Explaining Models of Mental Health via Clinically Grounded Auxiliary Tasks

Ayah Zirikly
Johns Hopkins University
`azirikly@jhu.edu`

Mark Dredze
Johns Hopkins University
`mdredze@cs.jhu.edu`

Abstract

Models of mental health based on natural language processing can uncover latent signals of mental health from language. Models that indicate whether an individual is depressed, or has other mental health conditions, can aid in diagnosis and treatment. A critical aspect of integration of these models into the clinical setting relies on explaining their behavior to domain experts. In the case of mental health diagnosis, clinicians already rely on an assessment framework to make these decisions; that framework can help a model generate meaningful explanations.

In this work we propose to use PHQ-9 categories as an auxiliary task to explaining a social media based model of depression. We develop a multi-task learning framework that predicts both depression and PHQ-9 categories as auxiliary tasks. We compare the quality of explanations generated based on the depression task only, versus those that use the predicted PHQ-9 categories. We find that by relying on clinically meaningful auxiliary tasks, we produce more meaningful explanations.

1 Introduction

Mental illness has a huge impact on the health and well-being of the United States and world populations. In the US, 25% of the population suffered at some point from mental illness [1]. The urgency to address the mental health crisis became even more critical with the COVID-19 pandemic and its negative impact on mental health, burdening kids and seniors especially (Loades et al., 2020). Depression is among the most prevalent mental disorders. In the United States alone, 21 million adults had at least one major depressive episode [2].

Computational linguistics and natural language processing (NLP) research on mental health has received increased attention in the last decade, with work on suicide risk assessment (Zirikly et al., 2019; Shing et al., 2018; De Choudhury et al., 2016; Coppersmith et al., 2018), anxiety prediction and classification (Osadchiy et al., 2020), and depression prediction and classification (Coley et al., 2021; De Choudhury et al., 2013), among many other tasks. Although clinical data was used for some models (Penfold et al., 2021), prior work also utilized other sources of data, such as social media to overcome challenges in data access and to better understand what influences mental health on a daily basis. The majority of the NLP research on depression classification is focused on improving performance to achieve state-of-the-art models. Such models typically act like a black box, and predictions are therefore not explainable. This results in poor integration of these models into clinical settings, given that clinicians need to understand why a patient is identified as depressed, so that they can make informed decisions in regards to diagnosis and evidence-based treatment (Zhou et al., 2015). Additionally, it has been shown that black-box models are not generalizable across different data genres or domains (Harrigian et al., 2020). This accentuates the need for explainable models, as they could help to troubleshoot and understand the transfer between datasets – e.g. within social media or from social media to electronic health records (EHR).

Recently, and with the proliferation of deep learning in particular, explainable AI (XAI) has attracted significant attention, with the field publishing multiple techniques that provide explanations for machine learning models. Techniques like LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) have been widely adopted and proven to work in different domains, mental health being one of them (Hu and Sokolova, 2021; Spruit et al., 2022; Uddin et al., 2022).

Clinicians rely on ongoing assessment of pa-

[1] https://www.nimh.nih.gov/health/statistics/mental-illness

[2] https://www.nimh.nih.gov/health/statistics/major-depression

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 30 - 39
July 15, 2022 ©2022 Association for Computational Linguistics

tient progress and well-being for therapeutic decisions. Many assessment instruments exist, including questionnaires such as the Patient Health Questionnaire (PHQ-9) for depression and the General Anxiety Disorder (GAD-7) screener for anxiety. PHQ-9 (Kroenke et al., 2001) is one of the most commonly used and validated depression assessment tools that mental health clinicians and primary care physicians use. The questionnaire addresses the presence and severity of nine symptoms or categories such as problems with sleep, eating, and self-harm to assess and monitor a patient's depression severity.

In this work, we leverage the availability of PHQ-9, a clinically accepted and interpretable tool to measure depression severity, and integrate its items into depression classification models as auxiliary classification tasks. We claim and prove that LIME explanations generated for models that use such clinically grounded auxiliary tasks are better and more informative than explanations on other *black-box* models that do not use these auxiliary models in the decision process.

We summarize our contributions as follows:

- We created a manually labeled dataset that highlights the most prominent terms in a tweet as the explanation for depression,
- designed a multi-task learning framework that uses PHQ-9 categories for depression classification, and
- showed that using auxiliary models (PHQ-9) improves the explainability of depression detection models, regardless of the complexity of the underlying model.

2 Related work

Depression classification has been an important area of focus in mental health NLP in social media data and electronic health records (EHR). To overcome the challenges of data access and to create community-based datasets, many initiatives started using Twitter and Reddit platforms to create depression annotated datasets. These datasets were collected using self-reported terms and regular expressions such as *I was diagnosed with depression* (Coppersmith et al., 2015), or in the case of Reddit, using mental health related subreddits (e.g. r/ADHD) as a proxy to retrieve relevant posts (Pirina and Çöltekin, 2018; Cohan et al., 2018; Yates et al., 2017). Many common techniques related to linguistic features are used to perform the classification task such as using LIWC in social media (Morales et al., 2017; Loveys et al., 2018) and EHR (Bittar et al., 2021). Researchers used a variety of machine learning techniques that range from conventional methods such as SVM (Tadesse et al., 2019; Yazdavar et al., 2017) and LR (Yazdavar et al., 2017; Karmen et al., 2015), to deep learning techniques such as feed-forward networks (Geraci et al., 2017), CNN and LSTM (Mumtaz and Qayyum, 2019; Kour and Gupta, 2022). Many recent work also explored the use of recent pre-trained language models to improve the depression classification task performance, such as BERT-CNN in (Rodrigues Makiuchi et al., 2019) and ALBERT (Owen et al., 2020). There has been a line of research that focused on predicting the symptoms (PHQ categories). (Delahunty et al., 2019) introduced a deep neural network model to predict PHQ-4 scores in Reddit depression dataset (Losada and Crestani, 2016) and DAIC-WOZ transcribed clinical interviews (Gratch et al., 2014). (Yadav et al., 2020) proposed identifying the presence of the depressive symptoms using the auxiliary task of figurative usage detection.

In the area of explainable AI (XAI), most the work that has been done focused on using explainable techniques to highlight the most important features in depression prediction. (Nemesure et al., 2021) used SHAP values (Lundberg and Lee, 2017) to highlight which features were most salient in the depression classification model. (Choi et al., 2020) used LIME to understand which features weighed the most in identifying college students at high risk of depressive disorder. In a recent work by (Nguyen et al., 2022), the authors showed the positive impact of using depression classifiers that are constrained by PHQ-9 symptoms, on their generalizability across different datasets.

3 Data

In this work, we focus on social media data because public access to clinical datasets is limited. The publicly available social media datasets that address depression classification only contain labels for depression (Coppersmith et al., 2015; Cohan et al., 2018), and it is challenging to find publicly available data that has annotations for both depression and PHQ-9 categories.

For our experiments, we use the **D**epression to (**2**) **S**ymptoms (D2S) dataset (Yadav et al., 2020). It is a collection of English only tweets that was crawled using depression-related terms that can be categorized into one of the PHQ-9 categories (symptoms): (S1) lack of interest, (S2) feeling down or depressed, (S3) trouble with sleeping, (S4) lack of energy, (S5) eating disorder, (S6) low self-esteem, (S7) concentration problems, (S8) hyper/lower activity, and (S9) self-harm.

The dataset contains the list of annotated tweet IDs, and a total of 3738 tweets labeled as depressive and 8417 as not depressive (control). The depressive tweets are further annotated with symptoms, where a label of 1 is assigned for $S9$ if the tweet has mentions of self-harm thoughts, 0 otherwise, and so forth for all 9 categories, where multiple categories can receive a 1 annotation. It is worth mentioning that the data, unlike PHQ-9 questionnaire, does not have scores for each category, but only a binary label. Additionally, the original dataset has annotations for sarcasm and metaphor labels for the depressive tweets, since Yadav et al. (2020) focused on the task of understanding how to classify PHQ-9 categories using the sarcasm and metaphor language labels. However, in our work we focus on the depressive and PHQ-9 symptoms annotations, with all annotations scoped at the tweet level, not at the user level as is the case in some other datasets.

We collected the tweets corresponding to the tweet IDs described in D2S using the Twitter API. Some tweets had become unavailable since the publication of D2S, resulting in a reduced dataset with 2132 depressive tweets and 5698 control tweets. Notwithstanding the change in dataset size, we adopt the train, dev, and test splits of D2S to maintain consistency. Table 1 shows the characteristics of the dataset splits, and the distribution of PHQ-9 annotations.

4 Depression classification models

Understanding the domain and the task should be the foundation in designing an NLP model, as opposed to simply applying NLP state-of-the-art models that are hard to interpret. This is especially true for clinical and mental health NLP, where a lack of explainability would result in poor integration in clinical settings. In our work, we aim to build models that mimic a clinical setting, where the clinician uses the scores from PHQ-9 questionnaires to screen if a patient is suffering from depression and to assess its severity.

In this section we discuss the approaches we used to build models that predict if a tweet is depressive or not. We propose three models; the first two, similarly to previous literature, focus on the depression classification task as a standalone problem, without considering how symptoms, in our case the PHQ-9 categories, can help interpret and influence the model's performance. The last model aims to study how predicting symptoms can help in classifying depressive tweets.

In the following subsections, we will describe our models and the balancing techniques we used to address the skewed distributions for the depressive and symptom labels.

4.1 Single task classification models

The task formulation for these models is as follows: given tweet t, classify if t is depressive (dep) or has any of the symptoms (PHQ-9 categories) enabled. For this task, we propose two simple models: logistic regression (LR) and multilayer perceptron (MLP). Both of these models take as input the pre-processed tweet. The preprocessing steps we have used include: lowercasing, tokenizing the tweet, normalizing the numbers (e.g. 123 -> 000), and removing tokens that occur less than 3 times in the training set. The preprocessed tweet text was then vectorized using a term frequency-inverse document frequency (TF-IDF) vectorizer with L2 regularization. We are aware of more sophisticated methods to build representations of the input text, such as applying and fine-tuning BERT contextual embeddings (Devlin et al., 2018; Brown et al., 2020), that could improve results. However, the focus of this paper is not to provide the best performance, but rather to show how using auxiliary models can help in providing better explanations for the depression classification models. Additionally, we believe simpler input forms can make the explainability process cleaner.

For each of the single task approaches, we build 10 different models that can address the classification tasks separately (dep + 9 symptoms).

Logistic Regression In this model we use logistic regression with a maximum of 50 iterations and L2 regularizer. For balancing the data, we apply a higher weight class for the *1:enabled* class for each of the depressive and symptom classes.

Table 2 shows the results of this model on the test data, where the symptoms models are trained

split	**control**	**dep**	S1	S2	S3	S4	S5	S6	S7	S8	S9
train	3989	1615	237	235	97	140	173	426	69	51	468
dev	570	140	16	19	5	5	26	53	6	4	28
test	1139	377	32	110	35	29	51	89	6	6	113
all	5698	2132	285	364	137	174	250	568	81	61	609

Table 1: Data statistics

	dep	S1	S2	S3	S4	S5	S6	S7	S8	S9
Precision	0.725	0.214	0.353	0.923	0.8	0.677	0.324	0	0	0.513
Recall	0.629	0.188	0.109	0.343	0.414	0.412	0.528	0	0	0.513
F1	0.673	0.2	0.167	0.5	0.546	0.512	0.402	0	0	0.513

Table 2: Logistic regression results for the single classification task

	dep	S1	S2	S3	S4	S5	S6	S7	S8	S9
Precision	0.249	0.093	0.081	0.308	0.186	0.16	0.14	0.008	0.007	0.25
Recall	1	0.625	0.518	0.571	0.552	0.726	0.652	0.167	0.167	0.558
F1	0.398	0.161	0.139	0.4	0.278	0.262	0.231	0.015	0.013	0.345

Table 3: MLP results for the single-task classification

	dep	S1	S2	S3	S4	S5	S6	S7	S8	S9
Precision	0.765	0.761	0.764	0.768	0.765	0.767	0.766	0.77	0.767	0.767
Recall	0.508	0.489	0.487	0.503	0.487	0.517	0.512	0.49	0.502	0.503
F1	0.611	0.595	0.595	0.608	0.595	0.618	0.614	0.595	0.607	0.608

Table 4: MTL results for the multitask classification

on the depressive only tweets and tested on all the test data (dep + control). We do not report accuracy given how skewed the dataset is.

Multilayer Perceptron Our multilayer perceptron model (MLP) is a three-layer fully connected feedforward network with a hidden layer of size 256. The best parameters obtained for this model on the dev data are: learning rate of 1e-3, a batch size of 32, dropout probability of 0.5, Adam optimizer (Kingma and Ba, 2014), and cross-entropy as the loss function. We minimize the impact of imbalanced data by balancing each batch separately for the 0/1 classes. Similarly to our LR model, the symptoms classifiers use only the depressive tweets for the training and development sets to minimize the imbalance, and because the non-depressive tweets are automatically given label 0 for each of the symptoms. However, for testing we use both depressive and control tweets to mimic the real-life scenario where we don't know the depression status of a patient. The results of our MLP model are depicted in table 3.

4.2 Multitask classification model

Our research question is based on studying the impact of using auxiliary models (symptoms) to generate better explanations for the depression classification model. Given that, we adopt a multitask learning (MTL) framework that classifies each tweet as depressive or not, in addition to each of the 9 PHQ-9 categories (symptoms), simultaneously. For comparability with our MLP model, we adopt the same neural network design choices. the MTL framework consists of multiple MLP networks, one for each of the tasks, with the same parameters in terms of dropout, learning rate, number of hidden layers, optimizer, and loss function. Table 4 shows the results of our MTL proposed model. Similarly to LR and MLP, the symptoms classification task uses only depressive tweets from the training and development sets.

5 Depression model explanations

It has been argued that depression classification models that use machine learning, and deep learning techniques in particular, have been hard to integrate into clinical settings due to the difficulty of interpreting and explaining their results (Sendak et al., 2019). In the literature, there are many initiatives to generate explanations for blackbox models such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), which are highly adopted and used. In our work, to test our hypothesis, we compare the explanations generated by LIME for each of the models listed in section 4 with our in-house gold annotated explanations dataset.

5.1 Explanations dataset

We randomly sampled 105 tweets from the test dataset that are depressive (*D2S-explain*), and manually annotated them. We had one annotator that is experienced in mental health research and its intersection with computational linguistics that read the tweets, and for each tweet identified the tokens that signal depression or that are most relevant to it. To evaluate the quality of the annotation, 25 randomly selected tweets from *D2S-explain* were checked by another annotator that is also an expert in mental health research with a degree in psychology. The first annotator is not a native English speaker, but has full professional proficiency in English, while the second annotator is a native English speaker. The second annotator had three options: accept, modify, or reject an explanation. This process was repeated until we reached 85% agreement for *accept* on the 25 tweets, after which the rest of *D2S-explain* was re-annotated by the first annotator.

Figure 1 shows an example tweet and its corresponding manual annotations [3]. Table 5 shows the details of the number of tweets that have any of the 9 symptoms enabled. Upon acceptance we plan to make the dataset publicly available under a DUA as discussed in 7.

```
I feel that existence is
pointless and everything is
hurting me to a point that I
can't sleep anymore.  My stomach
hurts every time I eat and I
feel that I need to throw up.

existence is pointless |
everything is hurting me
```

Figure 1: Example of manually labeling explanation terms in a tweet

5.2 Explanations evaluation

For each of the three models we developed, we employ LIME to generate explanations for the D2S-explain dataset. LIME is able to generate explanations by creating an interpretable model that is an approximation of the original model for each data point (tweet) from the dataset. The LIME explanations look like probability scores for all inputs (in our case, tokens) that indicate how much they are expected to have contributed to the output classification. By looking at the highest-probability tokens of a tweet, we can get a sense of what information the model has used to make its prediction for that tweet.

We identify the following three scenarios for generating and evaluating the explanations:

- (**D**) We generate explanations for each of the three models (LR, MLP, and MTL) for the **depressive** classification task. We rank the explanations (tokens) generated by LIME based on their top relevance probabilities and use the first ten tokens.

- (**S/S-comb**) In this scenario we generate the explanations for the **symptoms** prediction task for only the tweets that were predicted to have the corresponding symptom (S) enabled and that are correctly predicted by the model as depressive. Additionally, we combine all the explanations from the 9 symptoms models and rank the relevance/contributing probabilities of the tokens then pick the top ten tokens (S-comb).

- (**D+S**) In this scenario, we combine the explanations from the 9 symptoms models – same criteria as scenario S, with the explanations from scenario (D). Similarly to S-comb, we rank the relevance probabilities for all the explanations and pick the top ten.

The reason behind structuring the scenarios as proposed is to reflect the research question we formulated earlier and study the impact of using the explanations generated from the auxiliary tasks to help explain and interpret the depression models' outputs, as opposed to using the depression classification models alone.

For evaluation, we use the recall metric since we are mainly interested whether the models were able to generate explanation tokens that match the ones in the gold explanations. A prediction is considered

[3] All example tweets are paraphrased for privacy.

S1(1)	S2	S3	S4	S5	S6	S7	S8	S9
9	31	8	1	6	16	1	1	46

Table 5: Annotated test data sample stats

a true positive if the predicted explanation is fully or partially in the gold explanation. For instance, if the generated explanation is *lost hope* and the gold explanation is *lost hope in life*, the explanation is considered to be correct and the number of true positives increases by 1. However, this partial matching strategy only applies if the generated explanation contains more than only function words, stop words or pronouns; no credit is given for partial matches of that type. The reason behind the choice for a partial match evaluation is that it is sufficient for a clinician or mental health expert to see part of the term highlighted to understand why a model signaled depression.

D	Recall
LR	0.61
MLP	0.267
MTL	0.524

Table 6: Recall explanation results for scenario (D)

Tables 6, 7, and 8 show the recall performance for each of the scenarios listed above, which will be discussed in the next section.

6 Discussion

When we look at the results of the depression and symptoms classification task in tables 2, 3, and 4, we note that MLP yields the worst results across almost all the labels (dep and symptoms), whereas LR provides the best results for dep. However, its performance on the symptoms is poor, especially for concentration (S7) and activity (S8), where the number of positive instances is very limited. The MTL model, meanwhile, performs slightly worse than LR for the dep class, but is able to perform much better for all the symptoms and is not susceptible to the imbalanced nature of the data. For instance, the F1-scores for S7 and S8 in the MTL setting improve drastically. This observation supports the claim that using the symptoms with the depression labels can provide more reliable performance where we can think of the symptoms predictions as the first layer of explanations we can provide to the clinicians.

```
At certain times and without any
trigger, I think I am probably
not even mentally ill, but
rather just an attention seeking
sh**
mentally ill | attention seeking
sh** [gold annotation]
even, mentally, ..., seeking,
ill, attention [LR]
even, time, trigger, ...,
mentally, ill [MLP]
mentally, seeking, sh**,
trigger, mentally,..., ill,
attention [MTL]
```

Figure 2: Example of the explanations from LR and MTL

After applying LIME on each of the models, we note that the recall of the LR explanations is the highest among the three models at 0.61 (table 6). This is expected, given that LR performance on the dep class is the highest. When we qualitatively examined the explanations' output and compare the results between LR, MLP and MTL, we note that both LR and MTL explanations contain more relevant terms. Additionally, the fact that MLP performs much worse on correctly predicting the dep label affects the performance of the explanations recall. Figure 2 shows a paraphrased tweet example with the explanations generated from the three models.

The main research question we aimed to address is: does augmenting the explanations for depression models with those for PHQ-9 models provide more meaningful explanation to clinicians than those for depression models alone? To answer this, we need to check the recall performance for each of the three models for the D+S scenario in table 8. For the LR model, although the performance was poor for symptoms, the explanation recall increased 1.9% when augmenting with the symptoms' explanations. For MLP and MTL, the increase in recall performance is smaller with almost 1%. In MTL, we reason the smaller increase is caused by the fact that the MTL model already utilizes the symptoms to optimize the depression classification performance in its network design, thus MTL explanations produced for (D) reflects, to some extent,

S/S-avg	S1	S2	S3	S4	S5	S6	S7	S8	S9	S-comb
LR	0.057	0	0.01	0	0.019	0.076	0	0	0.152	0.152
MLP	0.267	0.267	0.267	0.267	0.267	0.267	0.267	0.267	0.267	0.267
MTL	0.533	0.533	0.495	0.524	0.533	0.524	0.514	0.533	0.533	0.533

Table 7: Recall explanation results for scenario (S/S-comb)

D+S	dep
LR	0.629
MLP	0.276
MTL	0.533

Table 8: Recall explanation results for scenario (D+S)

augmenting with the symptoms. We note that in the case of LR and MPL, the symptom models are independent from the depression model, so LIME explanations generated for those symptoms cannot technically be interpreted as having explained the depression model outputs. However, our results show that augmenting with explanations from these disjoint models improves recall of input tokens that would aid a clinician in evaluating tweets that get flagged as depressive, by focusing their attention on clinically relevant information.

To further support our claim and to make sure that ensembling multiple models will not also produce better results than D, we implement bagging techniques. We create 9 random samples to mimic the 9 symptoms sample size. For instance, sample 1 will randomly select 237 depressive and 1378 control tweets to mimic the size of the S1 dataset; the same technique would apply for each of the samples. We report the F1-score, in table 9, for the worst and best model based on which sample it has used. The results show a variance in performance which made us further investigate the recall performance of the explanations if we combined the explanations from (D) with the random 9 models explanations. The results are depicted in table 10 and show that (9samples+D) generates worse results than (D) and (D+S).

Limitations We understand that our work and results are limited in a number of ways. First, the D2S dataset is a Twitter dataset, which by itself can raise some questions about its reliability, however, we justify our decision due to the lack of clinical data access and this can be a proxy to prove our hypothesis using the symptoms models. We are also aware that the dataset is small and its distributions are skewed. In future work, we hope that we or other researchers can generate a large scale dataset for depression with PHQ-9 score annotations. Additionally, describing symptoms in tweets can be challenging due to the short text that cannot provide enough information about symptoms and/or depression. Another limitation is that the PHQ-9 annotations in D2S are binary, unlike the 4-point scale that is used in the PHQ-9 questionnaire, which allows to capture severity of symptoms. Choosing between 0 and 1 can be difficult in gray area cases, and degrades annotation quality. Finally, the manually annotated explanations in *D2S-explain* are only a proxy for what a clinician might find most informative in assessing tweets that are automatically flagged as depressive. Evaluating the informativeness of explanations in a true clinical setting would shed more light on this, but is beyond the scope of this paper.

7 Ethics statement

Although Tweets are publicly available, given the sensitivity of the task, we took the following extra measures, in light of what has been previously published by (Benton et al., 2017) and (Šuster et al., 2017).

- We obtained access to the D2S dataset after signing a data use agreement (DUA), and we followed all the agreements and instructions stated in the DUA. The dataset is stored on a secure server and not published with other researchers but those mentioned in the DUA and got approval.
- We did not obtain institutional review board (IRB) approval, since the dataset falls under *exempt determination* and not IRB approval, as stated in the code of federal regulations CFR 46.101(b)(4)[4] published by the United States Department of Health and Human Services (HHS).
- Our publicly available annotated explanations dataset will enforce a DUA that will respect

[4] https://www.hhs.gov/ohrp/sites/default/files/ohrp/policy/ohrpregulations.pdf

	worst model	best model	reported model
LR	0.645	0.684	0.692
MLP	0.384	0.42	0.398
MTL	0.668	0.692	0.611

Table 9: F1 performance results with bagging

	original model	9samples	9samples+D
LR	0.629	0.5	0.6
MLP	0.276	0.21	0.24
MTL	0.533	0.472	0.51

Table 10: Explanations recall performance with bagging

all the requirements of the D2S dataset DUA, in addition to any extra needed regulations and instructions.

8 Conclusion

Providing models that are explainable and adopt clinically grounded questionnaires is critical in building NLP solutions that can be integrated in clinical settings. In this work, we show that using auxiliary models, namely for PHQ-9 categories/symptoms, in combination with the depression classification models, allows us to generate explanations that are more meaningful and have higher recall when evaluated against a gold standard dataset of manually annotated explanations. This implies that we need to conduct more studies that can benefit from clinical practices and measures, and integrate it into the modeling design choices. To the best of our knowledge, we are the first to produce gold annotations of explanations for the depression classification task, and conduct a thorough analysis on how augmenting with the symptoms can improve the quality of the explanations.

References

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

André Bittar, Sumithra Velupillai, Angus Roberts, Rina Dutta, et al. 2021. Using general-purpose sentiment lexicons for suicide risk assessment in electronic health records: corpus-based analysis. *JMIR medical informatics*, 9(4):e22397.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bongjae Choi, Geumsook Shim, Bumseok Jeong, and Sungho Jo. 2020. Data-driven analysis using multiple self-report questionnaires to identify college students at high risk of depressive disorder. *Scientific reports*, 10(1):1–13.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

R Yates Coley, Jennifer M Boggs, Arne Beck, and Gregory E Simon. 2021. Predicting outcomes of psychotherapy for depression with electronic health record data. *Journal of affective disorders reports*, 6:100198.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Fionn Delahunty, Robert Johansson, and Mihael Arcan. 2019. Passive diagnosis incorporating the phq-4 for depression and anxiety. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 40–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph Geraci, Pamela Wilansky, Vincenzo de Luca, Anvesh Roy, James L Kennedy, and John Strauss. 2017. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evidence-based mental health*, 20(3):83–87.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788, Online. Association for Computational Linguistics.

YuanZheng Hu and Marina Sokolova. 2021. Explainable multi-class classification of the camh covid-19 mental health data. *arXiv preprint arXiv:2105.13430*.

Christian Karmen, Robert C Hsiung, and Thomas Wetter. 2015. Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods. *Computer methods and programs in biomedicine*, 120(1):27–36.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Harnain Kour and Manoj K Gupta. 2022. An hybrid deep learning approach for depression prediction from user tweets using feature-rich cnn and bidirectional lstm. *Multimedia Tools and Applications*, pages 1–37.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Maria Elizabeth Loades, Eleanor Chatburn, Nina Higson-Sweeney, Shirley Reynolds, Roz Shafran, Amberly Brigden, Catherine Linney, Megan Niamh McManus, Catherine Borwick, and Esther Crawley. 2020. Rapid systematic review: the impact of social isolation and loneliness on the mental health of children and adolescents in the context of covid-19. *Journal of the American Academy of Child & Adolescent Psychiatry*, 59(11):1218–1239.

David E Losada and Fabio Crestani. 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.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 78–87.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A cross-modal review of indicators for depression detection systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 1–12, Vancouver, BC. Association for Computational Linguistics.

Wajid Mumtaz and Abdul Qayyum. 2019. A deep learning framework for automatic diagnosis of unipolar depression. *International journal of medical informatics*, 132:103983.

Matthew D Nemesure, Michael V Heinz, Raphael Huang, and Nicholas C Jacobson. 2021. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific reports*, 11(1):1–9.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. *arXiv preprint arXiv:2204.10432*.

Vadim Osadchiy, Jesse Nelson Mills, Sriram Venkata Eleswarapu, et al. 2020. Understanding patient anxieties in the social media era: qualitative analysis and natural language processing of an online male infertility community. *Journal of Medical Internet Research*, 22(3):e16728.

David Owen, Jose Camacho Collados, and Luis Espinosa-Anke. 2020. Towards preemptive detection of depression and anxiety in twitter. *arXiv preprint arXiv:2011.05249*.

Robert B Penfold, Eric Johnson, Susan M Shortreed, Rebecca A Ziebell, Frances L Lynch, Greg N Clarke, Karen J Coleman, Beth E Waitzfelder, Arne L Beck, Rebecca C Rossom, et al. 2021. Predicting suicide attempts and suicide deaths among adolescents following outpatient visits. *Journal of affective disorders*, 294:39–47.

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 55–63.

Mark Sendak, Michael Gao, Marshall Nichols, Anthony Lin, and Suresh Balu. 2019. Machine learning in health care: a critical appraisal of challenges and opportunities. *EGEMs*, 7(1).

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

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

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87, Valencia, Spain. Association for Computational Linguistics.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Md Zia Uddin, Kim Kristoffer Dysthe, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2022. Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, 34(1):721–744.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: figurative language enabled multitask learning framework. *arXiv preprint arXiv:2011.06149*.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198.

Li Zhou, Amy W Baughman, Victor J Lei, Kenneth H Lai, Amol S Navathe, Frank Chang, Margarita Sordo, Maxim Topaz, Feiran Zhong, Madhavan Murrali, et al. 2015. Identifying patients with depression using free-text clinical documents. In *MEDINFO 2015: eHealth-enabled Health*, pages 629–633. IOS Press.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.