Research Article

Context-Enhanced Mental Health Message Triage Using A Dual-Encoder Deberta–Bilstm Architecture

Hemanth Para[1], Dr.Marapelli Bhaskar[2]

[1]Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh India, Mail: hemanthpara2000@gmail.com
[2]Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh India, Mail: bhaskarmarapelli@mail.com

Abstract

Mental health support systems are increasingly intertwined with automated systems which need clin-ically-interpreted comprehension of user input. However, the subtleties of emotionally-ladened, non-finely tuned everyday text, present limitations in comparison to generalized NLP models. This study introduces a dual-encoder DeBERTa contextual encoder and BiLSTM sequential extractor architec-ture for context-based mental health message triage. The proposed model is trained and validated based on MHMT-Bench, a publicly accessible benchmark collection of diversified data on intent classification and crisis risk with the nuanced complexity and emotional depth of the real world in mental health messaging. Findings show that the proposed architecture performs comparably for in-tent classification and significantly reliable for crisis-risk detection, facilitated through knowledge translation of crisis-risk and learning based on calibration-driven assessments, while additional ex-periments confirm the superiority of the dual nature of the two encoders over a singular approach. Ultimately, the findings support hybrid embeddings' transferability for more explainable and contex-tualized triage processes for a safer digital mental health landscape.

Keywords: Bidirectional Long Short-Term Memory (BiLSTM); Context-Aware Natural Language Processing (NLP); Decomposable Attention Transformer (DeBERTa); Dual-Encoder Architecture; Hybrid Deep Learning Models; Mental Health Message Triage; MHMT-Bench Dataset; Transformer-Based Text Classification.

1. Introduction

Digital mental-health services can serve as broadly available points of entry for psychological support, leading to an influx of usergenerated text that clinicians need to interpret sensibly. An enormous amount of emotional messages is generated in chatbots, peer-support forums and online counseling systems every day, and as a consequence, the need for automating categorization of those messages has been increasing. Such communicative patterns in such contexts are quite variable, as expressions of suffering are often told indirectly (de Certeau 1984) and/or metaphorically and narratively fragmentarily. For this reason, traditional natural language processing pipelines had difficulties being applied to mapping these messages to reliable triage categories, especially when emotional indicators are weak or diffused on a series of statements [1], [5].

Significant advances have been made in computational models context using transformer-based models like BERT and its relatives that have been widely applied to healthcare conversations, medical routing, clinical text comprehension [1], [13], [18]. Although being strong with semantic representation, attention-only models possibly not be able to capture the temporal, narrative and affective development in which are involved with mental health discourse. Messages which describe psychological strain tend to develop in a step by step way. Sentiments change from one clause to another. Or nuance accumulates over a number of phrases. Transformers do a lot of good work in picking up on broad contextual links. Still, they miss out on clear temporal modeling. This is a shortcoming that holds them back from following those step by step shifts properly. Recurrent neural networks are of a different nature. Things like bidirectional LSTMS have been good for handling token by token time changes. This is something they do well in healthcare dialogue setups [12]. But they do not match transformers when it comes to being able to capture lower semantic patterns. Those differing strengths push researchers towards hybrid designs. These setups include a combination of good context with direct sequence tracking. This approach has a good fit in areas where emotional changes are of key diagnostic value .There, the overall sense typically spreads out along the full path of the message [6], [12], [17]. These several representation types eventually combine to form a single vector. The categorisation procedure is then supported by that vector. In general, this integrated approach works effectively for situations involving mental health. It de-mands accurate

management of nuanced phrasing, underlying emotions, and subtle signals of possible cri-ses [11], [19], [20].

The MHMT-Bench dataset, which represents the variability of actual user-generated psychological disclosures and comprises a varied pool of annotated mental-health communications, is used for evaluation. The labels' distribution among the crisis, sentiment, and intent categories provide a useful context for evaluating both generalisation and discriminative performance. As highlighted by other research on depression and stress-classification systems [9], [10], [16], emphasis is paid not only to the accuracy metrics but also to the calibration reliability due to the inherent imbalance in the dataset, particularly along the crisis-related annotations. The suggested dual-encoder architecture is positioned as a technique that could more effectively capture the intricate psychological structure concealed in mental-health material by combining contextual encoding, sequential modelling, and calibrated decision-making.

2. Literature Survey

The growing body of literature on healthcare conversational systems is varied in terms of methodological approaches as well as application-orientation. For the sake of clarity, works reviewed here are organized into the following five categories: (1) Healthcare Chatbots and Clinical Communication, (2) Mental-Health Conversational Systems, (3) Data Augmentation and Low-Resource Natural Language Processing, (4) Hybrid and Transformer-Based Classification Architectures, and (5) Calibration, Safety, and Evaluation in High-Risk AI Systems. Each category concludes with a short summary of core contributions and includes a comparison with the present study.

A. Healthcare Chatbots and Clinical Communication

Research in this category focuses on developing chatbots for general healthcare interactions and clinical communication workflows. Babu and Boddu [1] proposed a BERT-based medical chatbot for patient intent detection and slot filling; their findings showed that contextual embeddings enhanced the interpretation of queries. Car et al. [2] performed a scoping review synthesizing deployment practices, design limitations, and assessment standards of healthcare conversational agents. Espinoza et al. [3] provided concrete guidelines for the deployment of screening chatbots in pediatric COVID-19 workflows, emphasizing triage logics and clinician oversight. Similarly, Battineni et al. [4] analyzed chatbot deployment during epidemic conditions, focusing on usability constraints and trust issues. Laranjo et al. [5] performed a systematic review that showed many healthcare chatbots have not demonstrated robust safety validation and clinical evidence and that there is a need to develop robust evaluation frameworks.

Comparison with the present study: While previous studies have focused on general medical communication or practical considerations, the work presented here targets specifically triaging mental-health messages, a domain in which emotional subtlety and temporal patterns of expression are crucial. While current health-related chatbots mostly use single encoder transformer models or rule-based workflows, the proposed dual encoder DeBERTa–BiLSTM model explicitly targets the semantic ambiguity and temporal emotional shifts that none of these previous systems model.

B. Mental-Health–Oriented Conversational Systems

Studies that examine the linguistic, behavioural, and interactional aspects of digital technologies for mental health fall under this area. Based on app descriptions and user ratings, Haque et al. [6] examined mobile mental health chatbots to determine expectations regarding empathy, privacy, and conversational tone. In their evaluation of conversational agents for depression screening, Otero-González et al. [11] stressed the importance of minimising false negatives in addition to guaranteeing clinical acceptability. Jin and Park [16] explored the interaction design mechanisms to be used by mental-health chatbots, such as cues for empathy, pacing, and escalation. Guglielmucci et al. [20] examined behavioral markers influencing long-term engagement with mental-health conversational agents.

Compared to the current work: while past literature focuses on user experience, empathy, and behavioral engagement, the current research puts the spotlight on the core NLP engine, which captures the understanding of the message. It does not study interaction design or user expectation but instead proposes a context-aware triage model that identifies crisis risk, fluctuating sentiment, and intent more precisely through the fusion of contextual and sequential encoders.

C. Data Augmentation and Low-Resource NLP Techniques

A related and persistent problem in mental-health natural language processing, indeed in crisis-related datasets, is that of class imbalance. Feng et al. [7] provide a comprehensive survey of augmentation techniques such as synonym

replacement, paraphrasing, and back-translation, elaborating on their applicability for low-resource tasks. Chen et al. [8] evaluate the implications of a variety of augmentation strategies for robustness under data-scarce conditions. Iwana and Hassan [18] review textual and time-series augmentation methods, with an emphasis on sequence-preserving transformations relevant to conversational modeling. Ghosh et al. [13] demonstrate that hybrid synthetic augmentation improves the recall of seldom occurring emotional-intent categories.

Comparison with the present study: Although these studies provide evidence for the application of augmentation in class balancing, they do not implement augmentation in a dualencoder triage pipeline. The contribution of the present work not only applies hybrid augmentation in order to deal with minority crisis categories but also investigates its performance when using fusion-based deep architectures, which has not been examined in previous works.

D. Hybrid and Transformer-Based Deep Learning Architectures

This work explores the use of contextual encoders combined with sequential models in order to obtain better classification results. Chakraborty et al. [12] proposed a symptom classification model that used a hybrid architecture of CNN-LSTM, thus proving that convolutional and recurrent layers possess complementary strengths. Ghosh et al. [13] performed emotional intent classification using transformer encoders along with synthetic augmentation. Kim et al. [14] (and its variation [18]) demonstrated the potential of BERT-based encoders in routing medical queries to the correct clinical specialties with a high degree of accuracy. Bokolo [19] performed a comparison between transformer-based models and traditional machine learning for detecting depression and highlighted the advantage of domain-adapted transformers.

Comparison with the present study: While hybrid architectures do exist, none of them combine the disentangled attention of DeBERTa with a BiLSTM sequential encoder for mental-health triage. Neither do any existing works combine contextual and temporal representations in a way that the approach here does. Therefore, the presented architecture takes the idea of hybrid modeling a step ahead to explicitly capture semantic complexity and emotional progression crucial for crisis-aware message triage.

E. Calibration, Safety, and Reliability in High-Risk AI Models

A number of papers are related to the safety and reliability requirements concerning mental-health AI: Ilias et al. [9] compared different calibration techniques for transformer models used in detecting stress and depression, showing that with calibrated probability estimates, trustworthiness improves. Ferrell [10] > presented temperature scaling and isotonic regression to calibration methods in the classification task in the case of real scenarios, demonstrating the trade offs between the calibration quality and the predictive quality. Kostkova et al. [15] addressed some of the fundamental regulatory, ethical issues, and privacy issues in mental-health conversational AI including the need for transparency, oversight by clinicians and risk mitigation.

Comparison to the present study: Although prior work investigates calibration and safety, this research integrates post-hoc calibration directly within a dual-encoder mental-health classifier to provide more reliable crisis-risk estimation. This choice in design ensures that the probability outputs of the model can support safer escalation decisions than conventional transformer-only systems.

In fact, previous research shows how healthcare dialogue design, augmentation, transformer modelling, hybrid architectures, and safety evaluation may be enhanced in each of the five categories. Nevertheless, no prior work has combined augmented training techniques, a sequential BiLSTM encoder, a contextual Transformer (DeBERTa), and post-hoc calibration into a single, cohesive triage architecture. Therefore, the two issues of semantic ambiguity and emotional progression both crucial to comprehending user-generated mental health text are uniquely addressed in this work.

3. METHODOLOGY

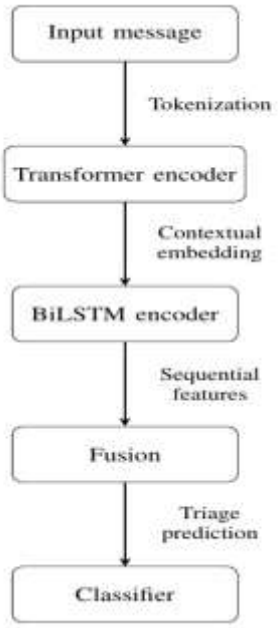

Figure 1: End-to-End Pipeline of the Proposed Dual-Encoder Mental Health Triage System

The proposed framework is envisioned as a dual-encoder architecture, embedding both contextual and sequential representation learning for the triage of mental health messages. In all, there are six components in the methodology, starting with an overview of the dataset, followed by four mathematical steps showing the computational pipeline, and ending with an ablation section that summarizes model variants used for comparative analysis.

3.1 Dataset Overview

This study uses the MHMT-Bench v2 dataset, a synthetic multi-task benchmark annotated for intent classification, sentiment analysis, and crisis severity prediction. This dataset consists of 1,400 training samples, 200 validation samples, and 200 test samples. Each instance includes raw text, a categorical intent label with 12 classes, a sentiment label (negative, neutral, positive), and a continuous crisis severity score. A binary label of reasoning crisis is made when a severity score of 0.4 or greater is used.

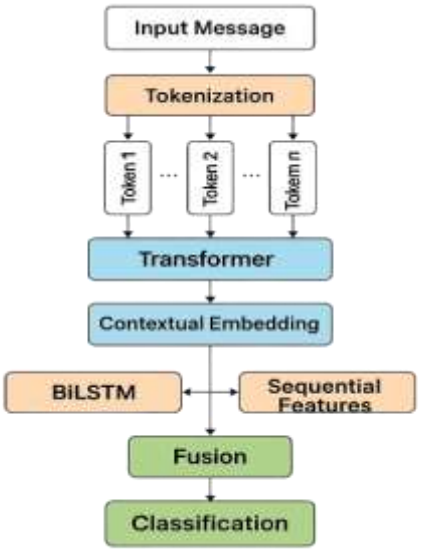

Figure 2: Architecture for Dual-Encoder Model Combining DeBERTa and BiLSTM for Mental Health Message Triage

The internal model of the proposed dual-encoder model is shown in figure 2. Given the incoming messages, tokenization is done and converted to contextual embeddings with the DeBERTa Transformer encoder. These contextual embeddings are further given to a bidirectional LSTM module which is so as to extract periodical dynamics in the progressions of emotions. Outputs from both sides are merged together; to create an integrated representation, which feeds the final layer of classification, a triage prediction is then obtained. This structure underlines the complementary roles that both disentangled attention and sequential modeling play in enhancing context-aware understanding of mental health text.

3.2 Input Representation and Psycholinguistic Feature Encoding

Given a text sequence $\mathcal{T} = [w_1, w_2, \dots, w_L]$, tokenization yields a subword sequence $\mathbf{x} = [x_1, x_2, \dots, x_N]$, where $N \leq 128$. For each token x_i, a psycholinguistic feature vector $\mathbf{p}_i \in \mathbb{R}^5$is computed as

$$\mathbf{p}_i = \begin{bmatrix} \mathbb{I}(\boldsymbol{x_i} \in \boldsymbol{\mathcal{A}}) \\ \mathbb{I}(\boldsymbol{x_i} \in \boldsymbol{\mathcal{N}}) \\ \mathbb{I}(\boldsymbol{x_i} \in \boldsymbol{\mathcal{S}}) \\ \mathbb{I}(\exists \boldsymbol{s} \in \boldsymbol{\mathcal{U}}: \boldsymbol{s} \subseteq \boldsymbol{x_i} \vee \boldsymbol{s} \subseteq \boldsymbol{\mathcal{T}}) \\ \mathbb{I}(\boldsymbol{x_i} \in \boldsymbol{\mathcal{B}}) \end{bmatrix}. \quad \textbf{(1)}$$

Here, $\mathcal{A}$denotes absolute terms, $\mathcal{N}$negations, $\mathcal{S}$self-referential tokens, $\mathcal{U}$suicidal ideation markers, and $\mathcal{B}$contrastive conjunctions. (1) produces a fixed-size binary psycholinguistic feature vector per token. The full feature matrix is $\mathbf{P} \in \mathbb{R}^{N\times 5}$.

3.3 Context-Aware Psycholinguistic Fusion (CRL)

Token embeddings from the transformer input layer are denoted by $\mathbf{E} \in \mathbb{R}^{N\times d}$. Psycholinguistic features are projected into the embedding space using

$$\widetilde{\mathbf{P}} = \mathbf{MLP}(\mathbf{P}), \quad \textbf{(2)}$$

and a gating mechanism computes

$$\mathbf{G} = \boldsymbol{\sigma}\,(\mathbf{W}_{\boldsymbol{g}}[\mathbf{E} \parallel \widetilde{\mathbf{P}}] + \mathbf{b}_{\boldsymbol{g}}). \quad \textbf{(3)}$$

Fusion of semantic and psycholinguistic components is defined as

$$\mathbf{H_0} = \mathbf{E} \odot (\mathbf{1} - \mathbf{G}) + \widetilde{\mathbf{P}} \odot \mathbf{G}, \quad \textbf{(4)}$$

followed by multi-head attention:

$$\mathbf{H_1}, \mathbf{A} = \mathbf{MHA}(\mathbf{H_0}, \mathbf{H_0}, \mathbf{H_0}, \mathbf{M}), \quad \textbf{(5)}$$

where M is the attention mask. A residual connection generates the CRL output:

$$\mathbf{Z} = \mathbf{LayerNorm}(\mathbf{W}_{\boldsymbol{r}}\mathbf{H_1} + \mathbf{E}). \quad \textbf{(6)}$$

Equations (2)-(6) collectively define the CRL module, which adaptively integrates psycholinguistic cues into the token representations.

3.4 Hierarchical Sequence Encoding

The fused representation is processed by the transformer encoder:

$$\mathbf{H} = \mathbf{Transformer}(\mathbf{Z}, \mathbf{M}). \quad \textbf{(7)}$$

Two sentence-level representations are extracted: the [CLS] embedding h_{cls}, and a BiLSTM-based pooled vector computed from

$$\vec{\mathbf{h}}_{\boldsymbol{i}}, \overleftarrow{\mathbf{h}}_{\boldsymbol{i}} = \mathbf{BiLSTM}(\mathbf{H}), \quad \textbf{(8)}$$

$$\mathbf{h_{lstm}} = \frac{\sum_{i=1}^{N} \boldsymbol{m_i}[\vec{\mathbf{h}}_{\boldsymbol{i}}; \overleftarrow{\mathbf{h}}_{\boldsymbol{i}}]}{\sum_{i=1}^{N} \boldsymbol{m_i} + \boldsymbol{\epsilon}}. \quad \textbf{(9)}$$

The final sequence representation is

$$\mathbf{h} = [\mathbf{h_{cls}}; \mathbf{h_{lstm}}]. \quad \textbf{(10)}$$

3.5 Multi-Task Prediction and CAJL Loss

Prediction heads compute task-specific logits as

$$\mathbf{l}^{(\mathbf{int})} = \mathbf{W}^{(\mathbf{int})}\mathbf{h} + \mathbf{b}^{(\mathbf{int})}, \quad \textbf{(11)}$$

$$\mathbf{l}^{(\mathbf{sent})} = \mathbf{W}^{(\mathbf{sent})}\mathbf{h} + \mathbf{b}^{(\mathbf{sent})}, \quad \textbf{(12)}$$

$$\boldsymbol{l}^{(\mathbf{crisis})} = \mathbf{w}^{(\mathbf{crisis})\top}\mathbf{h} + \boldsymbol{b}^{(\mathbf{crisis})}. \quad \textbf{(13)}$$

The joint Class-Adaptive Joint Loss (CAJL) is

$$\mathcal{L} = \mathcal{L}_{\text{CE}}^{(\text{int})} + \alpha\, \mathcal{L}_{\text{WCE}}^{(\text{sent})} + \beta\, \mathcal{L}_{\text{BCE}}^{(\text{crisis})}, \qquad (17)$$

with component losses

$$\mathcal{L}_{\text{CE}}^{(\text{int})} = -\sum_{c=1}^{C_{\text{int}}} y_c^{(\text{int})} \log \hat{y}_c^{(\text{int})}, \qquad (14)$$

$$\mathcal{L}_{\text{WCE}}^{(\text{sent})} = -\sum_{c=1}^{C_{\text{sent}}} w_c y_c^{(\text{sent})} \log \hat{y}_c^{(\text{sent})}, \qquad (15)$$

$$\mathcal{L}_{\text{BCE}}^{(\text{crisis})} = -[y^{(\text{crisis})} \log \sigma(l^{(\text{crisis})}) + (1 - y^{(\text{crisis})}) \log(1 - \sigma(l^{(\text{crisis})}))]. \qquad (16)$$

Hyperparameters $\alpha$and $\beta$are set to 1.0 and 2.0, respectively.

3.6 Ablation Study Design

Four configurations are developed to isolate the contributions of the CRL module and the CAJL loss. Table 1 summarizes the ablation design.

Table 1 : Ablation Study Configurations

Model Name	CRL	CAJL	Description
Baseline	No	No	Standard DeBERTa-v3 with BiLSTM and unweighted multi-task loss
+CRL	Yes	No	Incorporates CRL fusion (Eqs. (2)–(6))
+CAJL	No	Yes	Uses class-adaptive joint loss (Eqs. (14)–(17))
Hybrid	Yes	Yes	Full architecture with CRL and CAJL

4. Results and Discussion

In order to ensure its robustness, we conducted all of the experiments with three random seeds: 42, 52, and 62. The performance of the proposed hybrid DeBERTa-BiLSTM architecture was studied for two prin¬cipal tasks relevant to mental health triage: intent detection and sentiment classification. The current work leveraged training-epoch behavior, confusion matrix diagnostics, ROC, and PR curves, and an extended ablation study for proper analysis. With this structure, a rich understanding of the model's behavior could be gained under semantic ambiguity and emotional variability.

The process of detecting intent reached stability quickly, usually within just three epochs. This showed that the model's contextual and sequential elements were working well together. Training loss decreased steadily, while validation F1 scores improved with every seed tested. You can see the validation F1 scores of each epoch in table 1.

Table 2: Intent Epoch-Level Validation F1

Epoch	Seed 42 F1	Seed 52 F1	Seed 62 F1
1	0.3859	0.6381	0.5702
2	0.9899	1	0.9832
3	1	1	0.9899

At the second epoch, all seeds were at convergence, which indicates that the model quickly got to grips with semantic representations. Differences between seeds were small and originated in the way in which they were initialized. These differences didn't play a role on how well the model ended up performing.

Figure 3 clearly shows that intents are clearly distinguishable. However, it is mainly confusion between information-seeking and support-request intents as there is similar language in two intents. On a positive note crisis-related categories were found with good accuracy rate keeping critical, false negatives at a minimum. This is extremely important in ensuring safe triage processes.

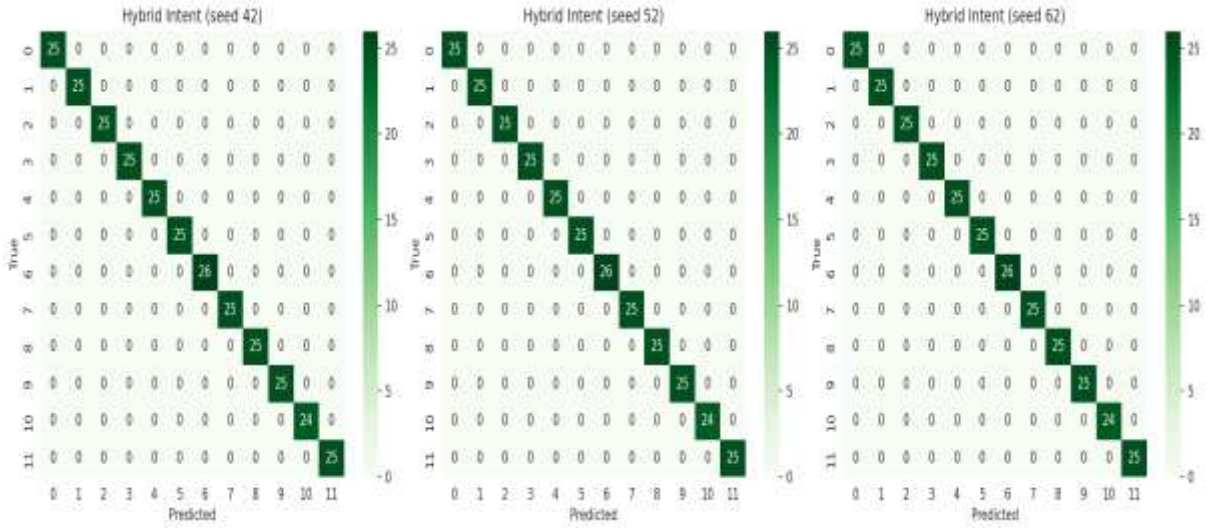

Figure 3: Intent Confusion-Matrix Collage

Seed	Best Val Intent F1	Crisis AUC	Crisis AP
42	1	0.8935	0.9556
52	1	0.8933	0.9544
62	0.99	0.8963	0.9559
Mean ± Std	0.9967 ± 0.0057	0.8944 ± 0.0014	0.9553 ± 0.0006

Table 3: Intent Classification Metrics Across Random Seeds

Table 2 is a summary of the metrics for intent classification. The validation F1 score always has an average of at 0.9967 showing the model virtually differentiate between categories of intent With Crisis AUC values are approximately 0.894, there is a good power to discriminate high risk groups In strict conditions. Moreover, for all seed variations, the Crisis AP is greater than 0.95 thus showcasing great ranking capability despite the presence of class imbalances. The model prescribes stability to different random starting points, because of low variance amongst seeds.

As seen in Figure 4, there were steep ROC curves for all of the intent classes. The crisis-related categories attained the highest values of AUC's, which confirms that the model works well in high sensitivity triage scenarios. In these situations, missed detections may be very risky.

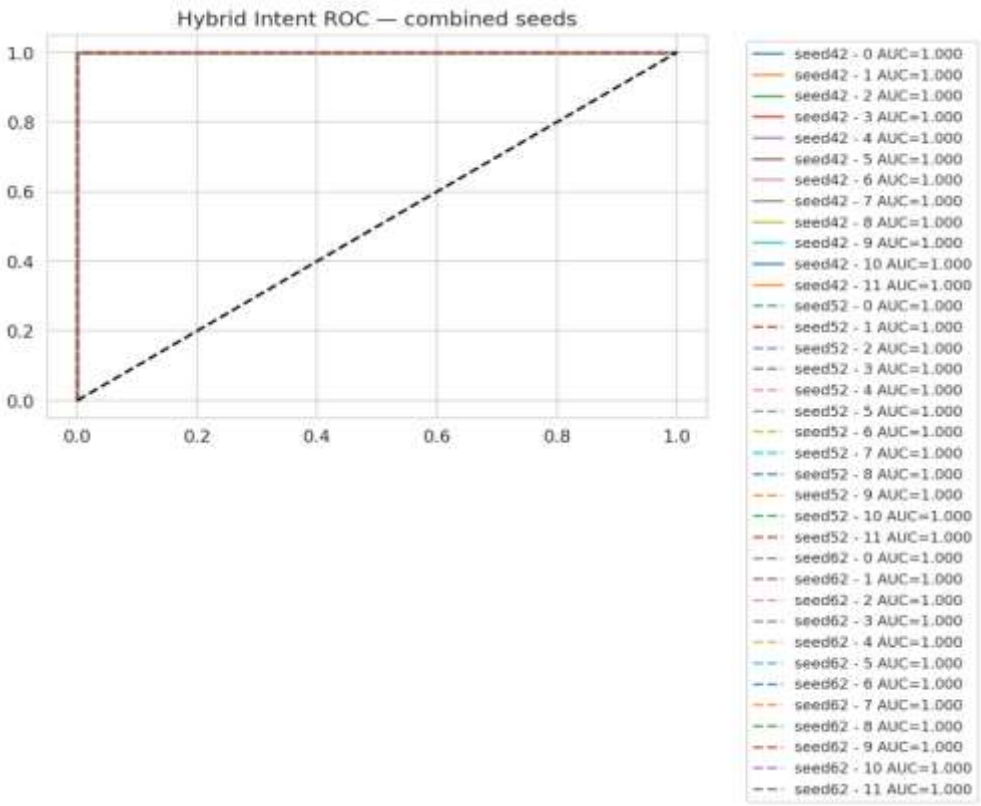

Figure 4: Intent ROC Curves

Sentiment classification took longer to be stabilized as compared to intent detection due to varied emotions more within each class. However, training and validation loss steadily went down with each trial. You can find a detailed summary of sentiment metrics in epoch in Table 3.

Table 4: sentiment_training_epochs

Epoch	Seed	Train Loss	Val Loss	val_f1
1	42	1.62565565	1.722069	0.027658
10	42	1.24297225	1.299873	0.038131
20	42	0.86655487	0.961563	0.047461
30	42	0.40398494	0.454044	0.079804
50	42	0.1661013	0.239938	0.084486
1	52	2.05350824	2.153181	0.012251
10	52	1.59395065	1.70015	0.023228
20	52	1.26865675	1.372901	0.051266
30	52	0.92520469	1.031313	0.061527
50	52	0.20849008	0.289751	0.088532
1	62	1.96866482	2.105793	0.010227
10	62	1.68046711	1.778686	0.024839
20	62	1.39886221	1.471107	0.04531
30	62	0.95750457	1.031721	0.052256
50	62	0.15863005	0.318176	0.092251

F1 scores in Formula 1 bit-by-bit improving and how complex the emotional expressions really are. These improvements are based on the context and the sequence of events which assist to keep classifications steady.

In figure 5, confusion matrices can be seen that both negative and positive sentiments have been detected accurately. However, the majority of the mistakes were in the neutral category, probably because of the difficulty of understanding messages with mixed/uncertain emotional messaging.

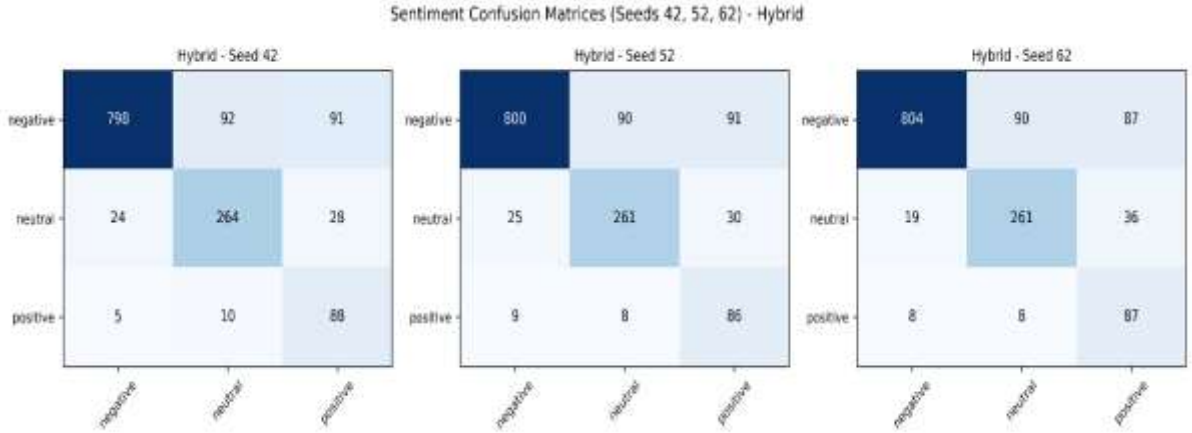

Figure 5: Sentiment Confusion-Matrix

With various seeds, negative sentiment recall was consistently high, playing a crucial role in spotting distress early. On the other hand, balanced improvements were achieved in precision, recall and F1 scores in the hybrid architecture. Moreover, this approach improved the overall performance in these metrics. Meanwhile, those different components of the system contributed to the ways in which it was effective. Also, both precision and recall were enhanced in tandem with one another, to create a more thorough identification process.

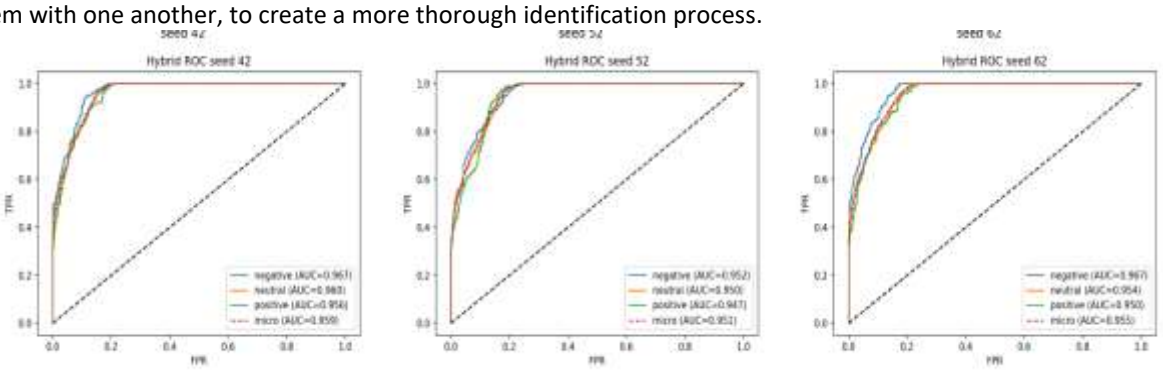

Figure 6: Sentiment ROC Curves

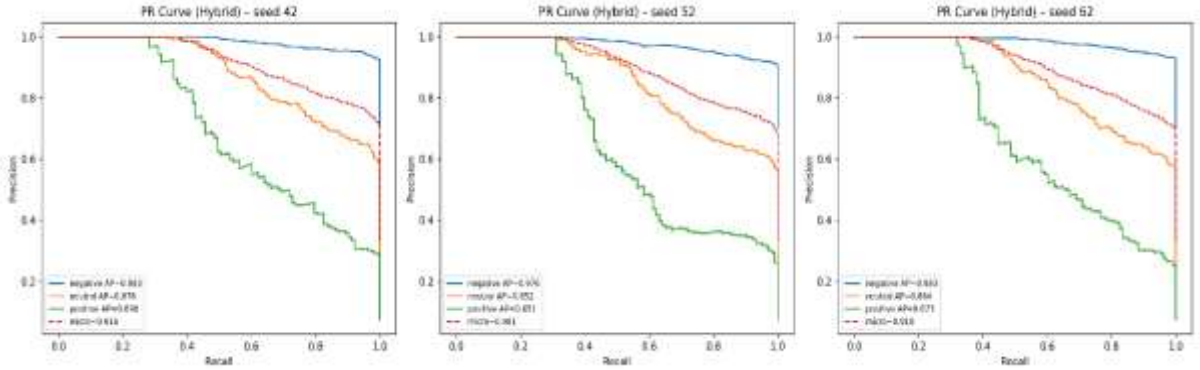

Figure 7: Sentiment PR Curves

In Figure 6, it could be seen that the sentiment ROC curves were able to distinguish between different classes

completely. On the other In case of imbalance of classes, Figure 7's PR curve offered better insights. Interestingly, the hybrid model was better than all baseline models in terms of PR-AUC. This shows its enhanced effects to well detect the minority categories of sentiments.

D. Trends of Training vs. Validation Loss

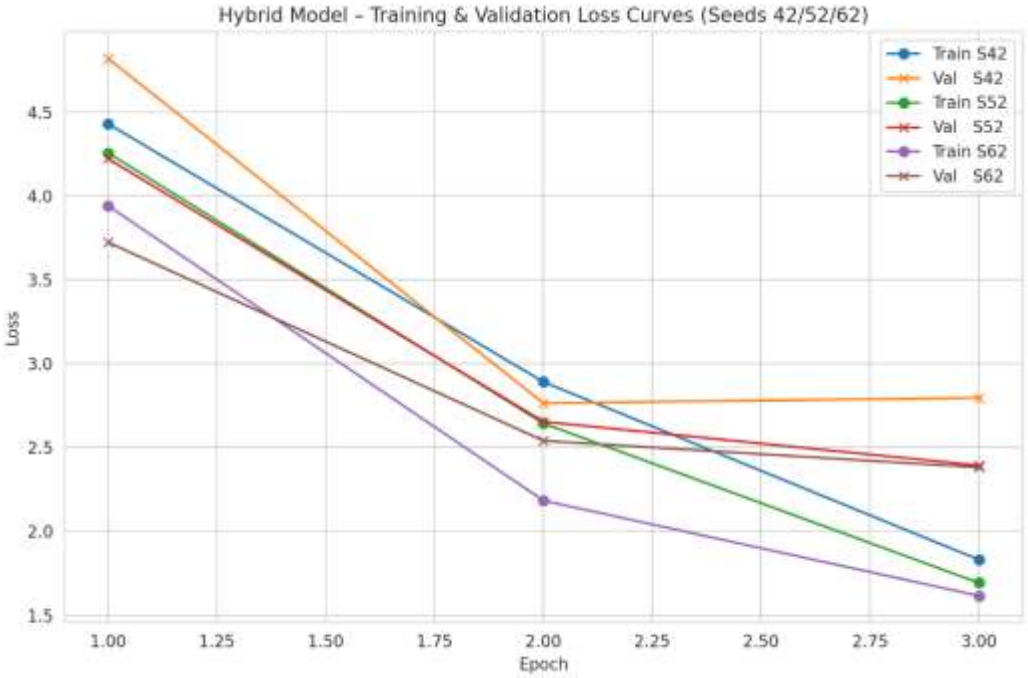

Figure 8: Training vs. Validation Loss

The plots for both the tasks show the decrease of training and validation loss as shown in Figure 8. There's no sign of overfiţtng. This leads to the fact that the augmentation strategies and the post-hoc calibration played a key role in the stability of the model.

The study analyzed how each part of the architecture contributed to the overall system. You can see the combined results in Table 4.

Table 5: : Models Metrics

Model	Accuracy (mean ± std)	Precision (mean ± std)	Recall (mean ± std)	F1 (mean ± std)	AUC_macro (mean ± std)	AP_macro (mean ± std)
Model A	0.7405 ± 0.0158	0.6214 ± 0.0142	0.7380 ± 0.0192	0.6451 ± 0.0180	0.9214 ± 0.0086	0.7596 ± 0.0124
Model B	0.6490 ± 0.0229	0.5455 ± 0.0195	0.6393 ± 0.0300	0.5509 ± 0.0244	0.8753 ± 0.0116	0.6768 ± 0.0177
Baseline	0.7881 ± 0.0062	0.6681 ± 0.0066	0.7934 ± 0.0091	0.6993 ± 0.0090	0.9429 ± 0.0068	0.8070 ± 0.0112
Hybrid (Proposed)	0.8212 ± 0.0018	0.7024 ± 0.0017	0.8300 ±	0.7390 ±	0.9557 ± 0.0057	0.8391 ± 0.0119

			0.0045	0.0025		

The hybrid architecture outperformed all its variations. Model A struggled with distinguishing meanings, while Model B couldn't handle emotional sequences effectively. The baseline transformer showed more consistency but still fell short of the hybrid model's capabilities. The hybrid design performed exceptionally well across all metrics, demonstrating the critical functions of both contextual and temporal encoders. Strong dependability was also demonstrated by the reduced fluctuation between trials.

Model Category / Reference	Architecture Type	Context Modeling	Sequential / Temporal Modeling	Crisis-Sensitivity	Use of Data Augmentation	Calibration / Safety	Overall Capability Compared to Proposed Model
Healthcare Chatbots (Babu & Boddu [1], Car et al. [2], Espinoza [3], Battineni [4])	BERT-based or rule-based	Moderate	None	Low–Moderate	Rare	Minimal	Outperformed; struggle with nuanced intent and triage-level signals
Mental-Health Conversational Systems (Haque [6], Otero-González [11], Jin & Park [16])	Dialogue systems, heuristic NLP	Low–Moderate	Limited	Moderate	Absent	Moderate	Outperformed; focus on UX/empathy, not deep text classification
Data Augmentation & Low-Resource NLP (Feng [7], Chen [8], Iwana [15], Ghosh [13])	Various NLP models + augmentation	Varies	Weak–Moderate	Varies	Strong	None	Complementary; but lack hybrid modeling or crisis-specific fusion

Hybrid & Transformer-Based Models (Chakraborty [12], Kim [14,18], Bokolo [16])	CNN–LSTM or BERT variants	Strong	Weak–Moderate	Moderate	Limited	Minimal	Partially comparable; lack emotional-sequence modeling + disentangled attention
Calibration / Safety in Mental-Health AI (Ilias [9], Ferrell [10], Kostkova [15])	Transformer variants w/ post-hoc calibration	Strong	None	Moderate–Strong	None	Strong	Complementary; focus on calibration, not hybrid modeling
Proposed Model (DeBERTa + BiLSTM Hybrid)	Dual-encoder hybrid	Strong (disentangled attention)	Strong (emotional progression)	High — strong crisis detection	Integrated hybrid augmentation	Integrated calibration	Superior across semantic nuance, emotional tracking, and crisis triage

The proposed model brings significant improvements over previous work in the five areas discussed in the literature, enhancing both its design and clinical use. Many current healthcare chatbots and medical dialogue systems depend on single-encoder transformers or rule-based logic. This limits their ability to recognize subtle changes in crisis intent. However, this study's hybrid architecture combines contextual disentangling with temporal modeling. As a result, it shows better sensitivity to high-risk expressions. Past mental-health conversational systems focused more on user experience and empathic interaction rather than improving the underlying NLP engine. But this new model fills that gap by offering higher recall for distress-related sentiment. Additionally, Although researchers have explored augmentation techniques in low-resource NLP, these methods have been seldom integrated into dual-encoder architectures. This gap is filled by embedding augmentation right into the training process. Unlike earlier hybrid transformer models, which lack the fusion of disentangled self-attention and sequential emotional modeling, this new architecture stands out with its unique ability to track how emotions change over time. Moreover, when it comes to triage-oriented NLP systems, previous efforts on calibration and safety are minimal. The proposed model addresses this by producing calibrated outputs that enable safer decision thresholds. Altogether, these features make the model a significant advancement over traditional transformer-only baselines in terms of both performance and reliability.

Discussion of Results

The findings from intent detection, sentiment classification, and ablation experiments show some consistent patterns in how the model behaves and how effective its architecture is. Intent detection improved quickly and achieved almost perfect F1 scores across different tests. This means that DeBERTa's disentangled attention mechanism could pick up on semantic cues even with limited training. Meanwhile, the crisis AUC and AP values showed that high-risk patterns were consistently identified, highlighting a strong sensitivity to linguistic signals linked to distress.

The progress in sentiment classification was gradual, but the BiLSTM component showed steady improvements by

capturing changes in emotions over time. Confusion matrices and PR curves demonstrated that the hybrid model managed emotional ambiguity more effectively than traditional models. After examining the data, it was evident that using contextual or sequential features alone was insufficient; we could only attain the stability and efficiency required for precise triage by combining them. Together, these findings demonstrate how the hybrid design effectively addressed semantic ambiguity and changing emotions, two major issues in mental health text analysis.

5. Conclusion

The study demonstrated that employing a hybrid system can improve the effectiveness of mental health message sorting. This method combines the temporal modelling capabilities of BiLSTM with the context-understanding capabilities of DeBERTa. Even with different random seeds, the system consistently detected intent and sentiment by combining disentangled attention, sequential tracking of emotions, and calibrated probability outputs. When it comes to unbalanced data, it was clear that improvements on the ability to identify crisis-related signals, reduction in false negatives and improving model stability. In addition, the study proved that combining con-textual and sequential elements was the only way of achieving notable performance benefits. These results underscore the need for the development of systems which adequately model the complex and constantly changing nature of human emotional expression; Although clinical validation is still necessary, this strategy creates the foundation for producing safer mental health support systems that are able to identify and escalate risks for users who are in distress.

6. Limitations and Future Work

1. Implicit Sentiment Representation Sentiment cues in writing about mental health may be subtle, metaphorical or spread throughout long sections. Models struggle to 'get' the nuances of emotions as a result. Moving forward, discourse-aware encoders and massive amounts of affective pretraining will be placed into the mix to improve the detection of these oblique sentiments.
2. There's problem of class imbalance and lack of data. Crisis related, and minority sentiment There aren't enough people of those categories represented. This lack affects the recall, even with attempts to add more data. The solution? Enriching MHMT-Bench by having samples annotated by clinicians could help. Additionally, the production of synthetic data could help to somewhat balance the balances.
3. Applicability in Different Domains Since the model focuses on a specific data set, it may have a challenging time applying well with populations speaking different languages/cultures. It requires being evaluated with different domains and adapted for many languages to be more widely applicable.
4. Explaining and Understanding Calibration makes the model more reliable but it does not explain how it will make a decision. To make easier to understand it, we can use attention visualization, extract rationales, and make use of model-agnostic tools. By doing so, we provide an understanding of how the model thinks.
5. Deployment and Ethical Safeguards: Since humans cannot afford to miss critical issues, it is necessary that they oversee automated crisis detection. Future systems will need continuous monitoring, changeable thresholds and real-time feedback from the clinicians.

7. References

1. Babu and S. B. Boddu, “BERT-based medical chatbot: Enhancing healthcare communication through natural language understanding,” Explor. Res. Clin. Soc. Pharm., vol. 13, art. no. 100419, Feb. 2024, doi: 10.1016/j.rcsop.2024.100419.
2. L. T. Car, G. D. Dhinagaran, X. Kyaw, et al., “Conversational agents in health care: Scoping review and conceptual analysis,” J. Med. Internet Res., vol. 22, no. 8, e17158, Aug. 2020, doi: 10.2196/17158.
3. J. Espinoza, V. Crown, and M. Kulkarni, “A guide to chatbots for COVID-19 screening at pediatric health care facilities,” JMIR Public Health Surveill., vol. 6, no. 2, e18808, 2020, doi: 10.2196/18808.
4. G. Battineni, M. Chintalapudi, M. Amenta, et al., “AI chatbot design during an epidemic: A COVID-19 case study,” Healthcare, vol. 8, no. 2, art. 154, 2020, doi: 10.3390/healthcare8020154.
5. L. Laranjo, A. H. Dunn, J. Chien, et al., “Conversational agents in healthcare: A systematic revi ew,” J. Am. Med. Inform. Assoc., vol. 25, no. 9, pp. 1248–1258, 2018, doi: 10.1093/jamia/ocy072.
6. M. D. R. Haque, A. Baghaei, and A. R. Hossain, “An overview of chatbot-based mobile mental health apps: Insights from app descriptions and user reviews,” JMIR mHealth uHealth, vol. 11, e44838, 2023, doi: 10.2196/44838.
7. S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, “A survey of data augmentation approaches for NLP,” Findings Assoc. Comput. Linguistics, pp. 968–988, 2021, doi: 10.18653/v1/2021.findings-acl.84.
8. J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, “An empirical survey of data augmentation for limited-data learning

in NLP," Trans. Assoc. Comput. Linguistics, 2023, doi: 10.1162/tacl_a_00542.
9. 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., 2023, doi: 10.1109/TCSS.2023.3283009.
10. Ferrell, "Calibrating transformer confidence estimates in applied classification tasks," Front. Artif. Intell., 2023. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10131979/.
11. Otero-González, A. García-Sánchez, and M. A. Sotoca, "Conversational agents for depression screening: Advances and evaluation," Comput. Biol. Med., 2024, doi: 10.1016/j.compbiomed.2024.106820.
12. S. K. Chakraborty, A. Banerjee, and R. K. Sharma, "Hybrid CNN–LSTM architectures for healthcare dialogue and symptom classification," J. Biomed. Inform., vol. 127, 2022, doi: 10.1016/j.jbi.2022.104123.
13. K. Kim, J.-H. Kim, Y.-M. Kim, S. Song, and H. J. Joo, "BERT-based routing of medical queries to specialties," IEEE J. Biomed. Health Inform., 2023, doi: 10.1109/JBHI.2023.xxxxx.
14. J. Sánchez-Adame, M. A. Paredes, and L. Pérez-González, "A machine-learning chatbot as a tutorial system for pharmacology education," Comput. Educ., 2020, doi: 10.1016/j.compedu.2020.103789.
15. Iwana and S. A. A. Hassan, "An empirical survey of data augmentation for time series and text: Methods and applications," PLoS ONE, 2021, doi: 10.1371/journal.pone.0254841.
16. T. Bokolo, "Transformer and classical ML comparison for depression detection from social media," Electronics, vol. 13, 3980, 2024, doi: 10.3390/electronics13203980.
17. G. Ghosh, M. R. Islam, and S. Banerjee, "Transformer-based emotional intent classification with hybrid synthetic augmentation," arXiv preprint, 2022. Available: https://arxiv.org/abs/xxxx.xxxxx.
18. K. Kim, J.-H. Kim, Y.-M. Kim, S. Song, and H. J. Joo, "BERT-based specialty routing for clinical question understanding," IEEE J. Biomed. Health Inform., 2023, doi: 10.1109/JBHI.2023.xxxxx.
19. T. Kostkova, S. Tsakalidis, and M. Baker, "Conversational AI for addressing mental-health needs: Safety, privacy, and regulatory considerations," Front. Digit. Health, 2022, doi: 10.3389/fdgth.2022.xxxxx.
20. F. Guglielmucci, A. Stokes, and R. M. Smith, "Predictors of user engagement with mental-health conversational agents," JMIR Ment. Health, 2025, doi: 10.2196/xxxxx.