American College *of* Neuropsychopharmacology

www.nature.com/dpn

ARTICLE OPEN

Check for updates

Using large language models as a scalable mental status evaluation technique

Margot Wagner[1]✉, Callum Stephenson[2], Jasleen Jagayat[2], Anchan Kumar[2], Amir Shirazi[2], Nazanin Alavi[2] and Mohsen Omrani[2]

Mental health care faces a significant gap in service availability, with demand for services significantly surpassing available care. As such, building scalable and objective measurement tools for mental health evaluation is of primary concern. Given the usage of spoken language in diagnostics and treatment, it stands out as a potential methodology. With a substantial mismatch between the demand for services and the availability of care, this study focuses on leveraging large language models to bridge this gap. Here, a RoBERTa-based transformer model is fine-tuned for mental health status evaluation using natural language processing. The model analyzes written language without access to prosodic, motor, or visual cues commonly used in clinical mental status exams. Using non-clinical data from online forums and clinical data from a board-reviewed online psychotherapy trial, this study provides preliminary evidence that large language models can support symptom identification in classifying sentences with an accuracy comparable to human experts. The text dataset is expanded through augmentation using backtranslation and the model performance is optimized through hyperparameter tuning. Specifically, a RoBERTa-based model is fine-tuned on psychotherapy session text to predict whether individual sentences are symptomatic of anxiety or depression with prediction accuracy on par with clinical evaluations at 74%.

NPP – Digital Psychiatry and Neuroscience; https://doi.org/10.1038/s44277-025-00042-z

LAY SUMMARY

Mental health services struggle to keep up with growing demand. This study explores how a RoBERTa large language model can help by analyzing text for signs of anxiety or depression. It uses online therapy transcripts and forum posts, learning to spot symptoms from what people write, even without tone or facial cues. With accuracy similar to human experts, this approach shows promise for making mental health evaluation more accessible and scalable.

INTRODUCTION

There is currently a mental health crisis as the demand for mental health services vastly outstrips the availability of quality care. In the United States and Canada, approximately 60 million people grapple with mental health concerns, yet, over two-thirds of those in need are unable to receive care despite a staggering $250 billion in healthcare expenditures [1, 2]. The prevailing model of care delivery relies on one-on-one interactions between clinicians and patients, which is labor-intensive and thus lacks scalability. The paucity of mental healthcare clinicians hampers the efficacy of this traditional format. The World Health Organization estimates around half the world's population lives in countries where there is one psychiatrist for 200,000 or more citizens [3]. In Canada, such inefficient utilization of resources has led to wait times of 8–16 months to see a psychiatrist from the initial referral [4]. Furthermore, the considerable degree of clinician involvement further inflates the cost of conventional treatment. Between 2011 and 2021, Ontario saw a 50% increase in mental health-related emergency department (ED) visits, often resulting from delayed access to outpatient psychotherapy [5, 6]. Nearly half of these ED visits were first-contact encounters, underscoring the need for improved triage and access to earlier-stage interventions [4, 7–9]. The higher cost of ED care ($500–$600/patient) compared to psychiatric outpatient care ($80–150/patient) adds significant pressure to the public health system [5, 10]. Additionally, untreated mental health problems can increase the cost of managing other chronic diseases by 1.5–3 times [11, 12], increase the length of hospital stay and chances of readmission [13–15], and yield poorer clinical outcomes [16]. Therefore, developing systems that improve resource utilization and access to comprehensive and timely care must be a priority [17]. The COVID-19 pandemic has exacerbated this mismatch in care by increasing demand for services, fostering mental health awareness, and diminishing the stigma associated with seeking care. This has created a persistent disparity between the supply and demand for mental healthcare, enduring even beyond the pandemic's cessation. To address the growing demand for mental health services, some healthcare systems have adopted stratified and stepped-care approaches. These models aim to optimize resource utilization by ensuring that limited and specialized services, such as psychiatrist appointments, are prioritized for individuals with the most severe or unstable conditions. This framework provides a

[1]Institute for Neural Computation, University of California San Diego, San Diego, CA, USA. [2]Department of Psychiatry, Queen's University, Kingston, ON, Canada.
✉email: mwagner@ucsd.edu

Received: 8 January 2025 Revised: 13 May 2025 Accepted: 16 July 2025
Published online: 13 November 2025

tiered system of care, where patients with less severe conditions can benefit from lower-intensity interventions, such as self-guided or therapist-assisted digital therapies, while those with more complex needs receive direct, specialized care (e.g. psychiatrist appointments). Despite these advancements, the identification process remains resource-intensive and time-consuming.

Currently, triage often involves the use of checklists and questionnaires, which rely on subjective measures by asking patients to report their symptoms and experiences on the intake triage day. While this approach provides useful insights, it can be limited by variability in patient responses, which may be influenced by transient mood or situational factors. A more effective model would integrate objective measures that go beyond a single point-in-time assessment, analyzing patterns of patient experiences over a period. Such an approach could leverage longitudinal data to provide a more accurate picture of the patient's needs, allowing for a tailored determination of the appropriate level of care. This would not only enhance the precision of triage decisions but also reduce reliance on resource-intensive processes. While LLMs are not required to interpret structured numerical scales, their utility lies in synthesizing narrative data beyond what checklists can capture. In systems with integrated electronic health records, LLMs could help extract longitudinal mental health information from previous clinical notes or patient messages, where available, to inform triage decisions. It is imperative to prioritize the development of innovative solutions that streamline and automate the delivery of mental healthcare to address this public health concern.

The field of mental health care has lagged behind broader innovation in the healthcare industry. This discrepancy primarily arises from the dearth of robust and precise quantitative measurements, biomarkers, and evidence-based practices in this domain compared to other medical specialties. Thus, it becomes essential for the field to center its efforts on the cultivation of novel data-driven measurement tools which can effectively evaluate the mental status of patients. Thought leaders, including Thomas Insel, have emphasized the need for diagnostic systems that integrate multiple levels of information [18, 19]. Frameworks such as the Research Domain Criteria (RDoC) [20] and the Hierarchical Taxonomy of Psychopathology (HiTOP) [21] address these gaps through dimensional, data-driven approaches to classification. In line with these efforts, this study aims to contribute to the development of scalable, objective tools that support comprehensive mental health assessment, thereby bridging the existing care gap and ensuring optimal treatment for individuals in need.

Establishing objective measurement tools for mental health evaluation is challenging due to their inherent complexity. Unlike other chronic diseases which typically utilize one or few biophysiological target variables, such as blood glucose in diabetes or blood pressure in hypertension, mental health disorders lack definitive biomarkers. As a result, there is a reliance on indirect measurements, including changes in sleep patterns or activity level [22] or changes in body posture [23] and speech tone [24]. While these biophysical symptom proxies hold significance in a clinician's decision-making process, the patient's verbalized thought content plays a central role in psychiatric assessment, particularly when integrated with affective cues, speech prosody, and structured clinical interviewing. The assessment of thought form and content, mood status, stressors and anxiety level predominantly rely on the patient's verbal expression. Accordingly, scalable technologies for mental status evaluation may benefit from focusing on patient language, while recognizing that speech is typically elicited in structured interviews and interpreted in the context of multimodal cues. However, patients' speech is unstructured data, which makes the process of extracting clinically relevant data a challenging endeavor. Clinicians interpret patient speech through a combination of attentive listening and professional judgment, a process that natural language processing (NLP) tools may one day support by assisting in identifying symptom-relevant language patterns.

To address the scalability and objectivity challenges outlined above, recent advances in deep learning, particularly in NLP, offer promising tools to support mental health evaluation [25–29]. Unlike traditional statistical and machine learning methods, deep learning methods do not heavily rely on feature engineering and can process longer, more complex sentences in a context-dependent manner. They also exhibit enhanced capabilities in learning language structures, allowing for effective transfer learning with limited data. Transformers, newer than convolutional and recurrent neural networks, show promise in handling sequential and textual data, making them suitable for mental health applications [28]. In this work, the focus is on fine-tuning a pretrained transformer model to detect symptomatic sentences related to depression and anxiety in a client's narrative. This paper investigates whether a fine-tuned transformer model can accurately detect depression- and anxiety-related language in psychotherapy narratives and benchmark its performance against existing approaches. This paper hypothesizes that a RoBERTa-based transformer model, fine-tuned on annotated psychotherapy narratives, can classify individual sentences as symptomatic or non-symptomatic with accuracy comparable to that of expert clinicians.

METHODS

Datasets

Training dataset. The non-clinical training data for this study was collected from online mental health forums where individuals share their personal experiences and challenges related to mental health. To prepare the data, the stories were segmented into sentences, and each sentence was carefully examined and labeled by an expert clinician (Expert A) as either neutral or exhibiting signs of anxiety and/or depression. In addition, sentences that relied on the previous or following sentence for context were flagged as dependent examples. An illustration of this dependency can be seen in the following pair of sentences: "Would I say my life is perfect and I am happy every day? No.". In isolation, each sentence does not provide a clear indication of symptoms, but when considered together, they reveal relevant information. Therefore, both sentences were marked as dependent examples. Furthermore, sentences that were unrelated, such as asides or emphatic statements, were also identified and flagged. These two categories of sentences–dependent and unrelated–were subsequently removed from the dataset as they were not suitable for labeling in the context of this research question. Out of the initial 3780 sentences, only 97 sentences (2.6%) were removed based on criteria, resulting in a dataset of 3683 sentences. The removal of non-emotional or context-dependent sentences was conducted manually to ensure labeling quality. While effective, this step limited scalability. Future iterations of this pipeline will implement a preliminary model to automate the identification of non-relevant content before classification. Sentences labeled as displaying signs of anxiety, depression or both were categorized as symptomatic (positive) examples while neutral sentences served as non-symptomatic (negative) examples. The resulting training dataset was 36% symptomatic and 64% non-symptomatic sentences. Once the training data was prepared, it was shuffled and split into a typical 80% training (2946 sentences) and 20% validation (737 sentences) dataset. The test set for evaluation purposes consisted of a clinical dataset. This sentence-level split may have allowed for stylistic overlap between sentences from the same user in both training and validation sets.

The training sentences used in this study had an average length of 19 words and 102 characters for symptomatic sentences, whereas non-symptomatic sentences had shorter averages of 16 words and 83 characters (Figs. 1, 2). In terms of data sources, each user contributed an average of 26 sentences. The highest number of sentences from a single patient in the dataset was 180 while the lowest contribution was 2 sentences.

Testing dataset. The clinical data utilized in this study was collected from a board-reviewed and ethically compliant online psychotherapy clinical trial conducted at Queen's University between 2020 and 2021. The study underwent a thorough review process by the Queen's University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board to ensure adherence to ethical standards (File #: 6020045). As part of their participation, patients provided written informed consent for the utilization of their anonymized data in academic research and publications. 55 subjects participated in the trial. During the trial, participants diagnosed with major depressive disorder (MDD) received 12 sessions of therapist-supported electronic cognitive behavioral therapy (e-CBT) in an asynchronous format. This asynchronous therapy involved engaging with weekly interactive online modules, which were delivered through a secure cloud-based online platform. During the initial week of the trial, participants were invited to share their narratives detailing their experiences with mental health challenges.

Participant narratives were segmented into 930 total sentences. Following the same inclusion criteria as the training dataset, 31 sentences (3.3%) were excluded, leaving 899 sentences for testing the algorithm performance. These sentences were similarly labeled as neutral or containing signs of anxiety and/or depression by two expert clinicians (Expert J and Expert M), who were different from the clinician involved in labeling the training dataset. Among the 899 sentences, Expert J considered 28% as symptomatic and 72% as non-symptomatic, while Expert M categorized 41.5% as symptomatic and 58.5% as non-symptomatic. Notably, this resulted in an inter-rater overlap (i.e. proportion of sentences having similar labels from both Experts J and M) of 76%, indicating a significant level of agreement between the two experts regarding sentence labeling. For testing purposes, the model performance was evaluated separately against the labels of each expert. No third expert was used to resolve disagreements, a consensus label was not set. This approach allowed for transparent comparisons between model predictions and individual clinician judgments. To assess label consistency across datasets, Expert J was also tasked with labelling the training dataset. The inter-rater overlap between Expert A and Expert J for the training dataset was 80%. This level of agreement reflects the inherent subjectivity in assessing symptom presence from isolated sentences and supports the use of expert consensus as a comparative benchmark in NLP model evaluation.

Model design. Due to the vast array of tasks in NLP, it is crucial to clearly define the specific task of interest before any modeling work. In this study, the task is a classification task, where the objective is to classify the text input based on a predefined label. Specifically, it is a binary sentence classification task, where the aim is to categorize a sentence input as one of two labels, symptomatic or non-symptomatic. Two particularly relevant subtasks in text classification are emotion recognition and sentiment analysis. Emotion recognition aims to assign a specific emotion (e.g. happy, sad, angry) to the input sentences. On the other hand, sentiment analysis focuses on capturing the overall attitude expressed in an input sentence (i.e. positive, negative, neutral). Given the nature of mental health, particularly anxiety and depression, these two subtask categories exhibit strong interrelationships and relevance to the current study.

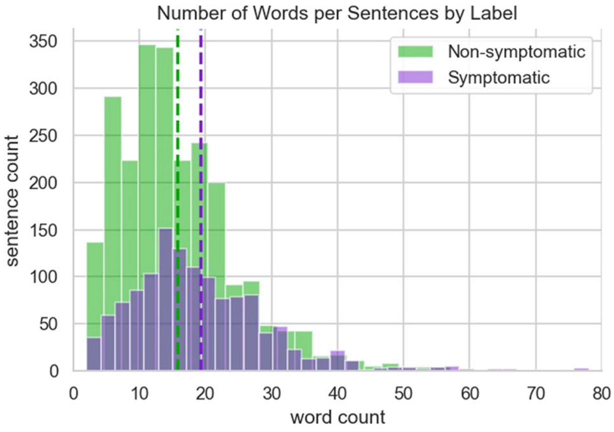

Fig. 1 The sentence length by label type, symptomatic and non-symptomatic where symptomatic sentences were less frequent and shorter.

Model training. The Transformer model class was selected due to its superior performance in NLP tasks related to emotion detection, surpassing previous models that lacked contextual understanding [30, 31]. Several models were selected from the HuggingFace transformer model library, including a standard Bidirectional Encoder Representations from Transformers (BERT) model as well as other BERT-based models including RoBERTa [32], DistilBERT [33], ALBERT [34], DeBERTa [35], and XLM-RoBERTa [36]. RoBERTa is a transformer-based language model that builds on BERT, optimized for downstream NLP tasks. BERT was chosen as it has established itself as the de facto and widely adopted baseline for NLP experiments [37]. In addition to training the baseline models, training was conducted on commonly employed baseline model variants, such as cased and uncased versions, as well as large-sized models, ensuring a comprehensive exploration of the model landscape.

A selection of transformer models that had undergone further fine-tuning for text classification and specific subtasks was also included in the initial model selection phase. While models from subtasks relevant to the current task (emotion recognition, sentiment analysis) were of particular interest, a wide range of subtasks were considered. These models were chosen based on their popularity within the HuggingFace community (as determined by the number of downloads) and their demonstrated performance in various tasks. A total of 44 unique models were tested using a standardized set of hyperparameters. Each model was trained for 5 epochs. The remaining training hyperparameters were set to their default values provided by HuggingFace.

Most models (75%) were baseline models that had been fine-tuned to another task before our training (Fig. 3). Among these fine-tuned models, 57% were specifically fine-tuned for a text classification task, 16% for token classification, and 2% for fill mask tasks. The token classification models were exclusively tuned for name entity recognition (NER) subtasks, while the fill mask task was for a biological subtask. The text classification models were primarily trained on subtasks of sentiment analysis (36%) and emotion recognition (32%) subtasks.

Before model training, the training examples underwent tokenization using HuggingFace's tokenizer class, which employs base model-specific (i.e. BERT, DistilBERT, RoBERTa) tokenization techniques. It is worth noting that the training data exhibits an imbalance, with only 36% of sentences meeting the criteria for being symptomatic. To address this imbalance, a weighted cross-entropy loss function was employed during model training, where the weights were determined based on the distribution of the two classes. Model accuracy was used as the sole criterion for model selection due to the sensitive nature of the given task, which necessitates maximizing correct predictions even if it comes at the expense of computational efficiency, such as model size or latency. Specifically, the F1 score was used as the model accuracy metric, as it maintains a balance between precision and recall. For the given task, it is crucial to try to accurately predict as many symptomatic cases as possible (recall) while also maintaining a high level of confidence in the positive predictions (precision). Therefore, both metrics were considered essential in evaluating the model performance.

Data augmentation serves as a valuable approach to enhance the training dataset by introducing additional examples through slight modifications of existing ones. While in image datasets this can be achieved through simple techniques such as scaling, rotation, or color manipulation, text datasets require methods that preserve sentence meaning. Simple text augmentations like random word swap or insertion are not sufficient for complex tasks such as emotion detection, where sentence meaning plays a pivotal role. Therefore, alternative methods that better preserve sentence meaning were considered and compared,

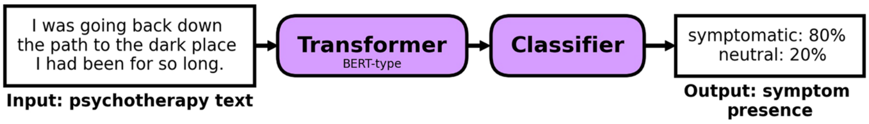
I was going back down the path to the dark place I had been for so long.
Input: psychotherapy text
Transformer
BERT-type
Classifier
symptomatic: 80%
neutral: 20%
Output: symptom presence

Fig. 2 Natural language processing mental health task process figure where a narrative sentence is input to a transformer model with a classifier to predict status as either symptomatic or non-symptomatic.

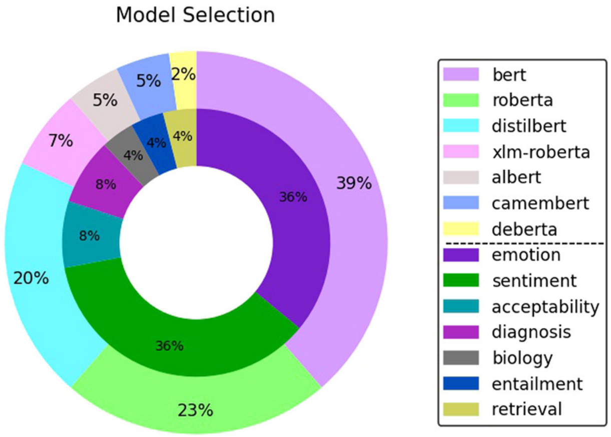

Fig. 3 Task and subtask distributions of models used in model selection. BERT models were pretrained on an emotion recognition task before the current study training.

including the use of back translation. These methods offer more effective ways to generate synthetic training examples that contribute meaningfully to the training process.

Back-translation, the technique employed in this study, involves translating the original sentence to another language and then back to English. It was implemented here with the NLPAug library for textual augmentation, utilizing HuggingFace transformer-based translation models. This approach is expected to maintain the sentiment of the sentence more accurately. The number and type of language intermediates used in the translation process were treated as additional tunable hyperparameters. Each available language intermediate on HuggingFace (as of 2022) was individually tested, effectively doubling the size of the training dataset with the inclusion of the back-translated synthetic counterparts for each sentence. During augmentation, the original class distribution (36% symptomatic, 64% non-symptomatic) was maintained to reflect the natural imbalance found in real-world clinical narratives to prioritize ecological validity. Moreover, various ratios of back-translated synthetic sentences to the original sentence were explored, along with different combinations of language intermediates. A comprehensive investigation was conducted across 20 individual languages, followed by experimentation with the top 11 performing models in different combinations on the complete dataset. The number of languages used was incrementally increased until a drop in performance was observed. For details on this analysis, consult Appendix A2.

Hyperparameter tuning was conducted using the Tree Parzen Estimation (TPE) method with HyperOpt, employing an Asynchronous Successive Halving (ASHA) scheduling algorithm implemented with Ray Tune [38, 39]. The selected hyperparameters and their respective ranges for tuning were chosen based on their ability to significantly enhance model performance in prior works [39]. The hyperparameters chosen included the number of training epochs, the random seed, the number of training examples per batch, and the learning rate. The final hyperparameters used can be found in Table 1.

The fine-tuned model trained on non-clinical data was then evaluated on the clinical dataset to assess its diagnostic capabilities. The model's performance was calculated and compared against the labels of Experts J and M, and the results are presented in Table 1. To limit the burden on clinical data and reduce computational costs, the test set evaluation on the best-performing model only was the focus, allowing for a direct comparison to human experts. The overall process of data augmentation, training, and tuning is outlined in Fig. 4.

RESULTS

Performance evaluation was conducted across 44 different models by comparing their F1 score (harmonic mean of precision and recall) and balanced accuracy (bA) on the training dataset. The top 11 models, based on their performance, were fine-tuned using the augmented training dataset. A detailed comparison of these models can be found in Table A1.1 in Appendix A1.

The model that achieved the highest performance in our evaluation was a fine-tuned RoBERTa model that was previously trained on the TweetEval [40] dataset before this dataset.[1] TweetEval is a dataset specifically designed for multi-class emotion recognition in Tweets, consisting of over 5000 Tweets categorized into four emotions: anger, joy, sadness, and optimism. By leveraging this dataset, the RoBERTa model demonstrated superior performance, outperforming word-based, context-free algorithms like fastText by approximately 10%. For a detailed analysis of each model's performance, please consult Appendix A1.

Subsequently, we assessed the performance of the best-performing model, trained on the public dataset, using the clinical dataset that was labeled by two additional experts. The model achieves an accuracy of approximately 74%. It is worth noting that even human experts do not exhibit complete consistency. The inter-rater overlap between Experts J and M was found to be 76%, which is comparable to the accuracy of the model. This suggests that the model approximates expert-level performance in identifying symptomatic content, at least within the context of sentence-level classification of psychotherapy narratives.

The results table displays results for the baseline model, which refers to the Emotion RoBERTa model. The Augmented column refers to the Emotion RoBERTa model applied to the augmented training dataset. Finally, the Tuned performance is after hyper-parameter tuning.

Both the F1 score (F1) and balanced accuracy (bA) are crucial metrics included in our evaluation. These metrics are specifically employed for classification tasks involving imbalance datasets, such as this. It is important to highlight that the F1 score maintains a balance between precision and recall, whereas the balanced accuracy assesses specificity and recall. The balanced accuracy metric directly considers true negatives, making it particularly useful when both true positives and true negatives are equally significant. On the other hand, the F1 score prioritizes the positive cases, emphasizing the accuracy of identifying positive instances. By utilizing both metrics, we ensure a comprehensive assessment of the model's performance in handling imbalanced classification scenarios.

DISCUSSION

In this project, we conducted fine-tuning on a variety of transformer models to identify symptomatic sentences in a client's mental health narrative, specifically those related to depression and anxiety. We aimed to compare the performance of these models and address the limited availability of labeled data by employing augmentation techniques to expand our dataset.

Our findings demonstrate that our most effective model achieved an accuracy of approximately 80% (F1: 79.3%, bA 84.0%) when distinguishing between symptomatic and non-symptomatic sentences, which is comparable to the performance of human experts. While these results are promising, they represent an incremental step toward scalable mental health evaluation. This study focuses on sentence-level classification for anxiety and depression using text alone and does not incorporate other clinical signals such as facial expression, prosody, or interactional nuance. As such, this is not a diagnostic system, but rather a proof-of-concept for language-based symptom detection. Alternative approaches, including multimodal learning, full narrative modeling, or temporal tracking of symptom expression across sessions, may further strengthen clinical utility. Despite being trained on a public dataset, this model showcased a similar level of accuracy when classifying sentences from a clinical dataset collected from 55 patients participating in a separate clinical trial (average F1 = 74%, bA = 73.5%). Notably, our model's

[1] The model has been made available on HuggingFace as margotwagner/roberta-psychotherapy-eval

Table 1. Performance of trained models across datasets.

	Hyperparameter-Tuned	Augmented Data	Fine-Tuned	fastText[a]
Public Dataset	**F1: 79.3** **bA: 84.0**	F1: 79.1 bA: 83.6	F1: 77.2 bA: 82.3	F1: 68.1 bA: 75.2
Clinical Dataset (Expert J)	F1: 73.0 bA: 72.0			
Clinical Dataset (Expert M)	F1: 75.0 bA: 75.0			

[a] All other non-DL models performed worse than fastText. For more details, please refer to Appendix A1.
Bold indicates best performing model.

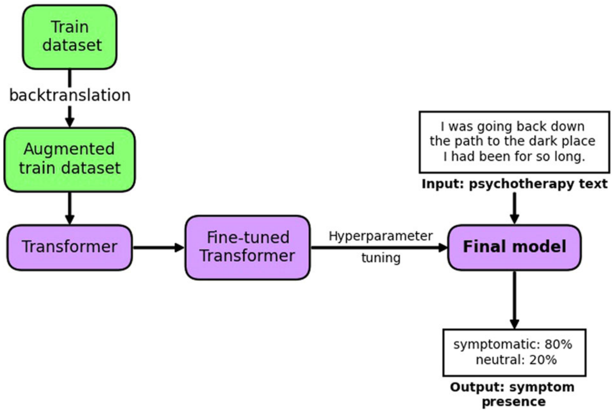

Fig. 4 Schematic of the overall training process. The training dataset is augmented and then used to fine-tune a transformer model. The model is optimized through hyperparameter tuning before the final model is attained.

performance is in line with the level of agreement between different expert raters, indicating its reliability.

Through the fine-tuning of various transformer models, augmentation techniques, and comparative analyses, our project has successfully developed a model capable of accurately classifying symptomatic sentences. Its accuracy on both public and clinical datasets, combined with its performance comparable to interrater agreement, highlights its effectiveness and potential for practical applications.

Our study aimed to identify the most effective model for the given task, and our analysis revealed that the RoBERTa model, fine-tuned on the TweetEval benchmark, outperformed the other models examined. RoBERTa is an enhanced version of the original BERT transformer model, benefiting from robust optimization during pretraining, which ultimately resulted in improved model performance [32, 40, 41]. Unlike BERT, RoBERTa underwent pretraining using an expanded dataset, comprising five English-language corpora that totaled over 160 GB of uncompressed text. These corpora include BOOKCORPUS [42], WIKIPEDIA, CC-NEWS [43], OPENWEBTEXT [44], STORIES [45].

The model was further fine-tuned on the TweetEval benchmark before our task. Specifically, it utilized the Emotion Recognition dataset, which contains over 5000 text statements sourced from Twitter. Each statement is associated with one of four emotions: anger, joy, sadness, and optimism. The exceptional performance of the RoBERTa model highlights the significance of emotion classification as a valuable precursor for mental health diagnostics. It is worth noting that the BERT base model, XLM-RoBERTa pretrained on a sentiment analysis task, and DistilBERT base models closely trailed in terms of performance.

By highlighting the superior performance of the RoBERTa model fine-tuned on Emotion TweetEval, our study highlights its strong performance relative to alternatives in this specific task. These findings suggest that emotion classification may be a useful component in supporting future mental health evaluation tools, though additional work is required before this approach can inform diagnostic decision-making.

One effective strategy for enhancing model training involves employing diverse data augmentation techniques to expand the training dataset. In our study, we explored the back-translation technique and observed a notable improvement in model performance as a result. Interestingly, we discovered that using intermediate languages from the Indo-European, Turkic, or Uralic language families yielded superior results compared to other language families, such as Sino-Tibetan, Japonic, Austronesian, or Afro-Asiatic. This can be attributed to the fact that languages belonging to the same language family as English tend to capture sentence structure and meaning more effectively due to their greater similarity.

To maximize the benefits of back-translation, we identified the combination of Turkish and Danish as particularly effective, allowing us to triple the size of the training dataset by generating two additional augmented sentences for each origin training sentence. This language combination produced the highest model performance, achieving an approximate 2% increase in the F1 score. These results also highlight that the model's performance is influenced by syntactic similarity to English. Individuals whose native language differs significantly from English, particularly non-Indo-European, may express symptoms in ways that the model struggles to interpret. This raises concerns about potential diagnostic bias in multilingual or culturally diverse populations. Future research should evaluate model performance across diverse linguistic and cultural groups to improve fairness and generalizability. For a detailed analysis of the language combination, corresponding performance, and the impact of increasing the ratio of synthetic sentences to original sentences, please refer to Appendix A2 Table A2.

By strategically implementing back-translation and specifically leveraging the Turkish and Danish languages, we successfully amplified the training dataset and enhanced the model's performance. The observed improvements validate the effectiveness of this approach for training models for the given task.

During the hyperparameter tuning process, we focused on optimizing several key hyperparameters, including the number of training epochs, random seeds, the number of training examples per batch, and the learning rate. While these hyperparameters were carefully tuned, it is worth noting that there may still be room for further optimization. After thorough experimentation, we observed only a modest improvement in the overall model performance, with the F1 scores increasing by a mere 0.2%. For a comprehensive list of the final set of hyperparameters employed in this study, please refer to Appendix A3. Moving forward, further exploration and optimization of hyperparameters may garner additional model success. It is an avenue that warrants future investigation and can potentially yield more substantial improvements.

The findings from this study highlight the promising potential of training a transformer model for a nuanced and intricate clinical task, specifically the detection of symptomatic language use, even when faced with limited labeled data. Furthermore, the transferability of the model's knowledge to diverse datasets collected in distinct clinical settings is a crucial outcome. Ultimately, these transformer models can revolutionize the field by enabling scalable and objective mental status evaluations based on patients' language usage.

As we move forward, it becomes imperative to envision future clinical trials that leverage these objective measurements to predict essential clinical outcomes. These outcomes could expand to include factors such as patient engagement, symptom reduction, or even relapse prediction. By incorporating these objective measurements into the design of future trials, we can potentially enhance our understanding of the complex interplay between language use and clinical outcomes.

Looking ahead, future research should explore the model's generalizability across different diagnostic domains, such as psychosis, posttraumatic stress disorder, obsessive-compulsive disorder, and eating disorders, and its potential integration into digital triage platforms and clinical workflows. Prospective trials should assess the model's utility in real-time settings, such as predicting therapy engagement, tracking symptom progression, and flagging early signs of relapse. These steps will help translate this proof-of-concept into a clinically actionable tool within a stepped-care framework.

By utilizing the power of transformer models and their ability to accurately analyze language patterns, we pave the way for more precise and comprehensive evaluations of mental health. This has the potential to significantly impact clinical practice and improve patient care. A logical next step is to strategically integrate these models into clinical research and care delivery, enabling insight-driven enhancements to mental health services.

REFERENCES

1. The state of mental health in America. Technical report. Alexandria, VA: Mental Health America;2023.
2. Mental health and substance use state fact sheets. Technical report. San Francisco, CA: Kaiser Family Foundation;2023.
3. World mental health report: transforming mental health for all. Technical report. Geneva: World Health Organization;2022.
4. Barua, B, Rovere, MC & Skinner, BJ. Waiting your turn: wait times for health care in Canada 2010 report. SSRN Electron J. 2011. https://doi.org/10.2139/ssrn.1783079
5. Canadian Institute for Health Information. Frequent emergency room visits for help with mental health and substance use. 2023. https://www.cihi.ca/en/indicators/frequent-emergency-room-visits-for-help-with-mental-health-and-substance-use Accessed 28 March 2025.
6. Simcoe Muskoka. Overall Mental Illness. 2023. https://www.simcoemuskokahealthstats.org/topics/mental-health/mental-illness/overall-mental-illness Accessed 28 March 2025.
7. Lavergne MR, Loyal JP, Shirmaleki M, Kaoser R, Nicholls T, Schütz CG, et al. The relationship between outpatient service use and emergency department visits among people treated for mental and substance use disorders: analysis of population-based administrative data in British Columbia, Canada. BMC Health Serv Res. 2022;22:477.
8. Saunders NR, Gill PJ, Holder L, Vigod S, Kurdyak P, Gandhi S, et al. Use of the emergency department as a first point of contact for mental health care by immigrant youth in Canada: a population-based study. CMAJ. 2018;190:E1183–E1191.
9. ICES | Incidence of access to ambulatory mental health care prior to a psychiatric emergency department visit among adults in Ontario, 2010-2018. *ICES* https://www.ices.on.ca/publications/journal-articles/incidence-of-access-to-ambulatory-mental-health-care-prior-to-a-psychiatric-emergency-department-visit-among-adults-in-ontario-2010-2018/ Accessed 9 August 2024.
10. College of Family Physicians of Canada. The value of continuity—investment in primary care. The College of Family Physicians of Canada. 2021.
11. Melek, S, Norris DT, Paulus J, Matthews K, Weaver A, Davenport S. Potential economic impact of integrated medical-behavioral healthcare: updated projections for 2017. Miliman Res Rep. 2018. https://www.psychiatry.org/getmedia/be47c6d5-12f7-4d07-8a43-15082b0a3029/Milliman-Report-Economic-Impact-Integrated-Implications-Psychiatry.pdf.
12. Cawthorpe D, Wilkes TCR, Guyn L, Li B, Lu M. Association of mental health with health care use and cost: a population study. Can J Psychiatry. 2011;56:490–4.
13. Doupnik SK, Lawlor J, Zima BT, Coker TR, Bardach NS, Hall M, et al. Mental health conditions and medical and surgical hospital utilization. Pediatrics. 2016;138:e20162416.
14. Hendrie HC, Lindgren D, Hay DP, Lane KA, Gao S, Purnell C, et al. Comorbidity profile and healthcare utilization in elderly patients with serious mental illnesses. Am J Geriatr Psychiatry. 2013;21:1267–76.
15. Siddiqui N, Dwyer M, Stankovich J, Peterson G, Greenfield D, Si L, et al. Hospital length of stay variation and comorbidity of mental illness: a retrospective study of five common chronic medical conditions. BMC Health Serv Res. 2018;18:498.
16. Goldman S, Saoulidi A, Kalidindi S, Kravariti E, Gaughran F, Briggs T, et al. Comparison of outcomes for patients with and without a serious mental illness presenting to hospital for chronic obstruction pulmonary disease: retrospective observational study using administrative data. BJPsych Open. 2023;9:e128.
17. Bellon J, Quinlan C, Taylor B, Nemecek D, Borden E, Needs P. Association of outpatient behavioral health treatment with medical and pharmacy costs in the first 27 months following a new behavioral health diagnosis in the US. JAMA Netw Open. 2022;5:e2244644.
18. Insel T. Transforming diagnosis. NIMH Director's BlogPosts from 2013. https://psychrights.org/2013/130429NIMHTransformingDiagnosis.htm.
19. Kapur S, Phillips AG, Insel TR. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol Psychiatry. 2012;17:1174–9.
20. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. 2013;11(1):126.
21. Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, et al. The hierarchical taxonomy of psychopathology (HiTOP): a dimensional alternative to traditional nosologies. J Abnorm Psychol. 2017;126(4):454–77.
22. Zhang Y, Folarin AA, Sun S, Cummins N, Bendayan R, Ranjan Y, et al. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: Multicenter longitudinal observational study. JMIR Mhealth Uhealth. 2021;9(4):e24604.
23. Gravenhorst F, Muaremi A, Bardram J, Grünerbl A, Mayora O, Wurzer G, et al. Mobile phones as medical devices in mental disorder treatment: an overview. Pers Ubiquitous Comput. 2015;19:335–53.
24. Kappen M, Vanderhasselt MA, Slavich GM. Speech as a promising biosignal in precision psychiatry. Neurosci Biobehav Rev. 2023;148:105121.
25. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med. 2022;5:46.
26. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res. 2021;23(5):e15708.
27. Burger F, Neerincx MA, Brinkman WP. Natural language processing for cognitive therapy: extracting schemas from thought records. PLoS ONE. 2021;16(10):e0257832.
28. Ahmed MS, Kornblum D, Oliver D, Fusar-Poli P, Patel R. Associations of remote mental healthcare with clinical outcomes: a natural language processing enriched electronic health record data study protocol. BMJ Open. 2023;13(2):e067254.
29. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C Machine Learning and Natural Language Processing in Mental Health: Systematic Review J Med Internet Res 2021;23:e15708.
30. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. Minneapolis, MN: ACL; 2019. pp. 4171–86.
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention is all you need. In: Advances in neural information processing systems, Conference on Neural Information Processing Systems. Long Beach, CA: NeurIPS; 2017.
32. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al.. Ro{bert}a: a robustly optimized {bert} pretraining approach. In: International Conference on Learning Representations, ICLR. ICLR: 2020.
33. Sanh V, Debut L, Chaumond J, Wolf T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019* arXiv:1910.01108v4 [Preprint]. 2020. Available from: https://arxiv.org/abs/1910.01108.
34. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: a lite bert for self-supervised learning of language representations. arXiv:1909.11942v6 In *ICLR 2026* [Preprint]. 2020. Available from: https://iclr.cc/virtual_2020/poster_H1eA7AEtvS.html.
35. He P, Liu X, Gao J, Chen W. Deberta: Decoding enhanced bert with disentangled attention. arXiv:2006.03654v6 *ICLR 2021* [Preprint]. 2021. Available from: https://openreview.net/pdf?id=XPZIaotutsD.
36. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzman F, et al. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, July 5 - 10, 2020. c 2020 Association for Computational Linguistics* arXiv:1911.02116v2 [Preprint]. 2020. Available from: https://aclanthology.org/2020.acl-main.747.pdf.

37. Rogers A, Kovaleva O, Rumshisky A. A primer in bertology: what we know about how bert works. Trans Assoc Comput Linguist. 2020;8:842–66.
38. Bergstra J, Yamins D, Cox DD. Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms. In Python in science, SciPy. Springer: Austin, TX;2013.
39. Li L, Jamieson K, Rostamizadeh A, Gonina K, Hardt M, Recht B, et al. Massively parallel hyperparameter tuning. in *the conference on machine learning and systems 2020* 2018. https://arxiv.org/abs/1810.05934.
40. Sun C, Qiu X, Xu Y, Huang X. How to fine-tune bert for text classification? In: Chinese Computational Linguistics. Cham: Springer;2019.
41. Barbieri F, Camacho-Collados J, Neves L, Espinosa-Anke L. Tweeteval: unified benchmark and comparative evaluation for tweet classification. arXiv:2010.12421v2 [Preprint]. 2020. Available from: https://aclanthology.org/2020.findings-emnlp.148/.
42. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. arXiv:1506.06724v1 [Preprint]. 2015. Available from: https://www.cvfoundation.org/openaccess/content_iccv_2015/papers/Zhu_Aligning_Books_and_ICCV_2015_paper.pdf.
43. Mackenzie J, Benham R, Petri M, Trippas JR, Culpepper JS, Moffat A. Cc-news-en: a large english news corpus. In *conference paper in the proceedings of the 28th international conference on information and knowledge management (CIKM 2020) in October 2020* 2020. https://jmmackenzie.io/publication/cikm20-resource/.
44. Gokaslan A, Cohen V. Openwebtext corpus. https://zenodo.org/records/3834942.
45. Trinh TH, Le QV. A simple method for commonsense reasoning. arXiv:1806.02847v2 [Preprint]. March 2023. Available from: https://arxiv.org/abs/1806.02847.

AUTHOR CONTRIBUTIONS

Margot Wagner, PhD*. Completed all analysis, algorithm, and model development. Wrote manuscript. Callum Stephenson, MSc. Edited manuscript. Helped with manuscript reviews. Jasleen Jagayat, MSc. Clinical data collection and dataset labelling. Anchan Kumar, MD. Clinical data collection and dataset labelling. Amir Shirazi, MD PhD. Advised model and algorithm development. Nazanin Alavi, MD. PI on the clinical trial. Reviewed and edits the manuscript. Mohsen Omrani, MD PhD. Developed and designed study. Advised model and algorithm development.

COMPETING INTERESTS

MO and NA have ownership stakes in a digital mental health company called OPTT Inc. No data, funding, or algorithms were used from the company in this study. All other authors report no biomedical financial interests or conflicts of interest.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44277-025-00042-z.

Correspondence and requests for materials should be addressed to Margot Wagner.

Reprints and permission information is available at http://www.nature.com/reprints