MentalHelp: A Multi-Task Dataset for Mental Health in Social Media

Md Nishat Raihan[1], Sadiya Sayara Chowdhury Puspo[1*], Shafkat Farabi[1*]
Ana-Maria Bucur[2,3], Tharindu Ranasinghe[4], Marcos Zampieri[1]

[1]George Mason University, Fairfax, VA, USA
[2]Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania
[3]PRHLT Research Center, Universitat Politècnica de València, Spain
[4]Aston University, Birmingham, UK
mraihan2@gmu.com

Abstract

Early detection of mental health disorders is an essential step in treating and preventing mental health conditions. Computational approaches have been extensively applied to social media to identify mental health conditions such as depression, PTSD, schizophrenia, and eating disorders. The interest in this topic has motivated the creation of various datasets. However, annotating such datasets is expensive and time-consuming, limiting their size and scope. To overcome this limitation, we present MentalHelp, a large-scale semi-supervised mental health dataset containing 14 million instances. The corpus was collected from Reddit and labeled in a semi-supervised way using an ensemble of three models: flan-T5, Disor-BERT, and Mental-BERT.

Keywords: Mental Health, Transformers, Social Media

1. Introduction

Mental Health Disorders are defined by the American Psychiatric Association as syndromes characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning. These disorders are usually associated with significant distress or disability in social, occupational, or other important activities. When severe, it can even lead to intense emotional distress, which may contribute to suicidal thoughts. According to the World Health Organization, suicide is the second leading cause of death among individuals aged 15 to 29 years.

Building automated systems for different mental disorders detection from text data has been substantially explored by researchers (Chancellor and De Choudhury, 2020). Early works (Bhat and Goldman-Mellor, 2017) depend on medical records. Since social media users express their thoughts and feelings, often discussing their mental health on different platforms, efforts have been made to train language models on data collected from social media (Malmasi et al., 2016; Aladağ et al., 2018; Coppersmith et al., 2015; Losada and Crestani, 2016; Wu et al., 2020) to analyze mental health. However, these datasets have to be annotated by mental health professionals, which ultimately limits their size. For example, Aladağ et al. (2018) collected over 500,000 Reddit posts, but only annotated 785 of them, making it difficult to train robust models using this data.

To facilitate training systems capable of modeling different mental disorders, we build *MentalHelp*, a large social media dataset containing more than 14 million instances. A semi-supervised approach is taken while labeling the dataset, similar to the work of Rosenthal et al. (2021), where the authors generate a large dataset of offensive tweets. We fine-tune multiple language models on small-sized publicly available annotated mental health datasets to create an ensemble of the best-performing ones and use them to label the entire MentalHelp dataset. In addition to predicted labels, we include the average of predicted probabilities for every instance. This grants a unique ability to filter the dataset based on the confidence level of the prediction, trading off between the size of training data and the accuracy of the labels.

The main contributions of this paper are:

1. The MentalHelp Dataset[1] - A large-scale semi-supervised dataset for mental health. The dataset includes confidence scores from multiple models, enabling researchers to filter subsets of data with high-quality labels.

2. A comprehensive evaluation of ten pre-trained and/or task fine-tuned models on eleven mental health benchmark datasets.

[*] Equal Contribution
WARNING: This paper contains examples that might depict symptoms of mental disorders.

[1] https://github.com/mraihan-gmu/MentalHelp

LREC-COLING 2024, pages 11196–11203
20-25 May, 2024.

Dataset	Reference	Category	Source	Instances
depsev	(Naseem et al., 2022)	Depression	Reddit	3,553
datd	(Owen et al., 2020)	Depression, Anxiety	Twitter	5,550
signdep	(Kayalvizhi and Thenmozhi, 2022)	Depression	Reddit	16,632
sdcnl	(Haque et al., 2021)	Suicide	Reddit	1,895
sid	(Mirza Ibtihaj et al., 2020)	Suicide	Twitter	1,385
dreaddit	(Turcan and McKeown, 2019)	Stress	Reddit	3,553
cds	(Hung Chia Yu, 2022)	Depression, PTSD, Anxiety	Reddit	18,006
swmh	(Ji et al., 2022a)	Depression, Anxiety, Bipolar, Suicide	Reddit	54,412
dr	(Pirina and Çöltekin, 2018)	Depression	Reddit, Blogs	5,984
sad	(Mauriello et al., 2021)	Stress	SMS	6,849
depTweet	(Yeow Zi Qin et al., 2022)	Depression	Twitter	46,020
merged	–	–	–	163,839

Table 1: Benchmark datasets used in MentalHelp.

2. Related Work

Transformer-based models have shown effectiveness in analyzing social media data to identify indications of mental disorders. For instance, they are utilized in the CLPsych Shared Task (Zirikly et al., 2019, 2022) for suicide risk prediction (Malmasi et al., 2016; Matero et al., 2019; Culnan et al., 2022), and in the eRisk Lab (Parapar et al., 2023) to address tasks such as depression, self-harm, and eating disorders detection (Jiang et al., 2020; Bucur et al., 2021; Martínez-Castaño et al., 2021).

With the success of pre-trained transformer-based models, several specialized models for clinical or mental health tasks have been released. For instance, Bio-BERT (Lee et al., 2020), Clinical-BERT (Yan and Pei, 2022), and Multitask-Clinical BERT (Mulyar et al., 2021) are pre-trained on biomedical corpora or clinical notes. Whereas Mental-BERT (Ji et al., 2022b), Disor-BERT (Aragon et al., 2023) and Suicidal-BERT [2] are pre-trained on the mental health domain on social media data.

While several datasets for mental disorders detection exist (Ji et al., 2018; Coppersmith et al., 2015; Losada and Crestani, 2016), most suffer from a small number of annotations due to the labeling being done by clinicians and other domain experts (Shing et al., 2018; Aladağ et al., 2018). Moreover, many datasets are not publicly available due to confidential information and the sensitive data they contain (Bhat and Goldman-Mellor, 2017; Coppersmith et al., 2015; Shing et al., 2018). Therefore, large and publicly available datasets in this domain are scarce. MentalHelp aims to fill this gap and facilitate the training and fine-tuning of models for mental disorders detection.

[2] https://huggingface.co/gooohjy/suicidal-bert

3. Benchmark Datasets

Several datasets have been released over the years for mental disorder detection and similar tasks. They often contain data related to depression, anxiety, bipolar disorder, and suicidal ideation. We gather eleven datasets widely used for such tasks (Table 1) and experiment with the merged dataset by combining all of them.

Most datasets contain posts from different subreddits since people tend to share anonymous thoughts related to specific topics via Reddit. However, some datasets contain texts from Twitter, such as the *sid* dataset[3], which includes 1,385 suicidal ideation tweets. The authors gathered the posts containing the word 'suicide' and manually labeled them into three classes: not suicidal, indicative of suicidal ideation, and indicative of a potential suicide attempt. Another dataset, *datd* (Owen et al., 2020), is a collection of 5,550 tweets manually annotated for depression and anxiety. Lastly, *depTweet*[4] is a larger dataset for depression detection that contains 46,020 posts from Twitter with binary labels (depressed, non-depressed).

Reddit datasets include *dreaddit* (Turcan and McKeown, 2019), a dataset for stress analysis, consisting of 3,553 Reddit posts with binary labels (stress, not stress) collected from subreddits related to anxiety, PTSD, and others. The *sdcnl* (Haque et al., 2021) dataset for suicide ideation contains 1,895 texts annotated in a semi-supervised manner. The dataset is created to distinguish between depression and suicide ideation and has binary labels. Another dataset for binary classification of depression, *dr* (Pirina and Çöltekin, 2018), contains 5,984 instances collected from Reddit and blogs.

For detecting the severity of depression, we use *depsev* (Naseem et al., 2022), comprised of 3,553

[3] https://github.com/M-Ibtihaj/Suicidal-ideation-detection

[4] https://huggingface.co/datasets/ziq/depression_tweet

Model Name	depsev	datd	signdep	sdcnl	sid	dread	cds	swmh	dr	sad	depTweet	avg
flan-T5	0.80	0.91	**0.90**	0.61	**0.86**	0.74	**0.69**	0.80	0.90	0.94	**0.94**	0.83$_{\pm 0.10}$
Disor-BERT	0.81	0.94	0.83	0.69	0.80	**0.86**	0.62	**0.81**	0.94	**0.95**	0.93	0.83$_{\pm 0.10}$
Mental-BERT	**0.82**	0.94	0.83	**0.73**	0.84	0.80	0.65	0.78	0.93	**0.95**	0.86	0.83$_{\pm 0.09}$
RoBERTa	0.81	0.94	0.83	0.71	0.80	0.81	0.61	0.78	0.**96**	**0.95**	0.83	0.82$_{\pm 0.10}$
BERT	0.81	0.94	0.83	0.68	0.81	0.78	0.62	0.78	0.92	0.94	0.82	0.81$_{\pm 0.10}$
Suicidal-BERT	0.81	0.93	0.83	0.69	0.84	0.72	0.61	0.77	0.92	0.95	0.92	0.82$_{\pm 0.11}$
Clinical-BERT	0.77	**0.95**	0.80	0.61	0.76	0.72	0.60	0.77	0.90	0.92	0.78	0.78$_{\pm 0.11}$
Bio-Clinical-BERT	0.79	0.94	0.82	0.62	0.75	0.76	0.60	0.76	0.90	0.94	0.78	0.79$_{\pm 0.11}$
Bio-BERT	0.81	0.93	0.82	0.64	0.79	0.75	0.60	0.77	0.92	0.94	0.81	0.80$_{\pm 0.10}$
GPT3.5-turbo (Few-Shot)	0.69	0.78	0.84	0.52	0.68	0.79	0.54	0.78	0.89	0.92	0.89	0.76$_{\pm 0.13}$
GPT3.5-turbo (Zero-Shot)	0.63	0.74	0.82	0.52	0.61	0.71	0.52	0.79	0.87	0.79	0.83	0.71$_{\pm 0.12}$

Table 2: Weighted F_1-score comparison of all the models on the benchmark datasets. The highest score(s) for each dataset are marked in bold.

data instances, and *signdep* (Kayalvizhi and Thenmozhi, 2022), with 16,632. The texts from *depsev* are divided into four severities: minimal, mild, moderate, and severe depression. For *signdep*, the Reddit posts are split into three classes: non-depressed, moderate, and severe.

Regarding datasets with labels for multiple mental disorders, we consider *cds*[5], which contains not only depression data, but also data collected from subreddit threads that are related to other mental disorders like anxiety and PTSD. Binary labeling for depression vs other disorders has been performed, which makes this dataset very challenging to work with. The largest labeled dataset from Reddit is the *swmh* dataset, gathered by Ji et al. (2022a), which contains 54,412 data instances collected from several subreddits related to suicide, depression, anxiety, and bipolar disorder. Additionally, we use another dataset *sad* by Mauriello et al. (2021), that contains SMS-like data for binary stress detection.

4. Performance Evaluation on the Benchmark Datasets

Models Transformer models are widely used for downstream text classification tasks in NLP, and several models have been specifically pre-trained for depression detection or similar tasks. We initially experiment with BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), since these models are widely used for most text classification tasks. We also experiment with models that are pre-trained on specific domains like Bio-BERT (Lee et al., 2020), which is pre-trained on a bio-medical corpus, and Clinical-BERT (Alsentzer et al., 2019), which is pre-trained on clinical notes. Bio-Clinical-BERT is later introduced by the authors of Clinical-BERT, where they combine both models and achieve better results in some scenarios.

We also use Mental-BERT (Ji et al., 2022b), which is pre-trained on a large Reddit corpus related to mental issues, and a double domain adaptation model called Disor-BERT (Aragon et al., 2023) that tends to perform very well for several mental disorder tasks. Disor-BERT is first adapted to the social media domain and then further adapted to the mental health domain, by two pre-training phases. We experiment with Suicidal-BERT, which was fine-tuned on the SuicideWatch[6] subreddit. Finally, we include flan-T5 (Raffel et al., 2020) in our experiments and fine-tune it for our task. The rationale for this choice is rooted in the model's text-to-text transfer capabilities.

Among LLMs, we use the GPT3.5-turbo (Brown et al., 2020) model. We do both zero-shot and few-shot prompting (5-shot).

Experimental Setup We use the Google Colab Pro+ platform for all the experiments. One Nvidia A100 GPU is used with 40 GB GPU RAM, 82 GB System RAM, and 140 GB storage available. All the models are evaluated using the same hyperparameters. The models were trained for three epochs with a batch size of 16 and a learning rate of 1e-5 with 0.01 weight decay. Most datasets mentioned in Table 1 have separate validation and test sets. The rest were split using a 60:20:20 ratio for training, validation, and testing. The models' performances are benchmarked based on their weighted F_1-scores.

Results

As shown in Table 2, flan-T5 achieves the highest weighted F_1 for most (four) datasets, including the *cds* dataset, which is quite challenging for all the models. Among the task fine-tuned models, Disor-BERT and Mental-BERT both come in second to flan-T5, being tied with three best scores. In contrast, other task fine-tuned models like Clinical-BERT, Bio-BERT and Bio-Clinical BERT perform less impressively, having only one best score among the three of them. While BERT

[5] https://huggingface.co/datasets/hungchiayu/cds-dataset2-depression

[6] https://www.reddit.com/r/SuicideWatch/

Instance	flan-T5	Disor-BERT	Mental-BERT	Average	Label
Anyone want to vent? Im a good listener... Just want to have someone to speak to online, nobody cares enough irl	0.010	0.001	0.001	$0.004_{\pm 0.005}$	NO
I think I died a long time ago. So what exactly am I doing here? It isn't living. I've been called lifeless and told I am wasting away, told I am wasteful, that I am more or less a zombie.	0.998	0.999	0.945	$0.981_{\pm 0.027}$	YES
When you cant just be yourself and youre so plagued by your damn mental illness.	0.999	0.999	0.995	$0.998_{\pm 0.002}$	YES
Didn think I would have live this long to see 2020.. Dont even know if this is considered an accomplishment.	0.044	0.001	0.995	$0.347_{\pm 0.561}$	NO
I want to disappear and stop being the burden I am to people. (I am not in danger)	0.997	0.003	0.882	$0.627_{\pm 0.543}$	YES

Table 3: Selected examples from the MentalHelp dataset. We present the confidence scores from flan-T5, Disor-BERT, and Mental-BERT. The confidence scores represent how confident a model is that a particular text depicts symptoms of mental disorders.

Model Name	Weighted F_1
flan-T5	**0.95**
Disor-BERT	0.92
Mental-BERT	0.91
RoBERTa	0.89
Suicidal-BERT	0.88
Clinical-BERT	0.85
GPT3.5-turbo (Few-Shot)	0.85
BERT	0.84
Bio-Clinical-BERT	0.84
Bio-BERT	0.83
GPT3.5-turbo (Zero-Shot)	0.82

Table 4: Weighted F_1-score comparison on the merged dataset.

Average confidence	# Instances	YES (%)	NO (%)
>0.9 or <0.1	7,522,421	57.52	42.47
>0.8 or <0.2	8,406,396	56.64	43.35
>0.7 or <0.3	9,514,001	55.04	44.95
>0.6 or <0.4	13,095,547	50.02	49.97
>0.5 or <0.5	14,097,946	49.97	50.02

Table 5: Class distribution in MentalHelp across different average confidence intervals.

performs better than some of the task fine-tuned models, it has no best F_1 score on any dataset. However, RoBERTa does well with two best scores. On the other hand, GPT3.5-turbo does not do well compared to any of the models when we take a Zero-Shot approach. Taking the Few-Shot approach improves the result, but still falls behind the other models.

These varied performances show that while some models are universally adept, others might have a predilection for specific datasets or tasks. This diversity highlights the importance of model choice, contingent on the clinical data's nature and the task's intricacies.

5. The MentalHelp Dataset

We present the steps taken to create MentalHelp.

Data collection from Reddit We collect data for our corpus from the social media platform Reddit. We carefully compile a list of subreddits (see Appendix A) manually, containing posts related to depression, suicide, and other mental health issues. We use PRAW, a Python Reddit API Wrapper[7] to extract data from relevant subreddits. We only collect the posts and not any user information. In total, we gather a total of 14,097,946 data instances. Selected examples are presented in Table 3.

Model training We first take all the datasets from Table 1 and merge them. For all the instances in the merged dataset that are labeled with any kind of mental disorders, we set their labels to *YES*, and for the rest, *NO*. We split the merged dataset in a 60:20:20 fashion to train, evaluate, and test our models and pick the best three based on the weighted F_1 score for the ensemble.

Generating the Labels with Democratic Co-training Our three best-performing models on the merged datasets are flan-T5, Disor-BERT, and Mental-BERT (see Table 4). We use each model to label every single instance in the whole corpus. When text data is processed through transformer models, it generates logits - raw and unprocessed output values for each class in the classification task. These logits are converted into probabilities using the softmax function. After obtaining confidence values from all three models,

[7] https://praw.readthedocs.io/en/stable/

we calculate their average and standard deviation. Data instances with an average confidence > 0.9 or < 0.1 are considered high quality. Finally, the labels are generated from the average confidence. If average confidence ≥ 0.5, the label is *YES*, else *NO*. A few samples from the MentalHelp dataset with their confidence scores are presented in Table 3. Table 5 presents the data distribution for different average confidence intervals.

6. Conclusion and Future Work

Detecting depression from text data in social media is an important task to improve individuals' well-being. While there have been several mental health datasets compiled thus far, a public large-scale dataset was not available. This motivated us to introduce MentalHelp, the large-scale dataset presented in this work. Large-scale datasets created using approaches similar to MentalHelp have resulted in training language models that provide state-of-the-art results in various datasets in related domains (Sarkar et al., 2021; Ranasinghe and Zampieri, 2023). Inspired by them, we make the dataset freely available to the community so that it can be used to train, test, and fine-tune models for multiple tasks. The average confidence of multiple models can be used to filter out the best-quality data from the dataset.

Future work includes fine-tuning models such as BERT and RoBERTa on MentalHelp to achieve competitive performance compared to the state-of-the-art models. We would also like to evaluate the performance of recently-introduced LLMs (e.g., GPT 4.0, Llama) in this task. Finally, we would also like to create a multilingual dataset on a similar scale to MentalHelp that can be used to train models to detect mental health issues in a multitude of languages.

Ethics Statement

As part of this research, we have not collected or processed writers'/users' information, nor have we carried out any form of user profiling to protect users' privacy. Researchers interested in downloading the dataset should adhere to the Reddit guidelines.

With MentalHelp, we encourage research on modeling and detecting mental disorders, which can ultimately improve the mental well-being of individuals. However, it is important to exercise caution when using this or any similar dataset in real-world clinical applications. We strongly recommend the careful supervision of mental health professionals.

Acknowledgments

We would like to thank the anonymous LREC-COLING reviewers for the constructive feedback provided. We further thank the creators of all datasets used in our experiments for making datasets available for this research.

Ana-Maria Bucur is partially supported by the POCIDIF project in Action 1.2. "Romanian Hub for Artificial Intelligence".

References

Ahmet Emre Aladağ, Serra Muderrisoglu, Naz Berfu Akbas, Oguzhan Zahmacioglu, and Haluk O Bingol. 2018. Detecting suicidal ideation on forums: Proof-of-concept study. *JMIR*.

Alsentzer, Murphy Emily, John, et al. 2019. Publicly available clinical BERT embeddings. In *Proceedings of CLPsych Workshop*.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. American Psychiatric Publishing.

Jayaraman Ananthakrishnan, Gayathri et al. 2022. Suicidal intention detection in tweets using bert-based transformers. In *Proceedings of ICCCIS*.

Mario Aragon, Adrián Pastor López Monroy, Gonzalez, et al. 2023. Disorbert: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of ACL*.

Harish S Bhat and Sidra J Goldman-Mellor. 2017. Predicting adolescent suicide attempts with neural networks. *arXiv preprint arXiv:1711.10057*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*.

Ana-Maria Bucur, Adrian Cosma, and Liviu P Dinu. 2021. Early risk detection of pathological gambling, self-harm and depression using bert. *CEUR workshop proceedings*.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*.

Glen Coppersmith, Dredze, et al. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of CLPsych Workshop*.

John Culnan, DY Romero Diaz, and Steven Bethard. 2022. Exploring transformers and time lag features for predicting changes in mood over time. Proceedings of CLPsych Workshop.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Joseph C Franklin, Jessica D Ribeiro, Fox, et al. 2017. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological bulletin*.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. In *Proceedings of ICANN*. Springer.

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

Shaoxiong Ji, Shirui Pan, Xue Li, et al. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*.

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, Fu, et al. 2022b. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of LREC*.

Zheng Ping Jiang, Sarah Ita Levitan, Zomick, et al. 2020. Detection of mental health from reddit via deep contextualized representations. In *Proceedings of LOUHI Workshop*.

Sampath Kayalvizhi and Durairaj Thenmozhi. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. In *Computational Intelligence in Data Science*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Proceedings of CLEF*. Springer.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. In *Proceedings of CLPsych Workshop*.

Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2021. Bert-based transformers for early detection of mental health illnesses. In *Proceedings of CLEF*. Springer.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, et al. 2019. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of CLPsych Workshop*.

Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, et al. 2021. Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *Extended abstracts of CHI*.

Andriy Mulyar, Ozlem Uzuner, and Bridget McInnes. 2021. Mt-clinical bert: scaling clinical information extraction with multitask learning. *Journal of the American Medical Informatics Association*.

Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2021. Classification of mental illnesses on social media using RoBERTa. In *Proceedings of LOUHI Workshop*.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of WWW*.

David Owen, Jose Camacho-Collados, and Luis Espinosa Anke. 2020. Towards preemptive detection of depression and anxiety in twitter. In *Proceedings of SMM4H Workshop*.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *Proceedings of CLEF*, pages 294–315. Springer.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Proceedings of SMM4H Workshop*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMIR*.

Tharindu Ranasinghe and Marcos Zampieri. 2023. A text-to-text model for multilingual offensive language identification. In *Findings of ACL: IJCNLP-AACL 2023*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of ACL: ACL-IJCNLP 2021*.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of ACL: EMNLP 2021*.

Saskia Senn, ML Tlachac, Ricardo Flores, and Elke Rundensteiner. 2022. Ensembles of bert for depression classification. In *Proceedings of IEEE EMBC*.

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 CLPsych Workshop*.

K Soumya and Vijay Kumar Garg. 2022. Named entity emotion intensity tagging for suicidal ideation detection from social media texts during mt. *International Journal of Intelligent Engineering & Systems*.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of LOUHI Workshop*.

Ana-Sabina Uban and Paolo Rosso. 2020. Deep learning architectures and strategies for early detection of self-harm and depression level prediction. In *CEUR workshop proceedings*.

World Health Organization. 2014. *Preventing suicide: A global imperative*. World Health Organization.

Jheng-Long Wu, Yuanye He, Liang-Chih Yu, and K Robert Lai. 2020. Identifying emotion labels from psychiatric social texts using a bi-directional lstm-cnn model. *IEEE Access*, 8:66638–66646.

Bin Yan and Mingtao Pei. 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of AAAI*.

Ayah Zirikly, Dana Atzil-Slonim, Maria Liakata, Steven Bedrick, et al. 2022. Proceedings of the eighth workshop on computational linguistics and clinical psychology. In *Proceedings of CLPsych Workshop*.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of CLPsych Workshop*.

A. Appendix - Subreddits used for Corpus Creation

/r/SuicideWatch
/r/SuicideNotes_
/r/realsuicidenotes
/r/depression
/r/depressed
/r/depression_help
/r/depressionregimens
/r/raisedbynarcissists
/r/TimeToGo
/r/CPTSD
/r/WeListenToYou
/r/trolldepression
/r/BPD_friends
/r/BreakUp
/r/BreakUps
/r/DepressionAndPTSD
/r/AvPD
/r/2meirl42meirl4meirl
/r/death
/r/mentalhealth
/r/MomForAMinute
/r/PepTalksWithPops
/r/breakupmusic
/r/breakup4k
/r/breakupbuddy
/r/offmychest
/r/AnxietyDepression
/r/depressionpartners
/r/anxiety
/r/adultdepression
/r/depressionolympics
/r/bipolar
/r/depression_de
/r/Tackle_depression
/r/mentalhealth
/r/DepressionNests
/r/DepressionMusic
/r/DepressionBuddies
/r/DepressionArt
/r/ADHD
/r/ResearchingDepression
/r/DepressionIsNotAJoke
/r/Relationships
/r/DepressionNotCensored
/r/Burnout_Depression
/r/psychology
/r/Postpartum_Depression
/r/depressionmemes
/r/depressionmeals
/r/stopdrinking
/r/getting_over_it
/r/BipolarReddit
/r/bipolar2
/r/DepressionDabs
/r/relationship_advice
/r/PsychoticDepression
/r/DepressionResearch
/r/MadeMeSmile
/r/todayilearned
/r/DepressionIndia
/r/DepressionPoems
/r/depression_awareness
/r/leaves
/r/TrueOffMyChest
/r/SuicideBereavement
/r/TrueSuicideWatch
/r/suicide_watch
/r/suicidebywords
/r/ComedySuicide
/r/SuicideBySuicide
/r/SuicideAnonymous
/r/SuicideSurvivor
/r/SoftwareSuicide
/r/AvoidSuicide
/r/suicideprevention
/r/TwoSentenceSadness
/r/sad
/r/teenagers
/r/SadHorseShow
/r/sadcringe
/r/sadposting
/r/ptsd
/r/diagnosedPTSD
/r/MedicalPTSD
/r/Veterans
/r/OCD
/r/OCDRecovery
/r/overeating
/r/Broken
/r/heartbreak
/r/therapy
/r/TalkTherapy
/r/OccupationalTherapy
/r/therapyabuse
/r/DeathPositive
/r/DeathBattleMatchups
/r/tackle_depression
/r/IncelsWithoutHate
/r/reasonstolive
/r/depression_help
/r/WeListenToYou
/r/SuicideBereavement
/r/MomForAMinute
/r/PepTalksWithPops
/r/tackle_depression
/r/suicidology
/r/reasonstolive
/r/TalesFromTheMilitary
/r/TransCommunity
/r/needadvice
/r/advice
/r/fosterit
/r/homeless
/r/babyloss
/r/LGBTForeverAlone
/r/LGBTeens
/r/relationship_advice