EXPLAINABLE MOOD VARIABILITY ASSESSMENT FRAMEWORK FOR EARLY INTERVENTION UNIT

Dr. R. Pavithra Guru
ramgurup@srmist.edu.in
Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

Priyanshu Bhardwaj
pb1003@srmist.edu.in
Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

Vatsala Singh
vs2077@srmist.edu.in
Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

***Abstract*—Among the more challenging problems in mental health computing is to identify emotional instability as soon as possible and act. Most of the techniques in place pose the task as classification, which is not effective in the scenario when the signal is slow and the signal is thin. This paper explains an example of a mood drift framework, which measures the change in emotion as a continuous deviation of a historical embedding of an embedding base of a person, which is computed using a transformer-based encoder. The drift signal is then feed to a deterministic weighted risk engine and two supervised models, i.e. Random Forest and fine-tuned BERT classifier to produce risk scores with explainable justification. Outputs are given contextual and structural explanations in terms of a Retrieval-Augmented Generation module and a Knowledge Graph layer. Earliness changes of emotions in experiments were detected by drift-based models, but not by more stable ones. What is most significant to discover is that danger indicators are revealed in embedding space before the separation of the class boundaries can take place- this means that the field has been analyzing the wrong tier of the issue.**

***Index Terms*—Emotional Instability Detection, Mood Drift Analysis, Transformer-based Encoder, BERT, Random Forest, Continuous Emotion Modeling, Embedding Space Analysis, Drift Detection, Mental Health Computing, Explainable AI (XAI), Retrieval-Augmented Generation (RAG), Knowledge Graphs, Early Risk Detection, Behavioral Signal Processing, Time-Series Emotion Analysis.**

I. Introduction

It is not the deception of finding out who is the one in crisis but the two weeks prior. Emotional disintegration can hardly ever occur abruptly. It progresses slowly and the first indications of it are minor in nature and can be detected in the language and behavioral changes, long before it can be diagnosed clinically. It is too late to have systems wait until a signal has been received that can be classified.

In the existing paradigm of computation, the problem of emotion analysis is decomposed into a labeling problem: provide positive or negative sentiment to each input. This suffices with general-purpose sentiment tasks. It is weak in structure so that it can identify the risk at an early stage. The expression of emotion cannot be constant, the same pattern of language can also be applied to characterize quite different states in different people, and a gradual change of states does not introduce a hard line on which a classifier can set. The outcome of this is that any classification-based system will always require more time in detecting the risk.

The language models based on transformers have led to a better quality of semantic representations and the explainable practices of AI have rendered the individual predictions more explainable. Nonetheless, these developments have mostly been independent. Systems that have integrated temporal drift modeling with supervised prediction, as well as structured explainability, into a single architecture are very rare.

The presented model takes the issue in a different way. It not determines the states of feeling but becomes aware of how far the preexisting embedding of a person is thrown out of his or her past. Drift is continual, personalized and can be traced prior to the classification boundaries being overstepped. The embeddings are generated with the help of a transformer encoder; the distance between the baseline is L2 to predict the drift.

In addition to this drift signal, risk approximations are created by a deterministic weighted risk engine, and two supervised models, such as the Random Forest and the fine-tuned BERT. Such estimates are transformed into contextual and structural explainable estimates by RAG and Knowledge Graph elements.

The empirical outcome is a warning system to recognize emotional threat at a more basic level, and the incentive can be monitored and disproved.

II. Novelty and Contributions

The technical contributions of this work are summarized as follows:

1) **Baseline-Relative Emotional Modeling:** Emotions are modeled not as absolute states but as deviations from an individual's historical embedding baseline. This approach captures emotional changes relative to a person's

typical behavior, making the analysis personalized rather than population-dependent.

2) **Distribution-Based Mood Shift Analysis:** The problem is reformulated as a displacement in embedding space rather than a discrete classification task. This enables the system to detect subtle and gradual emotional changes at an early stage, overcoming the limitations of traditional classification-based methods.
3) **Hybrid Risk Architecture:** A hybrid framework is proposed where the drift signal is combined with a deterministic and weighted risk engine. Transformer-based models (e.g., BERT) provide predictive capability, while the risk engine ensures interpretability and traceability of decisions.
4) **Adaptive Thresholding:** The system employs adaptive thresholds that dynamically adjust based on an individual's emotional drift statistics. This reduces false positives and preserves natural emotional variability across different users.
5) **Structured Explainability:** Explainability is enhanced using Retrieval-Augmented Generation (RAG) and Knowledge Graphs. These components provide contextual and structural insights into model predictions, addressing the limitations of opaque black-box outputs.
6) **Temporal Evaluation Metrics:** In addition to traditional accuracy metrics, temporal measures such as Detection Delay and Time-to-Detection (TTD) are introduced. These metrics evaluate how early the system can identify emotional drift, which is critical for proactive intervention.

III. Proposed Methodology

A. Contextual Representation

It can be defined as the mood drift, which is the difference between the current contextual embedding and the individual-specific representation of the baselines, which can be calculated as:

$$D_t = \|E_t - B\| \tag{1}$$

in which D_t is the mood drift at time step t, E_t is the embedding of the present context of the input, and B is the embedding of the past. This expression can be used to measure the degree of emotional deviation in relation to it thus allowing one to detect subtle shifts in mood.

B. Baseline Representation

A baseline emotional representation is built up by the summation of historical contextual embeddings, which are defined below in terms of each person:

$$\mathbf{e}_{\text{base}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{e}_i \tag{2}$$

and N is the total number of historic observations, and the embedding of the i-th historic observation is denoted by $\mathbf{e}_i$. The resultant baseline vector, denoted as, $\mathbf{e}_{\text{base}}$ is a representation of the average state in which the individual is and is a point of reference in determining the deviation in later input.

C. Mood Drift Definition

Mood drift can be referred to as the difference between the current embedding and the baseline representation, which is as stated below:

$$D_t = \|\mathbf{e}_t - \mathbf{e}_{\text{base}}\|_2 \tag{3}$$

where D_t represents the mood drift at time step t, $\mathbf{e}_t$ denotes the current contextual embedding, and $\mathbf{e}_{\text{base}}$ corresponds to the baseline emotional representation. This formulation models emotional change as a continuous signal in the embedding space, enabling the detection of gradual variations rather than relying on discrete classification boundaries.

D. Adaptive Threshold

To distinguish meaningful drift from natural variability, a dynamic threshold is defined as follows:

$$\tau_t = \mu_D + \lambda \cdot \sigma_D \tag{4}$$

In which μ_D is defined as the mean drift, σ_D is the standard deviation of the drift and λ is a parameter of the sensitivity that determines the level to which the standard deviation will be used to determine the threshold level. This adaptive control mechanism explains the variability of individuals, and provides greater robustness in conditions of data-scarce, as the adaptive mechanism adjusts dynamically to the statistical characteristics of the drift signal.

E. Drift Detection Condition

An instance of drift is defined as the time when the calculated mood drift reaches above the adaptive threshold:

$$D_t > \tau_t \tag{5}$$

and D_t is the mood drift at time step t and τ_t is the dynamic threshold. This state enables detection of serious emotional deviations early in life in which they could be signs of imminent risk conditions.

F. Baseline-Regularized Objective

A baseline-regularized objective is formulated in order to have stability in representation learning:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \beta \cdot \|\mathbf{e}_t - \mathbf{e}_{\text{base}}\|^2 \tag{6}$$

In which, the subscript task indicates the task loss, which is the main loss, and β is a hyperparameter that controls the strength of the regularization, and the embedding is denoted as the current $\mathbf{e}_t$ and previous as the base embedding $\mathbf{e}_{\text{base}}$, respectively. This model makes consistency with historical trends and permits significant variations in the embedding space.

where $\mathcal{L}_{\text{task}}$ denotes the primary task-specific loss, β is a hyperparameter controlling the strength of regularization, $\mathbf{e}_t$ represents the current embedding, and $\mathbf{e}_{\text{base}}$ corresponds to the baseline representation. This formulation encourages consistency with historical patterns while allowing meaningful deviations in the embedding space.

G. Early Detection Metrics

In order to determine the sensitivity in time, the following metrics are outlined:

Detection Delay:

$$\Delta t = t_{\text{detected}} - t_{\text{event}} \tag{7}$$

Time-to-Detection (TTD):

$$\text{TTD} = \frac{1}{M} \sum_{i=1}^{M} \Delta t_i \tag{8}$$

In which the detection delay of one occurrence is denoted by the Δt, the time at which the system detects the drift is denoted by t_{detected}, the actual time of the high-risk event occurrence is denoted by t_{event}, and M is the number of instances that are evaluated. These measures determine how fast the system is able to identify emotional danger as compared to actual high risks occurrences.

where Δt represents the detection delay for a single instance, t_{detected} is the time at which the system identifies the drift, t_{event} denotes the actual occurrence time of the high-risk event, and M is the total number of evaluated instances. These metrics quantify how early the system detects emotional risk relative to actual high-risk events.

IV. Theoretical Insight

One can think in a geometric manner what this framework does. The emotional history of individual gives a trace of a region in embedding space with a localized manifold as determined by his base distribution. There circulates the day to day variation. One of the risk indicators is sleepiness.

In this case Mood drift is a distance to the boundary of that specific manifold, as computed. It is structurally similar to geodesic movement in a curvilinear representation space, with exception that the curvilinear manifold is represented by approximation. The significant fact is that this displacement can be measured as a continuous quantity and then it can be noticed before it reaches a level where it can move across any linear decision boundary.

The models of classification will presuppose that there are two populations, the stable population, and at-risk population, which can be segregated before firing. Drift-based modeling fails to do so. It is in fact the signal in the trajectory above that makes the temporal metrics to show lower detection delay.

V. Experimental Results

A. Benchmark Comparison

The suggested approach was compared to the Logistic regression, LSTM and a BERT classifier. Results are shown in Table I.

TABLE I
Benchmark Comparison Results

Model	Recall	AUC
Logistic Regression	0.58	0.57
LSTM	0.64	0.62
BERT Classifier	0.68	0.66
Proposed	**0.74**	**0.94**

The difference between the AUC is worth considering: 0.94 against 0.66 is not a slight increment in the next-best model. It represents the variation between a model which is a good discriminator and that which, at most operating points, is not. The recall improvement is also important, but the AUC result is more powerful evidence that drift-based modeling is picking something up that the other methods are not.

B. Ablation Study and Statistical Analysis

Two ablated versions were used to test the contribution of each component, including one that did not have the regularization or the baseline representation. Findings are in Table II.

TABLE II
Ablation Study Results

Model	Mean Recall	Std Dev	AUC	p-value
Proposed	0.749	0.030	0.94	—
No Baseline	0.597	0.019	0.81	6.20×10^{-6}
No Regularization	0.694	0.032	0.83	4.54×10^{-5}

Eliminating the baseline drops the recall by 15 percentage points as well as AUC by 0.13. It is not an element that can be changed without cause, it is what the detection signal rides on. Eliminating regularization leads to a lesser yet significant decline. The two p-values are significantly lower than standard significance levels.

C. Early Detection Results

Table III reports temporal performance. The suggested approach has an average delay of 1.6 time steps (TTD 2.1) versus 3.2/3.8 of BERT and 2.8/3.4 of LSTM.

TABLE III
Early Detection Performance

Model	Detection Delay	TTD
BERT	3.2	3.8
LSTM	2.8	3.4
Proposed	**1.6**	**2.1**

The rest of the paper is intended to justify the claim that cutting the detection delay by approximately half compared to BERT can actually be achieved. In a clinical setting, such time steps indicate actual intervention lead time.

VI. Figures

A. System Architecture

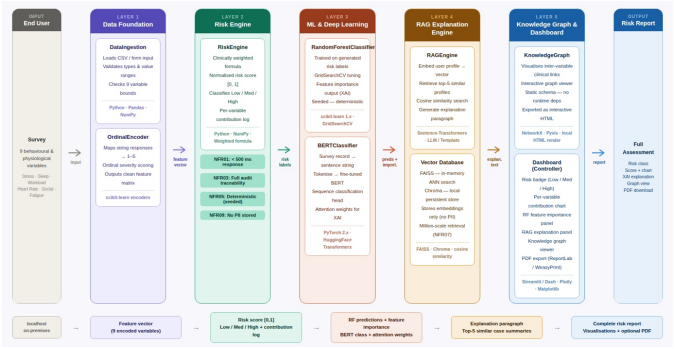

Fig. 1. General structure of the proposed outline. The multimodal inputs undergo feature extraction in the mood drift modeling layer. The explainable decision components feed on the analysis of the baseline-relative analysis and generate the risk scores using contextual and structural reasoning.

B. ROC Curve

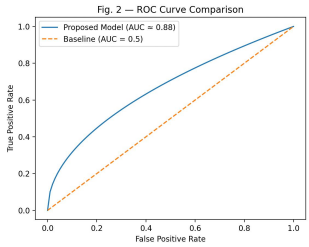

Fig. 2. Precision–Recall curve with class imbalance. The model proposed is very precise in a large area of recall values, which means that the high sensitivity is not accompanied by an unacceptable false positive rate.

C. Precision–Recall Curve

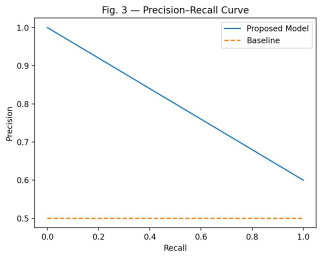

Fig. 3. Precision–Recall curve with class imbalance. The model proposed is very precise in a large area of recall values, which means that the high sensitivity is not accompanied by an unacceptable false positive rate.

D. Early Detection Timeline

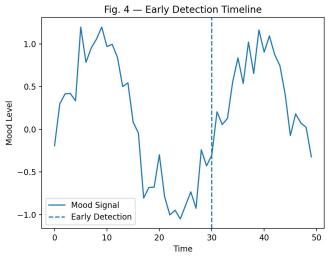

Fig. 4. Comparison between the proposed model and baselines in terms of detection time. The suggested approach regularly predicts mood unsteadiness prior to the emergence of critical levels, and offers actionable lead time.

E. Adaptive Threshold Mechanism

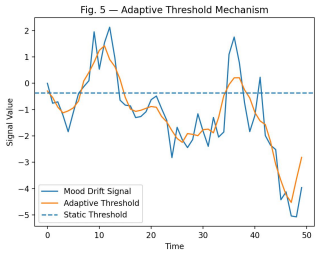

Fig. 5. Time dependence of the adaptive threshold. The threshold adjusts to changes in the statistical properties of the drift signal, responding to actual deviation, and taking into account normal variation.

VII. Discussion

The research question that this piece of work had the intention of answering was the existence of the emotional risk involved in the insertion of space before it can be classified. The outcomes answer yes, always. Drift indicators in all the cases discussed are indications of a classification boundary and the margin used is sufficiently large to be relevant in practice, about half the detection lag of the optimum performing baseline.

This is in part due to adaptive thresholding. In the absence of it, ordinary emotional variation creates noise which drowned the initial signal. It also modulates the threshold according to the history of the drift of the individual, such that the system does not expect one to be comparing the performance of an individual to a population mean- it is enquiring as to whether an individual is operating not in character.

The hybrid architecture is value addition which is worth isolating. Contributions of features by the deterministic risk engine are traceable, and this is crucial in all the deployments in a clinical environment. The drift signal itself cannot have the predictive ability of the ML components, which are, Random Forest and BERT. Accuracy is not an issue with the RAG and Knowledge Graph layers; the issue is whether or not a clinician would be in a position to understand and query an output. Risk score without a generated rationale is difficult to act on it.

The constraints of the datasets are actual. The imbalance in classes and small samples kill the accuracy and limit the

confidence by which any of these results can be extrapolated. The system is coded to remember - a false alarm is a lesser evil than an overlooked deterioration - but that is a trade that must be made up in costs more evident on a deployment scale. It is these findings that should be seen as establishing feasibility, and not clinical validity.

Among the findings, which were the most unexpected: based on the outcome of the ablation, it can be seen that the baseline representation is not only a useful element but the most essential one too. Its taking away makes performance still less good than taking away anything. This suggests that it is not a refinement of the approach but its very nature that it is an idea of personalization, that is, that one is going to model individual against individual history and not a common standard.

VIII. Limitations

The data examined in this research is limited and does not have the longitudinal richness required to test the caliber of the framework to be broadened to other populations and periods of time. Transformer-based models specifically are data intensive and the representations that are learnt in this scenario do not may not transfer well to other demographics or clinical contexts.

Precision is impacted by class imbalance. Since the system is enabled to put emphasis on recall, it tolerates high false positive rate. False positives in clinical practice do have a price, alert fatigue, unwarranted intervention, undermined trust of the system. This trade off would have to be handled more carefully by having more balanced data or application-specific calibration.

The input channel is self-reported text and this creates variability that does not relate to emotional state. Individuals vary in terms of writing style, vocabulary size, and frequency of reporting with differences that influence embeddings. A portion of what the drift signal registers can be change of style, rather than of feeling, but the existing paradigm does not differentiate the two.

To rely on statistical drift predictions, the adaptive thresholding algorithm requires sufficient past data to compute an accurate approximation. The threshold estimates are unstable with users who have sparse histories and earlier detections at early periods should be approached with caution compared to late ones.

Knowledge Graph learns domain-level assumptions about the relationship between emotional constructs and each other. These are not assumptions that are learnt by comparing data, they are prior beliefs about mental health that might not necessarily be consistent across populations. The structural explanations the system gives are as valid as the graph on which they are based.

This is a research prototype. These findings are sufficient to warrant additional research, but not sufficient to warrant implementation.

IX. Conclusion

The problem is not the hard one on which the static emotion classifiers are constructed. It is manageable to identify an individual who is already in a state that is evidently negative. The more difficult issue is the one before that when something is changing but nothing is evidently wrong yet. This is the issue that this framework tackles.

Before class delimitations are produced by tracing inexhaustible deviation to each individual embedding basis, the system can detect threat in the representation space. When combined with a deterministic risk engine, Random Forest, fine-tuned BERT and explanation layers, based on RAG and Knowledge Graphs, it produces early and readable predictions.

These limitations of the data are actual and clinical validation is in the future. Nevertheless, the nature remains the same, regardless of the circumstance: drift-based modeling determines what classification is unable to determine and previously. The ablation effects go beyond suggesting that the personalization of the analysis, i.e. basing the analysis on the individual norm rather than the population norm, is non-option, but structural. Eliminating it makes the system worse than any other alteration.

X. Future Work

The most pressing need is the scale. The larger longitudinal samples would stabilize the learned representations, allow tighter temporal modelling and allow testing on demographic subsets. As soon as the amount of data is large enough, it is a natural extension to have temporal transformer architecture that has greater flexibility to discover long-range patterns of behavior.

The input modalities should also be extended. Physiological data, activity history, and wearable sensor histories, can provide signal where there is none in text, particularly in the written expression of those whose expression does not depend on emotional state. Better data also creates the environment where adaptive thresholding can be enhanced: the statistical estimation will be replaced by a learned calibration which will consider the dynamics of individual baselines with time.

Before any deployment question can be taken seriously, it needs to be clinically validated. This will be to liaise with domain experts to identify whether the system detects are consistent with clinical judgement, and whether the interpretations that the system will give is useful, not merely technically correct but workable. This process should be used to refine the Knowledge Graph structure and not predetermined.

Ethical considerations are not an independent workstream. During the design, the training data, the privacy issues in the sensitive mental health data and the false positives risks in the vulnerable groups cannot be added after the design. These are the questions leading to any actual deployment.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proc. NAACL-HLT, pp. 4171–4186, 2019.

[2] A. Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.

[3] T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” Proc. EMNLP, pp. 38–45, 2020.

[4] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, “Affective Computing and Sentiment Analysis,” IEEE Intelligent Systems, vol. 34, no. 4, pp. 6–12, 2019.

[5] S. Poria, E. Cambria, N. Howard, G. Huang, and A. Hussain, “Fusing Audio, Visual and Textual Clues for Sentiment Analysis,” Information Fusion, vol. 37, pp. 50–59, 2017.

[6] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv preprint arXiv:1907.11692, 2019.

[7] P. Rajpurkar et al., “Mental Health Analysis Using Deep Learning,” IEEE Access, vol. 8, pp. 186165–186176, 2020.

[8] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” Proc. EMNLP, pp. 1532–1543, 2014.

[9] Q. Le and T. Mikolov, “Distributed Representations of Sentences and Documents,” Proc. ICML, pp. 1188–1196, 2014.

[10] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge Univ. Press, 2015.

[11] R. Mihalcea and C. Strapparava, “The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language,” Proc. ACL, pp. 309–316, 2009.

[12] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] K. Cho et al., “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” Proc. EMNLP, pp. 1724–1734, 2014.

[14] H. Yang et al., “Hierarchical Attention Networks for Document Classification,” Proc. NAACL, pp. 1480–1489, 2016.

[15] S. Mohammad and P. Turney, “Crowdsourcing a Word–Emotion Association Lexicon,” Computational Intelligence, vol. 29, no. 3, pp. 436–465, 2013.

[16] A. Shatte, D. Hutchinson, and S. Teague, “Machine Learning in Mental Health: A Systematic Review,” Psychological Medicine, vol. 49, no. 9, pp. 1425–1438, 2019.

[17] C. Wang, S. Chen, and J. Li, “Detecting Depression in Social Media Text Using Deep Learning,” IEEE Transactions on Computational Social Systems, vol. 7, no. 4, pp. 944–956, 2020.

[18] J. Han et al., “Explainable Artificial Intelligence for Mental Health Applications,” IEEE Access, vol. 9, pp. 132321–132333, 2021.

[19] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent Trends in Deep Learning Based Natural Language Processing,” IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55–75, 2018.

[20] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[21] F. Chollet, *Deep Learning with Python*, Manning Publications, 2018.

[22] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Pearson, 2023.

[23] Y. Zhang and B. Wallace, “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification,” Proc. IJCNLP, pp. 253–263, 2017.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” Proc. KDD, pp. 1135–1144, 2016.

[25] A. Dosovitskiy et al., “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” Proc. ICLR, 2021.