Micromodels for Efficient, Explainable, and Reusable Systems: A Case Study on Mental Health

Andrew Lee
University of Michigan
ajyl@umich.edu

Jonathan K. Kummerfeld
University of Michigan
jkummerf@umich.edu

Lawrence C. An
University of Michigan
lcan@umich.edu

Rada Mihalcea
University of Michigan
mihalcea@umich.edu

Abstract

Many statistical models have high accuracy on test benchmarks, but are not explainable, struggle in low-resource scenarios, cannot be reused for multiple tasks, and cannot easily integrate domain expertise. These factors limit their use, particularly in settings such as mental health, where it is difficult to annotate datasets and model outputs have significant impact. We introduce a *micromodel architecture* to address these challenges. Our approach allows researchers to build interpretable representations that embed domain knowledge and provide explanations throughout the model's decision process. We demonstrate the idea on multiple mental health tasks: depression classification, PTSD classification, and suicidal risk assessment. Our systems consistently produce strong results, even in low-resource scenarios, and are more interpretable than alternative methods.

1 Introduction

Systems in domains such as healthcare (Caruana et al., 2015) and finance (Heaton et al., 2016) often need to make difficult decisions that can lead to severe consequences. Building useful systems in these settings is difficult for two key reasons: data availability and the need for explanations. Raw data is often limited and annotating it requires specialized knowledge (Aguirre et al., 2021). When a dataset is available for a task, research on models will often overfit, developing optimizations that cannot be reused for other datasets or tasks (Guntuku et al., 2017; Matero et al., 2019; Chen et al., 2019). Attempts to reduce data needs by integrating domain knowledge often result in inefficient and expensive models (Yang et al., 2019; Liu et al., 2020; Xie et al., 2020). Integrating knowledge graphs is another alternative (Zhang et al., 2019), but poses challenges in domains in which domain knowledge is abstract or empirical (Deng et al., 2020). Without explanations of how these models reach their decisions, stakeholders cannot fully trust them. In fact, despite recent advances in neural networks, it has been found that medical experts prefer simpler logistic regression models because they are more interpretable (Caruana et al., 2015).

In this paper, we tackle these challenges – explainability and reusability of models, robustness under low-resource scenarios, and integration of domain knowledge by proposing a new paradigm called a *micromodel architecture*. In this approach, a system orchestrates a collection of specialized models to build easily interpretable feature vectors that integrate domain knowledge. Each micromodel is a binary classifier that represents a specific linguistic behavior. Simple aggregators combine the output of micromodels to form a feature vector. Finally, a task-specific model makes a prediction based on the feature vector. Our design provides explanations along every step of its decision making process, including global and local feature importance scores, and evidence of how the input text contributes to the model's decisions.

Training this type of system involves two phases. First, in order to build each micromodel, we introduce a data collection pipeline that uses pretrained language models such as BERT (Devlin et al., 2019). This training occurs once and then the micromodels can be reused across multiple tasks within a single domain. Second, the task-specific model is trained on the dataset of interest. During this phase the micromodels are not modified.

We demonstrate the benefits of micromodels in the important domain of mental health. Recent studies have shown a rapid increase in the prevalence of depression symptoms in various demographics (Ettman et al., 2020), along with elevated levels of suicidal ideation (Czeisler et al., 2020). Because our micromodels represent domain-level

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4257–4272
November 7–11, 2021. ©2021 Association for Computational Linguistics

linguistic patterns, they can be reused for multiple tasks within the same domain, while requiring only half or sometimes just a quarter of the task-specific annotation data, and also having the benefit of explainability across the entire pipeline.

The primary contributions of this paper are: (1) An efficient and reusable design using micromodels as modules to tackle various tasks within a domain by integrating domain knowledge; (2) A data collection pipeline to build datasets for micromodels; (3) An explainable procedure for our system's decision making process; and (4) An analysis of the reusability and efficiency of our approach under low-resource scenarios when applied to tasks such as depression classification, PTSD classification, and suicidal risk assessment.

2 Background and Related Work

We find inspiration in previous work that addressed explainability, reusability, efficiency under low-resource scenarios, and integration of domain expertise. We focus primarily on research that was carried out in the domain of mental health.

Explainability. Neural networks are black-box models that lack transparency and explainability. Structural analyses of neural networks (Vig et al., 2020), such as probing, has become a popular approach to investigate linguistic properties learned by language models (Wu et al., 2021; Chi et al., 2020; Belinkov et al., 2018; Hewitt and Manning, 2019; Tenney et al., 2018). However, these analyses do not explain how the models use their latent information for their tasks and how they reach their decisions. These drawbacks are especially problematic in the mental health domain (Carr, 2020). Linear models implemented with feature engineering can be analyzed via global feature importance scores, but they do not necessarily provide explanations at a query-level. Model-agnostic explanation frameworks such as SHAP or LIME values (Lundberg and Lee, 2017; Ribeiro et al., 2016) can provide query-level, or local, feature importance scores, but they are approximate explanations of the underlying model. Our approach provides (1) global and local feature importance scores, and (2) evidence from input text data that led to its output.

Reusability. Recent models in the mental health domain are often task-specific or data-specific. Examples include features extracted from metadata (Guntuku et al., 2017), or neural architectures that either fine-tune their embeddings (Orabi et al., 2018) or have task-specific layers (Matero et al., 2019). While task-specific designs can boost accuracy, they are difficult to extend to multiple applications. Furthermore, Harrigian et al. (2020) show that models trained for a task in the mental health domain do not generalize across test sets that originate from different sources. Because our micromodels are built on task-agnostic data, they are reusable for multiple applications within a domain.

Efficiency in Low-Resource Scenarios. Obtaining data in the mental health domain is difficult because of the sensitive nature of data and the need for expert annotators. While researchers have turned to proxy-based annotations, in which data is annotated using automated mechanisms (Yates et al., 2017; Winata et al., 2018), these datasets have caveats and biases (Aguirre et al., 2021; Coppersmith et al., 2015). These data limitations make it difficult to apply standard neural methods.

Integrating Domain Expertise. Psychologists have long studied effective methods for assessing patients for various mental health illnesses. Assessment modules such as the Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al., 2001) or PTSD Checklist (PCL) (Ruggiero et al., 2003) allow physicians to reliably screen for the presence or severity of various mental statuses.

Similarly, cognitive distortions are irrational or exaggerated thought patterns that can reinforce negative emotions, often exhibited by depressed patients (Beck, 1963). Recognizing and treating these negative thought patterns is the focus of cognitive-behavior interventions (Kaplan et al., 2017). The PHQ-9 and an example categorization of cognitive distortions can be found in the appendix.

While these assessment modules and methods are used in clinical settings, it has been unclear how to incorporate them into automated systems. In our work, we are able to represent responses to these questionnaires and instances of cognitive distortions using micromodels. This allows our models to leverage domain knowledge.

3 Micromodel Architecture

Our micromodel approach is inspired by recent work in microservice architectures—an organizational design in which applications are built from a collection of loosely coupled services (Nadareishvili et al., 2016). Each of these services

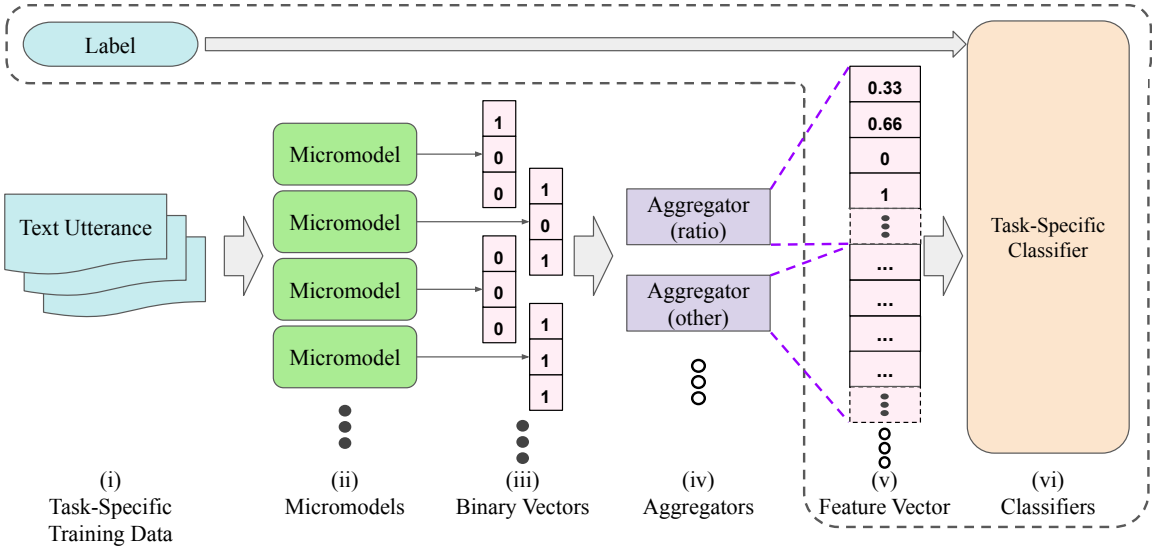

Figure 1: One training step for a task-specific classifier, given a collection of pre-built micromodels. The input (i) is a set of utterances and a single label. Our micromodels (ii) process each utterance to produce a set of binary vectors (iii). Here the vectors have three elements because our example contains three utterances. Aggregators (iv) summarize the binary vectors in a feature vector (v). A task specific classifier takes the feature vector as input and makes a prediction, which is compared to the true label to make an update. Note that the classifier only sees the feature vector (v) and its corresponding label for training.

typically has a fine-grained focus of responsibility. In a similar manner, we build a collection of micromodels, with each one responsible for identifying a specific linguistic behavior.[1]

3.1 Micromodels

A micromodel identifies a specific linguistic behavior. We use binary classifiers for their simplicity, but our architecture is general enough to allow for other representations. A micromodel can rely on any algorithm, from decision trees and heuristics to linear models and neural networks.

Each micromodel is responsible for representing a specific linguistic behavior. For mental health, we developed a set of micromodels that represent examples of cognitive distortions or responses to the PHQ-9 mental health questionnaire: one micromodel identifies expressions of apathy or lack of enthusiasm (PHQ-9 question 1), while another identifies examples of all-or-nothing thinking (cognitive distortion), and so on. We describe the process of constructing a micromodel in Section 3.3.

3.2 Architecture

Figure 1 shows our micromodel architecture. At the heart of the architecture is the collection of micromodels $M = \{mm_1, ..., mm_n\}$. Micromodels are pre-built using a task-agnostic dataset (see Section 3.3), and are not updated during task-specific training.[2] The six steps of our architecture are:

(i) Let (S_i, y_i) be one training data instance, where S_i contains multiple utterances $\{s_1, ..., s_k\}$ and y_i is the corresponding label for the whole set. For instance, imagine a task of predicting a Twitter user's mental status given their recent tweets. S_i would be the user's tweets, where each $s \in S_i$ is a single tweet, and y_i is the user's mental status. Note that there are no utterance-specific labels.

(ii, iii) Given (S_i, y_i), each micromodel $mm_j \in M$ produces a binary value for each utterance $s \in S_i$. A value of 1, or a "hit", indicates that utterance s is an example of the linguistic behavior that mm_j is looking for. We only use binary values in this work, but our architecture allows non-binary outputs too. The result is n binary vectors v of length k, one from each micromodel. Note that each binary vector v_j represents the indices in S_i where the target behavior of mm_j can be found.

(iv, v) Each binary vector is fed through a set of aggregators. Each aggregator maps the set of binary vectors into a feature vector that can be used

[1] This is where our term micromodel comes from – each model has a fine-grained focus of responsibility. We are not referring to each model's memory footprint.

[2] This is an intentional choice to prevent model drift. If we allowed updates to the task-specific models their model capacity may be repurposed to do something other than their original design intended.

for classification. An aggregator can perform any computation. For example, it could calculate the ratio of hits in the binary vector. The resulting feature values would then represent the proportion of utterances that demonstrate each specific linguistic behavior. We focus on one-to-one mappings between a micromodel and a feature value, but they can also be many-to-one or one-to-many operations. Together, steps (ii)-(v) could be considered a model that converts input text to a vector representation in an interpretable way.

(vi) The feature vector and its corresponding label y_i are passed to a task-specific classifier. We use explainable boosting machines (EBM) (Nori et al., 2019; Caruana et al., 2015), a type of generalized additive model (GAM) (Lou et al., 2012, 2013). These produce a prediction by adding together a set of functions of one or two input features. Each function is trained using bagging and gradient boosting. The result is a model that is more flexible than a linear model, while still being easy to interpret since it can be visualized as a set of graphs, one per function (see Section 5 for an example of this in practice).

While the above description was used for our experiments, our framework itself is more general.

First, our micromodels are not limited to binary values (iii). They can output continuous values, such as BERT similarity scores (Section 3.3), as long as the subsequent aggregators (iv) know how to process them. A simple example of such aggregation might be max-pooling the micromodel output vector (iii). In this example, the resulting feature value (v) would then represent the maximum similarity score that a micromodel identified in the task-specific training data (i).

Second, in our experiments, the task-specific classifier only sees the feature vector (v) during training, and not the original input text data. This is not a limitation of our architecture – other algorithms of choice could be used, including those that use neural features directly from the input text. This may improve accuracy, but at the cost of interpretability. Given the sensitive and high-risk domain of healthcare, where even the most accurate models become impractical without explainability (Caruana et al., 2015), we use EBMs in this work.

Third, researchers can give their own definition of "Text Utterances" (i). In the CLPsych 2015 Shard Task (Section 4.1), we define each "Text Utterance" to be a single tweet from a user. However,

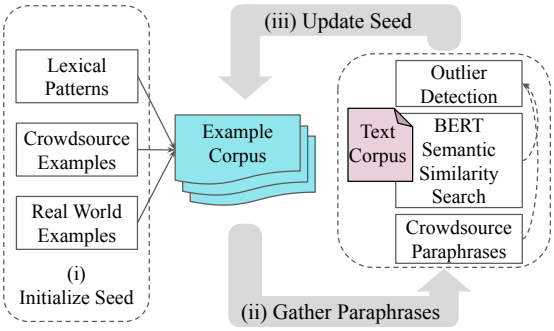

Figure 2: Data collection pipeline for each micromodel. Our approach is an iterative approach, in which the example corpus is updated with paraphrases. Optionally, an outlier detection module can be incorporated in order to find the sentences that would add the most diversity to the example corpus.

a different granularity could have been used, such as a set of tweets or all tweets from a user. Such grouping allows micromodels to capture contextual information from each training data instance.

Note that only the task-specific classifier's weights are updated during training. The micromodels are not updated – they are only used to extract the linguistic patterns that we care about. This is done for a few reasons: (1) We do this to avoid each micromodel's representation from shifting away from their intended meaning; (2) Fine-tuning each micromodel requires labels at a micromodel granularity, rather than task-level granularity. For instance, in the CLPsych 2015 Shared Task data (Section 4.1), this means instead of (# of users) annotations, we would need (# of micromodels) * (# of tweets per user) * (# of users) annotations; and (3) Not all micromodels have "weights", as they can also be arbitrary heuristics (Section 3.1).

3.3 Building Micromodels using BERT

Each micromodel is intended to detect a specific linguistic behavior. In order to build robust linguistic representations, it is critical to give each micromodel a diverse and representative sample of data. However, annotating data can be time consuming and expensive. We use BERT and Universal Sentence Encoders (Cer et al., 2018) to rapidly collect representative samples for each micromodel. Our approach is inspired by work on collecting data for dialogue systems. Specifically, Kang et al. (2018), Larson et al. (2019), Larson et al. (2020), and Stasaski et al. (2020) proposed ways to build a diverse dataset by iteratively collecting data, starting from a seed set and crowdsourcing paraphrases.

Figure 2 depicts our pipeline for building our micromodel datasets. For each micromodel, we build an example corpus and gather paraphrases. While crowdsourcing can be thought of a generative approach for paraphrasing, we take a retrieval approach by using a BERT model to search for semantically similar sentences in a separate corpus of unstructured text data. In particular, we use anonymized posts from the r/depression subreddit[3], a peer support forum for anyone struggling with a depressive disorder. While any corpus can be used to retrieve paraphrases, it is important that the linguistic phenomena that is of interest will be prevalent in the corpus. We used Sentence Transformers (Reimers and Gurevych, 2019)[4] and the "paraphrase-xlm-r-multilingual-v1" pre-trained model for our semantic similarity searches.

There are multiple ways to initialize the example corpus. One can build lexical queries by specifying patterns based on parsers or lexicons and apply them on a text corpus. For instance, to find examples of the labeling cognitive distortion (attaching a negative label to oneself), a lexical query might look for sentences that contain a first person pronoun with a nominal subject relation with a negative token according to the LIWC lexicon (Pennebaker et al., 2001).

While this may seem like an overly simple and generic pattern, because the lexical query is applied on a text corpus that pertains to depression, we are able to retrieve many examples of the target behavior, in this case the labeling cognitive distortion. It is important to consider which text corpus the lexical query is being applied to. To prevent micromodels from overfitting on these rule-based patterns, it is critical to run through multiple iterations of the BERT similarity search while updating the example corpus each round. This step will identify examples of the target linguistic behavior that do not match the lexical query.

Note that this step can be pseudo-automated in a couple of ways. One way is to apply a "negation" lexical query on the BERT results. For instance, in the example lexical query above, given new examples of the labeling cognitive distortion according to BERT, one might apply a lexical query for utterances that do not contain a first person pronoun or a negative LIWC token. This would identify semantically similar but syntactically diverse samples to be added back to the example corpus.

We also follow Larson et al. (2019) and use a Universal Sentence Encoder to identify outliers from our BERT results. This helps us identify utterances that would add the most diversity when added back to the example corpus. We use Snorkel[5] (Hancock et al., 2018; Ratner et al., 2017, 2016) to construct our lexical queries.

Note that given an example corpus, applying a BERT similarity search between an input sentence and the example corpus can also be a form of a micromodel. Once we have collected examples of a specific linguistic behavior, if the input sentence has a similarity score above a threshold value with any of the examples, our micromodel would return a value of 1, and a value of 0 otherwise. We call this a BERT query and use a handful of them for our experiments. These BERT queries are able to identify examples of nuanced concepts such as cognitive distortions or a response to a PHQ-9 question, allowing us to build *contextual* features that represent domain expertise. Note that a BERT query micromodel does not require training, as we only use its inference against an example corpus.

3.4 Discussion: Feature Engineering, Ensemble Models, and Micromodels

Prior to neural models, many NLP systems used linear models with manually defined input features. The process of defining these input features, sometimes called feature engineering, includes common features (e.g., unigrams, bigrams, trigrams) and domain-specific features (e.g., what time of day this tweet was posted). One appeal of neural networks is that they can automatically learn how to combine components of the input (e.g., unigrams, timestamps) to get informative features. While our approach has some similarities with feature engineering, there are several key differences.

First, micromodels are using *external* data (such as the r/depression subreddit - Section 3.3) to learn specific linguistic phenomena. This means they can learn things that cannot be learned from the task-specific data alone, particularly if data is limited.

Second, feature engineering typically produces a huge number of features, whereas we have on the order of tens of micromodels. This is critical for interpretability, as we can look at the output of all our micromodels and at the patterns learned

[3] https://www.reddit.com/r/depression/

[4] https://github.com/UKPLab/sentence-transformers

[5] https://github.com/snorkel-team/snorkel

by the EBMs. In contrast, it would be difficult to meaningfully interpret, for example, the weights assigned to all bigrams.

Third, the primary question for feature engineering is how to best summarize the available training data, while the primary questions for our approach are what data should be leveraged and what models should be built to understand and describe the training data. Another way to view this nuance is that feature engineering extracts task-level features that suit the data for a given task. Micromodels, on the other hand, build task-agnostic, domain-level features that can be applied on multiple tasks.

Lastly, features from prior work are typically syntactic, statistical, or derivative features, such as lexical term frequencies (Coppersmith et al., 2014), extractions from metadata (Guntuku et al., 2017), or sentiment analyses scores (Chen et al., 2019). In addition to these features, we are able to build *contextual* features using contextualized language models, which are able to capture more nuanced concepts reflecting domain expertise. While word embeddings have been used as features before (Mohammadi et al., 2019), they are often difficult to interpret. On the other hand, because the researcher defines the behavior of each aggregator, our resulting feature vector is easy to interpret.

Because a micromodel architecture orchestrates multiple models, it may appear similar to ensemble learning. The key difference is that every model in an ensemble learns the same task, while the micromodels each have a different aim. Micromodels are also intended to be used across tasks, whereas the models in an ensemble are task specific.

4 Evaluation

We evaluate our micromodel architecture in terms of accuracy, reusability, and efficiency under low-resource scenarios. We also address the explainability properties of our model in Section 5.

4.1 Data

CLPsych 2015 Shared Task (Coppersmith et al., 2015). This data contains tweets from 1,146 users labeled as Depression, PTSD, or Control. Users annotated as depressed or PTSD were based on self-identified diagnosis in tweets, which were removed afterwards. For each user identified as depressed or PTSD, an age- and gender-matched user was randomly sampled as a control user. For each user, up to 3,000 of their most recent public tweets were collected. The tasks include (1) classifying depression users versus control users (D vs. C), (2) classifying PTSD users versus control users (P vs. C), and (3) classifying depression users versus PTSD users (D vs. P).

CLPsych 2019 Shared Task (Shing et al., 2018; Zirikly et al., 2019). This data is from Reddit users who have posted in the r/SuicideWatch [6] subreddit, a peer support forum for anyone struggling with suicidal thoughts, and were annotated with 4 levels of suicidal risk (no risk, low, moderate, severe). A group of users who have never posted on r/SuicideWatch was used as a control group. The shared task includes 3 tasks: Task A is risk assessment looking *only* at the users' posts in r/SuicideWatch. Task B is also risk assessment, but also provides posts across other subreddits. Task C is about screening, with only posts that are *not* in r/SuicideWatch available, which removes self-reported evidence of risk.

4.2 Experimental Setup

We use 20 micromodels consisting of algorithms such as SVM, BERT queries, as well as heuristics. The choices for our micromodels were mainly motivated by existing tools commonly used by practitioners in the mental health domain, such as the PHQ-9 questionnaire and cognitive distortions. Out of the PHQ-9 questions and cognitive distortions, those with abundant examples in the r/depression subreddit were built as micromodels. Other linguistic behaviors that practitioners have studied (Zahn et al., 2015; Abraham and Fava, 1999; Levy and Deykin, 1989; Swearer et al., 2001; Cohan et al., 2018) were included as well. Details about each micromodel can be found in Table 1. For our SVM micromodel, we use a linear kernel and a bag of words feature representation [7]. Our Mental Illness, Antidepressants, Depression, and PTSD keyword micromodels use a carefully curated mapping of health conditions to n-grams[8], which were extracted from Benton et al. (2017), and simply return 1 if any corresponding keywords are found in the input utterance. Similarly, our LIWC micromodels return 1 when a keyword for each emotion is found according to LIWC. Each BERT query

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

[7] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[8] https://github.com/kharrigian/mental-health-keywords

Name	Algorithm	Category	Corpus Size	Reference that motivated this micromodel
All-or-Nothing Thinking	SVM	Cognitive Distortion	-	Beck (1963); Bridges et al. (2010)
Labeling	BERT Query	Cognitive Distortion	106	Beck (1963); Bridges et al. (2010)
Fortune-Telling Error	BERT Query	Cognitive Distortion	220	Beck (1963); Bridges et al. (2010); Sastre-Buades et al. (2021)
Loss of Concentration	BERT Query	PHQ-9	38	Kroenke et al. (2001)
Feeling Down, Depressed	BERT Query	PHQ-9	195	Kroenke et al. (2001)
Poor Appetite or Overeating	BERT Query	PHQ-9	49	Kroenke et al. (2001)
Self Harm	BERT Query	PHQ-9	54	Kroenke et al. (2001)
Feeling Worried, Nervous, Anxious	BERT Query	GAD-7	66	Spitzer et al. (2006)
Diagnosis	BERT Query	Other	55	
Self-Blaming	BERT Query	Other	37	Zahn et al. (2015)
Substance Abuse	BERT Query	Other	109	Abraham and Fava (1999); Levy and Deykin (1989)
Victimhood	BERT Query	Other	73	Swearer et al. (2001)
Mental Illness Keywords	Logic	Other	-	Cohan et al. (2018)
Antidepressants Keywords	Logic	Other	-	
Depression Keywords	Logic	Other	-	
PTSD Keywords	Logic	Other	-	
LIWC Sadness	Logic	Other	-	Cohan et al. (2018)
LIWC Anger	Logic	Other	-	Cohan et al. (2018)
LIWC Joy	Logic	Other	-	Cohan et al. (2018)
LIWC Fear	Logic	Other	-	Cohan et al. (2018)

Table 1: The micromodels we developed for this work.

Model	Expl?	Reuse?	D vs C n = 654	P vs C n = 492	D vs P n=573
LR	✓	✓	0.8	0.817	0.785
CNN		✓	0.79	0.85	0.87
UMD			0.86	0.893	0.841
WWBP			**0.904**	0.916	0.81
MM	✓	✓	0.821	**0.936**	**0.892**

Table 2: AUC scores for various approaches, where LR is a logistic regression model, CNN is a convolutional neural network, and MM is our micromodel approach. UMD is from Resnik et al. (2015), WWBP is from Preotiuc-Pietro et al. (2015) – these two systems were the only ones that reported AUC scores and are directly comparable to ours. We also indicate whether each approach is explainable and reusable.

micromodel has its own example corpus built using our data collection pipeline (Section 3.3), and uses a similarity score threshold value of 0.85. We use two aggregators. One is as described in Section 3.2, which returns the ratio of hits in a binary vector. The other aggregator looks for "windows": segments within each binary vector where many hits occur close to one another. These windows may represent temporal "episodes" – for instance, a period in which someone felt apathetic (PHQ-9 question 1), or a period in which someone had a sleeping disorder (PHQ-9 question 3) and so on.

4.3 Results and Analyses

Accuracy. We follow prior work (Resnik et al., 2015; Preotiuc-Pietro et al., 2015) and use ROC area-under-the-curve (AUC) to evaluate the accuracy of our approach, along with a wide range of baseline models and present them in Table 2. We include a logistic regression model, which has been a simple yet effective benchmark in similar tasks (Harrigian et al., 2020), as well as a convolution neural network (CNN) based on Orabi et al. (2018). Lastly we include any AUC scores that were available from system submissions from the shared task. Our approach consistently demonstrates high AUC scores, with the highest AUC scores for classifying PTSD users against control users and depression users against PTSD users.

Efficiency in Low-Resource Scenarios. Gathering and annotating data can be both time consuming and expensive, especially within the mental health domain. Our approach can work with relatively little task-specific data. The micromodels are not retrained, and the task-specific classifier can work with limited data because it is (1) a relatively simple model, and (2) informed by the micromodels. Figure 3 shows the AUC scores of our approach compared to our baseline models with various amounts of task-specific annotated data. We consider five sets; the first has a random sample of 1/16th of the available training data, and each subsequent set has twice as much data. We show results averaged over five runs of this data sampling process. Unlike the baseline models, our approach stays robust down to just 1/4th of the training data.

Reusability. Because micromodels are task agnostic, they can be reused for tasks within the same domain. This contrasts with the standard way of developing models, where the annotation scheme, embeddings, model structure, and so on, are carefully designed, curated, or fine-tuned per task.

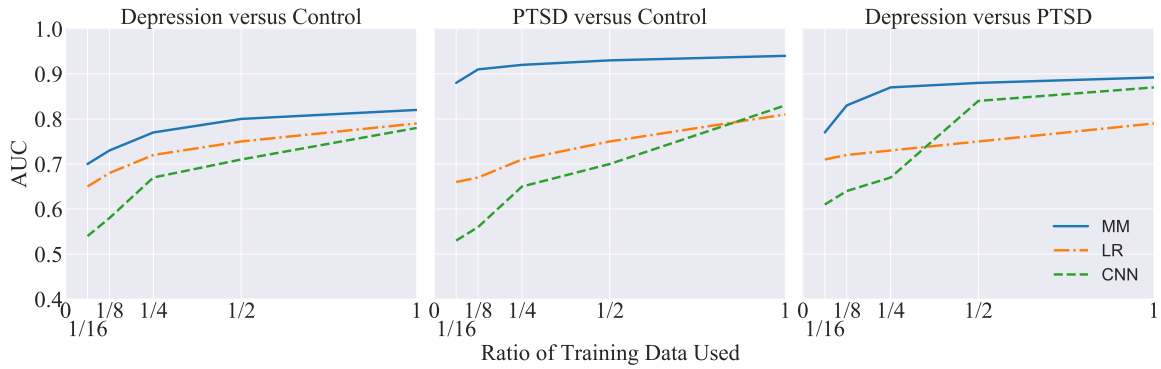

Figure 3: AUC scores for various models under low-resource scenarios. Each curve is an average of 5 runs, with random samples of training data for each run. Our micromodel approach (MM) converges in performance with half, and sometimes just a quarter of the task-specific annotation data. While logistic regression (LR) sees a more linear improvement and a convolutional neural network (CNN) sees sporadic jumps in performance, our approach flattens out early, indicating early convergence and less reliance on the annotated data.

In order to demonstrate the reusability of our micromodels, we also apply them to the CLPsych 2019 shared task. Note that none of the micromodels were updated – only the weights for the EBM classifier were learned using the annotated data. Table 3 shows the macro-F_1 scores of our approach amongst the systems submitted to the shared task[9]. Because we care about reusability, they are sorted by their average ranking across the three tasks.

There are a couple of observations to make from these results. First, despite not having any task-specific design in place, our approach ranks 3rd amongst the systems on average. Second, our approach is one of the best performing approaches for Task C. Unlike the first two assessment tasks, Task C is concerned with screening for suicidal risk given *none* of their posts from r/SuicideWatch. Because of the lack of self-reported evidence of any suicidal ideation, this task was considered the hardest task, as evident by the low F_1 scores. Since our suite of micromodels are built to identify various linguistic traits of depressive users, even without immediate signals of suicidal ideation, our approach is able to detect signs of depression, a precursor for suicide risk, and screen for users with potential risk of suicide. We believe this demonstrates our micromodels' ability to understand domain-level concepts, rather than task-specific patterns, thus allowing our micromodels to be reused in multiple tasks within the same domain.

Model	r/SuicideWatch Data? Only (Task A)	Yes (Task B)	No (Task C)
Mohammadi et al.	**0.481**•	0.339•	**0.268**•
Matero et al.	0.459•	**0.457**•	0.176
Micromodels	0.395	0.274	0.255
Ambalavanan et al.	0.477•	0.261	0.159
Ríssola et al.	0.291	0.311•	0.136
Morales et al.	0.178	0.212	0.165
Iserman et al.	0.402•	0.148	0.118
Bitew et al.	0.445•	-	-
Allen et al.	0.373	-	-
Hevia et al.	0.312	-	-
Ruiz et al.	-	0.370•	-
Chen et al.	-	0.358•	-

Table 3: Macro-F_1 scores of micromodels and system submissions from the CLPsych 2019 Shared Task. To understand the reusability of each system across the three tasks, they are sorted by the average of their rankings on each task. • indicates scores higher than that of our approach.

5 Step-wise Explanations

Our micromodel architecture provides various levels of explanations during each step. We first demonstrate the explanations provided by EBM classifiers before walking through each step.

EBMs are additive models in which a nonlinear function f_i is learned for each input feature i. One can calculate global feature importance scores by applying each feature function f_i on every point t in the training data. We then take the average of the absolute value of $f_i(t)$ for each feature i:

$$FeatureImportance_i = avg(abs(f_i(t))), t \in T \quad (1)$$

where T is our entire training data. Figure 4 shows the top 10 most important features for the three CLPsych 2015 shared tasks. Similarly, we can explain the model's decision for a specific instance $t \in T$ by simply applying $f_i(t)$ for each i.

Inspecting the plots of each f_i also provides a granular explanation of our classifier. Figure 5 con-

[9]We exclude systems without paper submissions

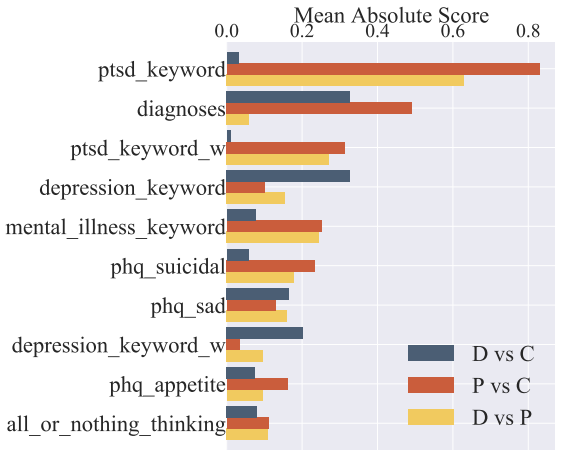

Figure 4: Ten most important features according to their average global feature importance scores on the three CLPsych 2015 shared tasks: depression versus control, PTSD versus control, and depression versus PTSD. Features ending in "w" are features from the aggregator that looks for windows of hits.

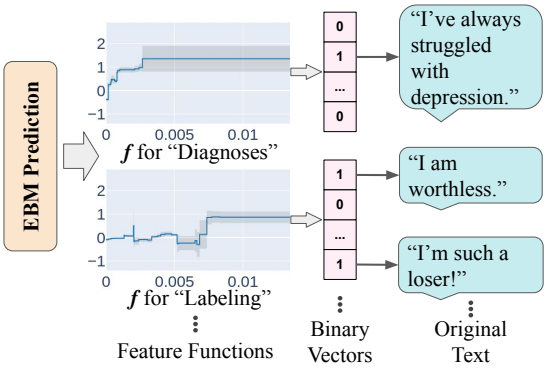

Figure 5: Explanations provided by various steps in the micromodel architecture. The feature functions f provide details on how each feature contributes to the classifier's decisions. The binary vectors indicate the location of each micromodel's hits in the input text data, allowing us to look them up as evidence.

tains examples of two of the feature functions from the depression detection task. The x-axis indicates the ratio of hits for each micromodel – in the context of this task, this represents the ratio of tweets per user that contain a specific linguistic behavior. While $f_{\text{Diagnoses}}$ produces a strong signal when a user contains *any* tweets that exhibit a diagnosis statement, f_{Labeling} produces a strong signal when more than roughly 0.75% of a user's tweets contain an example of the labeling cognitive distortion.

Other than the EBM classifier, our approach also provides explanations throughout each step. The first step consists of the micromodels, whose explainability depends on their underlying algorithms. The choice of these models likely involves a trade-off between accuracy and explainability.

The binary vectors produced by the micromodels indicate the utterance in which a specific linguistic behavior can be found. This provides provenance for our feature vector – we can use them to look up the sentences in the original input text before they were featurized. Figure 5 demonstrates this process [10]. Such text data provides evidence for the model's decisions. This text data can be combined with the feature importance scores to understand how they affected the model's decisions, or to uncover patterns in the users' behaviors.

As for aggregators, in this work we use simple and intuitive operations, making the resulting feature vector easy to interpret. Note that without an interpretable feature vector, the feature functions f_i also become difficult to understand as well.

[10] We use fabricated examples to protect the identity of Twitter users in the dataset.

6 Conclusion

In this paper, we introduced a new framework that uses a collection of micromodels to tackle various tasks within the mental health domain. Rather than directly applying contextualized language models to a task, we use them to rapidly collect diverse samples to build micromodels, which leads to a distributed-learning paradigm. Incorporating contextual language models in our data collection allows us to capture nuanced behaviors such as cognitive distortions. Furthermore, our pipeline allows us to leverage any amount of external data, rather than extracting features within the task domain.

The resulting micromodels allow us to build *contextual* features, each of which can represent linguistic behaviors or domain knowledge. Such a feature vector is intuitive to interpret while being effective for classifiers to learn from, even in low-resource scenarios in which not a lot of task-specific annotation data is available. Our approach provides explanations throughout the entire decision making process, including both global and local feature importance scores, as well as the exact locations of the text that contributed to the model's decisions. Because our micromodels are built in a task-agnostic manner, they can be reused for multiple tasks within the same domain.

The code for our micromodel architecture is publicly available at https://github.com/MichiganNLP/micromodels.git.

7 Ethical Considerations

While we believe our approach takes a step towards the application of intelligent systems to data-poor or sensitive domains such as mental health, it is important to discuss potential risks, harm, and limitations of our work.

Because our approach heavily relies on micromodels that represent linguistic behaviors or domain knowledge, it is critical that their representations are faithful. The authors responsible for building our micromodels were trained on cognitive behavior therapy and cognitive distortions. It is important to have trained experts heavily involved throughout our data collection process and guiding the evaluation of how accurate the micromodels are. This leads to a limitation of our work. While we evaluated our approach in an end-to-end manner for various tasks, we found it challenging to evaluate the micromodels in isolation. The difficulty in building test sets arise from not only the effort involved in gathering accurate annotations, but also from requiring high coverage and diversity of linguistic phenomena in the data as well.

Lastly, Aguirre et al. (2021) demonstrate that the CLPsych 2015 shared task dataset is not demographically representative. Our work is only a proof of a concept, and to be applied in a real world scenario, a non-biased dataset should be used.

Acknowledgements

We would like to thank Joseph Himle, Addie Weaver, and Anao Zhang from the School of Social Work at University of Michigan for the training on cognitive behavior therapy and cognitive distortions. We thank the members of the LIT lab at University of Michigan for constructive feedback. We thank the EMNLP reviewers for their helpful suggestions. This material is based in part upon work supported by a Google focus award, by the Precision Health initiative at the University of Michigan, and by DARPA (grant #D19AP00079).

References

Henry David Abraham and Maurizio Fava. 1999. Order of onset of substance abuse and depression in a sample of depressed outpatients. *Comprehensive Psychiatry*, 40(1):44–50.

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2932–2949.

Kristen Allen, Shrey Bagroy, Alex Davis, and Tamar Krishnamurti. 2019. Convsent at clpsych 2019 task a: Using post-level sentiment features for suicide risk prediction on reddit. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 182–187.

Ashwin Karthik Ambalavanan, Pranjali Dileep Jagtap, Soumya Adhya, and Murthy Devarakonda. 2019. Using contextual representations for suicide risk assessment from internet forums. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 172–176.

Aaron T Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, 9(4):324–333.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *arXiv preprint arXiv:1801.07772*.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162.

Semere Kiros Bitew, Ioannis Bekoulis, Johannes Deleu, Lucas Sterckx, Klim Zaporojets, Thomas Demeester, and Chris Develder. 2019. Predicting suicide risk from online postings in reddit: the ugent-idlab submission to the clpysch 2019 shared task a. In *CLPsych2019, the 6th Annual Workshop on Computational Linguistics and Clinical Psychology at NAACL-HLT 2019*, pages 158–161. Association for Computational Linguistics (ACL).

K Robert Bridges, Richard J Harnish, et al. 2010. Role of irrational beliefs in depression and anxiety: a review. *Health*, 2(08):862.

Sarah Carr. 2020. 'ai gone mental': engagement and ethics in data-driven technology for mental health.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Lushi Chen, Abeer Aldayel, Nikolay Bogoychev, and Tao Gong. 2019. Similar minds post alike: Assessment of suicide risk using a hybrid model. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 152–157.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

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.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

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.

Mark É Czeisler, Rashon I Lane, Emiko Petrosky, Joshua F Wiley, Aleta Christensen, Rashid Njai, Matthew D Weaver, Rebecca Robbins, Elise R Facer-Childs, Laura K Barger, et al. 2020. Mental health, substance use, and suicidal ideation during the covid-19 pandemic—united states, june 24–30, 2020. *Morbidity and Mortality Weekly Report*, 69(32):1049.

Changyu Deng, Xunbi Ji, Colton Rainey, Jianyu Zhang, and Wei Lu. 2020. Integrating machine learning with human knowledge. *Iscience*, page 101656.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Catherine K Ettman, Salma M Abdalla, Gregory H Cohen, Laura Sampson, Patrick M Vivier, and Sandro Galea. 2020. Prevalence of depression symptoms in us adults before and during the covid-19 pandemic. *JAMA network open*, 3(9):e2019686–e2019686.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Braden Hancock, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 1884. NIH Public Access.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (EMNLP)*, pages 3774–3788.

JB Heaton, Nicholas G Polson, and Jan Hendrik Witte. 2016. Deep learning in finance. *arXiv preprint arXiv:1602.06561*.

Alejandro González Hevia, Rebeca Cerezo Menéndez, and Daniel Gayo-Avello. 2019. Analyzing the use of existing systems for the clpsych 2019 shared task. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 148–151.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Micah Iserman, Taleen Nalabandian, and Molly Ireland. 2019. Dictionaries and decision trees for the 2019 clpsych shared task. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 188–194.

Yiping Kang, Yunqi Zhang, Jonathan K Kummerfeld, Lingjia Tang, and Jason Mars. 2018. Data collection for dialogue system: A startup perspective. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 33–40.

Simona C Kaplan, Amanda S Morrison, Philippe R Goldin, Thomas M Olino, Richard G Heimberg, and James J Gross. 2017. The cognitive distortions questionnaire (cd-quest): Validation in a sample of adults with social anxiety disorder. *Cognitive therapy and research*, 41(4):576–587.

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.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K Kummerfeld, Parker Hill Michael A Laurenzano Johann, and Hauswald Lingjia Tang Jason Mars. 2019. Outlier detection for improved data

quality and diversity in dialog systems. In *Proceedings of NAACL-HLT*, pages 517–527.

Stefan Larson, Anthony Zheng, Anish Mahendran, Rishi Tekriwal, Adrian Cheung, Eric Guldan, Kevin Leach, and Jonathan K. Kummerfeld. 2020. Iterative feature mining for constraint-based data collection to increase data diversity and model robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8097–8106.

Janice C Levy and Eva Y Deykin. 1989. Suicidality, depression, and substance abuse in adolescence. *The American journal of psychiatry*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. Clac at clpsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 34–38.

Michelle Morales, Prajjalita Dey, Thomas Theisen, Daniel Belitz, and Natalia Chernova. 2019. An investigation of deep learning systems for suicide risk assessment. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 177–181.

Irakli Nadareishvili, Ronnie Mitra, Matt McLarty, and Mike Amundsen. 2016. *Microservice architecture: aligning principles, practices, and culture*. " O'Reilly Media, Inc.".

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Daniel Preotiuc-Pietro, Maarten Sap, H Andrew Schwartz, and Lyle H Ungar. 2015. Mental illness detection at the world well-being project for the clpsych 2015 shared task. In *CLPsych@ HLT-NAACL*, pages 40–45.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29:3567.

Alexander J Ratner, Stephen H Bach, Henry R Ehrenberg, and Chris Ré. 2017. Snorkel: Fast training set generation for information extraction. In *Proceedings of the 2017 ACM international conference on management of data*, pages 1683–1686.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015. The university of maryland clpsych 2015 shared task system. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 54–60.

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.

Esteban A Ríssola, Diana Ramíirez-Cifuentes, Ana Freire, and Fabio Crestani. 2019. Suicide risk assessment on social media: Usi-upf at the clpsych 2019 shared task. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology: 2019 Jun 6; Minneapolis, Minnesota, USA. Stroudsburg: ACL; 2019. p. 167–71.* ACL (Association for Computational Linguistics).

Kenneth J Ruggiero, Kevin Del Ben, Joseph R Scotti, and Aline E Rabalais. 2003. Psychometric properties of the ptsd checklist—civilian version. *Journal of traumatic stress*, 16(5):495–502.

Victor Ruiz, Lingyun Shi, Wei Quan, Neal Ryan, Candice Biernesser, David Brent, and Rich Tsui. 2019. Clpsych2019 shared task: Predicting suicide risk level from reddit posts on multiple forums. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 162–166.

Aina Sastre-Buades, Susana Ochoa, Esther Lorente-Rovira, Ana Barajas, Eva Grasa, Raquel López-Carrilero, Ana Luengo, Isabel Ruiz-Delgado, Jordi Cid, Fermín González-Higueras, et al. 2021. Jumping to conclusions and suicidal behavior in depression and psychosis. *Journal of psychiatric research*.

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.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.

Katherine Stasaski, Grace Hui Yang, and Marti A. Hearst. 2020. More diverse dialogue datasets via diversity-informed data collection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4958–4968.

Susan M Swearer, Samuel Y Song, Paulette Tam Cary, John W Eagle, and William T Mickelson. 2001. Psychosocial correlates in bullying and victimization: The relationship between depression, anxiety, and bully/victim status. *Journal of Emotional Abuse*, 2(2-3):95–121.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2018. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*.

Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 2018. Attention-based lstm for psychological stress detection from spoken language using distant supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6204–6208. IEEE.

Zhaofeng Wu, Hao Peng, and Noah Smith. 2021. Infusing finetuning with semantic dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.

Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, and Shaojie Tang. 2020. A survey on domain knowledge powered deep learning for medical image analysis. *arXiv preprint arXiv:2004.12150*.

Zijiang Yang, Reda Al-Bahrani, Andrew CE Reid, Stefanos Papanikolaou, Surya R Kalidindi, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. 2019. Deep learning based domain knowledge integration for small datasets: Illustrative applications in materials informatics. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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 (EMNLP)*, pages 2968–2978.

Roland Zahn, Karen E Lythe, Jennifer A Gethin, Sophie Green, John F William Deakin, Allan H Young, and Jorge Moll. 2015. The role of self-blame and worthlessness in the psychopathology of major depressive disorder. *Journal of affective disorders*, 186:337–341.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

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.

A Feature Importance Scores

Figure 1 lists the global feature importance scores for the features used in classifying 1) depression versus control, 2) PTSD versus control, and 3) depression versus PTSD.

B Cognitive Distortions

Table 1 lists some common examples of cognitive distortions, along with their definitions and some examples.

C PHQ-9 Questionnaire

Table 2 lists the PHQ-9 Questionnaire.

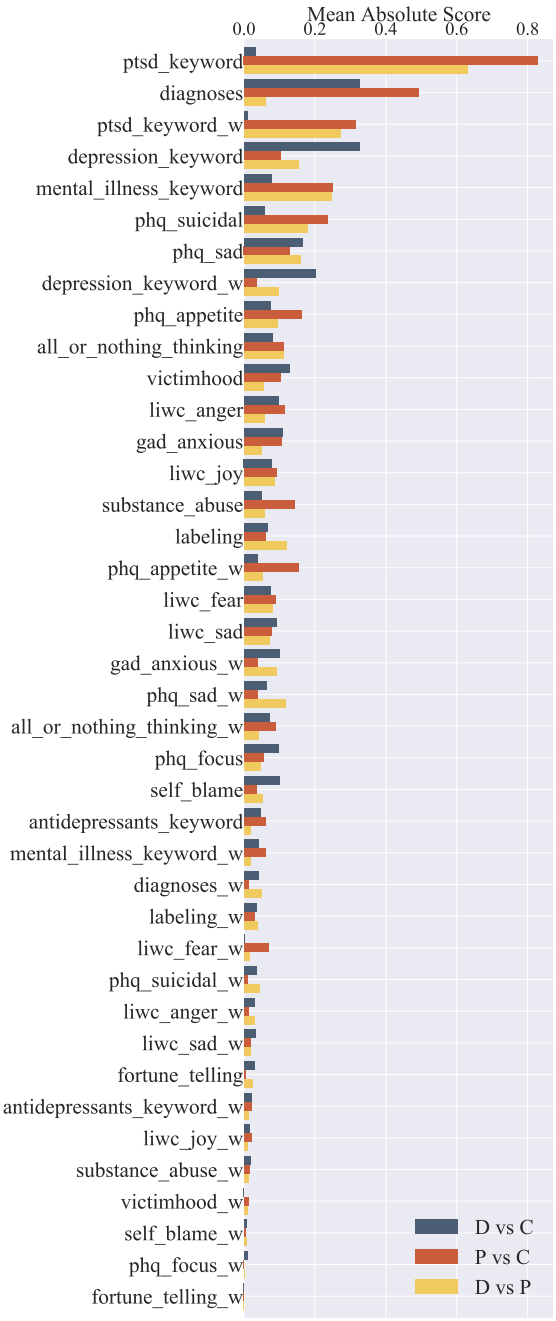

Figure 1: Global feature importance scores of the EBM classifiers trained on depression vs condition, PTSD vs condition, and depression vs PTSD. Features ending in "w" are features from the aggregator that looks for windows of hits.

Name	Description	Examples
All-or-Nothing Thinking	Seeing things in extreme, black-and-white categories. Thinking in absolutes such as "always", "never", or "every".	"I'm a *total* failure." "I *never* do *anything* right."
Overgeneralization	Seeing a single negative event as a never-ending pattern of defeat.	"She said no – I'm never going to get a date. I'll be lonely all my life." "I didn't get the job. I'll never find a job."
Labeling	Creating a completely negative self-image based on one's errors. Attaching a negative label to oneself.	"I'm an idiot!" "I'm a loser."
Fortune-Telling Error	Anticipating that things will turn out badly and feeling convinced that one's predictions are already-established facts.	"I'll make a fool of myself." "I'll never get better."
Disqualifying the Positive	Rejecting positive experiences by insisting they "don't count" for some reason or other.	(After a compliment) "They're just being nice." "That was a fluke."

Table 1: Definition and examples of common cognitive distortions according to Burns and Beck (1999)

PHQ-9 Questionnaire
1. Little interest or pleasure in doing things
2. Feeling down, depressed, or hopeless
3. Trouble falling or staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself or that you are a failure or have let yourself or your family down
7. Trouble concentrating on things, such as reading the newspaper or watching television
8. Moving or speaking so slowly that other people could have noticed.
Or the opposite being so figety or restless that you have been moving around a lot more than usual
9. Thoughts that you would be better off dead, or of hurting yourself

Table 2: PHQ-9 Questionnaire according to Kroenke et al. (2001)

References

David D Burns and Aaron T Beck. 1999. Feeling good: The new mood therapy.

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.