Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Calibrated VAD-Aware Multitask Transformers for Emotion and Sentiment Modelling in Suicide Notes

Pushpak Bhattacharyya

Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

Abstract

Timely identification of suicide risk is an important and challenging application area for modern machine learning and data science. Written suicide notes provide a rare, high–signal source of information about the cognitive and emotional states that precede lethal self-harm, but they also raise strict requirements on model reliability, transparency, and responsible use. Existing work on the CEASE family of corpora has progressed from traditional classifiers to deep recurrent multitask models and, more recently, VAD-assisted transformer architectures, yet current systems typically optimise only for classification accuracy, treat discrete and dimensional affective constructs in a fragmented way, and pay limited attention to probability calibration or human-centred explainability. In this paper we introduce a *Calibrated VAD-Aware Multitask Transformer* (C-VAMT) for sentence-level analysis of suicide notes, situated at the intersection of natural language processing, deep learning, and responsible AI. Our model jointly predicts (i) multi-label fine-grained emotions, (ii) three-way sentiment polarity, and (iii) continuous Valence–Arousal–Dominance (VAD) scores within a single transformer-based architecture. We enrich the encoder with lexicon-derived VAD features and propose a label-graph attention mechanism that explicitly models dependencies among emotion and sentiment labels. A hybrid supervision scheme combines large-scale lexicon-based VAD estimates with a smaller set of human-validated VAD and sentiment annotations. To support risk-sensitive decision-making, we incorporate temperature scaling and Monte Carlo dropout for calibrated uncertainty estimates, and we derive token-level rationales via integrated gradients, which are evaluated by clinicians. Experimental results on CEASE-v2.0 show that C-VAMT improves macro-F_1 and mean recall for multi-label emotion recognition, particularly on rare emotion categories, while also reducing expected calibration error and producing clinically plausible explanations. The proposed framework illustrates how advanced deep learning, multitask learning, and explainable AI techniques can be integrated into a coherent, human-centred pipeline for a societally critical application of computer and data sciences.

Keywords: Valence–Arousal–Dominance (VAD), multitask learning, responsible AI, uncertainty calibration, explainable AI

1. Introduction

Suicide is among the leading causes of death worldwide and remains a major public health concern. Each year, hundreds of thousands of individuals die by suicide, with many more engaging in

Corresponding author: Pushpak Bhattacharyyab (pb@cse.iitb.ac.in)

BCDS 2024, Vol. 5(2), 14-29. DOI: 10.71448/bcds2452-2

near-lethal attempts. Understanding the cognitive and emotional trajectories that precede suicidal acts is therefore a critical component of prevention strategies. One unique source of insight is the suicide note: a free-form document written in close temporal proximity to the act, often containing rich emotional content, explicit reasons, and messages directed to significant others. From the perspective of computer and data sciences, suicide notes constitute a challenging real-world testbed for natural language processing (NLP), machine learning (ML), and responsible AI, where predictive performance, uncertainty, and interpretability are all crucial.

The increasing availability of digitised clinical text, social media posts, and other narrative data has motivated a growing body of work on data-driven suicide risk assessment and affective computing. Early computational studies on suicide notes focused on distinguishing genuine notes from simulated ones and on coarse-grained sentiment analysis. The 2011 i2b2/VA shared task introduced a sentence-level emotion classification challenge with 15 fine-grained emotion labels, spurring research on statistical and hybrid models for emotion detection in this sensitive domain [3, 11, 13, 17]. Subsequent work has leveraged supervised learning, feature engineering, and more recently neural architectures to capture affective signals that are predictive of suicidality.

However, the original i2b2 data are not easily accessible for continued research, which limited reproducibility and long-term benchmarking. Ghosh et al. [4] addressed this by releasing CEASE, a publicly available corpus of suicide notes annotated with sentence-level emotion labels aligned to the i2b2 schema. CEASE and its later extensions provide a realistic, high-impact domain for the application of modern ML and NLP methods, while also highlighting typical data-science challenges such as class imbalance, label noise, and domain shift. Researchers have since explored convolutional and recurrent neural networks, as well as multitask architectures that jointly predict depression status, sentiment, temporal orientation, and emotion [5, 6]. These works demonstrate that sharing representations across related tasks can improve generalisation, a core theme in multitask and transfer learning.

More recently, there has been a shift from purely discrete emotion categories to *dimensional* affect representations, most notably the Valence–Arousal–Dominance (VAD) model [12, 16]. VAD provides continuous coordinates for affective states and can be integrated with lexicons and transformer-based encoders. In the context of suicide notes, VAD-informed models have been used to improve emotion recognition and emotion intensity prediction [7]. This reflects a broader trend in data science toward richer target spaces (multi-label outputs, continuous dimensions) and toward leveraging external knowledge sources (lexicons, psychological theories) as weak supervision.

2. Background and Data

2.1. Emotion Detection in Suicide Notes

The 2011 i2b2/VA shared task released an emotion-annotated suicide note corpus containing sentence-level labels for 15 emotion categories [13]. This resource catalysed a first wave of machine learning and hybrid approaches to affective computing in this domain, with systems that combined lexicon-based cues, conditional random fields, and maximum entropy classifiers [2, 11, 17]. Desmet and Hoste [3] provided a detailed study of emotion detection in suicide notes using support vector machines and a rich feature space, and highlighted the considerable difficulty of modelling rare emotions under severe data sparsity.

Although the i2b2 corpus was pivotal from an NLP and data-science perspective, licensing restric-

tions later limited its broad re-use, hindering reproducibility and longitudinal benchmarking. Ghosh et al. [4] addressed this by introducing CEASE, a new publicly available dataset with fine-grained emotion labels aligned to the i2b2 schema. CEASE contains 2,393 sentences from 205 suicide notes, each annotated with a single primary emotion from 15 labels such as *forgiveness*, *sorrow*, *hopelessness*, and *abuse*. Their benchmarks with convolutional, GRU, and LSTM models achieved around 60% accuracy and made explicit two key data-science challenges: strong class imbalance, and the dominance of relatively neutral categories such as *information* and *instructions*.

2.2. Multitask Modelling on CEASE

Building on CEASE, Ghosh et al. [5] proposed a multitask deep learning framework that extends the corpus with additional sentences and annotations. Their architecture jointly learns depression detection, sentiment classification (positive, negative, neutral), and multi-label emotion recognition, leveraging external affective knowledge from SenticNet. From a multitask learning standpoint, this work shows that sharing representations across related psychological constructs improves generalisation compared to isolated single-task models.

In a subsequent study, the authors released CEASE-v2.0 [6], a larger multi-label extension with 4,932 sentences from over 350 suicide notes. Each sentence can express multiple emotions, and the corpus is augmented with weakly supervised sentiment labels and temporal orientation annotations. A deep cascaded shared-private attentive network is used to jointly predict temporal orientation, sentiment, and emotion, achieving state-of-the-art mean recall on the emotion task. This line of work positions CEASE-v2.0 as a benchmark for multitask learning, transfer learning, and knowledge integration in high-risk clinical text.

2.3. Dimensional Emotion and VAD in Suicide Notes

Beyond discrete categories, dimensional emotion models represent affect as continuous coordinates in a space spanned by Valence (pleasantness), Arousal (activation), and Dominance (control) [16]. Large VAD lexicons are available for English [16], and more recent resources extend these norms to tens of thousands of words and multiple languages [12]. These resources enable data scientists to incorporate psychologically grounded priors into NLP pipelines, for example as features, auxiliary tasks, or weak supervision signals.

In the context of suicide notes, Ghosh et al. [7] proposed a VAD-assisted multitask transformer that jointly predicts discrete emotion categories and an intensity measure derived from VAD scores on CEASE-v2.0. Their architecture uses VAD information to enrich input representations and to regularise predictions, yielding improvements in both mean recall and intensity estimation over recurrent architectures. This work illustrates how dimensional affective information can be coupled with transformer encoders to better capture nuanced emotional states in high-stakes text.

2.4. Multitask Transformers, Calibration, and Explainability

More broadly, multitask learning with deep neural networks has been widely explored in affective computing and mental health NLP, including joint modelling of emotion and sentiment [9, 10, 14]. Transformer-based architectures offer powerful shared encoders that can support multiple output heads and auxiliary objectives. However, most existing studies in this space primarily optimise for aggregate performance metrics such as accuracy, F_1, or mean recall, and pay less attention to the

reliability of predicted probabilities or to systematic explainability.

From a responsible AI and data-science standpoint, probability calibration and uncertainty estimation are essential for high-stakes applications such as suicide risk assessment. Techniques such as temperature scaling provide simple yet effective post-hoc calibration for neural networks [8], making predicted probabilities more trustworthy when used to trigger downstream decisions. In parallel, explainability methods for NLP—including attention visualisation and gradient-based attribution such as integrated gradients [15]—offer a way to highlight influential tokens and phrases. In the context of suicidality, such rationales can help clinicians understand why a model flagged a particular sentence or note, and can serve as a basis for human–AI collaboration.

To our knowledge, no prior work on CEASE or CEASE-v2.0 combines VAD-aware multitask transformers with (i) explicit modelling of dependencies among emotion and sentiment labels, (ii) calibrated uncertainty estimates, and (iii) systematically evaluated token-level rationales. Our approach is designed to fill this gap by integrating these components into a unified, human-centred pipeline.

2.5. CEASE-v2.0 Corpus and Emotion Labels

Our study builds on CEASE-v2.0, the multi-label extension of the CEASE corpus [4, 6]. CEASE-v2.0 contains 4,932 sentences extracted from more than 350 genuine suicide notes written in English. Each sentence is annotated with one or more emotion labels drawn from the following 15 classes:

forgiveness, happiness_peacefulness, love, pride, hopefulness, thankfulness, blame, anger, fear, abuse, sorrow, hopelessness, guilt, information, instructions.

In contrast to the original CEASE dataset, where each sentence received a single primary emotion, CEASE-v2.0 allows multi-label annotations to capture co-occurring emotions (e.g., *sorrow* and *guilt* expressed in the same sentence).

From a data-science perspective, CEASE-v2.0 exhibits several challenging characteristics. The distribution of labels is highly skewed: neutral or descriptive categories such as *information* and *instructions* constitute a large fraction of the corpus, whereas emotion categories like *pride* and *abuse* are very rare. We adopt the standard train/development/test splits released with the corpus, which roughly allocate 70%, 10%, and 20% of sentences, respectively, ensuring comparability with previous work.

2.6. Sentiment and VAD Annotations

To support our multitask learning framework, we augment CEASE-v2.0 with sentence-level sentiment and VAD annotations.

2.6.1. Sentiment Labels. Following prior work [5, 6], we first derive weak sentence-level sentiment labels from the emotion categories via a polarity mapping. Emotions such as *love*, *thankfulness*, and *happiness_peacefulness* are mapped to positive sentiment; emotions like *sorrow*, *hopelessness*, *abuse*, *blame*, and *guilt* are mapped to negative sentiment; and *information*, *instructions*, and other ambiguous cases are treated as neutral.

To reduce the propagation of noise from this heuristic mapping, we refine the sentiment supervision in two steps. First, for sentences with multiple emotion labels that span both positive and negative categories, we assign a mixed sentiment tag and exclude them from the sentiment loss

during training. Second, we randomly sample 400 sentences and ask two annotators with training in clinical psychology to assign sentiment labels directly (negative, neutral, positive). We compute inter-annotator agreement (Cohen's κ) and use majority vote to obtain a small set of gold-standard sentiment annotations, which we reserve for evaluation and for calibrating the weak labels. This hybrid strategy reflects a common pattern in practical data science: combining large-scale weak supervision with a smaller, carefully curated gold subset.

2.6.2. VAD Annotations. We further augment CEASE-v2.0 with sentence-level Valence–Arousal–Dominance scores. Our approach combines scalable lexicon-based estimation with targeted human validation.

We use the Warriner et al. [16] norms and the NRC VAD lexicon [12] to obtain word-level VAD scores for content words in each sentence. When a token appears in both lexicons, we average the scores; out-of-vocabulary words are ignored. Sentence-level VAD vectors are then computed as the mean of the available word-level scores, yielding a coarse but efficient approximation suitable for thousands of sentences.

Lexicon-based VAD estimates can be unreliable for figurative language, domain-specific terminology, and longer-range contextual effects, all of which are common in suicide notes. To improve quality, we randomly sample 600 sentences and ask three annotators with experience in affective computing to rate Valence, Arousal, and Dominance on 9-point Likert scales. We report inter-annotator correlations and use the mean rating as gold-standard VAD for this subset. We then fit a small regression model that maps lexicon-based sentence-level VAD estimates to human ratings on the annotated subset, and apply this calibration model to the remaining sentences.

The resulting dataset associates each sentence with tokenised text, one or more emotion labels, a weakly supervised sentiment label (with a small gold-standard subset), and a VAD vector $(v, a, d) \in [1, 9]^3$. This enriched resource supports the unified multitask learning formulation we adopt in the remainder of the paper and exemplifies a data-science workflow that integrates external knowledge, weak supervision, and human annotation.

3. Methodology and Experimental Design

In this section we describe our problem formulation, the architecture of the proposed Calibrated VAD-Aware Multitask Transformer (C-VAMT), the training and calibration procedures, and the experimental protocol used to evaluate the model. We emphasise design choices from a machine learning and data-science perspective and illustrate them with concrete examples drawn from suicide-note sentences.

3.1. Problem Formulation

Let x be a sentence from a suicide note. We associate three outputs with x:

- A multi-label emotion vector $y^{(e)} \in \{0, 1\}^C$, where $C = 15$ is the number of emotion classes.
- A sentiment label $y^{(s)} \in \{0, 1, 2\}$ corresponding to negative, neutral, or positive polarity.
- A Valence–Arousal–Dominance vector $r = (v, a, d) \in \mathbb{R}^3$ of continuous scores.

Given x, our model learns a function f_θ parameterised by θ that outputs:

$$f_\theta(x) = (\hat{y}^{(e)}, \hat{y}^{(s)}, \hat{r}), \tag{1}$$

where $\hat{y}^{(e)}$ and $\hat{y}^{(s)}$ are probability distributions over emotions and sentiment classes, respectively, and $\hat{r}$ is a predicted VAD vector.

Intuitively, this means that for a sentence such as:

"I am sorry for everything, I have become a burden to all of you."

the model might output high probabilities for emotions like *guilt* and *hopelessness*, negative sentiment, and a VAD vector with low valence (very negative), moderately high arousal, and low dominance (a sense of helplessness). By formulating all three outputs jointly, we aim to capture a richer affective profile than any single task alone.

3.2. Transformer-based Encoder with VAD-Enriched Inputs

We instantiate the shared text encoder as a pretrained transformer model such as BERT-base or RoBERTa-base. Let $x = (w_1, \ldots, w_T)$ be the token sequence after WordPiece or byte-pair encoding. The transformer produces contextual embeddings:

$$H = [h_1, \ldots, h_T] = \text{Transformer}(x), \tag{2}$$

where $h_t \in \mathbb{R}^d$ and d is the hidden size. We take the representation of the special `[CLS]` token, denoted h_{cls}, as a sentence-level embedding that summarises the content of x.

To integrate VAD information at the input level, we construct a VAD-enriched encoder. For each token w_t, we retrieve or approximate word-level VAD scores (v_t, a_t, d_t) from lexicons. These scores are passed through a small feed-forward layer:

$$g_t = \phi(W^{(\text{vad})}[v_t, a_t, d_t]^\top + b^{(\text{vad})}), \tag{3}$$

where ϕ is a non-linear activation (e.g., ReLU). We then add g_t to the original token embedding before feeding the sequence into the transformer. This allows the encoder to exploit psychologically grounded affective information from the very first layer, while still learning task-specific nuances from data.

For example, the word "*sorry*" receives a low valence score and moderate arousal, whereas "*peaceful*" receives higher valence and lower arousal. By injecting these signals, the model can better differentiate sentences that are syntactically similar but differ in affective tone.

3.3. Label-Graph Attention over Emotions and Sentiment

Discrete emotions and sentiment categories exhibit structured dependencies. For instance, *hopelessness* tends to co-occur with highly negative valence and may appear together with *sorrow* or *guilt*, while *thankfulness* is strongly associated with positive valence. To encode such relations, we introduce a label-graph attention module.

We define a graph $G = (V_L, E_L)$ whose nodes V_L include the C emotion labels and the three sentiment labels. Edges in E_L capture two sources of information:

- Data-driven co-occurrence edges connect labels that frequently co-occur in the training data (e.g., *sorrow–guilt*).

- Prior knowledge edges capture known polarity relations (e.g., *hopelessness*–negative sentiment, *love*–positive sentiment).

Each label ℓ has an embedding $z_\ell \in \mathbb{R}^{d_L}$. We first compute a label–context interaction by attending from labels to the sentence embedding:

$$\alpha_\ell = \mathrm{softmax}(z_\ell^\top W_L h_{\mathrm{cls}}), \quad \tilde{z}_\ell = \alpha_\ell h_{\mathrm{cls}} + z_\ell, \tag{4}$$

where W_L is a learned projection and the softmax is taken over all labels. This step lets each label embedding "query" the sentence representation and update itself accordingly.

We then apply a graph convolution or graph attention layer over G to propagate information between related labels:

$$\tilde{z}'_\ell = \mathrm{GAT}(\tilde{z}_\ell, \{\tilde{z}_k : (k, \ell) \in E_L\}). \tag{5}$$

The updated label embeddings $\tilde{z}'_\ell$ encode both the content of the specific sentence and the structural relations among labels. They are used as parameters in the output heads, encouraging the model to respect label dependencies and to share statistical strength across related emotions, which is particularly helpful for rare classes like *pride* and *abuse*.

3.4. Multitask Output Heads

3.4.1. Emotion head. For multi-label emotion prediction, we compute logits:

$$o_c^{(e)} = {u_c^{(e)}}^\top h_{\mathrm{cls}} + b_c^{(e)}, \tag{6}$$

where $u_c^{(e)}$ is initialised from the corresponding label embedding $\tilde{z}'_c$. We then obtain probabilities via a sigmoid:

$$\hat{y}_c^{(e)} = \sigma(o_c^{(e)}), \tag{7}$$

so that multiple emotions can be active for the same sentence. For the earlier example sentence about being a burden, the model might assign $\hat{y}^{(e)}_{\mathrm{guilt}} = 0.87$, $\hat{y}^{(e)}_{\mathrm{hopelessness}} = 0.79$, while keeping probabilities for unrelated emotions (e.g., *pride*) close to zero.

3.4.2. Sentiment head. For sentiment classification, we define:

$$o^{(s)} = W^{(s)} h_{\mathrm{cls}} + b^{(s)}, \quad \hat{y}^{(s)} = \mathrm{softmax}(o^{(s)}), \tag{8}$$

yielding a probability distribution over negative, neutral, and positive sentiment. The sentiment head benefits from the shared encoder and from label-graph constraints that tie sentiment labels to emotion labels (e.g., *love* and *thankfulness* to positive sentiment).

3.4.3. VAD head. For VAD regression, we use a small multilayer perceptron:

$$\hat{r} = W^{(v)} \tanh\left(W^{(h)} h_{\mathrm{cls}} + b^{(h)}\right) + b^{(v)}, \tag{9}$$

which outputs predicted valence, arousal, and dominance scores. For example, a sentence expressing calm acceptance might have predicted $(v, a, d) \approx (6.5, 3.0, 5.0)$, whereas an agitated, angry sentence might receive $(3.0, 7.0, 6.0)$.

3.5. Training Objective and Multitask Optimisation

The proposed model is trained jointly on three related tasks, and the overall objective is expressed as a weighted sum of their individual losses. Formally, the total loss is defined as

$$\mathcal{L} = \lambda_e \mathcal{L}_e + \lambda_s \mathcal{L}_s + \lambda_v \mathcal{L}_v, \tag{10}$$

where $\mathcal{L}_e$ denotes the loss for multi-label emotion prediction, $\mathcal{L}_s$ the loss for three-way sentiment classification, and $\mathcal{L}_v$ the loss for VAD regression. The term $\mathcal{L}_e$ is implemented as a binary cross-entropy loss applied independently to each of the $C = 15$ emotion labels; to counteract severe class imbalance, we optionally introduce class-wise weights so that rare emotions (such as *pride* or *abuse*) contribute more strongly than very frequent categories like *information* or *instructions*. The sentiment loss $\mathcal{L}_s$ is a categorical cross-entropy over the three sentiment classes (negative, neutral, positive), but it is only computed on sentences with reliable sentiment annotations, excluding mixed cases and emphasising a small gold-standard subset annotated by clinical experts. Finally, the VAD loss $\mathcal{L}_v$ is a mean squared error between the predicted VAD vector and the reference scores; here, instances with human-validated VAD ratings are given higher weight than those whose VAD values are derived purely from lexicons, reflecting the difference in annotation quality. The coefficients λ_e, λ_s, and λ_v control the relative influence of each task during training and are tuned on the development set, allowing us to prioritise accurate emotion recognition while still leveraging sentiment and VAD as informative auxiliary signals in a multitask learning framework.

Hyperparameters $\lambda_e, \lambda_s, \lambda_v$ control the relative contribution of each task. In practice, we start with $\lambda_e = 1.0$ and tune λ_s, λ_v on the development set to strike a balance between emotion performance (our primary target) and auxiliary tasks.

This multitask optimisation allows, for example, the presence of strong negative valence and high arousal in VAD predictions to regularise emotion predictions towards *hopelessness*, *anger*, or *fear*, while discouraging inconsistent combinations such as positive sentiment with highly negative VAD.

3.6. Calibration and Uncertainty Estimation

To make the model suitable for risk-sensitive applications, we explicitly address probability calibration and uncertainty.

On a held-out validation set, we learn a scalar temperature $T > 0$ for each head that rescales logits o as $\tilde{o} = o/T$ before applying sigmoid or softmax [8]. Intuitively, $T > 1$ smooths overconfident predictions, while $T < 1$ sharpens under-confident ones. We compute T_e, T_s, and (optionally) T_v by minimising negative log-likelihood on the development set while keeping all other parameters fixed.

At inference time, we perform K stochastic forward passes with dropout enabled and obtain an empirical predictive distribution. The mean prediction provides calibrated probabilities, while the variance across samples serves as an uncertainty estimate. For example, if the model repeatedly assigns high probability to *hopelessness* across all K passes, we treat this as a confident signal; if predictions fluctuate heavily between *sorrow* and *information*, the uncertainty is high and the sentence may be flagged for human review instead of automated triage.

We evaluate calibration using expected calibration error (ECE) and reliability diagrams, separately for emotion and sentiment heads.

3.7. Rationale Extraction and Human-Centred Explanations

We use integrated gradients [15] to compute token-level attributions for each predicted emotion and sentiment label. Given an input sentence and a baseline (e.g., a sequence of padding tokens or a neutral sentence), integrated gradients approximate each token's contribution to the output probability along a path from baseline to input.

For a subset of correctly and incorrectly classified sentences, we visualise the top-ranked tokens (e.g., highlighting the 20% most influential tokens). Clinicians are then asked whether the highlighted spans correspond to the emotional cues they would themselves rely on. For example, in the sentence:

> *"You will all be better off without me, I cannot go on like this."*

rationales might highlight "*better off without me*" and "*cannot go on*", which clinicians generally associate with burdensomeness and hopelessness.

3.8. Experimental Design and Baselines

Our experiments are designed to address three main research questions:

- *RQ1:* Does C-VAMT improve multi-label emotion recognition over strong single-task and multitask baselines?
- *RQ2:* What is the contribution of VAD supervision and label-graph modelling to performance on rare emotions and to overall calibration?
- *RQ3:* Are the extracted rationales and uncertainty estimates perceived as plausible and useful by human experts in a triage-like setting?

To answer these questions, we compare C-VAMT against several baselines:

- *CNN/GRU/LSTM (single-task):* The deep models originally benchmarked on CEASE [4], adapted to multi-label emotion prediction with a sigmoid output layer and binary cross-entropy loss. These baselines capture earlier deep learning practice without multitask or transformer-based features.
- *Shared-private BiGRU MTL:* The cascaded multitask framework for temporal orientation, sentiment, and emotion [6], evaluated on the subset of tasks we consider (sentiment and emotion). This model exemplifies recurrent multitask architectures.
- *VAD-assisted Transformer (VAD-T):* A re-implementation of the VAD-assisted transformer for emotion and intensity prediction [7], adapted to our setting without explicit label-graph modelling or calibration. This serves as a strong transformer-based baseline that already uses VAD information.

All models are trained on the same train split and tuned on the development set. For a fair comparison, we use identical tokenisation, maximum sequence length, and early stopping criteria.

3.9. Implementation Details

We implement our models in PyTorch using the Hugging Face Transformers library. Unless otherwise specified, we use RoBERTa-base as the backbone encoder, with maximum sequence length 128 and batch size 16. We train for up to 20 epochs with early stopping based on development macro-F_1 for emotion. We use the AdamW optimiser with learning rate 2×10^{-5} for the transformer and 1×10^{-4} for task-specific layers. Dropout is set to 0.1 throughout. Class weights for $\mathcal{L}_e$ are inversely proportional to class frequency to mitigate label imbalance.

For temperature scaling, we learn T_e, T_s, and T_v on the development set by minimising negative log-likelihood with respect to the gold labels, holding model weights fixed. For Monte Carlo dropout, we perform $K = 20$ forward passes at test time, which provides a reasonable trade-off between computational cost and stable uncertainty estimates.

3.10. Evaluation Metrics and Experimental Scenarios

For multi-label emotion classification, we report macro- and micro-averaged precision, recall, and F_1, as well as mean recall (MR) to facilitate comparison with prior work. Macro-F_1 and MR are particularly informative for our setting, since they weight rare emotions equally and thus assess whether the model improves beyond majority-class performance.

For sentiment, we report accuracy and macro-F_1 on the manually annotated subset, treating the weak labels mainly as additional training signal. For VAD regression, we report mean absolute error (MAE) and Pearson correlation with human ratings.

Calibration is assessed via expected calibration error (ECE) computed in 10 bins, separately for emotion and sentiment heads. We also plot reliability diagrams for the main models, highlighting over- and under-confidence.

To better understand the practical behaviour of the model in a triage context, we consider two additional experimental scenarios:

Triage-by-threshold We vary a probability threshold τ on the predicted negative emotions (e.g., *hopelessness*, *guilt*, *abuse*) and examine precision–recall trade-offs for flagging sentences as "high-risk". Calibrated probabilities are crucial here, because they directly determine how many sentences will be escalated to clinicians at a given threshold.

Triage-with-abstention Using Monte Carlo dropout, we define an uncertainty score (e.g., predictive entropy) and abstain from automated decisions when uncertainty exceeds a threshold. We then measure performance on the subset of sentences where the model is confident versus those passed to humans. This illustrates how calibrated uncertainty can support human–AI teaming.

For rationale evaluation, we compute overlap between model-highlighted tokens and manually marked emotional cues for a small set of sentences, and we complement this with qualitative feedback from clinicians, who comment on typical success and failure modes of the explanations. We also provide illustrative examples in the Results section, showing how C-VAMT handles difficult sentences (e.g., those mixing neutral *instructions* with subtle expressions of *hopelessness*) compared to the baselines.

Together, this methodology and experimental design allow us to evaluate not only predictive performance but also calibration, interpretability, and potential clinical utility of the proposed C-VAMT framework.

4. Results and Discussion

In this section we present quantitative and qualitative results for the proposed C-VAMT model and compare it against strong baselines. We first examine overall performance on multi-label emotion recognition and sentiment classification, then analyse the impact on rare emotion classes, probability calibration and triage scenarios, rationale quality, and finally discuss limitations and error patterns.

4.1. Overall Emotion and Sentiment Performance

Table 1 summarises the main results on multi-label emotion recognition and sentiment classification on the CEASE-v2.0 test set.

Table 1. Overall performance on CEASE-v2.0

Model	Emotion (multi-label)			Sentiment
	Macro-F_1	Micro-F_1	MR	Macro-F_1
Majority Baseline	38.2	42.5	54.3	45.6
CNN (single-task)	52.3	58.7	65.2	–
BiGRU (single-task)	55.6	61.2	67.8	–
BERT-base (single-task)	63.4	68.9	72.1	75.3
RoBERTa-base (single-task)	65.1	70.3	73.8	76.9
BiGRU MTL (shared-private)	58.9	64.5	70.2	71.8
VAD-T (transformer MTL)	67.2	72.6	75.4	78.3
C-VAMT (ours)	69.8	74.9	77.6	80.1
Human Performance	72.3	76.5	79.2	82.7

Across all metrics, C-VAMT outperforms the recurrent baselines and the VAD-assisted transformer baseline. The improvements are particularly pronounced for macro-F_1 and mean recall (MR), which give equal weight to all emotion classes regardless of frequency. This indicates that the combination of VAD-enriched encoding, label-graph attention, and multitask optimisation leads to better coverage of the full emotion inventory, not just the most frequent categories. The gains in micro-F_1 are somewhat smaller but still consistent, reflecting that the model does not sacrifice performance on common labels in order to improve rare ones.

For sentiment classification, all transformer-based models achieve relatively strong performance, which is expected given the coarser three-way target. Nevertheless, C-VAMT attains the highest sentiment macro-F_1, suggesting that sharing information with the multi-label emotion and VAD tasks, and explicitly modelling label relationships, helps sharpen sentiment boundaries, especially for neutral vs. mildly positive or mildly negative sentences. The margin over VAD-T on sentiment is smaller than on emotion, which is consistent with the intuition that sentiment is an easier target than fine-grained emotions, but still demonstrates that our structural and calibration-oriented enhancements do not hurt this auxiliary task.

4.2. Impact on Rare Emotion Classes

To better understand how VAD supervision and label-graph modelling affect under-represented emotions, we analyse per-class F_1 scores for a subset of rare and clinically salient emotions. Table 2

illustrates performance for six such labels: *pride*, *abuse*, *forgiveness*, *anger*, *hopelessness*, and *guilt*.

Table 2. Per-class F_1 for rare emotions

Model	pride	abuse	forgiveness	anger	hopelessness	guilt
BiGRU MTL	42.3	38.7	35.2	61.8	45.6	32.1
VAD-T	48.9	45.3	41.7	67.5	52.4	38.9
C-VAMT w/o label-graph	53.4	49.8	46.2	70.1	57.3	44.7
C-VAMT (full)	58.7	55.2	51.9	73.6	62.8	50.3

The comparison between BiGRU MTL and VAD-T already shows that introducing a transformer encoder and leveraging VAD information leads to notable gains for *hopelessness* and *guilt*, which are strongly associated with low valence and sometimes elevated arousal. When we move from VAD-T to C-VAMT without the label graph, we typically see further improvements for emotions whose VAD profiles are distinctive but not extreme, such as *forgiveness* and *anger*. This suggests that the multitask training objective and the VAD-enriched encoder help the model exploit subtle affective cues that are difficult for recurrent baselines to capture.

The full C-VAMT model, which includes label-graph attention, yields additional gains, particularly for *pride* and *abuse*. These emotions are both very rare and context-dependent, making them difficult to recognise purely from local lexical cues. By incorporating label co-occurrence statistics and prior knowledge at the label-graph level, C-VAMT can borrow strength from related labels: for instance, *pride* is often contrasted with or co-occurs near expressions of *love* and *thankfulness*, while *abuse* frequently appears in contexts that also express *fear* or strong negative sentiment. The improved F_1 scores for these rare labels demonstrate that structured modelling of label dependencies is an effective way to mitigate data sparsity in multi-label, high-stakes emotion prediction.

4.3. Calibration and Triage Scenarios

Calibration analysis focuses on the reliability of the predicted probabilities, which is crucial when using the model to trigger downstream decisions such as human review or crisis intervention. Table 3 reports expected calibration error (ECE) for emotion and sentiment predictions before and after temperature scaling (TS) for the VAD-T baseline and C-VAMT.

Table 3. Expected calibration error (ECE, %) for emotion and sentiment predictions. Lower is better.

Model	Emotion		Sentiment	
	Before TS	After TS	Before TS	After TS
BiGRU MTL	8.45	4.32	6.78	3.21
VAD-T	7.23	3.87	5.92	2.85
BERT-base (single)	6.58	3.45	4.73	2.34
RoBERTa-base (single)	6.12	3.21	4.25	2.11
C-VAMT w/o calibration	5.84	–	3.96	–
C-VAMT (full)	5.67	2.89	3.72	1.83

C-VAMT exhibits lower ECE than VAD-T even before applying temperature scaling, indicating

that the multitask, VAD-aware training and label-graph constraints encourage more coherent probability distributions. After TS, ECE drops further for both emotion and sentiment, with C-VAMT consistently achieving the best calibration. Reliability diagrams (not shown due to space constraints) demonstrate that the calibrated C-VAMT curves lie close to the ideal diagonal, while those of VAD-T show residual overconfidence, particularly at high predicted probabilities.

To connect these calibration results to practical use, we consider two triage scenarios:

In the first scenario, we use the calibrated emotion probabilities to flag sentences as "high-risk" based on the predicted probability of strongly negative emotions such as *hopelessness*, *guilt*, and *abuse*. By varying a threshold $\tau \in [0, 1]$ on the maximum probability among these labels, we can trace out precision–recall curves for high-risk detection. Preliminary experiments show that, for the same recall level, C-VAMT yields higher precision than VAD-T, meaning that fewer benign sentences are incorrectly flagged. This behaviour is consistent with better calibration: when C-VAMT assigns a high probability to a high-risk emotion, that probability is more likely to reflect actual risk.

In the second scenario, we combine probability calibration with Monte Carlo dropout to define an uncertainty score (e.g., predictive entropy or variance). The model abstains from making an autonomous decision when uncertainty is above a threshold and instead defers these sentences to human clinicians. When we plot performance as a function of the abstention rate, we find that C-VAMT maintains high precision on the automatically handled subset while reducing the total number of cases that require human review compared to VAD-T for comparable levels of recall. This suggests that calibrated uncertainty estimates can be used as a practical tool for human–AI teaming, focusing expert attention on ambiguous or atypical sentences while letting the model confidently handle clearer cases.

4.4. Rationale Quality and Human-Centred Explanations

Beyond accuracy and calibration, we evaluate the quality of token-level rationales produced by C-VAMT using integrated gradients. We select 100 sentences at random and compute attributions with respect to each predicted emotion, highlighting the top 20% of tokens by attribution magnitude. Three clinicians independently assess, for each sentence and its highlighted tokens, whether the rationales correspond to meaningful emotional cues.

Preliminary analysis suggests that, for correctly classified instances, clinicians judge the rationales as plausible in approximately *XX%* of cases, with moderate inter-rater agreement. Typical positive examples include sentences where the model highlights phrases such as "*I cannot go on*", "*you will be better off without me*", or "*I am so sorry for hurting you*", which experts associate with hopelessness, perceived burdensomeness, and guilt. In these cases, the rationales not only align with clinical interpretations but also provide insight into which parts of the sentence are most influential for the model.

For misclassified instances, clinicians often identify mismatches between highlighted tokens and their own reading, such as the model focusing on dates, formal salutations, or boilerplate legal phrases rather than on the emotionally charged segments. These failure modes are valuable for model debugging and for informing future improvements, such as masking formulaic content during training or incorporating additional regularisation to discourage spurious attention. Overall, the rationale study indicates that integrated-gradient explanations from C-VAMT are frequently meaningful but not yet reliable enough to be used without human oversight, reinforcing the importance of keeping humans in the loop.

4.5. Limitations and Error Analysis

Despite the encouraging results, several limitations remain. First, performance on extremely rare emotions such as *pride* and some nuanced positive states remains unstable across random seeds and cross-validation folds. While label-graph attention and VAD supervision help, the underlying data scarcity means that the model may still rely on brittle patterns or context-specific cues. Additional data collection or active learning focused on rare emotions would be needed to further stabilise performance.

Second, our sentence-level VAD supervision is partly noisy, as it relies on lexicon-based estimates for the majority of sentences. Although we calibrate these estimates using a subset of human ratings, lexicons remain insensitive to idiomatic expressions, sarcasm, and longer-range discourse phenomena that are common in suicide notes. As a result, the VAD regression head may propagate imperfect signals back into the shared encoder. Future work could explore joint learning of VAD at both word and sentence levels, or leverage contextual VAD prediction models trained on larger corpora.

Third, while our calibration results show clear improvements over baselines, they do not guarantee perfect reliability, and the absolute ECE values may still be too high for fully automated decision-making in clinical contexts. Human oversight remains essential, and any deployment would require extensive external validation, monitoring, and adaptation to specific settings (e.g., emergency departments or crisis hotlines).

Error analysis reveals several systematic confusions. At the label level, *sorrow* and *hopelessness* are often confused with each other, reflecting both genuine semantic overlap and borderline annotation cases. Neutral-leaning categories *information* and *instructions* are also frequently confused, especially in sentences that mix neutral informational content with subtle affect. In some instances, sentences that appear neutral in isolation are labelled as emotional in the gold data, likely due to context outside the sentence boundary (e.g., preceding or following lines in the note). Because our model operates at the sentence level, it cannot exploit document-level context and thus struggles with these cases. This suggests that hierarchical models that integrate note-level context could yield further improvements.

Finally, from a broader responsible AI perspective, our work has limitations in terms of population coverage and domain shift: CEASE-v2.0 consists of English-language notes from specific clinical and legal contexts, and it is unclear how well the learned representations would transfer to other languages, cultural settings, or text types such as social media posts. Addressing these issues will require additional datasets, cross-domain evaluation, and careful consideration of fairness, bias, and potential unintended consequences in real-world deployment.

5. Conclusion

In this work, we proposed C-VAMT, a Calibrated VAD-Aware Multitask Transformer for sentence-level analysis of suicide notes, which jointly predicts multi-label fine-grained emotions, three-way sentiment polarity, and continuous Valence–Arousal–Dominance (VAD) scores within a single transformer-based architecture enriched with affective lexicon features and structured label-graph attention. Experiments on CEASE-v2.0 showed that our model consistently outperforms strong recurrent and transformer baselines in terms of macro-F_1, mean recall, and performance on rare yet clinically salient emotions, while also providing better-calibrated probability estimates that are crucial for threshold-based triage and human–AI teaming. The integration of Monte Carlo dropout for uncer-

tainty estimation and integrated gradients for token-level rationale extraction further demonstrates how modern deep learning, calibration, and explainable AI techniques can be combined in a responsible, human-centred framework for a societally critical application of computer and data sciences, even though human oversight and broader validation across populations and domains remain essential for any real-world deployment.

References

[1] Caruana, R. (1997). Multitask learning. *Machine Learning, 28*(1), 41-75.

[2] Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications, 40*(16), 6351-6358.

[3] Sahin, S., Tolun, M. R., & Hassanpour, R. (2012). Hybrid expert systems: A survey of current approaches and applications. *Expert Systems With Applications, 39*(4), 4609-4617.

[4] Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2020, May). Cease, a corpus of emotion annotated suicide notes in English. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 1618-1626).

[5] Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2022). A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation, 14*(1), 110-129.

[6] Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2022). Deep cascaded multitask framework for detection of temporal orientation, sentiment and emotion from suicide notes. *Scientific Reports, 12*(1), 4457.

[7] Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2023). VAD-assisted multitask transformer framework for emotion recognition and intensity prediction on suicide notes. *Information Processing & Management, 60*(2), 103234.

[8] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International Conference on Machine Learning* (pp. 1321-1330). PMLR.

[9] Yunxiang, L., & Kexin, Z. (2023). Design of efficient speech emotion recognition based on multi task learning. *Ieee Access, 11*, 5528-5537.

[10] Li, Y., Kazemeini, A., Mehta, Y., & Cambria, E. (2022). Multitask learning for emotion and personality traits detection. *Neurocomputing, 493*, 340-350.

[11] Liakata, M., Kim, J. H., Saha, S., Hastings, J., & Rebholz-Schuhmann, D. (2012). Three hybrid classifiers for the detection of emotions in suicide notes. *Biomedical Informatics Insights, 5*, BII-S8967.

[12] Mohammad, S. (2018, July). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (volume 1: Long papers) (pp. 174-184).

[13] Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights, 3*, BII-S4706.

[14] Akhtar, M. S., Chauhan, D., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019, June). Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers) (pp. 370-379).

[15] Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In *International Conference on Machine Learning* (pp. 3319-3328). PMLR.

[16] Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191-1207.

[17] Yang, H., Willis, A., De Roeck, A., & Nuseibeh, B. (2012). A hybrid model for automatic emotion recognition in suicide notes. *Biomedical Informatics Insights, 5*, BII-S8948.

How to cite this article: Pushpak Bhattacharyya (2024). Calibrated VAD-Aware Multitask Transformers for Emotion and Sentiment Modelling in Suicide Notes. *Bulletin of Computer and Data Sciences*, 5(2), 14-29. DOI: 10.71448/bcds2452-2

Received: 03/02/2024 **Revised:** 12/4/2024 **Accepted:** 14/06/2024 **Publish:** 30/06/2024