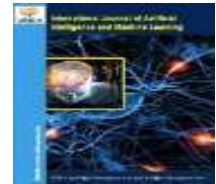




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Fine-Tuned RoBERTa Transformers for Multi-Category Mental Health Status Recognition in Written Text

Dondapati Kalyan¹, Ch Satyananda Reddy²

¹Department of Computer Science & Systems Engineering Andhra University College of Engineering (A) Visakhapatnam, India. E-mail: dondapatikalyan123@gmail.com

²Department of Computer Science & Systems Engineering Andhra University College of Engineering (A) Visakhapatnam, India. E-mail: satyanandau@yahoo.com

Abstract

Detecting early signs of anxiety, depression, and stress in regular texts may prove tremendously helpful for the triage process and for reducing the clinician's case-managing load. This paper describes the creation and tuning of a transformer-based classifier on short self-reported statements sorted into seven mental-health categories: Anxiety, Bipolar, Depression, Normal, Personality Disorder, Stress, and Suicidal. We use the pretrained RoBERTa-base weights as an initializer for a transformer-based classifier with a linear classification head on top of the base model. We trained the classifier on a balanced set of 105000 annotated statements, 15000 for each category, and evaluated its performance on a separate held-out sample of 9767 previously unseen statements. It demonstrated outstanding results, achieving 93.87% accuracy on the test set with a mean F1-score of 0.9408 and a per-category F1-score varying from 0.808 (for Depression) to 0.997 (for Personality Disorder). Upon a closer inspection of misclassified texts, we noticed an intriguing pattern: for 80.6% of misclassified cases, the true category was in the vicinity of Depression. Our investigation suggested that this might be explained by either the confounding presence of similar features in both categories or by the classifier's bias. For a closely related task, a classical ML pipeline with gradient-boosted trees was reported by other authors to achieve around 84.2% accuracy on a coarser, three-category version of this problem. Using the same underlying data but more than doubling the classification resolution to seven categories, our fine-tuned transformer achieved a much better overall accuracy. Finally, a proof of concept web-based application was created to showcase the potential of such technology. Overall, our experiments demonstrated that there is a realistic possibility of using a fine-tuned transformer for mental health screening.

Keyword: Mental Health, Text Classification, Natural Language Processing, Transformer, RoBERTa, Deep Learning, Machine Learning, Sentiment Analysis, Flask, Web Application.

INTRODUCTION

Anxiety, depression, and stress are common, and the opportunity to intervene successfully at an early stage is one of the most cost-effective ways to help such a patient. Structured instruments, such as the PHQ-9 depression severity questionnaire, have demonstrated their ability to routinely assess the progression of the disease in a primary-care setting [1]. The challenge is to apply similar principles to a more informal and frequent medium – the written text. The natural language processing field offers methods to automatically mine such data for patterns that can be associated with a particular disease or disorder, thus replacing the time-consuming task of manual screening by clinicians.

Previous approaches to this task focused on traditional machine learning algorithms that used either handcrafted

features such as Bag-of-Words statistics, N-Grams, or other representation such as Word Embeddings. These pipelines get results up to a point, but they treat words largely in isolation, and closely related psychological states often turn on exactly the contextual cues that a bag-of-words style representation discards. Transformer models upended that by learning to let a word's meaning shift with its surrounding context directly from raw text at scale, a shift that has produced measurable gains in text classification across many domains, healthcare among them.

This paper addresses the following refined research question: can a fine-tuned transformer achieve adequate performance in distinguishing between seven mental-health categories in text, as opposed to the classical three-class anxiety/depression/stress classification tasks attempted in prior work? The investigation will involve fine-tuning a RoBERTa-base model on the obtained class-balanced corpus, assessing performance with standard classification metrics, one-vs-rest ROC/AUC analysis, and a detailed

analysis of misclassified cases. Moreover, a simple web UI will be provided to experiment with the classifier in real time.

The rest of the paper is structured as follows: Section II reviews related work in the field of machine learning applied to mental-health text classification. Section III defines the problem and discusses the research gap motivating this study. Finally, Section IV describes the dataset, the data preprocessing pipeline, and the model architecture and training parameters. Section V explains how the browser-based demo was implemented. Section VI presents the experimental findings — per-category scores, ROC-AUC results, and the error breakdown. Section VII presents the results along with the earlier classical-ML approach, while the final section (VIII) concludes the paper and suggests directions for future research.

RELATED WORK

Approaches to mental health text classification can be divided into two groups based on whether they use pre-engineered textual features or learn representations of the text directly from the text. Most closely related to the present study, Juanita et al. developed a machine-learning pipeline to sort clinicians' written replies within online health consultations into one of three topic labels — anxiety, depression, or stress — using consultation transcripts sourced from Indonesian-language clinical services [2]. The researchers employed five different data classifiers: Naïve Bayes, K-Nearest Neighbor, Random Forest, Neural Network Algorithm, and Gradient Boosting. In addition, they applied the bagging classifier with Genetic Optimization. Each of these models was tested with the use of four different feature extraction methods, including Word2Vec, Bag of Words, GloVe, and N-Gram. The highest accuracy rate (84.2%) was reported for the Gradient Boosting algorithm with the N-Gram feature extraction method. This result was tied by the bagging classifier with gradient boosting. Given that it addresses a closely adjacent classification task through a rigorous classical pipeline, this earlier study is used here as the principal point of comparison in Section VII.

Beyond this single consultation-style setting, a broader body of work applies machine learning and NLP to mental health more generally. Le Glaz et al. conducted a systematic review charting where such techniques are already in use — diagnosis, risk flagging, and treatment monitoring — spanning both clinical records and everyday text sources [3]. Chancellor and De Choudhury offer a more critical perspective focused on inferring mental-health status from social media, discussing not only what has worked but also where the approach falls short, particularly regarding label reliability and generalizability [4]. Chung and Teo step back to organize the field as a whole, categorizing prior work by data source, modeling choice, and target condition, and note that multi-class, multi-condition classification remains relatively unexplored [5]. Haque et al. demonstrate that a similar text-classification approach extends beyond the adult populations most studies focus on, successfully detecting depression in children's writing [6]. Moving past consultation transcripts entirely, resources such as the Reddit-derived SMHD corpus enable multi-condition classification directly from users' self-reported diagnoses [7], while earlier work by Coppersmith et al. established that measurable linguistic markers of depression and bipolar disorder appear in everyday social-media posts, not just clinical text [8].

Zooming out, text classification as a field has moved away from the fixed embeddings of Word2Vec [9] and GloVe [10] — the very two feature-extraction schemes tested in the base study [2] — toward transformer architectures, precisely because a transformer allows a word's representation to shift with context rather than staying pinned to one vector no matter how the word is used. That change traces back to the self-attention mechanism introduced by Vaswani et al., which does away with recurrence entirely and lets any two tokens in a sequence interact directly [11]. Devlin et al. built on that mechanism to create BERT, a bidirectional pretrained encoder that raised the bar across a wide swath of NLP benchmarks [12], and Liu et al.'s RoBERTa took the same recipe further still — dropping the next-sentence-prediction objective and training longer on far more data — to wring out extra downstream accuracy [13]. This paper takes that RoBERTa encoder as its starting point and fine-tunes it specifically for multi-category mental-health text classification.

PROBLEM STATEMENT AND RESEARCH GAP

A. Problem Formulation

Screening written statements for mental-health-related content is fundamentally a fine-grained, multi-category text classification problem rather than a simple binary flag. Given a short, self-written statement, the task is to classify it into one of seven clinically motivated classes – Anxiety, Bipolar, Depression, Normal, Personality Disorder, Stress, and Suicidal – based only on the text of the statement without utilizing the additional context of an associated clinical history.

$$D = \{(x_i, y_i)\}_{i=1}^N, x_i \in X, y_i \in C, |C| = 7 \quad (1)$$

where X is a space of possible statements and C is the set of class labels, as defined by the problem statement and D is the training data set used to learn the function f that maps between X and C . The learning objective is to fit a parametric function f_θ that maps an input statement to a probability distribution over C , from which the predicted category is taken as the most probable class:

$$\hat{y} = \text{f}\theta(x) = \text{argmax}_{c \in C} P\theta(y = c | x) \quad (2)$$

Two aspects of the task make it particularly challenging compared to the standard topic classification. First, the topics are not always lexically distinct from each other – in the case of depression and suicidal ideation, both have a substantial overlap in lexemes related to hopelessness and self-blame; thus, the distinction between some categories may be determined by more subtle differences. Second, the available datasets, such as the one from the Kaggle Sentiment Analysis for Mental Health competition, are not balanced, have a limited number of entries for some categories, and include lexemes that might introduce significant label noise if not preprocessed. A proper choice of the classifier should be motivated by the necessity to learn the decision boundary that minimizes the distance between similar classes, operate on a balanced and sufficiently large training set, and remain lightweight enough for deployment as a real-time service rather than a research-only tool.

B. Research Gap

Earlier work on this problem falls into two broad camps, and neither fully closes the gap this study targets. Classical machine-learning pipelines built on hand-engineered features, exemplified by Juanita et al.'s Gradient-Boosting-plus-N-Gram system [2], report competitive accuracy but only over a coarse three-way split of anxiety, depression, and stress, and their bag-of-words-style features cannot represent the contextual cues that separate closely related psychological states. Later transformer-based studies close part of that representational gap but each leave a different piece unaddressed: Ameer et al. [14] fine-tune RoBERTa but cover a different five-category label set (ADHD, Anxiety, Bipolar, Depression, and PTSD, alongside a separate None class) that only partially overlaps with the seven categories used here, and do not report per-category ROC-AUC or a breakdown of where the model's errors concentrate; Sánchez Rodríguez et al. [15] fine-tune BERT/MentalBERT over four categories (Anxiety, Depression, Bipolar, and Borderline Personality Disorder) with a focus on discourse-level interpretability rather than deployable classification accuracy; and Verma et al. [16], while also using seven categories, substitute PTSD for Personality Disorder in their label set and depend on auxiliary sentiment and severity features fused into BERT, without providing a deployable, real-time interface.

C. Research Objectives

Motivated by these observations, the current work aims to achieve the following goals: (i) fine-tune a RoBERTa-base transformer on a class-balanced, 7-category corpus of 105,000 statements, as opposed to using handcrafted features or a lower-resolution set of labels; (ii) report performance in terms of per-class precision, recall, and F1-score, one-vs-rest ROC-AUC, and a confusion matrix-based error analysis, to obtain a more detailed description of the recognition performance than a single overall accuracy figure; (iii) compare the results to the classical approach of Juanita et al. [2] and similar works [14]–[16] using the same metrics; and (iv) provide a lightweight local web interface to the trained model, in order to demonstrate practical applicability beyond theoretical research.

II. METHODOLOGY

A. Dataset Description

The data used for this study comes from publicly available dataset on Kaggle, which contains aggregated and cleaned statements from various mental-health resources categorized under seven status tags: Anxiety, Bipolar, Depression, Normal, Personality Disorder, Stress, and Suicidal. The imbalanced nature of the original dataset motivated the use of only a subset of the whole corpus. Specifically, ten percent of each category — 105,000 statements in total, 15,000 per category — was sampled in accordance with a down-sampling strategy. Down-sampling was preferred over over-sampling techniques such as SMOTE [17], which were designed for structured, lower-dimensional feature spaces rather than raw text.

B. Data Preprocessing

The entries with missing statement text were removed, and the statements with fewer than six tokens were filtered out to minimize label noise caused by insufficient context in the input samples. This reduced the total number of balanced 105,000 statements to those that eventually constituted the training and validation sets (94,203 and 9,767), as described in Section IV-D. The resulting statements have been tokenized using RoBERTa's default byte-pair-encoding tokenizer, with each sequence padded or truncated to 128 tokens per input. The seven category names were then mapped to numeric labels via a label encoder fitted specifically on those names.

C. Model Architecture

RoBERTa-base — a 12-layer transformer encoder holding roughly 125 million parameters — forms the backbone of the classifier. Above the encoder is a linear classifier, which generates the final logits using the mapping between the pooled representation and labels of seven mental-health-related categories. Then, the logits will be transformed through a softmax layer that converts the final logits to probabilities over the labels, with the maximum probability used as the predicted result. Fig. 1 shows the overall system design, including the data acquisition, text preprocessing module, dataset pipeline, RoBERTa's embedding layer, transformer encoder and classification head, the training pipeline, and evaluation/deployment layer.

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) V \quad (3)$$

where Q, K, V belong to $d_k \times d_k$ matrices, are the query, key, and value projection matrices, respectively, and d_k is the key dimension; using attention in such a way over all twelve transformer layers allows the model to learn contextual information about the input tokens.

$$h = \text{RoBERTa}(x) \in \mathbb{R}^{768} \quad (4)$$

The pooled representation h corresponding to the sequence-start token is passed through a linear classification head with weight matrix $W \in \mathbb{R}^{7 \times 768}$ and bias $b \in \mathbb{R}^7$ to produce one logit per mental-health category, which softmax then converts into a probability distribution:

$$z = Wh + b, P(y = c | x) = \exp(z_c) / \sum_{j=1}^7 \exp(z_j) \quad (5)$$

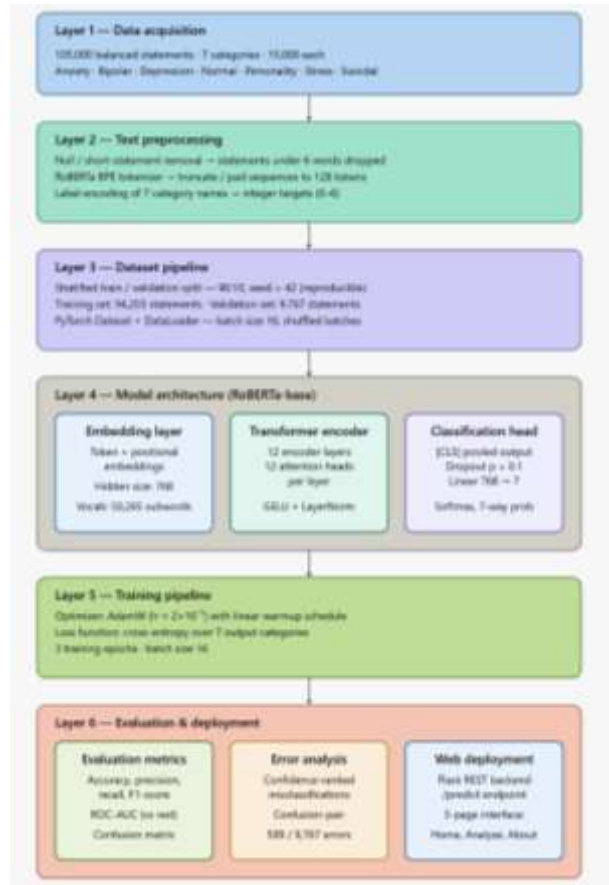


Fig. 1. Six-level scheme of the detailed study: data acquisition, pre-processing, dataset pipeline, RoBERTa model architecture, training pipeline, and evaluation/deployment.

D. Training Configuration

The data was divided into a training/testing set using a stratified split of 94203 training and 9767 validation samples (90.6%/9.4%), which was done according to the same class distribution (random seed: 42). Fine tuning of the model was performed using the AdamW optimization algorithm [18], a learning rate of $2e-5$, and a warmup + linear decay schedule, through PyTorch [19] and HuggingFace Transformers [20]. A batch size of 16 and three epochs with cross entropy loss were used to train the model:

$$\text{loss} = -(1/N) \sum_{i=1}^N \sum_{c=1}^7 y_{i,c} \log P\theta(y = c|x_i) \quad (6)$$

AdamW updates each model parameter ω using a decoupled weight-decay term λ in addition to the usual bias-corrected moment estimates:

$$\omega_{t+1} = \omega_t - \eta(\hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) + \lambda\omega_t) \quad (7)$$

E. Evaluation Metrics

Validation performance of the model was assessed in terms of overall accuracy and individual category-wise precision, recall and F1 scores using scikit-learn [21], following best practices in the interpretation of classification metrics for multi-class problems [22]. One-vs-rest approach was employed to calculate receiver operating characteristics (ROC) curves and area under the curve (AUC) values for each category separately. In addition, the confusion matrix along with the ranking of mistakes made by the model was constructed. The following formulas are used to calculate P_c , R_c , and $F1_c$ for each category c based on the number of true positive (TP_c), false positive (FP_c) and false negative (FN_c) cases:

$$P_c = \text{TP}_c / (\text{TP}_c + \text{FP}_c), R_c = \text{TP}_c / (\text{TP}_c + \text{FN}_c), F1_c = 2P_c R_c / (P_c + R_c) \quad (8)$$

F1 macro average assigns the same importance to all seven classes, while F1 weighted average takes into account support of each class n_c , and therefore, the classes having larger number of validation cases contribute more to the final metric:

$$F1_macro = (1/7) \sum_{c=1}^7 F1c, F1_weighted = \sum_{c=1}^7 (nc/N) F1c \quad (9)$$

On the contrary, ROC-AUC score is calculated for each class c in the framework of one-vs-rest classification scheme as the area under the curve defined by different decision thresholds τ and corresponding TPR and FPR for this category:

$$TPRc(\tau) = TPc(\tau) / (TPc(\tau) + FNC(\tau)), FPRc(\tau) = FPc(\tau) / (FPc(\tau) + TNc(\tau)) \quad (10)$$

SYSTEM IMPLEMENTATION

To demonstrate the practical usage of the trained model other than just testing, a simple web app was developed, that integrates the model as a service using the Flask framework. It implements a basic client-server architecture – upon launch, the server loads the fine-tuned RoBERTa weights and initializes the Flask backend that provides a single endpoint for inference prediction. The frontend is a simple HTML form with embedded CSS and JS that lets the user input their statement and see the resulting category and confidence for all 7 classes. The form submission URL is set to /predict, which takes the input text, tokenizes it, and runs the same prediction process as the model did during training. The received logits are then converted to probabilities using softmax and returned to the frontend as JSON. The page then displays the predicted label and a ranked bar chart of all seven probabilities. All of this happens on the local machine – no statement a user types ever leaves the device for an external server – which matters given how sensitive this kind of text can be.



Fig. 2. System interface for loading real held-out test examples directly from the validation set for evaluation.



Fig. 3. Example system output: a statement correctly classified as Anxiety with 99.92% confidence, shown alongside the full 7-category probability breakdown.



Fig. 4. Example system output: a statement correctly classified as Depression with 99.62% confidence.

RESULTS

F. Overall Performance

Using the test dataset which had 9,767 utterances, the performance of the RoBERTa model was measured at an accuracy of 93.87% in addition to the macro-averaged F1-score of 0.9408 and weighted-averaged F1-score of 0.9383. Fig. 5 shows this overall validation accuracy at a glance.

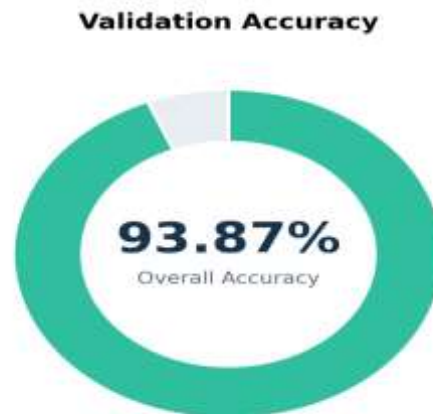


Fig. 5. Overall validation accuracy (93.87%).

G. Per-Category Performance

Table I demonstrates the precision, recall, F1-score, and support for all the categories. Except for two classes, all the other categories had an F1-score of 0.96 or higher, which proved a strong ability of the model to distinguish between different classes. In contrast, the F1-scores of the Depression (0.808) and Suicidal (0.839) classes are much lower compared to those for the other classes and will be elaborated upon in Section VI-E.

TABLE I. PER-CATEGORY CLASSIFICATION PERFORMANCE

Category	Precision	Recall	F1-Score	Support
Anxiety	0.9930	0.9937	0.9934	1,435
Bipolar	0.9877	0.9979	0.9928	1,449
Depression	0.8425	0.7767	0.8083	1,487
Normal	0.9825	0.9510	0.9665	1,122
Personality Disorder	0.9941	1.0000	0.9970	1,340
Stress	0.9776	1.0000	0.9887	1,443
Suicidal	0.8140	0.8659	0.8391	1,491



Fig. 6. Per-category F1-score, ranked, with macro-average shown as a dashed line.

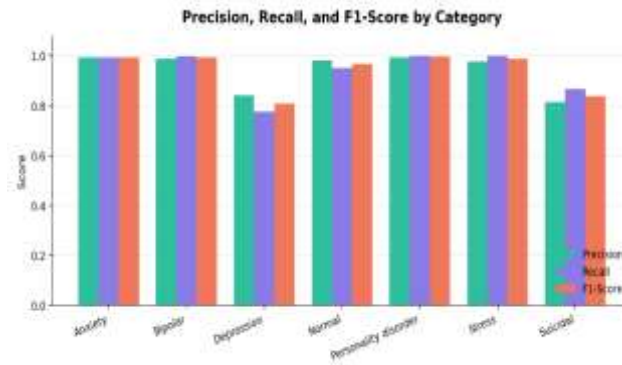


Fig. 7. Precision, Recall, and F1-Score comparison side by side for all seven classes.

H. Confusion Matrix Analysis

To see how predictions line up against true labels across all seven categories, a confusion matrix was generated (Fig. 8). In line with the per-category numbers in Table I, it shows that the great majority of categories are assigned with only a handful of off-diagonal (incorrect) predictions, whereas Depression and Suicidal show a clear cluster of mutual misclassification.

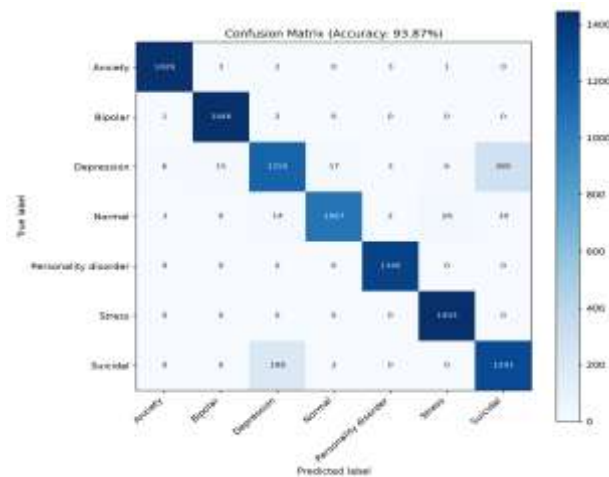


Fig. 8. Confusion matrix across all seven mental health categories.

I. ROC-AUC Analysis

To test the degree of separation between the classes regardless of decision threshold used, one-vs-rest ROC curves were drawn for all of the classes (Fig. 9). Classes of Anxiety, Bipolar, Personality Disorder, and Stress scored a perfect AUC value of 1.000, followed by Normal at 0.999; Depression and Suicidal lagged behind with 0.977 and 0.984, respectively. While both the classes of Depression and Suicidal have better performance compared to the random guessing, their lower AUC values and the wider curve compared to other classes support the results shown previously in F1-score and confusion matrix: Depression and Suicidal are two classes that create more difficulties for the classification process. The macro average AUC for all seven classes was 0.9943.

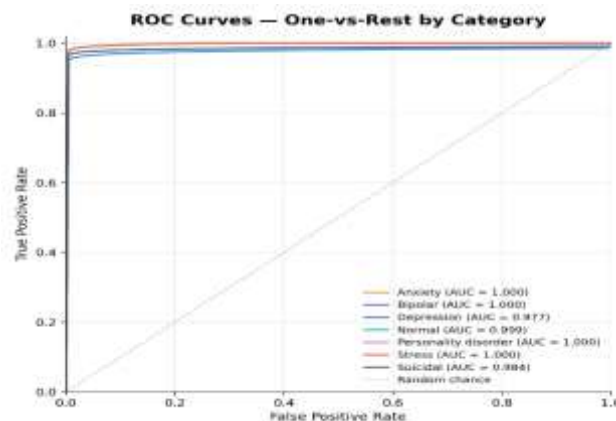


Fig. 9. One-vs-rest ROC curves with AUC score for each category.

J. Error Analysis

In all, out of the 9,767 validation statements analyzed, 599 (or 6.13%) were classified incorrectly. Further, when these errors are broken down in pairs, one can see how localized they actually are – 285 Depression statements were misclassified as Suicidal, while 198 Suicidal statements were misclassified as Depression statements. Combined, these two misclassifications amount to 483 statements (out of 599 total errors), making up some four out of every five errors (80.6%). All the rest of the possible confusion pairs (such as Normal/Stress, Depression/Bipolar, Normal/Depression, etc.) make less than 30 errors each. Instead of being dispersed across multiple categories, then, the mistakes made by the model seem to be localized to one specific, linguistically motivated confusion pair – statements about hopelessness, worthlessness, and self-criticism are used in both Depression and Suicidal cases, with the only difference being that of degree or stated intent.

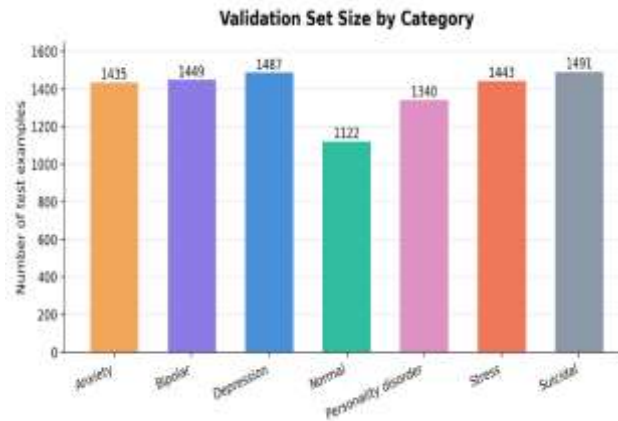


Fig. 10. Number of validation examples per category.

COMPARISON WITH BASE STUDY

Table II compares the proposed approach against four related studies that classify mental-health-related text: the classical machine learning study by Juanita et al. [2], which serves as the principal baseline for this work, alongside three further transformer-based studies — Ameer et al. [14], Sánchez Rodríguez et al. [15], and Verma et al. [16] — that were identified as covering a similar task on comparable social-media or clinical text. It is important to note that none of these studies is perfectly equivalent to the present work: they differ in category count (from three to seven), in the specific labels used even when the count matches (Ameer et al. and Verma et al. include ADHD and/or PTSD rather than Personality Disorder, Stress, and Suicidal), source platform (Indonesian-language clinical consultations versus English-language Reddit posts), and label granularity (doctor-authored versus patient/user-authored text). Nonetheless, taken together they offer a broader point of reference for evaluating the relative benefit of transformer-based fine-tuning over both classical feature-engineered pipelines and other transformer-based approaches on related mental-health text classification tasks.

TABLE II. ACCURACY COMPARISON WITH BASE STUDY

Study	Model	Categories	Accuracy
Juanita et al. [2]	Gradient Boosting + N-Gram	3	84.20%
Ameer et al. [14]	Fine-tuned RoBERTa	5	83.00%
Sánchez Rodríguez et al. [15]	BERT / MentalBERT	4	82.00%
Verma et al. [16]	Fine-tuned BERT + sentiment fusion	7	91.70%
This Study	Fine-tuned RoBERTa-base	7	93.87%

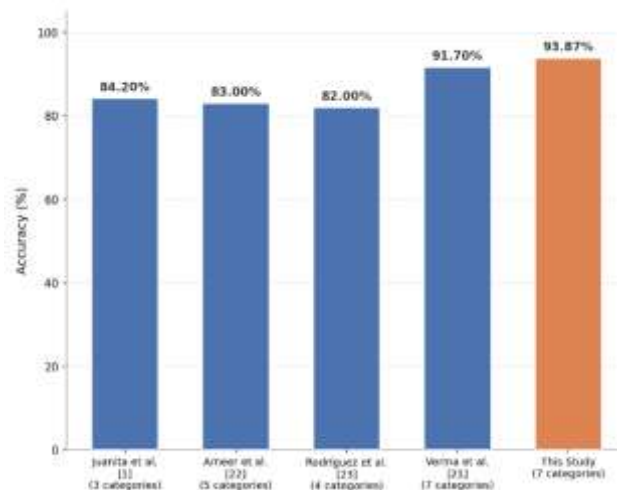


Fig. 11. Accuracy comparison across the base study and three additional related studies.

Even while handling more than twice as many categories, the transformer-based approach proposed here comes out 9.67 percentage points ahead of the base study's strongest classical model in accuracy, a gap that fits a pattern seen elsewhere in NLP research: contextual representations learned by transformers generally beat hand-engineered, context-blind features like N-Gram and Bag-of-Words, especially on tasks that hinge on fine-grained semantic distinctions between classes. The comparison against other transformer-based studies is more nuanced. Ameer et al. [14] and Sánchez Rodríguez et al. [15] fine-tune RoBERTa and BERT/MentalBERT, respectively, on Reddit posts spanning four or five mental-health categories that only partly overlap with the present label set (Ameer et al.'s five categories share only Anxiety, Bipolar, and Depression with this study, substituting ADHD, PTSD, and a None class for Personality Disorder, Stress, and Suicidal), and report accuracies of 83.00% and roughly 82%; the present work still comes out ahead by 10–12 percentage points despite covering a comparably sized or larger label space, though the underlying datasets, label sources, and category definitions differ enough that this gap should be read as suggestive rather than a controlled comparison. Verma et al. [16], by contrast, fine-tune BERT on a Reddit corpus spanning seven categories close in number to those used here, additionally folding in sentiment features and severity levels to reach 91.70% accuracy — the closest result to this study's 93.87%. Their label set substitutes PTSD for Personality Disorder, so the two studies are not working from an identical seven-class definition; even so, the narrower remaining gap is a useful reminder that once category count and modeling approach are broadly comparable, the choice of backbone architecture and auxiliary features accounts for a smaller share of the difference than the gap against classical, feature-engineered baselines. In general, the trend seen in all four comparison experiments is supported by the larger body of research, whereby transformer fine-tuning is better than classical feature engineering, but differences between transformer models tend to be more affected by the makeup of the dataset, categorizations used, and auxiliary information rather than the specific architecture used.

CONCLUSION AND FUTURE WORK

The goal of this paper was to sort written statements into seven mental-health categories by fine-tuning RoBERTa-base on a balanced pool of 105,000 labeled statements, and the results back that approach up: 93.87% accuracy and a macro-F1 of 0.9408 on held-out data, with near-perfect scores across five of the seven classes. What errors remain are not scattered at random — they concentrate almost entirely on distinguishing Depression from Suicidal statements, two categories whose wording genuinely overlaps, rather than pointing to any broad weakness in the model. Compared against an earlier classical-ML study built around a coarser, three-category version of this task, the method presented here reaches higher accuracy while covering a substantially larger label space, and a Flask-based demo confirms the model can run locally in real time.

Future work could address the Depression–Suicidal confusion directly, for example through a hierarchical classification scheme that first distinguishes severity level before assigning a specific category, or through additional training data that more explicitly captures the linguistic markers of suicidal intent. Other promising directions include validating the model on clinically sourced or multilingual text rather than social-media-style statements, incorporating attention-based explainability techniques to improve clinical trust and interpretability, and evaluating ensemble approaches that combine transformer-based and classical models. Finally, given the sensitive nature of this application domain, future deployments should undergo appropriate clinical validation and ethical review before any use beyond research and educational demonstration.

REFERENCES

- [1] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.
- [2] S. Juanita, A. H. Daeli, M. Syafrullah, W. Anggraeni, and M. H. Purnomo, "A machine learning approach for text pattern diagnosis in mental health consultations," *Decision Analytics Journal*, vol. 15, p. 100572, 2025, doi: 10.1016/j.dajour.2025.100572.
- [3] A. Le Glaz et al., "Machine learning and natural language processing in mental health: Systematic review," *J. Med. Internet Res.*, vol. 23, no. 5, p. e15708, 2021, doi: 10.2196/15708.
- [4] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *npj Digital Medicine*, vol. 3, p. 43, 2020, doi: 10.1038/s41746-020-0233-7.
- [5] J. Chung and J. Teo, "Mental health prediction using machine learning: taxonomy, applications, and challenges," *Applied Computational Intelligence and Soft Computing*, vol. 2022, p. 9970363, 2022, doi: 10.1155/2022/9970363.
- [6] U. M. Haque, E. Kabir, and R. Khanam, "Detection of child depression using machine learning methods," *PLoS ONE*, vol. 16, no. 12, p. e0261131, 2021, doi: 10.1371/journal.pone.0261131.
- [7] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions," in *Proc. 27th Int. Conf. Computational Linguistics (COLING)*, 2018, pp. 1485–1497. [Online]. Available: <https://aclanthology.org/C18-1126/>
- [8] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 51–60, doi: 10.3115/v1/W14-3207.
- [9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. on Learning Representations (ICLR) Workshop*, 2013, doi: 10.48550/arXiv.1301.3781.
- [10] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
- [11] A. Vaswani et al., "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008, doi: 10.48550/arXiv.1706.03762.
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [13] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint*, 2019, doi: 10.48550/arXiv.1907.11692.
- [14] I. Ameer, M. Arif, G. Sidorov, H. Gómez-Adorno, and A. Gelbukh, "Mental illness classification on social media texts using deep learning and transfer learning," *arXiv preprint arXiv:2207.01012*, 2022, doi: 10.48550/arXiv.2207.01012.
- [15] I. Sánchez Rodríguez, J. Bianchi, F. Pinelli, F. Panizza, E. Ricciardi, P. Pietrini, and M. Petrocchi, "Understanding mental health discourse on Reddit with transformers and explainability," *Scientific Reports*, vol. 16, art. 6796, 2026, doi: 10.1038/s41598-026-35918-3.
- [16] I. Verma, S. Tiwari, L. Jain, and A. Chandra, "MINDMETRICS — Mental health text classification using BERT with sentiment fusion and explainable AI," *International Journal on Advanced Computer Engineering and Communication Technology*, vol. 15, no. 1, pp. 318–324, 2026, doi: 10.65521/ijacect.v15i1.2571.
- [17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019, doi: 10.48550/arXiv.1711.05101.
- [19] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 8024–8035, doi: 10.48550/arXiv.1912.01703.
- [20] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP: System Demonstrations*, 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.
- [21] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [22] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.