Transformer Framework for Early Mental Health Risk Detection from Multi-Platform Text with Resource Recommendations

Madhu C S[1], Disha Jaipal[2], Ganavi R[3], Debojeet Paul[4] and Gautham R[5]

[1 2 3 4 5] Department of CSE, Don Bosco Institute of Technology, Bengaluru, Karnataka, India

Abstract. Our paper explores transformer-based techniques used for early detection of mental health risks from social media text. With the increasing trend of individuals sharing their emotional and psychological distress online, researchers have come to use natural language processing (NLP) and deep learning to identify signs of issues such as depression, anxiety, and anger. We focus on reviewing the papers using model BERT and its variants for pretrained transformers paired with traditional machine learning classifiers, to understand their methodologies, effectiveness and challenges. Building these insights from the literature, we propose a hybrid transformer-based framework that integrates contextual embeddings with machine learning classifiers for improved emotional state detection. The proposed design also includes a severity grading mechanism and a personalized recommendation engine that suggests relevant resources based on the intensity of detected distress. This approach aims to support early, scalable, and mental health interventions, moving beyond detection toward a meaningful digital support for individuals at risk.

Keywords: Mental Health Detection, Transformer Models, Contextual Embeddings, Emotional State Classification

1 Introduction

Declining Mental health is a global concern. Despite a large number of people suffering, there is still stigma, misinformation, ignorance and limited access to professional support continue to prevent many individuals from seeking timely help. As psychological distress, anxiety, and mood disorders become more prevalent leading to violence on in online space and repercussions of this growing unchecked issue in online spaces. This has created a growing need for systems that offer proactive, compassionate, and context-aware intervention. Social media platforms like Twitter and Reddit have become online arenas where individuals share their thoughts, feelings, and everyday experiences on a regular basis. This stream of content generated by users offers great insight into personal states of mind and overall emotional patterns. Yet, despite some websites using automated software to identify possible distress or mark dangerous content for removal, their interventions are frequently confined to simple censorship or bland support messages. Such interventions, while having good intentions, fall short of the contextual understanding and individualization to effectively support individuals in distress. Recent advancements in Natural Language Processing (NLP), particularly through transformer-based models such as BERT and RoBERTa, have shown substantial potential in extracting complex emotional cues and detecting mental health risks from online text.

But achieving consistent accuracy in these models is a challenge, for this growing necessity and due to the nuanced, layered nature of emotional expression on social media. Emotionally rich and context-sensitive content often requires models to interpret not just isolated words or phrases but their meaning within temporal, behavioral, and situational contexts. A multi-layered model for the detection of emotions, that utilizes a variety of linguistic, context, and behavioral cues, promises to improve the accuracy and reliability of such predictions. Such accurate predictions are critical to developing systems that can provide the appropriate type of support to the appropriate person at the appropriate time. To explore this potential, this survey reviews recent research that leverages transformer-based architectures for mental health risk detection from social media. By synthesizing findings from existing studies, we aim to shed light on the strategies that improve emotional recognition, explore the role of multimodal and context-aware modeling, and assess the feasibility of building more reliable and responsive mental health support tools.

2. Related Works

2.1 BERT-RF for Depression Detection

Methodology

The study employed a benchmark dataset comprising 20,000 English-language Twitter user profiles labeled as either depressed or non-depressed [6]. To prepare the data, standard preprocessing techniques were applied, including the removal of stop words using the NLTK library, text normalization, and stemming to reduce words to their root forms.

The tweets were then converted into indexed token sequences and passed through a pre-trained BERT model to extract contextualized embeddings while preserving semantic relationships [6]. Rather than utilizing the high-dimensional BERT embeddings directly for classification, the authors introduced an intermediate feature engineering step.

A Random Forest classifier was used to generate probabilistic meta-features from the embeddings, effectively enhancing feature discrimination. These engineered features were subsequently used to train five different classifiers: Random Forest (baseline), Multilayer Perceptron (MLP), K-Nearest Neighbors Classifier (KNC), Logistic Regression (LR), and Long Short-Term Memory (LSTM) [6]. Among these, Logistic Regression achieved the highest performance, with an accuracy, precision, recall, and F1-score of 99%. The evaluation strategy included an 80:20 train-test split, as well as k-fold cross-validation to ensure result robustness [6].

Compared to using BERT embeddings alone, which resulted in significantly lower classification performance (like, LR F1-score = 0.56)—the BERT-RF hybrid framework demonstrated substantial performance improvements, showcasing the effectiveness of integrating contextual embeddings with probabilistic feature transformation.

Challenges

The model essentially relies on identifying certain linguistic patterns that signal melancholy. Still, it's important to recognize that people often present themselves differently in casual versus formal environments—one's behavior in a professional context might not match how they express themselves in private or on social media. This complicates the interpretation of melancholic cues and raises questions about the model's generalizability across contexts [6].

Ethical considerations are significant here, particularly in terms of mental health profiling on social media. Issues such as user consent and privacy must be addressed thoroughly to ensure responsible use of such sensitive information [6]

Additionally, since the approach leverages BERT-based embeddings, it's crucial to further scrutinize the clarity and reliability of the model's predictions. Without careful evaluation, there's a risk of misinterpretation or overconfidence in the results [6].

2.2 Hugging Face BERT Models for Depression Detection

This paper presents a new approach to tackle problems in current mental health diagnosis by combining cutting-edge transformer models with real-world social media information [4]. The researchers use public Twitter content, including tweets and bios, to train and adapt pre-trained BERT models to spot depression. The research compares transformer-based models from Hugging Face with traditional machine learning methods. The goal is to find out which approach gives the most accurate and efficient way to predict mental health risks, while setting the stage for more effective and scalable Approaches.

Methodology

This study employed four pre-trained transformer models from the Hugging Face library: distilbert-base-uncased- finetuned-sst-2-english (DBUFS2E), bert-base-uncased (BBU), mental-bert-base-uncased (MBBU), and distilroberta-base (DRB). These models had been previously fine-tuned on large corpora including reviews, tweets, and various other text sources [4] The dataset used was Autodep, automatically collected via the Twitter API. It included over 11.8 million tweets and 553 bio-descriptions, representing users who had disclosed their mental health struggles [4]. Data preprocessing was started with tweets and bios first filtered for English text and stripped of retweets, mentions, URLs, emojis, and special characters. Followed by the text being converted to lowercase and extra spaces were removed. Tokenization of the lowercased text was performed using the NLTK library, with common stop words were removed using NLTK's built-in list. And lemmatization was applied via the WordNetLemmatizer to standardize inflected words into their base forms. Additionally, stemming was tested, although lemmatization produced slightly better outcomes [4] Following preprocessing, each dataset (tweets and bios) was divided into training and testing subsets using 10-fold cross-validation. Each fold was trained for five epochs using Hugging Face's Trainer method. Model-specific tokenizers were utilized to convert processed text into

sequences of tokens that could be used as model inputs. The training process was optimized to improve models for binary classification, i.e., distinguishing between depressed and non-depressed classes. Performance was measured based on accuracy, F1 score, and AUC metrics, which were computed during training through an evaluation function [4]. These measurements allowed us to gauge precision, recall, and how well the model performed overall in classification. We developed the final depression detection models by fine-tuning these BERT variants on the cleaned Twitter data. DBUFS2E outperformed all other models reaching an AUC of 0.98 on tweets and 0.96 on bios. This shows it has a strong ability to pick up on signs related to depression [4].

Challenges

It should be noted that additional user data, including images, emojis, and hashtags, can enhance predictions. Future research can explore combining data types to improve the performance of the models.

2.3 Emotional Mental Intelligence- Digital Twin Framework

In the paper, they aim to address the barriers of stigma, accessibility, and affordability in mental healthcare by designing and developing a dialogue system that analyses individuals' mental status.[3] They built a conversational AI chatbot that can detect the severity of depression and mental health symptoms using a digital twin–inspired framework, improving early detection, access, and personalization in mental healthcare.

Methodology

The study used the Extended Distress Analysis Interview Corpus (E-DAIC), a semi-clinical corpus of 219 transcripts of interviews between a virtual interviewer, "Ellie," and patients with varying levels of depression, classified based on PHQ-8 score.
[3] The data contained transcriptions of verbal conversation and metadata like PHQ scores, gender, and history of mental illness.
Preprocessing involved choosing interview statements of over 50 characters and their classification into five degrees of depression severity using PHQ-8 thresholds. In the interest of classification efficiency, these were subsequently combined into three classes: no symptoms, mild, and moderate/severe depression [3]. For solving the problem of data imbalance, the authors used contextual data augmentation strategies involving bi-directional language models and BERT-style embeddings to create diversified training samples of less prominent classes, specifically [3].
The cleaned and tokenized text data was mapped to token embeddings using a pre-trained BERT model. The contextual embeddings were fed into a custom classification layer, which produced probability distributions over the three depression severity classes [3] The model was embedded inside the Rasa framework, which controlled both intent classification and dialogue flow using BERT-based NLU components and fallbacks [3].

Model training was base in an 85:15 train-validation split. BERT was trained using Adam W optimizer (learning rate = 1e−5), with training/validation batch sizes of 4 and 32, and a soft max activation to produce class probabilities [3]. In ambiguous user input, a fallback classifier was utilized to invoke default responses. This providing robustness in case of ambiguity [3].
The testing was carried out with baseline measures, which are accuracy, precision, recall, and F1-score. The authors also created specialized rules and dialogue "stories" to test response generation and intent routing along different conversational paths [3].
In analysis of the E-DAIC dataset, the model had a cumulative accuracy of 69% with moderate F1-scores for all three severity levels [3]. This was superior to that of comparable applications, such as Ada Health, with a top-5 diagnosis accuracy of 51% only [3].
In 20 university student field trials, EMI chatbot attained 40% top-1 and 65% top-2 accuracy for depression severity prediction against PHQ-9 scores [3] Participants gave the chatbot a positive usability rating with a mean System Usability Scale (SUS) score of 84.75%—well above the industry benchmark of 68% [3].

Challenges

The model's classification accuracy was constrained by the limited and imbalanced dataset, as well as the lack of multimodal emotional cues such as facial expression and voice tone, which the chatbot could not interpret [3]. The sample size in real-user testing was also small (n=20), limiting statistical generalizability [3].

2.4 Transformer Calibration for Stress and Depression

The primary aim of this study is to improve the reliability and performance of transformer-based models for identifying stress and depression in social media posts by injecting linguistic features into BERT/Mental BERT models and applying label smoothing to calibrate the models' prediction confidence. This is the first study to jointly address both classification performance and calibration for mental health detection from text, while also incorporating novel linguistic cues like LIWC, Top2Vec, and GOSS features derived from LDA topic modelling [2].

Methodology

Three benchmark datasets, Dreaddit(stressful v. non-stressful posts), Depression-Mixed (depressive v. non-depressive), and Depression-Severity (minimum, mild, moderate, severe depression), were downloaded from Reddit and other similar discussion boards [2].

Preprocessing consisted of tokenization of input posts and feature extraction of different extra-linguistic features. These were:
(i)

(i) emotion-based ratings based on the NRC lexicon (10-dim), (ii) LIWC 2022 features capturing psychological and linguistic word categories (117-dim), (iii) 25-dimensional LDA topic probabilities with corresponding GOSS scores of topic deviation, and
(iv) 512-dimensional Top2Vec semantic embeddings [2].
The preprocessed text was then run through either BERT or Mental BERT to obtain contextual token embeddings. These were then combined with the linguistic feature vectors through a multimodal adaptation gate followed by shifting. This gave enhanced word embeddings, which were then normalized and fed into the transformer models for final classification [2].
For model calibration, the authors used label smoothing, a method that alters the hard target labels by adding small probability mass to all classes, assisting in overconfidence reduction and uncertainty estimation improvement [2] The ultimate architecture, referred to as M-BERT or M- Mental BERT (e.g., M-BERT(LIWC)), was made up of the transformer encoder with a dense ReLU layer and soft max output [2].
Training was performed with the Adam optimizer (learning rate = 0.001), Step LR scheduler, and batch size of 8. Depression-Severity used early stopping and fivefold cross-validation, whereas 80:20 split was performed on the other sets. The models were trained for a maximum of 30 epochs on a Tesla P100 GPU using the PyTorch and Transformers library [2].
Performance was measured with standard metrics: accuracy, precision, recall, and F1-score. Calibration was measured with Expected Calibration Error (ECE) and Adaptive Calibration Error (ACE), comparing predicted confidence to accuracy in bins [2].

Challenges

Although the given M-BERT model was effective, certain limitations of the model make it less useful in real-world applications. The lack of explainability functionalities like SHAP and attention visualization limiting trust and usability in clinical settings; the current study plans to overcome these limitations by integrating explainable AI functionalities and clearer user feedback systems [2].
Additionally, the model was trained with little hyperparameter tuning and without repeated statistical testing on two datasets, which can affect generalizability. Our project mitigates this by systematic tuning and multiple rounds of evaluation [2].
Finally, while linguistic features such as NRC and LIWC were utilized, their varying impacts suggest the need for more user-specific, personalized features—an area our framework addresses by incorporating personalized input cues and adaptive feedback [2].

2.5 MIRoBERTa: Mental illness text classification

This work is targeted at enhancing the contextualization and efficiency of transformer-based models in the area of mental health screening on social media. To this aim, the authors propose domain-adaptive versions of two pre-trained models BERT and RoBERTa by further training them on Reddit data pertaining to mental health disorders.

These trained models, or MIBERT and MIRoBERTa, are designed to make more accurate classifications in the multi-class case for many mental disorders such as depression, anxiety, OCD, PTSD, and others [5].

Methodology

The data utilized in this research was gathered from 11 mental health subreddits such as r/depression, r/anxiety, r/OCD, and r/ADHD, using the Reddit Pull-Push API. At least 10 upvotes from posts were used to guarantee relevance and interest of the posts. The dataset eventually comprised more than 54,000 labeled posts, and an additional 8,207 posts labeled as "none" (not mental health) [5].

During the preprocessing, the authors applied standard cleaning techniques: lowercasing, stop word removal, and lemmatization. Surprisingly, the authors preserved stop words since transformer models can find meaning in stop words based on their contextual embeddings [5].
The article contrasted conventional and deep learning approaches (e.g., TF-IDF with SVM, BiLSTM with GloVe) but solely used transformer models. Apart from the conventional pretrained BERT and RoBERTa, the authors pretrained their domain-specific counterparts—MIBERT and MIRoBERTa—on Reddit mental health text with Masked Language Modelling (MLM) to learn domain-specific patterns. This was followed by fine-tuning on the target task [5].
Fine-tuning employed the AdamW optimizer with learning rates ranging from 1e-5 to 3e-5, a batch size of 16, and three epochs of training. Pretraining RoBERTa and MIRoBERTa took about 9 hours and pretraining BERT variants took about 8.5 hours. The experiments were conducted on the Hugging Face Transformers library [5].
Model performance was measured with standard metrics: accuracy, precision, recall, and F1-score (all weighted to address class imbalance). For better interpretability, SHAP (SHapley Additive exPlanations) was applied to find out which words impacted each model's predictions most [5].
The domain-adaptive models outperformed all the baseline comparisons. Out of the individual models, MIRoBERTa worked the best with accuracy and F1-score of 0.847 across a learning rate of 1e-5. The best overall performance was achieved using an ensemble of MIRoBERTa and RoBERTa with accuracy, precision, recall, F1-score of 0.851 respectively [5].
In per-class classification, MIRoBERTa had the best recall (0.841) for the anxiety class and RoBERTa ensemble had an F1-score of 0.804 for depression classification [5].

Challenges

Though the MIRoBERTa model achieves considerable quantitative performance, the prevailing limitations are ascertained by the authors. Failure to distinguish between co-occurring mental health syndromes (anxiety and depression) and no clinical validation for the labeling of subreddits taints the model's validity [5]. Second, although SHAP

was used to provide some interpretability, the model is not adaptive to incorporate user feedback or personalized learning aspects—areas in which our project seeks to bring innovation. Lastly, the absence of multilingual or multilabel extensions limits its scalability across heterogeneous user groups [5].

2.6 Depression Detection on Social Media Posts using BERT and Support Vector Machine

The objective of the current research is to evaluate the effectiveness and feasibility of employing BERT embeddings with Support Vector Machine (SVM) classifiers for depression detection in social media content. By analyzing the effect of different SVM kernel functions on Reddit and Twitter-based datasets, the study aims to establish the best configuration for depression detection using text-based indicators. The authors point out that this approach provides a computationally less demanding solution than deep learning methods, capitalizing on the strong contextual embeddings of BERT [7].

Methodology

Two publicly available datasets were used: one from Reddit and one Twitter, both of which were obtained through Kaggle. The Reddit dataset contains around 7,000 English language posts from subreddits with depression as the topic, and the Twitter dataset contains labeled tweets (depressed, non-depressed) with a natural class imbalance towards tweets that are non-depressed [7].
Data preprocessing was done with NLTK through typical cleaning processes: removal of URLs, mentions, hashtags, lowercasing, and stop word removal. All the above were done to achieve textual consistency prior to tokenization and embedding [7].
Text that had been preprocessed was tokenized with BERT's tokenizer with a sequence length of 128 tokens. The bert-base-uncased model was employed in order to get contextual embeddings by averaging the output of the hidden state of each token, which provided a 768-dimensional vector per post. These vectors were utilized as SVM features [7].
There were four various SVM kernel functions considered:
• Linear: a default for linearly separable data
• Polynomial: to accommodate complex, non-linear relationships
• Radial Basis Function (RBF): used in mapping to higher-dimensional spaces.
• Sigmoid: approximates neural network performance, but more vulnerable to hyperparameters
The research focused on the tuning of crucial hyperparameters, including the regularization parameter and kernel-specific gamma values, to achieve high performance [7].
The models were compared based on accuracy, precision, recall, and F1-score, which gave a balanced estimation of classification performance for imbalanced classes [7].

The testing showed that Polynomial and RBF kernels worked better than others on both platforms. On the Twitter dataset, both kernels had an accuracy of 87.5%, higher than linear (85.5%) and sigmoid (76.4%) setups [7]. On the Reddit dataset, the Polynomial kernel had the highest accuracy of 94.4%, followed closely by RBF at 94.2%, linear at 93.5%, and sigmoid at 93.3% [7].
The results show the language used for depression on Reddit is more explicit and allows for classification more accurately. The tweets are briefer and more ambiguous and thus are slightly harder to detect [7].

Challenges

Although the BERT + SVM model produced solid performance, the paper had some limitations. Firstly, it was purely English-language text-based, and therefore restricting model applicability across languages [7]. Secondly, it was noisy datasets, especially the Twitter dataset, which was not manually checked and had imbalanced samples—something that could potentially affect the results' generalizability [7]. Most significantly, the model does not have any explainability mechanisms, i.e., SHAP, attention-based visualization, which are particularly important in mental health use cases to provide interpretability and ethical usage. Our project seeks to address this by using explainable AI methods and exploring multilingual and multimodal data (e.g., audio, images) to enhance robustness [7].

2.7 Enhanced word embedding

The primary objective of this research is to improve the identification of mental health disorders and alcohol/substance abuse problems, such as depression, suicidal thoughts, eating disorders, and alcoholism, through the application of word embeddings from specific domains to social media text. The authors seek to maximize the semantic space so that words from specific classes are semantically close to condition-specific anchor words (or pivot words) but well-distanced from noise words. This is intended to enhance classification performance in small, domain-specific corpora through the creation of embeddings that more accurately capture language associated with specific conditions [1].

Methodology

The study utilized Reddit data from five condition-related subreddits-r/depression,----r/Suicide Watch, r/alcoholism, r/eating disorders, and r/bulimia—alongside 18 general- interest subreddits like r/books, r/science, and r/law to serve as control data [1]. Only comments containing self-referential material were selected, and a filtering mechanism was used to remove those with co-morbid conditions in an attempt to maintain the labels intact [1].
Text preprocessing consisted of stop words removal and predictor bigrams being combined into a single token to preserve semantic intent (e.g., keeping "mental health" as a single token). The researchers identified each category's predictive words

through statistical tests—i.e., Chi-squared and Mann-Whitney U—thus creating term pairs and discovering pivot terms that portrayed each condition [1]. These terms guided the creation of enhanced word embeddings by modifying the training objective so that semantically related words were pulled toward the pivot vector and semantically unrelated words were pushed away.

Several variations of embeddings were evaluated, including basic GloVe, enhanced GloVe, GloVe + retrofitting, and concatenated models.For the classification task, four models were used: logistic regression, random forest, a CNN-based deep learning model (DL1), and Distil BERT. Surprisingly, the best performance came from the CNN model with enhanced GloVe and retrofitting (Embedding Model 4) [1].
Training was conducted using a 70:30 train-test split, with hyperparameter tuning applied via grid search and fivefold cross-validation for the traditional models. The convolutional neural network (CNN) model employed multiple filter sizes (2, 3, and 5), max pooling layers, and softmax output. Accuracy, precision, recall, and F1-score were used as the evaluation metrics. The authors

also employed cosine similarity and PCA-based visualization to examine the semantic clustering behavior of the embeddings in the vector space [1].
The enhanced embeddings reflected a dramatic superiority compared to standard baseline approaches. In particular, Embedding Model 4, GloVe-initialized, retrofitted, and augmented, recorded an F1-score of 86% for multi-class classification with the CNN model, surpassing normal Word2Vec, GloVe, and even Distil BERT by as much as 13.97% in F1 score [1]. Furthermore, the analysis of cosine similarity confirmed that the upgraded embeddings resulted in more cohesive intra-class clustering and more distinct inter-class differentiation [1].

Challenges

Despite its strong performance, the study presents several limitations that affect its applicability to real-world mental health detection. First, although the enhanced embeddings improved semantic representation for class-specific language, the model lacks interpretability mechanisms such as SHAP values or attention-based visualizations, which are essential in clinical and sensitive decision-making environments [1]. Second, the dataset was explicitly filtered to exclude comorbid posts—i.e., instances where users exhibited multiple mental health conditions—to maintain label purity. This reduces the model's generalizability, especially given that comorbidity is common in real mental health scenarios [1]. Third, the embedding-based approach used in the study is static; it does not utilize contextualized word representations from transformer models like BERT or RoBERTa, which limits its capacity to capture nuanced meanings based on context. Finally, the models operate on fixed embeddings and lack mechanisms for user-specific adaptation or continuous learning, which may hinder their effectiveness in evolving or personalized mental health applications.

3. Discussions

The current body of research clearly points to an increasing reliance on transformer-based models for detecting mental health risks within social media data. Most studies indicate that contextual embeddings derived from models such as BERT or RoBERTa consistently deliver strong performance in text classification tasks related to this domain. [4], [6], [7]. Abbas et al. [6], for instance, introduced a BERT-RF feature engineering system that performed better than baseline classifiers with only BERT by considering probabilistic features, with an F1-score of 0.99 on logistic regression. This system had a significant effect on the design of our system, in particular on the blending of contextual representations and traditional machine learning classifiers.
However, most other research focused only on primary classification without the inclusion of modular structures or personalization aspects. For example, Shoaib et al. [7] combined BERT embeddings with support vector machine classifiers and experimented with kernel functions in Twitter and Reddit contexts; however, their model did not comprise behavioral analysis or user-based feedback systems. Similarly, Pourkeyvan et al. [4] performed benchmarks on Hugging Face models (e.g., DistilBERT and MentalBERT) on Twitter tweets and bios without performing any follow-up severity measures or customized support systems.
A few like MIRoBERTa [5] and the Digital Twin system [3] have addressed domain adaptation and personalization to some degree. MIRoBERTa pretrained in a domain-specific manner using Reddit mental health subreddits with better classification performance. However, the absence of multilingual capabilities and the lack of extensive personalization (e.g., user feedback loops) restrict its adaptability. Abilkaiyrkyzy et al.[3]. implemented a Digital Twin chatbot with a secondary conversational layer for severity estimation but with the limitation that it was founded on dialogue data (E-DAIC) and not ideal for passive social media analysis.
Some other papers emphasized improving model reliability or feature quality. Ilias et al. [2] integrated calibration techniques to ensure predictions are more reliable in MentalBERT models based on other linguistic and topical features. Ramírez-Cifuentes et al. [1] also experimented with improving word embeddings using Reddit data and showed improvement in substance abuse and suicidal ideation detection. These methods did not incorporate multi-label classification or platform-adaptive behavior monitoring, though.
Three further studies were scanned for completeness but not incorporated into our system directly. Tay et al. [13] conducted a general transformer classification survey of scalability, accuracy, and cost across NLP tasks but not mental health or social media posts. Aina and Fatumo [14] developed a hybrid vision-language system for emotion detection from faces using YOLOv8 and Vision Transformers, orthogonal to our text-based system. Finally, Pattun and Kumar [15] used GPT-3.5 embeddings for generic emotion classification using the ISEAR dataset. Their pipeline worked well, but it did not have the domain grounding, platform integration, and severity-sensitive response mechanisms needed for early mental health risk detection.
In short, while the majority of research papers offer valuable insights towards model construction, feature engineering, or dataset usage, not many systems bring together multi-platform input, emotional severity evaluation, and customized resource supply to a unified platform. This contribution builds on these set foundations and mitigates certain shortcomings by presenting a scalable, transformer-based hybrid system that connects detection with action.

The following table summarizes and contrasts the methodologies, datasets, and results from various transformer-based models and frameworks discussed in the literature for mental health detection on social media.

Study/ Model	Technique / Framework	Dataset Used	Key Features	Best Accuracy / F1-Score	Limitations	Reference
BERT-RF	BERT embeddings + Random	Twitter depression dataset	Preprocessing + BERT → RF-	99% (Logistic	Limited to formal language; lacks	[6]

	Forest meta-features	(20,000 user profiles)	generated features → traditional ML classifiers	Regression)	interpretability and multimodal cues	
Hugging Face BERT variants	DBUFS2E, BBU, MBBU, DRB transformers	Twitter (11.8M tweets + 553 bios)	Fine-tuned BERT variants + cross-validation	AUC: 0.98 (tweets)	Ignores non-textual cues like emojis, hashtags, and user images	[4]
EMI Digital Twin	Conversational AI chatbot + Rasa framework	E-DAIC interviews (PHQ-8 labeled)	BERT embeddings + severity classification + chatbot dialog system	69% (E-DAIC), SUS: 84.75	Small user testing size; lacks multimodal feedback (e.g., tone, facial expressions)	[3]
Transformer Calibration (M-BERT)	Mental BERT + LIWC, GOSS, Top2Vec features	Reddit (3 benchmark datasets)	Label smoothing + linguistic fusion + calibration	High F1 (varies)	No explainability, limited hyperparameter tuning, no user personalization	[2]
MIRoBERTa	Domain-adapted RoBERTa trained on Reddit	11 mental health subreddits	MLM pretraining + SHAP-based interpretability	Accuracy/F1: 0.851 (ensemble)	Overlapping symptoms hard to classify; lacks user feedback & multilingual support	[5]
BERT+ SVM	BERT embeddings + SVM (Polynomial, RBF)	Reddit + Twitter (Kaggle)	768-dim BERT embeddings → SVM with kernel tuning	Reddit: 94.4%, Twitter: 87.5%	No explainability; English-only and noisy datasets	[7]
Enhanced Word Embedding	GloVe + Retrofitting + Class-pivoted enhancement	Reddit (mental health + control subreddits)	Custom embeddings + CNN + semantic space optimization	F1: 86%	Static embeddings; no contextual semantics or personalization; no comorbidity handling	[1]

4. Conceptual Framework

The proposed system is a modular pipeline designed for the early detection of emotional distress and mental health risks through the analysis of user-generated social media content. The architecture integrates data collection, preprocessing, feature extraction, emotional classification, severity assessment, and resource recommendation, along with optional feedback mechanisms for continuous learning.

A. Data Collection Module

This module is responsible for acquiring raw data from social media platforms such as Twitter and Reddit, either through public APIs or curated datasets. The collected content primarily includes textual posts, tweets, and comments that are likely to reflect users' emotional states.

Preprocessing pipeline:

- Noise Removal: Elimination of URLs, emojis, hashtags, and special characters.
- Normalization: Conversion to lowercase, lemmatization or stemming for lexical uniformity.
- Language Filtering: Retention of only English- language content to ensure model compatibility.
- Stopword Removal: Removal of high-frequency, low-information words to re duce noise.

B. Feature Extraction Module

This module generates semantically rich representations of textual data using transformer-based architectures.

- Contextual Embeddings: Utilizes pretrained models such as BERT, RoBERTa, or Distil BERT to derive token-level and sentence-level embeddings.
- Augmented Features Incorporates additional linguistic or psycholinguistic features (e.g., LIWC categories, sentiment scores, word counts) to enhance model interpretability and performance.

C. Classification Module

A hybrid classification framework is deployed to detect and categorize user emotions.

- Classifier Architectures: Combines deep contextual embeddings with traditional features using classifiers such as Support Vector Machines (SVM), Random Forests (RF), or Multilayer Perceptrons (MLP).
- Emotion Categories: Trained to detect emotional states including depression, anxiety, anger, and general distress.
- Confidence Scoring: Each prediction is accompanied by a probabilistic confidence score, enabling downstream modules to interpret prediction certainty.

D. Severity Assessment Module

To quantify the intensity of the detected emotional state, a severity grading system is employed.

- Severity Levels: Emotional states are categorized as Low, Moderate, or High.

Assessment Criteria:

- Model confidence in negative classifications.
- Frequency and persistence of distress signals over time.
- Co-occurrence of multiple severe emotion labels (e.g., depression + suicidal ideation).

E. Personalized Resource Recommendation System

Based on the detected emotion and its severity, the system suggests relevant mental health resources.

- Low/Moderate Severity: Directs users to forums, online support communities, and self-guided therapy content.
- High Severity: Recommends immediate professional help, including crisis helplines and licensed therapy platforms.

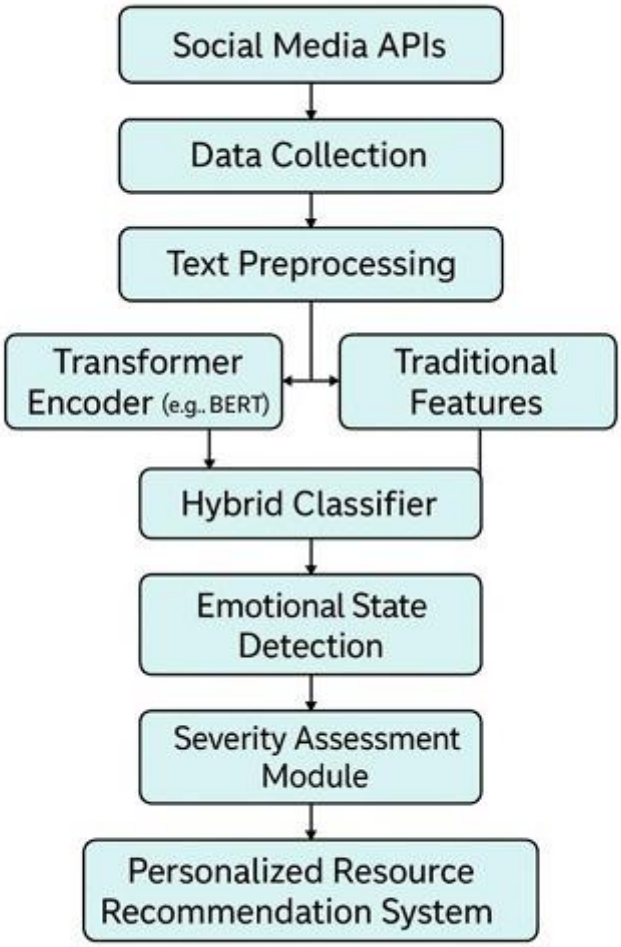


Figure 1: Proposed Architecture of our solution

5. Conclusion

This study presents a robust and modular framework for the early detection of mental health conditions, with a focus on depression and emotional distress, using user-generated content from social media platforms. By combining advanced transformer-based models such as BERT and RoBERTa with traditional machine learning classifiers, the system effectively captures contextual emotional cues and classifies the severity of mental health risks. The integration of a severity assessment module and a personalized recommendation engine ensures users receive appropriate support based on their emotional state. Furthermore, thesystem's design supports adaptability and scalability,

making it suitable for real-world applications such as chatbot-based screening tools or digital health platforms. The optional feedback and retraining loop provides a mechanism for continuous learning, enabling the model to stay updated with evolving language patterns and user behaviour. Overall, this work contributes to the growing field of AI-assisted mental healthcare and lays the foundation for accessible, proactive, and personalized support systems that can help bridge existing gaps in traditional mental health services.

References

D. Ramírez-Cifuentes, C. Largeron, J. Tissier, R. Baeza-Yates, and A. Freire, "Enhanced word embedding variations for the detection of substance abuse and mental health issues on social media writings," IEEE Access, vol. 9, pp. 130449–130470, Sep. 2021.

L. Ilias, S. Mouzakitis, and D. Askounis, "Calibration of transformer- based models for identifying stress and depression in social media," IEEE Trans. Comput. Social Syst., vol. 11, no. 2, pp. 1979–1991, Apr. 2024.

Abilkaiyrkyzy, F. Laamarti, M. Hamdi, and A. E. Saddik, "Dialogue system for early mental illness detection: Toward a digital twin solution," IEEE Access, vol. 12, pp. 2007–2024, Jan. 2024.

Pourkeyvan, R. Safa, and A. Sorourkhah, "Harnessing the power of Hugging Face transformers for predicting mental health disorders in social networks," IEEE Access, vol. 12, pp. 28025–28038, Feb. 2024.

M. Sao and H.-J. Lim, "MIRoBERTa: Mental illness text classification with transfer learning on subreddits," IEEE Access, vol. 12, pp. 197454– 197470, Dec. 2024.

M. A. Abbas, K. Munir, A. Raza, N. A. Samee, M. M. Jamjoom, and Z. Ullah, "Novel transformer based contextualized embedding and probabilistic features for depression detection from social media," IEEE Access, vol. 12, pp. 54087–54100, Apr. 2024.

H. Shoaib, R. Ali, S. M. S. Khan, and M. Adil, "Depression detection on social media posts using BERT and support vector machine," in Proc. 14th IEEE Int. Conf. Commun. Syst. Netw. Technol. (CSNT), Apr. 2025, pp. 708– 713.

Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive?," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), Dec. 2020, pp. 4795–4810.

S. M. Dhamdhere and A. P. Patil, "Mental health safety and depression detection in social media text data: A classification approach based on a deep learning model," in Proc. IEEE PuneSect. Int. Conf. Electr., Comput. Commun. (PuneCon), Dec. 2023, pp. 1–6.

Al-Mutairi and D. AlShamrani, "Text mining and emotion classification on monkeypox Twitter dataset: A deep learning–natural language processing approach," in Proc. 16th Int. Conf. Develop. e-Syst. Eng. (DeSE), Oct. 2023, pp. 151–156.

S. Aina and S. O. Fatumo, "A hybrid learning-architecture for mental disorder de tection using emotion recognition," in Proc. IEEE 5th Int. Conf. Artif. Intell. Comput. Appl. (ICAICA), Oct. 2023, pp. 457–462.

H. S. Naik,
M. B. Dudhamal, and H. D. Tanna, "Mental disorder indication detection withtransformer model architecture," in Proc. 8th Int. Conf. Inventive Comput. Technol. (ICICT), Mar. 2024, pp. 1560–1565.

G. Pattun and P. Kumar, "Emotion classification using generative pre-trained embedding and machine learning," in Proc. IEEE Int. Conf. Machine Learning and Applied Network Technologies (ICMLANT), Dec. 2023, pp. 389–394.

A Subsection Sample
Please note that the first paragraph of a section or subsection is not indented. The first paragraphs that follows a table, figure, equation etc. does not have an indent, either.
Subsequent paragraphs, however, are indented.

Sample Heading (Third Level). Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Sample Heading (Forth Level). The contribution should contain no more than four levels of headings. The following Table 1 gives a summary of all heading levels.

Table 1. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

$$x + y = z \qquad (1)$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

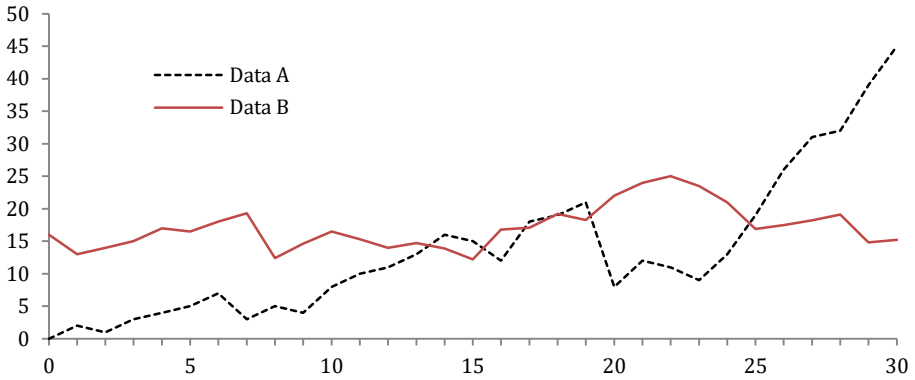


Fig. 1. A figure caption is always placed below the illustration. Short captions are centered, while long ones are justified. The macro button chooses the correct format automatically.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], as well as a URL [5].

References

1. Author, F.: Article title. Journal 2(5), 99–110 (2016).
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).
4. Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010).
5. LNCS Homepage, http://www.springer.com/lncs, last accessed 2016/11/21.

Copyright & License:
© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.