels of risks. We include the control users and transform the label space into three classes according to the level of risks. In addition to data from Reddit, we also evaluate the performance of data collected from Twitter. We use the Twitter dataset with tweets expressing suicidal ideation and normal posts as the control group, which is collected by (Ji et al., 2018; Ji et al., 2022). We term this dataset as T-SID.

**Other Mental Disorders**   We also evaluate the performance of classifying other mental disorders such as stress, anxiety, and bipolar. Dreaddit (Turcan and McKeown, 2019) is a dataset for stress analysis with posts collected from five different forums of Reddit[5]. Specifically, it considers three major stressful topics, i.e., interpersonal conflict, mental illness, and financial need, and collects posts from ten related subreddits, including some mental health domains such as anxiety and PTSD. This dataset consists of a total of 3,553 posts split into train and test sets. Many factors may cause stress. We then use another dataset for recognizing everyday stressors called SAD, which contains 6,850 SMS-like sentences (Mauriello et al., 2021).

The SAD dataset derives nine stress factors from stress management articles, chatbot-based conversation systems, crowdsourcing, and web crawling. The specific stressor categories include work, health, fatigue, or physical pain, financial problem, emotional turmoil, school, everyday decision making, family issues, social relationships, and other unspecified stressors. Lastly, we use a dataset called SWMH (Ji et al., 2022) that contains Reddit posts with various mental disorders, including stress, anxiety, bipolar, depression, and suicidal ideation. Note that this dataset uses weak labels during the annotation process.

### 3.2.   Baselines

We compare our pretrained language models for mental health with various existing pretrained models in different domains. They are BERT and RoBERTa pretrained with general corpus, BioBERT pretrained in the biomedical domain, and ClinicalBERT pretrained with clinical notes. Note that the aim of this paper is not to achieve the state-of-the-art performance but to demonstrate the usability and evaluate the performance of our pretrained models, though we have achieved competitive performance in some datasets when compared with the state of the art.

### 3.3.   Results and Discussion

We evaluate the model performance by comparing the recall and F1 scores. Mental disorder detection is usually a task with unbalanced classes, leading to using the F1 score metric. It is also essential to reduce the false negatives, i.e., to ensure as few cases as possible that the detection model misses people with mental disorders. Thus, we also report recall scores.

**Results of Depression Detection**   We first compare the performance of depression detection. Table 2 reports the results on three depression dataset collected from Reddit. MentalRoBERTa archives the best performance on eRisk and CLPsych datasets, and MentalBERT is the second best model on the Depression_Reddit dataset.

**Results of Classifying Other Mental Disorders**   We then compare the performance of classifying other mental disorders and suicidal ideation. Table 3 shows the performance on various datasets with different mental disorder classification tasks. In T-SID, SAD, and Dreaddit, MentalRoBERTa is the best model with the highest recall and F1 scores. The MentalBERT has the highest F1 score in the UMD dataset, while its F1 score is not competitive to other models. While for the SWMH dataset with several mental disorders, the MentalRoBERT obtained the best F1 score.

**Discussion**   When comparing the domain-specific pretrained models for mental health with models pretrained with general corpora, MentalBERT and Men-

Table 2: Results of depression detection. The bold text represents for the best performance.

| Model | eRisk T1 | | CLPsych | | Depression_Reddit | |
|---|---|---|---|---|---|---|
| | Rec. | F1 | Rec. | F1 | Rec. | F1 |
| BERT | 88.53 | 88.54 | 64.67 | 62.75 | 91.13 | 90.90 |
| RoBERTa | 92.25 | 92.25 | 67.67 | 66.07 | **95.07** | **95.11** |
| BioBERT | 79.16 | 78.86 | 65.67 | 65.50 | 91.13 | 90.98 |
| ClinicalBERT | 76.25 | 75.41 | 65.67 | 65.30 | 89.41 | 89.03 |
| MentalBERT | 86.27 | 86.20 | 64.67 | 62.63 | 94.58 | 94.62 |
| MentalRoBERTa | **93.38** | **93.38** | **70.33** | **69.71** | 94.33 | 94.23 |

Table 3: Results of classifying other mental disorders including stress, anorexia, suicidal ideation. The bold text represents for the best performance.

| Model | UMD | | T-SID | | SWMH | | SAD | | Dreaddit | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Rec. | F1 | Rec. | F1 | Rec. | F1 | Rec. | F1 | Rec. | F1 |
| BERT | 61.63 | 58.01 | 88.44 | 88.51 | 69.78 | 70.46 | 62.77 | 62.72 | 78.46 | 78.26 |
| RoBERTa | 59.39 | **60.26** | 88.75 | 88.76 | **70.89** | 72.03 | 66.86 | 67.53 | 80.56 | 80.56 |
| BioBERT | 57.76 | 58.76 | 86.25 | 86.12 | 67.10 | 68.60 | 66.72 | 66.71 | 75.52 | 74.76 |
| ClinicalBERT | 58.78 | 58.74 | 85.31 | 85.39 | 67.05 | 68.16 | 62.34 | 61.25 | 76.36 | 76.25 |
| MentalBERT | **64.08** | 58.26 | 88.65 | 88.61 | 69.87 | 71.11 | 67.45 | 67.34 | 80.28 | 80.04 |
| MentalRoBERTa | 57.96 | 58.58 | **88.96** | **89.01** | 70.65 | **72.16** | **68.61** | **68.44** | **81.82** | **81.76** |

talRoBERTa gain better performance in most cases. Domain-specific pretraining in the biomedical or clinical domain turns out to be less helpful than pretraining on the target domain of mental health. Those results show that continued pretraining on the mental health domain improves prediction performance in downstream tasks of mental health classification.

## 4. Related Work

**Contextualized Text Embeddings** Contextualized embeddings have been intensively studied in NLP. Self-supervised large-scale pretraining facilitates the learning of semantic and contextual information and benefits various downstream applications such as text classification (Chaturvedi et al., 2016), sentiment analysis (Cambria et al., 2022), personality detection (Mehta et al., 2020), and social media analysis (Grassi et al., 2011). There are also many domain-specific variants of pretrained contextualized text embeddings. Embeddings in specific domains aim to encode domain-specific information to boost the performance of a specific domain. For example, BioBERT (Lee et al., 2020) pretrained the BERT model in the biomedical domain using research articles from PubMed, which was applied to many biomedical tasks such as biomedical relation extraction and named entity recognition. ClinicalBERT (Alsentzer et al., 2019) used clinical notes as the pretraining corpus to continue the pretraining of the BERT model. Those domain-specific variants also foster variable downstream applications by fine-tuning pretrained embeddings (Lin et al., 2019; Ji et al., 2020).

**NLP for Mental Healthcare** Mental healthcare research in social media is increasingly applying NLP techniques to capture users' behavioral tendencies. Various methods are implemented for labeling, i.e., identifying emotions, mood, and profiles that might indicate mental health problems (Calvo et al., 2017). One of the most representative tasks is mental health detection that categorizes given social posts into different classes of mental disorders such as depression (Tadesse et al., 2019). Mental state understanding requires effective feature representation learning and complex emotive processes. Resnik et al. (Resnik et al., 2013) applied topic modeling, an unsupervised approach that reduces the input of textual data feature space to a fixed number of topics to feature engineering in depression detection.

Feature engineering-based machine learning method designs manual features and builds classifiers for mental health detection (Shatte et al., 2019; Abd Rahman et al., 2020). Various features such as sensor signals from personal devices (Mohr et al., 2017) and EEG signals (Gore and Rathi, 2019) have been applied. For detection from textual data in particular, text features include word counts, TF-IDF (Campillo-Ageitos et al., 2021), topic features (Shickel et al., 2020) and sentiment traits and lexicons (Yoo et al., 2019; Ansari et al., 2022). Lee et al. (Lee et al., 2021) used micromodels trained with BERT to create a different path towards domain adaptation for mental healthcare. Severe mental disorders but without intervention may lead to suicidal ideation (Windfuhr and Kapur, 2011).

Many machine learning-based methods have been applied for suicidal ideation detection (Ji et al., 2021). Recent work applies deep representation learning methods, which enable automatic feature learning to solve the early mental disorder identification task. Those methods typically build text embeddings and feed the embeddings into neural architectures such as convolutional neural networks (Rao et al., 2020), recurrent networks (Bouarara, 2021), self attention-based Transformers, hybrid architectures like CNN-LSTM (Kang et al., 2021) and more other deep learning architectures (Su et al., 2020). Recent works use pretrained language models and fine-tune the model for mental health tasks (Jiang et al., 2020; Martínez-Castaño et al., 2021; Bucur et al., 2021). However, there are no existing pretrained language models trained with mental health-related text to benefit the domain application directly.

## 5. Conclusion and Future Work

This paper trains and releases two masked language models, i.e., MentalBERT and MentalRoBERTa, on the domain data of mental health collected from the Reddit social platform. This paper is the first work that trains domain-specific language models for mental healthcare. Our pretrained models are publicly available and can be reused by the research community. Besides, we conduct a comprehensive evaluation on the performance for downstream mental health detection tasks, including depression, stress, and suicidal ideation detection. Our empirical results show that continued pre-training with mental health-related corpus can improve classification performance.

Our paper is a positive attempt to benefit the research community by releasing the pretrained models for other practical studies and with the hope to facilitate some possible real-world applications to relieve people's mental health issues. However, we only focus on the English language in this study since English corpora are relatively easy to obtain. In the future work, we plan to collect multilingual mental health-related posts, especially those less studied by the research community, and train a multilingual language model to benefit more people speaking languages other than English.

## 6. Social Impact

The paper trains and releases masked language models for mental health to facilitate the automatic detection of mental disorders in online social content for non-clinical use. The models may help social workers find potential individuals in need of early prevention. However, the model predictions are not psychiatric diagnoses. We recommend anyone who suffers from mental health issues to call the local mental health helpline and seek professional help if possible.

Data privacy is an important issue, and we try to minimize the privacy impact when using social posts for model training. During the data collection process, we only use anonymous posts that are manifestly available to the public. We do not collect user profiles even though they are also manifestly public online. We have not attempted to identify the anonymous users or interact with any anonymous users. The collected data are stored securely with password protection even though they are collected from the open web. There might also be some bias, fairness, uncertainty, and interpretability issues during the data collection and model training. Evaluation of those issues is essential in future research.

## 8. Bibliographical References

Abd Rahman, R., Omar, K., Noah, S. A. M., Danuri, M. S. N. M., and Al-Garadi, M. A. (2020). Application of machine learning methods in mental health detection: a systematic review. *IEEE Access*, 8:183952–183964.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Ansari, L., Ji, S., Chen, Q., and Cambria, E. (2022). Ensemble hybrid learning methods for automated depression detection. *IEEE Transactions on Computational Social Systems*.

Bouarara, H. A. (2021). Recurrent Neural Network (RNN) to Analyse Mental Behaviour in Social Media. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, 13(3):1–11.

Bucur, A.-M., Cosma, A., and Dinu, L. P. (2021). Early risk detection of pathological gambling, self-harm and depression using bert. *arXiv preprint arXiv:2106.16175*.

Calvo, R. A., Milne, D. N., Hussain, M. S., and Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Cambria, E., Hussain, A., Durrani, T., Havasi, C., Eckl, C., and Munro, J. (2010). Sentic computing for patient centered applications. In *IEEE ICSP*, pages 1279–1282.

Cambria, E., Benson, T., Eckl, C., and Hussain, A. (2012). Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. *Expert Systems with Applications*, 39(12):10533–10543.

Cambria, E., Liu, Q., Decherchi, S., Xing, F., , and Kwok, K. (2022). SenticNet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *LREC*.

Campillo-Ageitos, E., Fabregat, H., Araujo, L., and Martinez-Romo, J. (2021). NLP-UNED at eRisk 2021: self-harm early risk detection with TF-IDF and linguistic features. *Working Notes of CLEF*, pages 21–24.

Chaturvedi, I., Ong, Y.-S., Tsang, I., Welsch, R., and Cambria, E. (2016). Learning word dependencies in text by means of a deep recurrent belief network. *Knowledge-Based Systems*, 108:144–154.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015). Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on CLPsych*, pages 31–39.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., and Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. In *CHI*, pages 2098–2110. ACM.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Gore, E. and Rathi, S. (2019). Surveying machine learning algorithms on eeg signals data for mental health assessment. In *2019 IEEE Pune Section International Conference (PuneCon)*, pages 1–6. IEEE.

Grassi, M., Cambria, E., Hussain, A., and Piazza, F. (2011). Sentic web: A new paradigm for managing social media affective information. *Cognitive Computation*, 3(3):480–489.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*.

Huang, K., Altosaar, J., and Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Ji, S., Yu, C. P., Fung, S.-f., Pan, S., and Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018.

Ji, Z., Wei, Q., and Xu, H. (2020). BERT-based ranking for biomedical entity normalization.

*AMIA Summits on Translational Science Proceedings*, 2020:269.

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., and Huang, Z. (2021). Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8:214–226.

Ji, S., Li, X., Huang, Z., and Cambria, E. (2022). Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*.

Jiang, Z. P., Levitan, S. I., Zomick, J., and Hirschberg, J. (2020). Detection of mental health from reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156.

Kang, M., Shin, S., Jung, J., and Kim, Y. T. (2021). Classification of mental stress using cnn-lstm algorithms with electrocardiogram signals. *Journal of Healthcare Engineering*, 2021.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lee, A., Kummerfeld, J. K., An, L. C., and Mihalcea, R. (2021). Micromodels for efficient, explainable, and reusable systems: A case study on mental health. *arXiv preprint arXiv:2109.13770*.

Lin, C., Miller, T., Dligach, D., Bethard, S., and Savova, G. (2019). A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Losada, D. E. and Crestani, F. (2016). A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.

Martínez-Castaño, R., Htait, A., Azzopardi, L., and Moshfeghi, Y. (2021). BERT-Based Transformers for Early Detection of Mental Health Illnesses. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 189–200. Springer.

Mauriello, M. L., Lincoln, T., Hon, G., Simon, D., Jurafsky, D., and Paredes, P. (2021). SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems. In *Ex-*

tended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Mehta, Y., Majumder, N., Gelbukh, A., and Cambria, E. (2020). Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53:2313–2339.

Mohr, D. C., Zhang, M., and Schueller, S. M. (2017). Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47.

Pirina, I. and Çöltekin, Ç. (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.

Rao, G., Zhang, Y., Zhang, L., Cong, Q., and Feng, Z. (2020). MGL-CNN: A hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, 8:32395–32403.

Resnik, P., Garron, A., and Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1348–1353.

Shatte, A. B., Hutchinson, D. M., and Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine*, 49(9):1426–1448.

Shickel, B., Siegel, S., Heesacker, M., Benton, S., and Rashidi, P. (2020). Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280. IEEE.

Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., and Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on CLPsych*, pages 25–36.

Su, C., Xu, Z., Pathak, J., and Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10(1):1–26.

Tadesse, M. M., Lin, H., Xu, B., and Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Turcan, E. and McKeown, K. (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.

Windfuhr, K. and Kapur, N. (2011). Suicide and mental illness: a clinical review of 15 years findings from the uk national confidential inquiry into suicide. *British Medical Bulletin*, 100(1):101–121.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., 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.

Yoo, M., Lee, S., and Ha, T. (2019). Semantic network analysis for understanding user experiences of bipolar and depressive disorders on reddit. *Information Processing & Management*, 56(4):1565–1575.