

Research and Applications

Using large language models to detect outcomes in qualitative studies of adolescent depression

Alison W. Xin, BS¹, Dylan M. Nielson, PhD¹, Karolin Rose Krause, PhD², Guilherme Fiorini, PhD³, Nick Midgley, PhD³, Francisco Pereira, PhD¹, Juan Antonio Lossio-Ventura, PhD^{1,*}

¹Machine Learning Core, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, United States, ²Centre of Research in Epidemiology and Statistics (CRESS UMR 1153), Université Paris Cité, Paris 75004, France, ³Department of Clinical, Educational and Health Psychology, University College, London WC1E 6BT, United Kingdom

*Corresponding author: Juan Antonio Lossio-Ventura, PhD, Machine Learning Core, National Institute of Mental Health, National Institutes of Health, 10 Center Dr, Suite 3D41, Bethesda, MD 20892, United States (juan.lossio@nih.gov)

Abstract

Objective: We aim to use large language models (LLMs) to detect mentions of nuanced psychotherapeutic outcomes and impacts than previously considered in transcripts of interviews with adolescent depression. Our clinical authors previously created a novel coding framework containing fine-grained therapy outcomes beyond the binary classification (eg, depression vs control) based on qualitative analysis embedded within a clinical study of depression. Moreover, we seek to demonstrate that embeddings from LLMs are informative enough to accurately label these experiences.

Materials and Methods: Data were drawn from interviews, where text segments were annotated with different outcome labels. Five different open-source LLMs were evaluated to classify outcomes from the coding framework. Classification experiments were carried out in the original interview transcripts. Furthermore, we repeated those experiments for versions of the data produced by breaking those segments into conversation turns, or keeping non-interviewer utterances (monologues).

Results: We used classification models to predict 31 outcomes and 8 derived labels, for 3 different text segmentations. Area under the ROC curve scores ranged between 0.6 and 0.9 for the original segmentation and 0.7 and 1.0 for the monologues and turns.

Discussion: LLM-based classification models could identify outcomes important to adolescents, such as friendships or academic and vocational functioning, in text transcripts of patient interviews. By using clinical data, we also aim to better generalize to clinical settings compared to studies based on public social media data.

Conclusion: Our results demonstrate that fine-grained therapy outcome coding in psychotherapeutic text is feasible, and can be used to support the quantification of important outcomes for downstream uses.

Key words: large language models, BERT, Llama 2, Llama 3, adolescent depression, depression outcomes, mental health.

Background

Globally, in adolescents aged 10–19 years, the prevalence of major depressive disorder and dysthymia is estimated at 8% and 4%, respectively.¹ Advancing understanding of treatment outcomes is critical in addressing this public health problem. In clinical trials and routine specialist care, around 40% of youth leave treatment without showing meaningful improvement in depressive symptoms, which include low mood, anhedonia, sleep disruption, suicidality, or irritability, defined by the DSM and ICD-11. Less is known about the impact of treatment on other outcomes, such as relationships or quality of life. Between 2007 and 2017, a systematic review of clinical studies on depression found that 94% of studies measured depressive symptoms, 52% measured general functioning, and less than 10% measured any other outcomes.²

Previously, clinical researchers performed a post-hoc analysis of interview transcript data from the qualitative study IMPACT-My Experience (IMPACT-ME³), a substudy nested

within the Improving Mood with Psychoanalytic and Cognitive Therapies (IMPACT) study of the psychological treatment of adolescent depression.^{4,5} Using qualitative content analysis, they produced a systematic and comprehensive framework of adolescent depression treatment outcomes, identifying 7 broad outcome domains and 29 specific outcomes of interest.⁶ Analysis of these outcomes in qualitative data could complement traditional quantitative measurement of symptom change, providing a more holistic impression of how treatment affects depression.

However, manual qualitative analyses of large volumes of qualitative data are time-intensive and may not always be feasible. Recent developments in natural language processing (NLP), particularly improvements in large language models (LLMs), can help address this challenge by automating the analysis of large volumes of data. A recent survey demonstrated that NLP enables automated screening for symptoms of several mental disorders from text data,⁷ though many of these studies only address a single binary classification

(eg, depressed and non-depressed) or regression (eg, severity). Additional limitations are the use of social media data, which complicates clinical integration, and lack of work focused on adolescents.

In this paper, we demonstrate the feasibility of using LLM embeddings as part of models for detecting fine-grained psychotherapeutic outcomes, as well as higher-level domain labels. We compare the performance of models operating on text embeddings produced with various LLMs used in mental health research, including the recently released Llama 3. We limit ourselves to open-source LLMs deployable within our own servers to eliminate concerns with protected health information or personally identifiable information.

Related work

Various NLP techniques have been used to detect mental health disorders by automating the analysis of large volumes of data. This analysis often entails screening text data for mentions of symptoms, as described in a recent survey of nearly 400 articles.⁷ Notably, social media posts are the predominant data source used (81%),⁸ followed by interviews (7%), EHRs (6%), screening surveys (4%), and narrative writing (2%).⁷ NLP techniques transform text into numerical representations, which may include specific linguistic features, language representation features, and others. NLP uses both traditional machine learning (ML) and deep learning-based methods for tasks related to depression, such as risk assessment, symptom detection, and more.

Deep learning-based methods have garnered significant attention due to their superior performance compared to traditional ML methods.⁷ In particular, LLMs have become foundational tools for transforming text inputs into quantitative vector representations known as embeddings. In contrast to traditional ML, embeddings are learned from data, using neural networks or other approaches, without requiring explicit human expertise to define them. These embeddings can then be used as inputs for classification models that predict annotations, such as the presence of specific depression markers. There are many embedding approaches, such as GloVe,⁹ word2vec,¹⁰ and transformer-based models like BERT¹¹ and RoBERTa,¹² and many have been used for identifying depression markers.^{13,14} Deep learning methods are generally categorized into convolutional neural network-based, recurrent neural network-based, and transformer-based approaches.^{7,15}

Transformer-based LLMs, including BERT, RoBERTa, Llama,^{16–18} Mistral,¹⁹ and the GPT* series,²⁰ incorporate attention mechanisms that manage long-range dependencies between segments of text. Traditional NLP methods require extensive feature engineering and considerable amounts of labeled data. In contrast, LLMs pre-trained on large datasets can use transfer learning to perform well on new tasks with minimal additional training. Given the limited annotated data for adolescent depression, this property is particularly useful for our work. Transformers can be fine-tuned for prediction and classification tasks in a variety of domains, including depression detection.^{21–26} Transformer-based models excel at text classification and sentiment analysis, which may translate to reliability and accuracy in identifying symptoms and outcomes. Using LLMs aligns our study with cutting-edge NLP developments, highlighting the relevance and impact of our findings on adolescent depression.

Many previous studies tackle broad binary classification problems (ie, depression and control group) or an existing set of clinical symptoms. However, the lack of interpretability in many models prevents clinicians from relying on the outcomes of automated screening techniques. The scientific community has initiated several efforts to improve the clinical applicability of ML studies, including the Early Risk Prediction on the Internet (eRisk) workshop, which has been part of the Conference Labs of the Evaluation Forum since 2017. In 2023, eRisk featured a depression-related task (Task 1)²⁷ that involved ranking sentences based on their relevance to each of the 21 symptoms of depression derived from the Beck Depression Inventory-II (BDI-II).²⁸ Symptoms included pessimism, thoughts about suicide, or sleep problems, rated on a severity scale from 0 to 3. Outside of eRisk, other studies aggregate symptoms from different questionnaires (eg, BDI-II^{29,30} and PHQ-9³¹) and transformer-based models (eg, BERT) to screen for depression in patients.³² These initiatives mainly rely on social media data, which limits clinical applicability due to differences with content targeted or elicited in therapeutic settings. Additionally, limited research focuses on adolescent participants, whose data can be different in aspects ranging from vocabulary to the specific symptoms and problems mentioned.

Our work differs from previous research on depression in several key ways. First, it focuses on detecting fine-grained symptoms, outcomes, and impacts of depression, implementing a framework that aims to capture more nuance than diagnosis or clinical symptoms. Furthermore, the labels in the study are defined to be particularly relevant to adolescents. Finally, we use a dataset from a psychiatric study, rather than social media. Additionally, given the sensitivity of the dataset, our approach was developed using the latest open-source LLMs, such as Llama, rather than commercial ones such as GPT or Claude. This measure provides both security and affordability to the research community.

Materials and methods

Data

IMPACT-ME interviews

Interviews were taken from IMPACT-ME,⁶ a qualitative study within the IMPACT trial.⁴ IMPACT examined the efficacy of Brief Psychosocial Intervention (BPI), Cognitive Behavioral Therapy (CBT), and Short-Term Psychoanalytic Psychotherapy (STPP) for adolescents aged 11–17 diagnosed with unipolar Major Depressive Disorder. In IMPACT-ME, research psychologists conducted semi-structured interviews with patients, parents, and therapists at treatment start, end, and 1-year follow-up, exploring therapy experiences and observed changes.⁶

Qualitative analysis and annotation

Krause et al. conducted a secondary qualitative analysis on these interviews to explore the range of treatment outcomes relevant to patients. Only interviews from the end of treatment were analyzed. These interviewers were transcribed verbatim and included pauses, filler words, interruptions, and typos. Participants were excluded if any of the interviews from patient, parent, or therapist were missing, if treatment ended within the first 3 sessions, or if they were referred to inpatient care. Of the remaining 34 cases (9 BPI, 9 CBT, and 16 STPP participants; 102 interviews), the average age was

16.2 years ($s = 1.5$, range=12-19), and 21 (61%) were female. To categorize outcomes, Krause et al. first designed an a priori coding framework based on existing taxonomies of treatment outcomes. During annotation, outcome-relevant passages were extracted, and the coding framework was further modified to incorporate new themes. The final framework contained 29 specific outcome categories within 7 high-level domains,⁶ listed and described in Table 1. All annotations were performed by 1 researcher.

Ethical considerations

The original study protocols for the IMPACT trial and the IMPACT-ME study were approved by Cambridgeshire 2 Research Ethics Committee, Addenbrooke's Hospital, Cambridge, UK (REC Ref: 09/H0308/137) and were performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All participants above the age of 16 provided informed written consent. Parental consent and youth assent were obtained for younger adolescents.

Dataset splitting

We split the dataset of 34 subject cases into a training set of 26 subjects and a test (holdout) set of 8 subjects. Transcripts were grouped by subject (ie, by triplets of interviews relating to each adolescent participant) to ensure that models were not trained and tested on data referring to the same adolescent. We conducted 4-fold cross-validation (CV) on the training set of 26 cases. In each fold, we excluded all text from several cases and trained the model on the remaining cases. The test set, consisting of 8 cases, was manually balanced to include both positive and negative examples for all specific outcomes. The test set was not used in this paper and is reserved for future evaluation. Each case generated multiple text blocks, ranging from approximately 10 000 to 25 000 depending on the text segmentation used (Section "Transcript Segmentations"). A text block may be positive for multiple labels. When all positive labels were combined, there were 1543, 732, and 840 positive examples in the Original, Monologue, and Turns segmentations, respectively, with around 10 times as many negative examples. The numbers of positive examples across all labels are provided in [Supplementary Material, Table S1](#).

Preprocessing

Conversion to labeled text blocks

Before analysis, empty lines and header information, such as subject ID and interviewer ID, were removed from the transcription files. We then divided the transcripts into chunks of text called "speaker blocks." Each speaker block represents a portion of the text spoken by 1 person. These blocks were indicated by the start of a new paragraph in the transcripts. The original annotations by Krause et al. (from the IMPACT-ME study) were made by highlighting specific parts of the transcript that were relevant to a particular outcome. These highlighted excerpts could begin or end anywhere within a speaker block, meaning that a single block could contain multiple highlighted excerpts related to different outcomes. For our analysis, we labeled an entire speaker block as "positive" if any part of it contained text that was flagged as relevant (positive) to an outcome. This means that even if only a small portion of the speaker block was relevant, the whole block was considered positive for that outcome.

Transcript segmentations

Original

The initial segmentation of text generated from annotations, containing 32 520 blocks, included various uninformative text segments. Outcome-relevant dialogue would often be interspersed with interjections, acknowledgments, or requests for elaboration, for example, an interviewer saying "okay" or "yes" to encourage a patient would be included within the excerpt and labeled as positive in our dataset. To address these uninformative text blocks, we created 2 additional segmentations of the transcript, Monologue and Turns, described below.

Monologue

We discarded all interviewer speech and blocks with 12 or fewer characters. We manually determined the cutoff by examination of the labeled text in the training set. By only retaining non-trivial interviewee text, we aimed to produce "monologues" about the study experience, although some interviews, such as those conducted jointly with both parents of a patient, retained multiple interviewees interacting in dialogue. Of the original 32 520 blocks, 12 941 were retained in this filtration.

Turns

We partitioned blocks at each interview utterance, grouping together sequential pairs of utterances by interviewer and interviewee into "turns" of the conversation. By concatenating blocks, the Turns segmentation kept informative interviewer questions together with short interviewee responses that were otherwise uninformative (eg, "I: How has your mood been?" "P: Fine..."). For interviews with multiple interviewees, all utterances between interviewer utterances were concatenated into the same turn. This process produced 16 139 blocks of text. Table 2 illustrates how segmentations might be created from a passage. However, this example lacks many of the interview transcripts' idiosyncrasies, such as inclusion of hesitations and filler words.

The training set contained 25 852 blocks in the Original, 10 008 blocks in Monologue, and 12 814 blocks in Turns. Full counts of the number of positive examples for each label in the complete and training set can be found in [Supplementary Material, Table S1](#).

Methods

Our approach consisted of 2 stages. First, interview transcripts from IMPACT-ME were converted into quantitative embedding representations using LLMs. Next, logistic regression models were trained to predict the text labels. The 2 subsections describe each stage in more detail.

LLM embeddings

In the first stage, text from each block was passed to transformer-based LLMs to be converted into *embeddings*. Text was converted into a sequence of *tokens*—words or word fragments—using model-specific *tokenizers*. Sequences of tokens were passed to the LLMs, and the embedding was extracted from the last hidden layer using the PyTorch implementation of the Hugging Face Transformers python library.^{33,34} The final layer is a tensor with dimension $b \times t \times d$, where b is the batch size, t is the number of tokens, and d is the hidden dimension. Here, $b = 1$ because we pass each block alone, t depends on input sequence length and the

Table 1. Names of the labels and a brief description.^a

Abbrev.	Name	Description
Any		$A \cup \dots \cup G$
A	Symptom change	$A_1 \cup \dots \cup A_8$
A1	Mood and affect	Less low and depressed; low mood is more fleeting, less overwhelming.
A2	Anger and aggression	Less angry, irritable, aggressive; fewer outbursts; better able to manage temper.
A3	Appetite	Healthier appetite and weight.
A4	Sleeping and energy	Healthier sleep patterns and energy levels.
A5	Self-harm	Less self-harm (eg, cutting, trichotillomania)
A6	Suicidality	Reduced suicidal ideation and behavior
A7	Anxiety	Fewer fears, worries, panic attacks; less social anxiety; engaging in activities
A8	Other comorbidities	For example, substance abuse or obsessive-compulsive symptoms
B	Coping and self-management	$B_1 \cup B_2 \cup B_3$
B1	Behavioral activation	More active; returning to hobbies or engaging in new activities; sense of purpose, routine, and structure
B2	Coping and resilience	Specific coping strategies, understanding of feelings, thoughts, and behaviors; anticipating and managing challenges; more resilient, greater self-efficacy, sense of control
B3	Cognition and behavior	Challenges negative automatic thoughts, more flexible thinking styles
C	Functioning	$C_1 \cup C_2 \cup C_3 \cup C_4$
C1	Global functioning	Better function across range of life domains, engages in typical adolescent activities
C2	Executive functioning	Able to get things done; improved concentration, motivation, planning, organization
C3	Academic and vocational functioning	Attends school more regularly; works more effectively in school, achieves better results
C4	Social functioning	More outgoing and talkative, more present within friendship groups, more socially connected; easier to make conversation, relate to others, be mindful of others' feelings
D	Personal growth	$D_1 \cup \dots \cup D_6$
D1	Assertiveness	Better able to stand up for needs and opinions, overcome urge to please, can express disagreement or disapproval when appropriate
D2	Autonomy and responsibility	More independent, takes responsibility for life and actions
D3	Identity	Finding out who they are and how to be themselves around other people; less idealized self-images that can accommodate both positive and challenging personality traits; positive and negative feelings
D4	Processing past and present	Making sense of challenging past or ongoing experiences such as bereavement, parental divorce, or family conflict
D5	Confidence and self-esteem	More confident, less insecure, less vulnerable to judgment, higher self-regard
D6	Feeling seen and seeing differently	Feeling listened to, understood, or cared for; experiences of being worth of another's attention; new perspectives; opportunity to release feelings, thoughts or memories
E	Relationships	$E_1 \cup \dots \cup E_5$
E1	Ability to talk	More able to talk about feelings and thoughts, which helps deepen relationships; stronger support network facilitates opening up
E2	Family functioning and relationships	Getting on better with their family: less conflict, better understanding from family; easing of entrenched tensions between family members; families communicate more openly; role within the family system clarified
E3	Friendships	Reactivation or deepening of existing friendships, expanding friendship groups or changing friends by turning toward more supportive friendships
E4	Peer relationships	Getting on better with peers in school
E5	Romantic relationships	Getting on better with romantic partner
F	Wellbeing	$F_1 \cup F_2 \cup F_3$
F1	Peace of mind	Calmer, more balanced, relaxed, and carefree; feeling as if a weight had been lifted off their shoulders; more accepting of things that cannot be changed
F2	Optimism	More positive and optimistic outlook into their lives and the future
F3	Future orientation	Can make plans for the future and have goals
G	Parental support and wellbeing	$G_1 \cup G_2$
G1	Parental support	Parents are better able to understand their child's difficulties and more aware of how their parenting practices may contribute to these difficulties; parents learn to support and parent their child more effectively
G2	Parental wellbeing	Parents feel less guilty, isolated, stressed, and worried; parents feel reassured, supported, and able to express their own frustrations and issues

^a Unless otherwise indicated, assume descriptions refer to changes with the adolescent patient.

tokenizer, and d varies by model, described in Table 3. Averaging across tokens produces a final d -dimensional embedding vector representing each text block. Aside from the source corpus, the main differences between models are

(1) *maximum sequence length*, or the number of tokens a model can effectively represent before truncation, (2) the *hidden dimension*, or dimensionality of the internal embedding vectors, and (3) the number of parameters,

Table 2. A comparison between how blocks would be formed between the Original, Monologue, and Turns segmentation.^a

Original	Monologue	Turns
I: How are you?		I: How are you?
P: Ok...		P: Ok...
I: Just ok?		I: Just ok?
P: Not feeling the best about school.	P: Not feeling the best about school.	P: Not feeling the best about school.
I: Why?		I: Why?
P: I hate this group project.	P: I hate this group project.	P: I hate this group project.
I: Mhmm.		I: Mhmm.
P: They always ignore me so	P: They always ignore me so	P: They always ignore me so
I have to work alone.	I have to work alone.	I have to work alone.

^a A change in text color indicates the boundary of the input block. The example text is not based on any interview in the dataset.

Table 3. Maximum sequence length (max. seq.), hidden dimension size (hidden dim.), and millions of parameters (params) for the LLMs used to generate embeddings.

Model	Max. seq.	Hidden dim.	Params (10 ⁶)
BERT	512	768	110
MentalBERT	512	768	110
MentalLongformer	4,096	768	102
Llama 2-7B	4,096	4,096	7,000
Llama 3-8B	8,192	4,096	8,000

Abbreviations: BERT = Bidirectional Encoder Representations from Transformers; LLM = large language model.

roughly proportional to the number of internal transformations before reaching the final embedding vector. We describe the LLMs we used below, with details summarized in Table 3.

Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT)³⁵ is widely employed in various NLP tasks, including text classification and named-entity recognition. By conditioning on both the left and right context of each token, BERT can generate high-quality embeddings that encode rich semantic information. We used the base pre-trained uncased (case-insensitive) BERT model provided by the original authors, which contains 12 layers, 768 hidden dimensions, and 110 million parameters.

MentalBERT

MentalBERT is a domain-specific model initialized from a general BERT model and then further pretrained on text relevant to mental health.³⁶ The pretraining dataset includes a range of subreddits within the mental health domain, such as “r/depression,” “r/SuicideWatch,” “r/Anxiety,” “r/off-mychest,” “r/bipolar,” “r/mentalillness,” and “r/mentalhealth.” Evaluation indicates that MentalBERT outperforms BERT and ClinicalBERT (pretrained on PubMed data) in classifying mental health conditions, including depression, stress, and anorexia.³⁶ We used the base uncased MentalBERT model, which uses the same architecture as BERT base uncased.

MentalLongformer

MentalLongformer is a domain-specific model based on Longformer.³⁷ Longformer is a modified transformer architecture designed to effectively handle long text sequences. Its modified self-attention mechanism scales linearly rather than

quadratically with sequence length, combining local windowed attention with task-motivated global attention. Upon its release, it outperformed previous models in autoregressive language modeling tasks involving long sequences.³⁸ MentalLongformer was pretrained using the same mental health dataset as MentalBERT, described above.³⁷

Llama 2

Large Language Model Meta AI (LLaMA, later Llama) is a series of autoregressive LLMs developed by Meta AI.^{16,17} Llama models are state-of-the-art open-source LLMs for general applications. Llama 2¹⁷ inherits its pre-training configurations and model structure from Llama 1.¹⁶ Llama 2 incorporates RMSNorm for pre-normalization, uses the SwiGLU activation function, and integrates rotated position embeddings. Llama 2 diverges from LLaMA 1 by extending the context length from 2048 to 4096 and introducing Grouped Query Attention (GQA). Llama 2 weights are available only upon request to Meta AI, unlike the 3 publicly available models above. We generated the embeddings using Llama 2-7B, the smallest base model in the series.

Llama 3

Compared to Llama 2, Llama 3 uses a tokenizer with a vocabulary of 128K tokens, enhancing language encoding efficiency and resulting in improved model performance.¹⁸ To increase the inference efficiency of Llama 3 models, GQA was implemented in both the 8B and 70B versions. The models were trained on sequences of 8192 tokens, extending the context length from 4096 in Llama 2. A mask was used to ensure self-attention remains within document boundaries. We used the Llama 3-8B base model to generate the embeddings.

Training classification models

To classify labels, we trained logistic regression models on the d -dimensional averaged embedding vector for each passage. Logistic regression models were trained with Python’s scikit-learn using L2 penalty, balanced class weighting, and the Limited-memory Broyden–Fletcher–Goldfarb–Shanno solver.³⁹ Models were trained and evaluated within a 4-fold CV loop. For each of the 4 test folds, the C hyperparameter of logistic regression was tuned with inner 3-fold CV, using the same fold partitions as the outer 4-fold CV. We searched 16 possible values of C: (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0, 50.0, 100.0, 500.0, 1000.0, 5000.0). Data were grouped by subject ID and stratified by label. Models for labels A8, E4, and E5 could not be trained because fewer than 4 subjects were present in the training

averaged F1 scores are described in [Supplementary Material, Table S3](#).

Classification performance for each label

In the Original segmentation, D1 performs the best across all models. In Monologue, the best performer was one of D3 or F2. In Turns, A3 and D3 performed well for all models, but the top performer for MentalLongformer was F2. For any combination of model and segmentation, the lowest performer tended to be D2 or G1, with A2 performing poorly in Original. Many labels were inconsistent across models and segmentations. For example, A5 was in the top 4 for all models in the Original segmentation, but underperformed in Monologue and Turns in non-Llama models. The worst classified labels tended to have high variance in performance across embedding models, though the relative rankings are consistent. For every embedding, the averaged ROC AUC for models of “Any” outcome were between 0.75 and 0.85. Further details on relative classification performances can be found in [Supplementary Material, Table S2](#), and macro-

Our second goal was to determine whether embeddings from a particular LLM had consistently better performance than others. For our 28 specific outcomes, we tested the null hypothesis of no difference in model performance with the Friedman test.⁴¹ We excluded aggregate labels, that is, the 7 domain labels and the “Any outcome” label, to avoid double-counting. Friedman test results were $Q_3 = 8.571, p = 0.0356$ for Original; $Q_3 = 13.16, p = 0.00431$ for Monologue; and $Q_3 = 12.56, p = 0.00570$ for Turns. At significance level $\alpha = 0.05$, the Friedman test results supported rejection of the null hypothesis of model equivalence, and we proceeded with the post hoc Bayesian comparison tests.^{42,43}

The Bayesian post hoc test indicated that both Llama models had probability ≥ 0.94 of outperforming any other non-Llama model (Figure 2). BERT had a < 0.04 probability of outperforming any model except for MentalLongformer, where the probability of BERT being better was 0.14, 0.15, and 0.08 for Original, Monologue, and Turns, respectively. Llama 2-7B and Llama 3-8B had a 0.84, 0.71, and 0.92 probability of practical equivalence for Original, Monologue, and Turns. MentalLongformer and MentalBERT as well as BERT and MentalBERT had practical equivalence probabilities of 0.19-0.38 in Original and Turns. All other pairwise model comparisons returned ≤ 0.06 probability of practical equivalence.

We identified 3 primary reasons for misclassifications and provide examples in [Table 5](#). These cases do not partition all errors, as some text blocks may be misclassified for multiple reasons. First, misclassifications occurred with short text blocks. Short sentences often lack the contextual richness

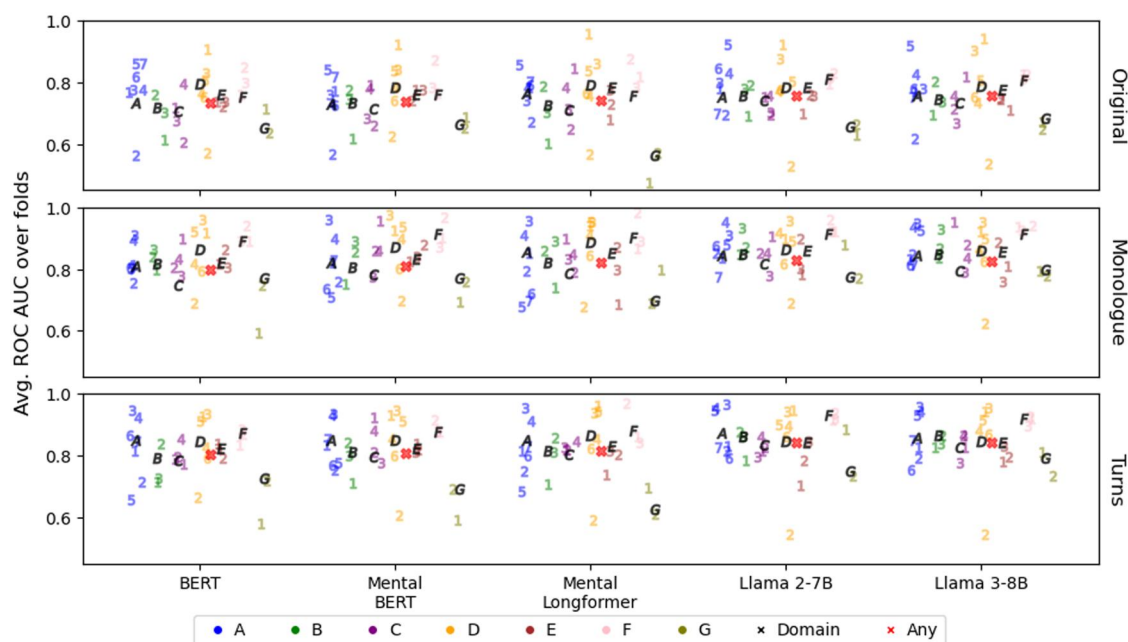


Figure 1. Average ROC AUC performance for logistic regression models. Labels are grouped vertically by domain (and color) and horizontally by segmentation (right axis). Numbers indicate the performance of each specific outcome within a domain, black letters the domains, and the red X represents “Any” of the outcomes.

Table 4. Averaged ROC AUC across outer *k*-fold cross-validation.

Sgmnt.	Original					Monologue					Turns				
	BERT	MBERT	MLong	L2-7B	L3-8B	BERT	MBERT	MLong	L2-7B	L3-8B	BERT	MBERT	MLong	L2-7B	L3-8B
Label															
Any	0.735	0.737	0.742	0.758	0.757	0.799	0.811	0.822	0.829	0.826	0.806	0.809	0.817	0.844	0.843
A	0.732	0.730	0.766	0.753	0.755	0.811	0.822	0.856	0.844	0.846	0.850	0.852	0.851	0.876	0.858
A1	0.767	0.772	0.782	0.780	0.773	0.815	0.820	0.851	0.836	0.831	0.815	0.834	0.815	0.832	0.830
A2	0.562	0.566	0.671	0.695	0.618	0.757	0.758	0.794	0.851	0.849	0.715	0.754	0.750	0.815	0.792
A3	0.777	0.760	0.741	0.798	0.779	0.910	0.962	0.956	0.909	0.950	0.947	0.937	0.955	0.968	0.961
A4	0.777	0.731	0.785	0.831	0.827	0.895	0.897	0.908	0.954	0.937	0.923	0.932	0.914	0.955	0.945
A5	0.861	0.843	0.856	0.923	0.919	0.801	0.707	0.678	0.880	0.925	0.657	0.780	0.684	0.949	0.932
A6	0.818	0.725	0.791	0.845	0.780	0.806	0.736	0.721	0.873	0.812	0.866	0.771	0.800	0.794	0.755
A7	0.861	0.819	0.802	0.699	0.752	0.804	0.829	0.696	0.773	0.812	0.846	0.854	0.831	0.827	0.851
B	0.717	0.734	0.726	0.757	0.746	0.816	0.808	0.819	0.850	0.846	0.795	0.811	0.816	0.864	0.866
B1	0.615	0.619	0.604	0.690	0.696	0.798	0.752	0.739	0.846	0.836	0.715	0.711	0.710	0.786	0.829
B2	0.762	0.777	0.786	0.790	0.802	0.840	0.854	0.861	0.873	0.869	0.839	0.844	0.860	0.879	0.868
B3	0.702	0.746	0.703	0.780	0.738	0.863	0.890	0.890	0.933	0.929	0.727	0.800	0.814	0.864	0.842
C	0.708	0.714	0.710	0.739	0.733	0.749	0.782	0.787	0.816	0.793	0.784	0.798	0.804	0.834	0.828
C1	0.720	0.791	0.845	0.751	0.817	0.899	0.956	0.898	0.907	0.953	0.772	0.926	0.819	0.815	0.830
C2	0.607	0.659	0.647	0.697	0.716	0.805	0.858	0.791	0.850	0.881	0.796	0.814	0.822	0.829	0.865
C3	0.675	0.683	0.715	0.704	0.669	0.777	0.776	0.833	0.778	0.790	0.788	0.779	0.826	0.799	0.776
C4	0.795	0.781	0.785	0.758	0.755	0.832	0.861	0.850	0.852	0.837	0.853	0.881	0.846	0.863	0.868
D	0.796	0.785	0.790	0.782	0.787	0.863	0.870	0.888	0.864	0.859	0.848	0.850	0.865	0.846	0.847
D1	0.907	0.922	0.956	0.923	0.942	0.918	0.925	0.942	0.900	0.922	0.928	0.933	0.962	0.949	0.944
D2	0.573	0.624	0.569	0.528	0.537	0.690	0.696	0.679	0.689	0.626	0.666	0.608	0.591	0.547	0.546
D3	0.830	0.841	0.861	0.876	0.903	0.962	0.974	0.953	0.957	0.954	0.937	0.946	0.939	0.943	0.955
D4	0.765	0.785	0.758	0.771	0.733	0.818	0.898	0.909	0.919	0.865	0.829	0.855	0.854	0.896	0.882
D5	0.807	0.838	0.839	0.802	0.805	0.921	0.936	0.951	0.892	0.900	0.911	0.915	0.943	0.900	0.917
D6	0.752	0.742	0.766	0.775	0.752	0.793	0.804	0.846	0.819	0.826	0.794	0.801	0.822	0.864	0.869
E	0.761	0.762	0.782	0.784	0.776	0.820	0.832	0.851	0.859	0.855	0.826	0.823	0.833	0.845	0.843
E1	0.737	0.775	0.680	0.699	0.709	0.816	0.824	0.687	0.782	0.811	0.841	0.840	0.739	0.705	0.782
E2	0.721	0.740	0.731	0.760	0.747	0.863	0.879	0.871	0.900	0.887	0.792	0.820	0.804	0.785	0.797
E3	0.737	0.776	0.776	0.757	0.750	0.805	0.842	0.798	0.802	0.758	0.825	0.814	0.815	0.846	0.834
F	0.754	0.761	0.758	0.812	0.806	0.892	0.915	0.901	0.914	0.914	0.876	0.880	0.881	0.933	0.920
F1	0.759	0.762	0.819	0.795	0.825	0.890	0.901	0.868	0.922	0.937	0.835	0.879	0.862	0.918	0.922
F2	0.848	0.871	0.877	0.830	0.828	0.940	0.967	0.985	0.961	0.941	0.869	0.916	0.972	0.940	0.930
F3	0.798	0.784	0.785	0.816	0.810	0.890	0.870	0.892	0.923	0.924	0.888	0.877	0.843	0.921	0.899
G	0.653	0.665	0.565	0.658	0.683	0.772	0.769	0.697	0.775	0.799	0.727	0.692	0.627	0.752	0.794
G1	0.715	0.687	0.476	0.629	0.672	0.595	0.695	0.797	0.879	0.792	0.580	0.592	0.695	0.885	0.814
G2	0.636	0.653	0.571	0.663	0.648	0.747	0.760	0.691	0.772	0.780	0.719	0.693	0.611	0.736	0.736

Abbreviations: L2-7B = Llama 2-7B; L3-8B: Llama 3-8B; MBERT = MentalBERT; MLong = MentalLongformer; Sgmnt. = segmentation.

necessary for accurate classification, leading to incorrect labeling. Second, misclassifications appeared when there were no mentions of the symptom within the text block. This issue typically resulted from segmentation errors, where relevant information was contained in a previous conversation block but was not carried forward to the current segment. Finally, another source of error was the use of subtle, vague, or indirect language. For example, people may discuss depressive experiences without referring to the patient's own experience, making it challenging for the model to correctly identify and classify the content. Additionally, there are label-specific errors. We found that labels G1: Parental Support and G2: Parental Well-being consistently performed poorly across models and segmentations (Supplementary Material, Table S2). This error may arise due to data processing. When creating embeddings, we concatenate all interviewers and do not incorporate additional information about the subject ID or the interview type (adolescent patient, parent, or therapist). This measure avoids data leakage but also impedes the model from disentangling when parents are referring to themselves or being referred to.

Discussion and conclusion

The current objective for this research project is to develop robust models that can detect fine-grained symptoms and outcomes in interviews or autobiographical text. The ultimate goal, however, is to apply the models to text from other studies and estimate how much symptoms and outcomes not captured in standard questionnaires drive the impact of depression, as well as capturing changes as a result of therapy that may go beyond traditional symptom-based outcome measures. The results suggest that detection is feasible across a wide range of outcomes and at a level of performance that would make it reasonable to use model predictions as a complementary measure to standardized questionnaires. To make this possible, we plan to release the trained models and make them available to the wider research community. This would allow validation over different patient populations and for more clinical researchers to provide input on the quality and robustness of the models across outcomes.

We find that LLM embeddings allow for effective classification of outcomes, even for labels with few positive examples, and could be useful for future work on understanding

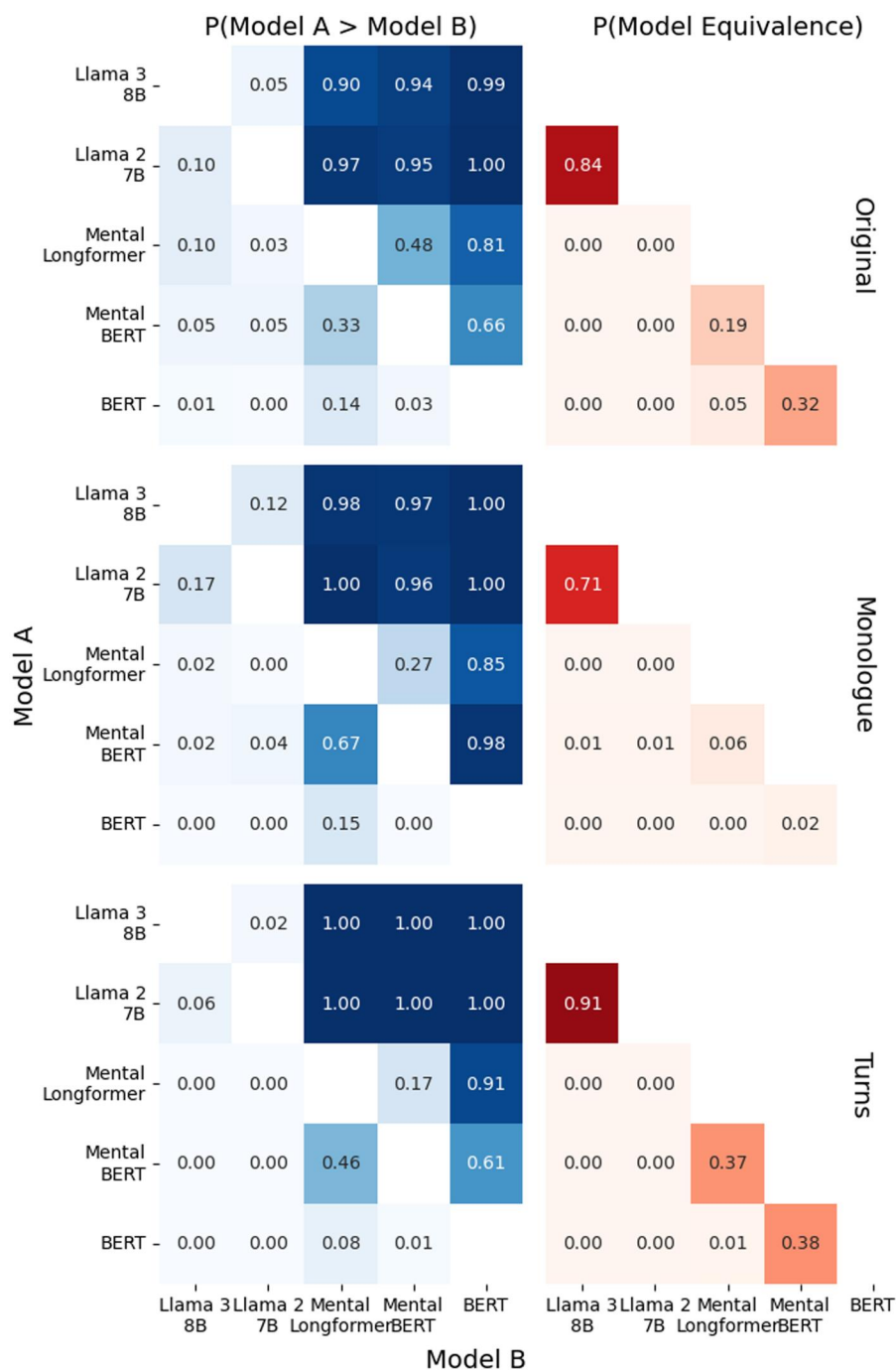


Figure 2. Bayesian model comparison test, with a region of practical equivalence of 0.01. Results for each segmentation are grouped by row. (Left, blue) Probability that model A (y-axis) outperforms model B (x-axis). (Right, red) Probability of model A and model B being practically equivalent.

the holistic experience of depression and its treatment. Across the 36 labels considered, performance was never below an average ROC AUC of 0.60 for every model within a segmentation, and generally much higher than that. Given these results, we believe that classifiers using LLM embeddings as inputs could prove useful for detecting fine-grained outcomes. On the other hand, although we propose a few broad reasons for errors in labeling, it is still unclear why specific labels were easier or harder to classify. Even within the same domain, specific outcomes can run a wide range, for example, D3: Identity being best overall while D2: Autonomy is worst overall. Although we can identify patterns in these labels,

such as D3 appearing in the more consistent terminology of therapists vs D2 appearing in more diverse or indirect terms from parents, we have not developed robust clinical interpretations of these results that are consistent across all label differences. The relative performances of aggregated labels, such as labels for high-level domains A through G, tend to be consistent across models within a segmentation, suggesting that some variability may be due to the small number of positive examples.

The Bayesian comparison test between models suggests that, of the models investigated, Llama models produce more informative embeddings for classification. Llama 2-7B and

Table 5. Examples of misclassified text blocks, separated by 3 observed reasons and 2 types of errors.

Errors	Reason	Label	Transcript
FP	Short	A1: Mood and affect	I: Yeah. P: So yeah so.
FN	Short	E1: Ability to talk	I: Right... P: But we haven't got that anymore...
FP	No mention	B2: Coping and resilience	I: Okay... so is that sort of following your sort of feeling quite emotional... P: Yeah...
FN	No mention	F2: Optimism	I: yeah... in what way... P: I've sort of got like a different... mindset and... I've looked back on it and just thought a lot I guess...
FP	Subtle mention	A7: Anxiety	I: for the first one... P: just sort of like erm... just mainly to do with the feelings and emotions so like... you know say the disappointment of um... not... learning in GCSEs and then going for the exam mark could affect me in a certain way, or the fact that I wouldn't be able to do sports bothered me in a certain way...
FN	Subtle mention	F3: Future orientation	I: oh great... P: she can see a result that she'll be a qualified something at the end and then she can earn money...

FP = false positives; FN = false negatives.

Llama 3-8B have a high probability of practical equivalence, which is expected considering their architectural similarities. The non-negligible probability of practical equivalence for MentalBERT and MentalLongformer is also unsurprising, considering the models are pre-trained with the same set of mental health data. MentalBERT, though fine-tuned on domain-specific data, only has a high probability of outperforming BERT on Monologue (0.98) and has a moderate probability of practical equivalence on Original and Turns (0.32, 0.38). Additionally, although the Bayesian comparison test suggests that Llama models outperform the other models tested, the advantage is not very large. Other concerns, such as resource usage, may be a deciding factor in choosing an embedding model for different tasks. Llama 2-7B and Llama 3-8B, for example, require a GPU for inference, while BERT, MentalBERT, and MentalLongformer can run on a few CPU cores.

The k -fold CV results provide reasonable estimates of model performance in new participants, given that we have reported on all the experiments that we have carried out. Nevertheless, the final analysis of variance and generalization of the models should be performed on the test set, which we are currently withholding to allow for further model development on the training set. Testing on the holdout will produce performance estimates for sparse labels that we could not model during the k -fold CV step. The dataset used in this work is unique because of the fine granularity of the coding framework and the effort required for annotation. Thus, we do not make claims on generalization of labeling performance beyond this population. Instead, we present this work as a demonstration that combining state-of-the-art text embeddings and logistic regression is a robust approach that can effectively handle fine-grained labels.

A limitation of our work is that results may be biased due to the influence of the annotator or the LLM. As previously mentioned, annotation was performed by a single researcher, who also led the research for developing the coding framework. An additional rater would be able to provide additional perspective or verify that the coding guidelines lead to reproducible results. However, this work is primarily concerned with developing label classifiers rather than testing the reliability of the coding framework, and thus steps to reduce rater bias were not in scope. Future work with this framework may include multiple annotators to reduce the influence of any single researcher's perspective. In addition to bias from human labeling, bias may also be introduced by LLMs.

Because LLMs are usually trained on written text, they may not be effective at embedding transcribed conversational speech. LLMs may fail to accurately encode relevant semantic information or misinterpret vernacular language, leading to incorrect or inconsistent labeling. It is also unclear how reported performance reflects idiosyncrasies of this dataset. The small sample size, both in total positive labels and number of subjects, may impede applying these particular models to other datasets. Additionally, verbatim transcription of dialogue is not common in written text or in machine transcription, which often exclude pauses, filler words, and accent indicators.

Given the vulnerability of adolescents with depression, or mental health disorders in general, special care is warranted in future work on this topic. We do not recommend use of third-party language models with this or any similar dataset, unless there are contractual guarantees satisfying requirements about the handling of personally identifying information or private health information. Additionally, special care should be taken to ensure that any model concerned with classification or identification of sensitive health information does not produce biased or discriminatory results. This requires explicit effort to ensure equitable data collection, both so that training sets do not introduce bias, and so that results can be rigorously evaluated for potentially harmful outcomes before being broadly applied.

Future work will aim to improve the extraction of information from text and increase labeling accuracy. Text representation can be improved through, for example, embeddings from more advanced models or by fine-tuning. We also plan to make models more generalizable through supplementing the training data with additional synthetic examples paraphrased by LLMs from existing annotations. Additionally, while current work focuses on labeling using numeric model embeddings, labels could also be generated using text-based instruction models, such as Llama 3 Instruct. Appropriate prompt engineering may allow models to better attend to relevant features in the transcripts and produce accurate labels. However, prompt engineering zero-shot or few-shot responses requires additional data engineering (eg, through generating artificial examples) to prevent data leakage. Furthermore, calculating ROC AUCs comparable to our current models requires the production of estimates of LLM prediction confidence, which require in turn extensive development and validation efforts. Therefore, this line of research is outside the scope of our current work.

- eds, CEUR Workshop Proceedings. Vol 2454. CEUR-WS.org; 2019:331-342.
15. Squires M, Tao X, Elangovan S, et al. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Inform.* 2023;10:10.
 16. Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models. arXiv:230213971. 2023, preprint: not peer reviewed. <https://doi.org/10.48550/arXiv.2302.13971>
 17. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. arXiv:230709288. 2023, preprint: not peer reviewed. <https://doi.org/10.48550/arXiv.2307.09288>
 18. Meta A. Llama 3 model card. 2024. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
 19. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. arXiv:231006825. 2023, preprint: not peer reviewed. <https://doi.org/10.48550/arXiv.2310.06825>
 20. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst.* 2020;33:1877-1901.
 21. Jiang Z, Levitan SI, Zomick J, Hirschberg J. Detection of mental health from Reddit via deep contextualized representations. In: Holderness E, Jimeno Yepes A, Lavelli A, Minard AL, Pustejovsky J, Rinaldi F, eds. *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics; 2020:147-156. <https://aclanthology.org/2020.louhi-1.16>
 22. Malviya K, Roy B, Saritha S. A transformers approach to detect depression in social media. In: *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE; 2021:718-723.
 23. Heston TF. Safety of large language models in addressing depression. *Cureus.* 2023;15:e50729.
 24. Aragon M, Parapar J, Losada DE. Delving into the depths: evaluating depression severity through BDI-biased summaries. In: Yates A, Desmet B, Prud'hommeaux E, et al., eds. *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*. Association for Computational Linguistics; 2024:12-22. <https://aclanthology.org/2024.clpsych-1.2>
 25. Wang Y, Inkpen D, Kirinde Gamaarachchige P. Explainable depression detection using large language models on social media data. In: Yates A, Desmet B, Prud'hommeaux E, et al., eds. *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*. Association for Computational Linguistics; 2024:108-126. <https://aclanthology.org/2024.clpsych-1.8>
 26. Xu X, Yao B, Dong Y, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. *Proc ACM Interact Mob Wearable Ubiquitous Technol.* 2024;8:1-32. <https://doi.org/10.1145/3643540>
 27. Parapar J, Martín-Rodilla P, Losada DE, Crestani F. Overview of eRisk 2023: early risk prediction on the internet. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*. Springer-Verlag; 2023:294-315. https://doi.org/10.1007/978-3-031-42448-9_22
 28. Dozois DJ, Dobson KS, Ahnberg JL. A psychometric evaluation of the Beck Depression Inventory-II. *Psychol Assess.* 1998;10:83-89.
 29. Zhang Z, Chen S, Wu M, Zhu KQ. Psychiatric scale guided risky post screening for early detection of depression. In: Raedt LD, ed. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, AI for Good*. International Joint Conferences on Artificial Intelligence Organization; 2022:5220-5226. <https://doi.org/10.24963/ijcai.2022/725>
 30. Pérez A, Warikoo N, Wang K, Parapar J, Gurevych I. Semantic similarity models for depression severity estimation. In: Bouamor H, Pino J, Bali K, eds. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2023:16104-16118. <https://aclanthology.org/2023.emnlp-main.1000>
 31. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* 2001;16:606-613.
 32. Nguyen T, Yates A, Zirikly A, Desmet B, Cohan A. Improving the generalizability of depression detection by leveraging clinical questionnaires. In: Muresan S, Nakov P, Villavicencio A, eds. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Vol 1 (Long Papers). Association for Computational Linguistics; 2022:8446-8459. <https://aclanthology.org/2022.acl-long.578>
 33. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Vol 32. 2019:8026-8037.
 34. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics; 2020:38-45. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
 35. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol 1 (Long and Short Papers). 2019:4171-4186. <https://doi.org/10.18653/v1/N19-1423>
 36. Ji S, Zhang T, Ansari L, et al. MentalBERT: publicly available pre-trained language models for mental healthcare. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, eds. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association; 2022:7184-7190. <https://aclanthology.org/2022.lrec-1.778>
 37. Ji S, Zhang T, Yang K, Ananiadou S, Cambria E, Tiedemann J. Domain-specific continued pretraining of language models for capturing long context in mental health. arXiv:2304.10447. <http://arxiv.org/abs/2304.10447>, 2023, preprint: not peer reviewed. <https://doi.org/10.48550/arXiv.2304.10447>
 38. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv:2004.05150. <http://arxiv.org/abs/2004.05150>, 2020, , preprint: not peer reviewed. <https://doi.org/10.48550/arXiv.2004.05150>
 39. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.
 40. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett.* 2006;27:861-874.
 41. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc.* 1937;32:675-701.
 42. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res.* 2006;7:1-30.
 43. Benavoli A, Corani G, Demšar J, Zaffalon M. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *J Mach Learn Res.* 2017;18:1-36. <http://jmlr.org/papers/v18/16-305.html>